Structural studies of the endotoxin sensing protein Factor C

Karen McClymont

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

Institute of Molecular, Cell and Systems Biology
College of Medical, Veterinary and Life Sciences
University of Glasgow

March, 2017

Abstract

Human exposure to Gram-negative bacterial endotoxin can result in life-threatening conditions including sepsis and toxic shock syndrome. Given the importance of avoiding exposure to endotoxin, a sensitive test is required to ensure that medical devices and injectable medicines are not contaminated. *Limulus* amebocyte lysate (LAL), produced from the blood of horseshoe crabs, contains an endotoxin-sensing protein, Factor C, that initiates a coagulation cascade. This lysate is used to test for contamination of medical products, with formation of a clot indicating the presence of endotoxin in the sample.

Alternative methods to the LAL test have been explored to overcome problems with the test. These include the detrimental effects that harvesting has on the wild population of horseshoe crabs. More recently, tests relying on the direct detection of enzymatically active Factor C have been introduced, some of which utilize Factor C produced recombinantly. Despite the importance of Factor C to the health industries, no detailed explanation of its mechanism of endotoxin recognition has been published. The aim of this project was therefore to develop a better understanding of how Factor C is activated by endotoxin binding and to identify conformational changes that take place as a result.

Two main strategies were employed for structural characterisation and functional analysis of Factor C. First, fragments of Factor C protein were produced recombinantly in *Escherichia coli* to determine the structures of the individual domains and identify the endotoxin binding region more precisely. Second, expression of the whole Factor C protein in eukaryotic expression systems was attempted to produce material for experiments that would shed light on any overall conformational change that occurs.

Factor C fragments were successfully expressed in *E. coli* and purified. Circular dichroism spectroscopy identified that the majority of these fragments were folded, and changes in the spectra of some fragments in the presence of lipopolysaccharide identified potential binding sites. A preliminary structure of the first pair of complement control domains has been determined by nuclear magnetic resonance spectroscopy. Comparison of this structure with other LPS binding proteins will pave the way for future work to characterise the endotoxin binding site of Factor C.

Contents

List of tables

List of figures

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr Brian Smith and Dr Sharon Kelly, without whom none of this would have been possible. Brian, your infinite knowledge of all things science never fails to amaze me. Thank you for guiding me over the past few years, I really appreciate everything you've done for me. Sharon, thanks for always looking out for me, it's been a privilege to work alongside you and I literally owe you my life!

To all members of the Smith lab past and present, it's been a joy. Special thanks to Ellen, Kate, Vibhuti and Gisela for helping to keep me sane, for squash, for the international dinners and for all the other good times we've shared. Thanks also to Donald Campbell for all his technical support.

To my family, for never giving up on me (even though it was the biggest surprise of your life – Mum!). To my friends, for always being there no matter what, especially Michelle, Hayley and Josie. And to Kieran, for all the little things, that really do mean a lot.

This PhD would not have been possible without the support and funding from Medical Research Scotland and Marine Biotech Ltd., for which I will be forever grateful. Thanks in particular to Scott Johnstone for all his guidance.

Author's Declaration

I declare that, except where explicit reference is made to the contributions of others, this thesis is my own work and it has not been submitted for any other degree, in whole or part, at the University of Glasgow or any other institution.

Karen McClymont

September 2016

Definitions/Abbreviations

AA	Amino acid
ALF	Anti-lipopolysaccharide factor
ARIA	Ambiguous restraints for iterative assignment
BAC	Bacterial artificial chromosome
BLAST	Basic local alignment search tool
BSA	Bovine serum albumin
CCP	Complement control protein
CD	Circular dichroism
cDNA	Complementary DNA
CIP	Alkaline phosphatase, calf intestinal
CMPS	Carboxymethylated (1→3)-β-D glucan
cr	*Carcinoscorpius rotundicauda*
CysEGF	Cys-rich and EGF-like domains together
Cys	Cysteine
dATP	Deoxyadenosine triphosphate
DANGLE	Dihedral angles from global likelihood estimates
dCMP	Deoxycytidine monophosphate
dCTP	Deoxycytidine triphosphate

dGMP	Deoxyguanosine monophosphate
dGTP	Deoxyguanosine triphosphate
DNA	Deoxyribonucleic acid
DsbC	Disulphide bond C
DTT	Dithiothreitol
dTTP	Deoxythymidine triphosphate
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EGF	Epidermal growth factor
EGFP	Enhanced green fluorescent protein
FASTA	Text-based format for representing nucleotide sequences
FC	Factor C
FRET	Fluorescence resonance energy transfer
HCl	Hydrochloric acid
HRV 3C	Human rhinovirus 3C
HSQC	Heteronuclear single quantum coherence spectroscopy
IBPP	Inclusion bodies protein purification
IMAC	Immobilized metal ion affinity chromatography
INEPT	Insensitive nuclei enhanced by polarisation transfer

IPTG	Isopropyl β-thiogalactopyranoside
LAL	*Limulus* amebocyte lysate
LCCL	*Limulus* Factor C, Coch-5b2 and Lgl1
LIC	Ligation independent cloning
lp	*Limulus polyphemus*
*lp*FC	*Limulus polyphemus* Factor C
LPS	Lipopolysaccharide
MBL	Marine Biotech Limited
mRNA	Messenger RNA
NEB	New England biolabs
NHGRI	National human genome research institute
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
NOESY	Nuclear Overhauser effect spectroscopy
OD_{600nm}	Optical $density_{600nm}$
PAMP	Pathogen-associated molecular pattern
PCR	Polymerase chain reaction
phiLOV	Photostable light, oxygen or voltage domain
PRR	Pattern recognition receptor

RNA	Ribonucleic acid
RP-HPLC	Reversed-phase high performance liquid chromatography
SDS-PAGE	Sodium dodecyl sulfate – Polyacrylamide gel electrophoresis
SGC	Structural Genomics Consortium
snRNPs	Small nuclear ribonucleoprotiens
SPR	Surface plasmon resonance
TAE	Tris base, Acetic acid and EDTA
TBIO	Thermodynamically balanced inside-out PCR-based gene synthesis
TNF-α	Tumour necrosis factor α
TEV	Tobacco etch virus
TFA	Trifluoroacetic acid
TOCSY	Total correlation spectroscopy
tPA	Tissue plasminogen activator
TROSY	Transverse relaxation-optimised spectroscopy
Trx	Thioredoxin
tt	*Tachypleus tridentatus*
WUSTL	Washington University, St Louis

1 Introduction

For over 40 years, horseshoe crabs have been important for the testing of all parenteral pharmaceuticals and medical devices for human use, for the detection of bacterial endotoxin. They have played a crucial role in preventing the spread of Gram-negative bacterial infections from drugs and other medical products, and the features that form the basis of this test have resulted in extensive research into the biological process behind this discovery.

1.1 Bacterial Endotoxin

Gram-negative bacteria are found in almost every life-supporting environment. Gram-negative bacteria include the model organism Escherichia coli and several pathogenic bacteria strains, that are responsible for a number of infectious diseases. Bacterial endotoxin or lipopolysaccharide (LPS) is the main lipidic constituent of the outer leaflet of the outer cell membrane of Gram-negative bacteria. In humans/mammals, its presence is often the first sign of a Gram-negative bacterial infection which, if undetected or untreated, can result in 'intravascular coagulation, amebocytopenia, incoagubility of the blood and death' (Levin and Bang, 1968).

Human exposure to endotoxin, not even necessarily from live bacteria can result in life threatening medical conditions, including toxic shock syndrome (TSS) and sepsis, which can lead to multiple organ failure and ultimately death. This is a result of an undamped stimulation of the host defence mechanisms, leading to the secretion of an excess amount of inflammatory cytokines from monocytes and macrophages, including tumour necrosis factor α (TNF-α), a number of interleukins (e.g. IL-1 and IL-8) and interferons (Ceramil and Beutler, 1988; Karima *et al.*, 1999; Kreutz *et al.*, 1997). Given the importance for humans of avoiding exposure to LPS, we need a sensitive test to ensure that medical devices and injectable medicines are not contaminated.

LPS molecules are built of three structural components, each of which display distinctive genetic, biochemical and antigenic features. These are the variable O-antigen polysaccharide chain, the core-oligosaccharide and the highly conserved hydrophobic Lipid A (Figure 1-1). Lipid A is composed of two phosphorylated glucosamine sugars attached to a number of fatty acids. These fatty acids can vary in number and length, and while a number of them carry a hydroxyl group, others are not hydroxylated. The overall

structure of lipid A is conserved among different bacteria, but variation arises from differences in the acylation pattern, the length of the fatty acid residues and the number of fatty acids (Park and Lee, 2013). Lipid A is recognised as a pathogen-associated molecular pattern (PAMP) by immune cells and is accountable for the bioactivity of endotoxin (Wang and Quinn, 2010). Core oligosaccharides show more diversity and can be split into two regions that show structural differences to each other: the inner core covalently bonded to lipid A and the outer core that is attached to the O-antigen (Wang and Quinn, 2010). The inner core is made up of sugars found to be unique to bacteria: 2-keto-3-deoxyoctonic acid (KDO) and heptose, whereas the outer core comprises common sugars (Van Amersfoort *et al.*, 2003). LPS variability is determined by the O-antigen that is made up of a varying number of oligosaccharide units, attached to the outer core. The O-antigen is immunogenic and responsible for the numerous serotypes as a result of the varying composition and lengths. If undefended, upon lysis of the bacterial cells, LPS circulates into the blood stream resulting in systemic exposure. It was shown that removal of the O-antigen and core-oligosaccharide regions had a limited effect on LPS activity, indicating that these regions do not play a significant role in detection by the host immune receptor (Raetz and Whitfield, 2002; Trent *et al.*, 2006).

When horseshoe crabs become infected by LPS from Gram-negative bacteria, aggregation and degranulation occurs within the blood stream. This suggests that the coagulation cascade observed in horseshoe crabs plays a significant role for the defence of the species against Gram-negative bacteria invasions.

In addition to the Factor C protein, there are a number of other proteins that recognise LPS. These include human receptors such as Toll-like receptor 4 (Chow *et al.*, 1999), the ferric hydroxamate uptake receptor (FhuA) found on the surface of *E. coli* (Ferguson *et al.*, 2000) and the anti-lipopolysaccharide factor (ALF-Pm3) from the black tiger shrimp Penaeus monodon (Yang *et al.*, 2009). Structural studies to determine the binding site of LPS in FhuA and ALF-Pm3 indicated a highly-conserved region mainly consisting of positively charged and hydrophobic residues that are able to interact with the hydrophilic and phosphate groups of lipid A, through electrostatic and hydrophobic interactions.

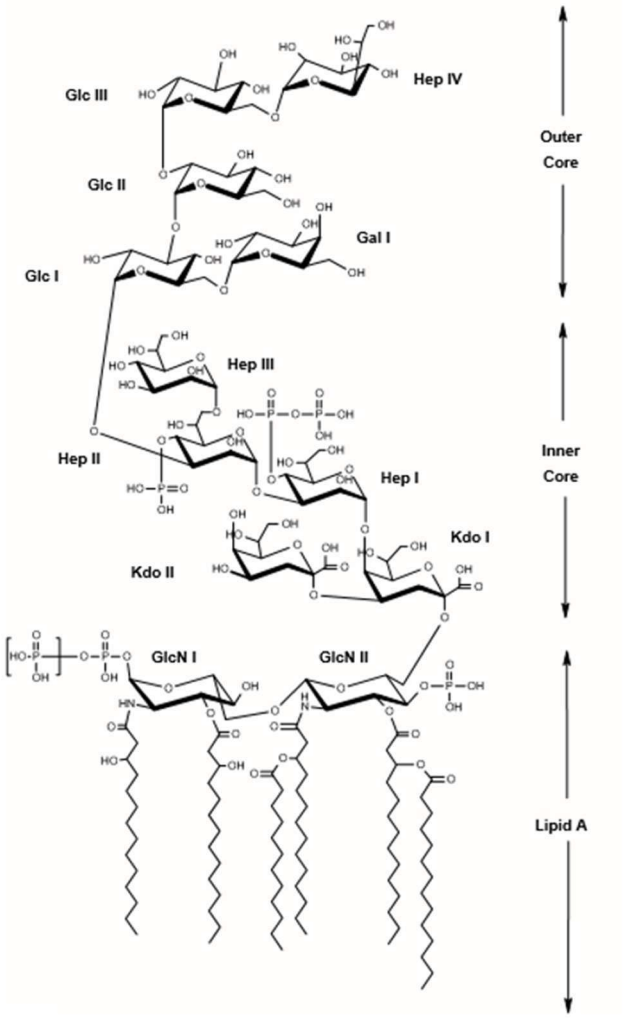

Figure 1-1: Chemical structure of lipopolysaccharide (LPS) from E. Coli K-12. Lipid A glucosamines (GlcN) are linked to KDO sugars from the core oligosaccharide, which are linked to heptose sugars. The outer core is shown to consist of hexose moieties, including D-galactose (Gal) and D-glucose (Glc), and a disordered heptose group. The O-antigen is attached to Glc from the outer core (not shown). Figure reproduced from Ferguson *et al.* (Ferguson *et al.*, 2000).

1.2 Horseshoe Crabs

There are four species of horseshoe crab: *Carcinoscorpius rotundicauda*, *Limulus polyphemus*, *Tachypleus gigas* and *Tachypleus tridentatus*. *Limulus* is the Atlantic species and is found off the East coast of North America. The other three species originate from Southeast Asia. Divergence studies have established that separation between the *L. polyphemus* and *T. tridentatus* species occurred around 135 million years ago, between the Asian species and *T. gigas*, 52.5 million years ago and between *T. tridentatus* and *C. rotundicauda*, 36.3 million years ago (Kawabata *et al.*, 2003).

The Atlantic horseshoe crab (*Limulus polyphemus*) has played an important role in several industries including the fishing industry, the agricultural industry and, most significantly,

the biomedical industry. In addition, they are preyed on by a variety of birds that attempt to feed on their eggs and also on stranded live, or deceased adult crabs (Botton and Loveland, 1989). In spite of the difficulties faced by horseshoe crab populations, they have survived for millions of years. However, as a result of the number of functions they serve, and thus the demand for them, the future of the horseshoe crab's existence is under threat.

For many years, a common use of horseshoe crabs was as soil fertiliser and livestock feed. However, as a result of a decline in numbers of the horseshoe crab population along with increased competition from alternative fertilisers, there is no longer a use for the horseshoe crabs in this industry (Berkson and Shuster, 1999). Subsequently, the commercial fishery industry identified the use of horseshoe crabs as bait to catch American eels, whelks (also known as conch) and, to a lesser extent, catfish. They have played a role in vision research, allowing for better understanding of the physiological properties of the visual system (Liu and Passaglia, 2009). Most notably, however, the horseshoe crab has played an important role in the medical industry, after the discovery by W.H. Howell in 1885 of the clotting ability of *Limulus* blood (Novitsky, 1984).

Figure 1-2: Limulus polyphemus. Credit: Mark Thiessen/National Geographic Creative

Further studies into this unique characteristic resulted in the discovery by Frederik Bang that an extract of their blood cells, *Limulus* amebocyte lysate (LAL), was sufficient to recapitulate blood clotting in the horseshoe crabs, and further that the cause of this clotting was due to exposure to Gram-negative bacteria (Bang, 1956). This feature is still exploited, with horseshoe crabs being harvested to obtain samples of their blood for use in the biomedical industry (Figure 1-2). Up to 400 ml of blood is taken from each crab per bleed, depending on their size (Armstrong and Conrad, 2008). Although blood regeneration of surviving crabs is relatively quick (3 – 7 days), it can take up to 4 months for an individual's amebocyte levels to be restored. Initial studies reported mortality levels as a result of this bleeding to be around 10% (Rudloe, 1983). However, this figure is believed to have risen to around 30% (Leschen and Correia, 2010).

Attempts to breed *Limulus polyphemus* in captivity have proven difficult. This may be due in part to the fact that it can take them 9 - 11 years to reach sexual maturity, together with their moulting habits, which see them shedding their exoskeleton up to 16 or 17 times as they grow, each time becoming increasingly physically demanding and time consuming. They also appear to have a very specific mating ritual dependent on tidal and lunar patterns, with a need for particular temperatures, salinity levels and substrate type (Brockmann, 1990; Okun, 2012). As a result of this, along with the increasing negative impact on the species, habitat destruction, extensive fishing and pollution, the supply of blood from which to make LAL has come under threat and as such there is an urgent need to guarantee there is a sustainable supply of horseshoe crab amebocyte, or develop new assays to make use of the highly-sensitive endotoxin-sensing feature of the LAL test.

The innate immune systems of vertebrates and invertebrates consist of multiple mechanisms by which Gram-negative bacterial infections, amongst others, can be recognised and eradicated from the whole organism. Three general qualities of these defence system are: the recognition of pathogen-associated molecular patterns (PAMPs) by pattern recognition receptors (PRRs); transcriptional activation by intracellular signalling cascades of primary response genes encoding, for example, cytokines, chemokines and effector molecules; and destruction of the infecting bacteria by phagocytosis and by antimicrobial peptides (Koshiba *et al.*, 2007; Stuart and Ezekowitz, 2005). Lacking an adaptive immune system, invertebrates rely solely on the innate

immune system for protection against pathogens and their incredibly efficient blood-clotting systems are a central factor of this innate immunity.

1.2.1 Amebocytes

Horseshoe crabs are harvested to obtain samples of their blood (hemolymph), which contains hemocyanin (a copper containing protein) and an abundance of one cell type, the amebocyte. The multifunctional hemocyanin carries molecular oxygen and its presence is evident by the blue colour of the serum. Upon exposure to bacterial endotoxin, the horseshoe crab's innate immunity serves to eradicate the infection by activating a clotting process as a result of interaction with a component in the blood. Experiments conducted by Bang determined hemocyanin did not play a role in the formation of the gel as a result of endotoxin contamination, and thus prompted further investigation into the role of the amebocyte (Bang, 1956).

The amebocyte is a granular cell found in the hemolymph of horseshoe crabs and makes up over 99% of blood cells in this species. These circulating cells contain the *Limulus* coagulation system, which plays a key role in the crab's defence against endotoxin infection. In 1968, Levin and Bang discovered that cell free plasma is incoagulable, even in the presence of LPS, establishing that the amebocytes, as the only supply of clottable protein in the blood of *Limulus*, play an important role in LPS-mediated coagulation (Levin and Bang, 1968). It was found that non-cellular components do not have any influence on the process.

The secretory granules found within the amebocytes are exocytosed to release a number of proteins and proteases to work in the defence against the bacteria and upon exposure to endotoxin, the amebocyte degranulates. There are two types of granules: large (L) and small (S), that have been shown to contain a number of proteins and peptides that act as defence molecules against endotoxin infections (Toh *et al.*, 1991). These are outlined in Table 1-1 and include coagulation factors, protease inhibitors, antimicrobial substances and lectins (Iwanaga, 2002; Iwanaga *et al.*, 1998).

The process of coagulation involves aggregation of amebocytes followed by shrinkage of the aggregated mass, before degranulation and appearance of a liquid phase that undergoes gelation upon exposure to bacterial endotoxin (Levin and Bang, 1968). The speed of gelation and concentration of endotoxin present are related implying that an

enzymatic reaction takes place. Factor C has been identified as being located in the large granules of the amebocyte and after degranulation, in addition to its role in LPS-binding, may contribute to several defence mechanisms including cell adhesion as a means by which clots and foreign material are removed.

L-granules		**S-granules**	
Coagulation Factors	Factor C	Antimicrobial Substances	Tachyplesins
	Factor G		Polyphemusins
	Proclotting enzyme		Big defensin
	Coagulogen		Tachycitin
Protease Inhibitors	LICI-1		Tachystatins
	LICI-2		
	LICI-3		
	LEBP-PI		
	Limulus cystatin		
	α_2-Macroglobulin		
Antimicrobial substances	Anti-LPS factor		
	Big defensin		
	Factor D		
	Tachylectin-1		
	Tachylectin-2		
	Tachylectin-3		
	Limunectin		
	18K-LAF		
Others	8.6 kDa protein		
	Pro-rich protein		
	L1		
	L4		

Table 1-1: Secretory granule components.

The immobilisation of bacteria as a result of the disruption caused to amebocytes by bacterial endotoxin highlights the main way in which horseshoe crabs defend themselves against the constant exposure to gram-negative bacteria and the sensitivity of this method is emphasized by the fact that picograms of endotoxin is enough to elicit the gelation response.

1.2.2 Coagulation Cascade

Further investigation of *Limulus* coagulation was motivated by the resemblance of this reaction to a number of previously established processes including the Schwartzman reaction in rabbits, intravascular coagulation in humans as a result of gram negative sepsis and the resemblance between amebocytes and mammalian platelets that function in mammalian coagulation. In particular, the similarities between amebocytes and mammalian platelets appear to be of significance and include the fact they both contain granules, aggregation of the cells followed by loss of granules occurs during their coagulation processes, they both aggregate at the site of injury and they also both display bactericidal activity (Levin and Bang, 1968). A difference of note however, between the horseshoe crab clotting system and that of mammalian coagulation, is that positive feedback activation does not appear to take place in the horseshoe crab system.

The coagulation cascade comes about as a result of sequential enzymatic activation of serine protease zymogens after exposure of the amebocyte to bacterial endotoxins. Autocatalysis activates Factor C (Factor **Ċ**), which in turn activates Factor B (Factor **Ḃ**) and thus activates proclotting enzyme to give clotting enzyme, each by limited proteolysis. Clotting enzyme transforms coagulogen to coagulin gel resulting in a clot that acts to destroy the endotoxin infection. Tai *et al.* suggested the cleavage of a single peptide bond between an Arg and Lys gives rise to the production of this clot (Tai *et al.*, 1977). Figure 1-3 illustrates the coagulation cascade pathway. Due to the indisputable similarities, there are speculations surrounding the idea that the evolution of mammalian blood coagulation may have arisen from this primitive mechanism (Young *et al.*, 1972).

In 1981, Kakinuma *et al.* discovered that carboxymethylated (1→3)-β-D-glucan (CMPS), a polysaccharide with anti-tumour properties found on the surface of fungi, also activates the *limulus* coagulation system, similarly resulting in the formation of a clot (Figure 1-3) (Kakinuma *et al.*, 1981). It was determined that CMPS did not have any effect on the previously discovered factors involved in the endotoxin-mediated pathway (Factor C, Factor B and proclotting enzyme), indicating a separate component, sensitive to CMPS, must be present to activate the proclotting enzyme. This factor, termed Factor G, was found to differ significantly from the others in that it gave no response in the presence of endotoxin. This confirmed the notion that two distinct pathways exist as a result of

PAMPs; one mediated by endotoxin and the other by (1→3)-β-D-glucan, both of which catalyse the production of coagulin gel.

The coagulation cascade forms the basis of the *Limulus* amebocyte lysate (LAL) test, which was developed by Levin and Bang and is used for the detection of endotoxin (Levin and Bang, 1968).

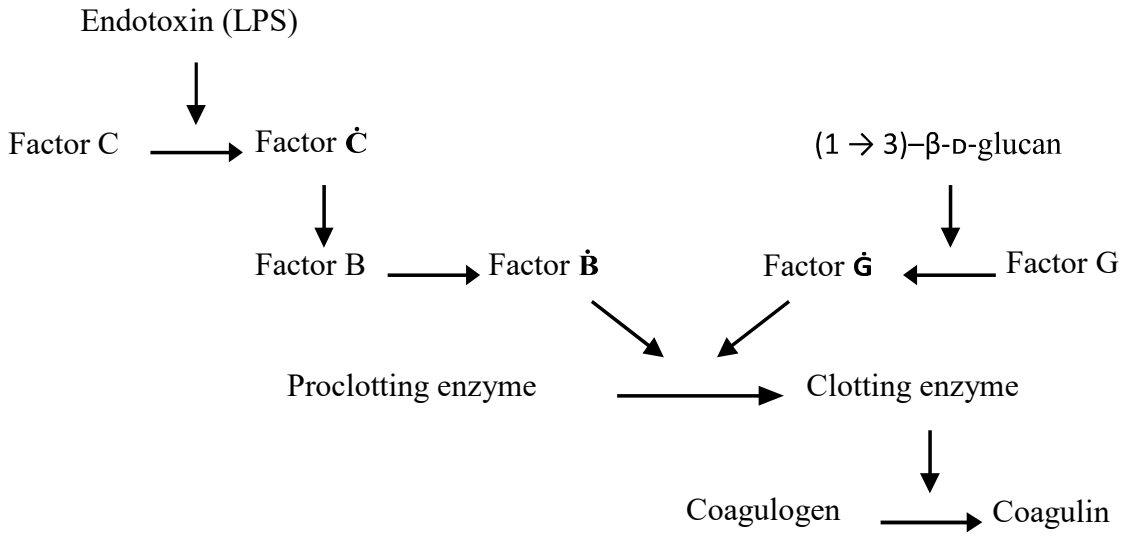

Figure 1-3: The coagulation cascade. Endotoxin and (1→3)-β-D-glucan-mediated pathways are shown to produce the same result of blood coagulation via activation of two different factors.

1.2.3 *Limulus* Amebocyte Lysate Test

Prior to the discovery of the LAL test, a number of methods existed for the detection of endotoxin that involved the use of whole or parts of living animals. The pyrogen test in rabbits was the most commonly used, where test solutions were injected into a rabbit's bloodstream and its temperature was monitored over a specific time frame. If a high temperature was detected, the solutions were deemed to be contaminated with endotoxin and the pharmaceutical devices were discarded (Novitsky, 1984). The discovery of the endotoxin-sensing protein from horseshoe crab blood allowed for the development of a new endotoxin detection test meaning the rabbit pyrogen test was no longer necessary. This test, termed *Limulus* amebocyte lysate (LAL), was discovered by Levin and Bang when they established that Gram-negative bacteria resulted in rapid gelation of the lysate (Levin and Bang, 1968).

The LAL test has been used for many years for the screening of pharmaceutical products including implants and vaccinations. It is simple and reliable and in high demand due to the lack of a more sensitive test. Horseshoe crabs are bled by cardiac puncture. Amebocytes are separated from hemocyanin by low speed centrifugation. Intact amebocytes are lysed by addition of pure distilled water and the clotting factors are found in the supernatant after high-speed centrifugation. This supernatant is the LAL, which can be activated by the addition of ions such as sodium, calcium or magnesium salts, that have been found to be necessary for the activation of proclotting enzyme (Tai and Liu, 1977). Sensitivity can be enhanced by performing solvent extraction. To perform the test, equal volumes of LAL and the test product are mixed in a glass test tube, and incubated at 37°C for one hour. The formation of a clot indicates endotoxin is present in the sample. There are a number of ways the test can be employed for the analysis of pharmaceutical products before their distribution, to detect trace amounts of LPS in solution (Novitsky, 1984).

Alternative testing methods to the LAL test have been explored to overcome the drawbacks of this process, which include the possibility of fungal activation, fluctuations in the sensitivity between batches of LAL and the detrimental effects the process has on the horseshoe crab population. Fluctuations in pH or temperature, or the presence of substances such as divalent cations or chelating agents, may interfere with the LAL reagent, decreasing its capability to respond to endotoxin, thus giving rise to false negative and positive results. A number of biomedical companies have attempted to produce alternative tests based on purified or recombinant Factor C, with varying results, in an effort to bring an end to the bleeding process of horseshoe crabs. Examples of these tests include the LAL-based detection kits from Lonza (PYROGENT™ Gel Clot Assays) and Charles River (Endosafe® Endotoxin Testing Systems) and the recombinant Factor C test from Hyglos (EndoZyme® recombinant Factor C (rFC) Assay). However, false-positives can arise from a substrate other than Factor C cleaving the chromogenic substrate and the tests are prone to false negative results due to sample conditions being inadequate for Factor C activation, such as a too high or too low pH. This supports the need for further investigation into the method behind LAL in order to develop highly sensitive and specific reagents for detection of endotoxin.

1.3 Factor C

The glycoprotein Factor C was discovered by Nakamura *et al.* after further studies into Factor B were performed in the hope that a better understanding of the horseshoe crab clotting process would be achieved (Nakamura *et al.*, 1985). They found that activation of Factor B required the action of another component (Factor C) to initiate the clotting cascade and therefore activate the proclotting enzyme. All three components were deemed necessary for coagulation. Factor C was originally purified from *Tachypleus tridentatus* in 1986 by Nakamura *et al.* and since then, research has been undertaken to elucidate its structure, function and interactions within the coagulation cascade (Nakamura *et al.*, 1986).

Factor C is a serine protease zymogen that works as a 'biosensor' in response to LPS. The pre-pro Factor C is synthesised as a single polypeptide, which has a molecular weight of 123 kDa, is subsequently cleaved to form a two-chain glycoprotein consisting of two disulphide-linked polypeptide chains: an 80 kDa heavy chain and a 43 kDa light chain (Ariki *et al.*, 2004; Muta *et al.*, 1991; Nakamura *et al.*, 1986). Investigations into the membrane anchored form believed to act as a sensor on the surface of amebocytes determined a 123 kDa Factor C antigen that corresponds to the single-chain zymogen (Ariki *et al.*, 2004). It was revealed that binding to LPS results in the autocatalytic activation of two-chain Factor C (Factor **Ċ**) resulting in cleavage of the light chain into two fragments of molecular weight 7.9 kDa and 34 kDa, designated the A chain and B chain, respectively. The resulting Factor **Ċ** is composed of the three disulphide-linked chains: the heavy chain, A chain and B chain. Tokunaga *et al.* deduced that hydrolysis of a Phe-Ile bond in the light chain sequence -Pro-Phe-Ile-Trp-Asn-Gly- activates the Factor C zymogen, giving rise to the A and B chains, and initiating the sequential activation of the clotting factors (Tokunaga *et al.*, 1987). It is thought that a conformational change takes place upon Factor C binding to LPS, which brings about this cleavage of the Phe-Ile bond and that this process is comparable to streptokinase-complexed plasminogen molecules' activation of plasminogen (Reddy and Markus, 1972).

The location of the LPS-binding region was first investigated by Nakamura *et al.* using Salkylation (Nakamura *et al.*, 1988a). The two-chain intermediate form of ttFC was reduced by addition of dithiothreitol (DTT, 0.25 mM), then incubated with 4-vinylpyridine to induce S-pyridylethylation (PE). After dialysis, the PE-FC was separated

by Reversed-Phase High Performance Liquid Chromatography (RP-HPLC), into the heavy and light chains. Incubation of Factor C with LPS and either the S-alkylated heavy chain or light chain determined whether inhibition of activation took place. The heavy chain, when S-alkylated, inhibited Factor C activation by LPS, however, the light chain when S-alkylated failed to inhibit the activation, indicating that the amino-terminal heavy chain of Factor C is likely to have the site of the LPS-binding. Consequently, it can be presumed that LPS recognition is affiliated with the heavy chain and catalytic activity can be attributed to the light chain.

Nakamura *et al.* showed that the B chain of Factor C is a serine protease domain, while Tokunaga *et al.* determined the A chain had sequence similarity with the family of tandem repeats present in a number of factors, in particular, human complement factor B (Mole *et al.*, 1984; Nakamura *et al.*, 1988a; Nakamura *et al.*, 1988b; Tokunaga *et al.*, 1987). However, it was Muta *et al.* that determined the mosaic structure of Factor C domains and revealed its similarities to components of the mammalian complement cascade (Muta *et al.*, 1991). They performed a search for proteins with sequence similarity to Factor C and identified that alongside the serine protease domain, Factor C comprises: five complement control proteins (CCPs) also known as 'sushi' domains or short consensus repeat domains – four in the H chain and one in the A chain; an epidermal growth factor (EGF)-like domain in the H chain; a lectin-like domain, which appears between CCP3 and CCP4; a Cys-rich domain found at the H chain's amino terminus; a proline-rich region located at the carboxy-end of the H chain; and an LCCL module. A schematic representation of the domain organisation of Factor C can be seen in Figure 1-4.

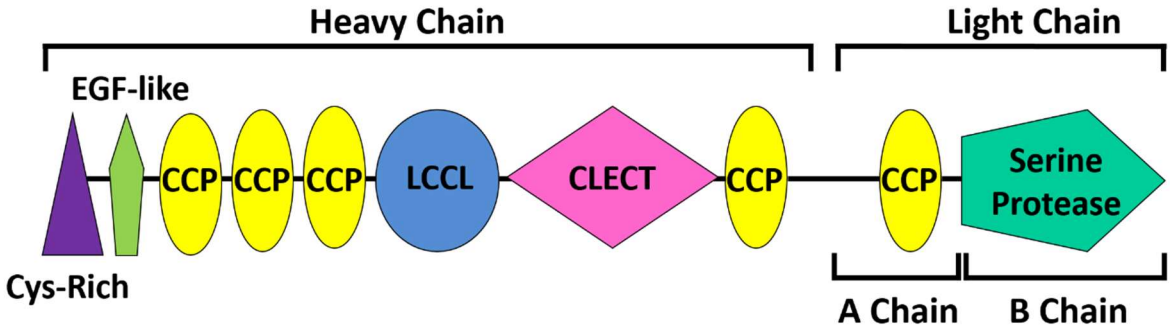

Figure 1-4: Schematic representation of the domain structure of Factor C. The positions of the heavy and light chains are indicated, along with the A and B chains. CCP = Complement Control Protein. LCCL = Limulus Factor C, Coch-5b2, LgII. CLECT = C-Type Lectin Domain.

Complement control proteins have been identified over 140 times in more than 20 human proteins that are most often associated with the mammalian complement system. Factor C was the first protein identified in invertebrates shown to exhibit these domains. They are said to have a 'sushi'-like disulphide folding structure, which was discovered by Lozier *et al.* when studying the β2-glycoprotein 1 (Lozier *et al.*, 1984). The domain has a compact hydrophobic core enclosed within six beta strands with four conserved cysteine residues forming disulphide bonds in the pattern Cys1 – Cys3 and Cys2 – Cys4. Four of the β-strands appear to have a variable orientation, while β-strand 2 and β-strand 4 have a more conserved structural orientation. The variability tends to occur primarily at the interface between adjacent domains and Norman *et al.* speculated that interactions occur between these (Norman *et al.*, 1991). CCPs have a number of additional highly conserved residues including a tryptophan and other mainly hydrophobic residues. Their role in other proteins is to facilitate protein-protein interactions, which suggests that the CCPs of Factor C could interact with other proteins during the clotting cascade, although it also likely that some CCPs play an entirely structural role (Day *et al.*, 1987; Gaboriaud *et al.*, 2000; Norman *et al.*, 1991).

The EGF-like domain comprises a sequence of around 30 – 40 amino acids from the epidermal growth factor (EGF). It is commonly found in proteins such as coagulation factors VII, IX, X and XII, decoagulants such as tissue plasminogen activator (tPA), and complement factors. It has a total of six cysteine residues that are shown to be involved in disulphide bond formation following the pattern Cys1 – Cys3, Cys2 – Cys4 and Cys5 – Cys6 (Savage *et al.*, 1973). These disulphide bonds force the structure into three loops, with the main structure forming two antiparallel β-strands that occur between cysteine 3 and 4 (Appella *et al.*, 1988). Most EGF-like domains appear to have a conserved proline before Cys1 and a proline or a glycine before Cys2. EGF-like domains found in other proteins often play a role in ligand interactions.

The Cys-rich region encompasses 11 out of a total of 56 cysteine residues present in the Factor C sequence. It is unknown whether this is a single domain, as typically, Cys-rich repeats are formed of around 40 amino acids with four or six conserved cysteine residues in each. Disulphide bonds are known to be conserved between Cys1 – Cys2, Cys3 – Cys5 and Cys4 – Cys6, however, the Cys3 – Cys5 bond is at times missing, due to the absence of one of the Cys residues (Bazan, 1993; Mallett and Barclay, 1991). This could explain

the odd number of Cys residues found in the Factor C Cys-rich region as it is unknown whether the unpaired Cys residue forms an inter-chain disulphide bond. Folded single Cys repeats are thought to form loops defined by the Cys1 – Cys2 and Cys4 – Cys6 disulphide bonds, and a conserved Tyr or Phe residue enables correct packing of the loops.

The identification of a lectin-like domain was surprising in that Factor C was the first protease where a domain of this type was discovered. The domain appears to be a member of the C-type lectin superfamily, that often function as calcium-dependent carbohydrate binding molecules (Drickamer, 1988). However, lectins are in fact diverse in terms of structure and function, with a number of C-type lectins being identified that fulfil different roles, including protein interactions with other proteins, lipids or nucleic acids (Kilpatrick, 2002). The pattern of disulphide bond formation appears to match that of acorn barnacle lectins, where Cys1 – Cys2, Cys3 – Cys6 and Cys4 – Cys5 form disulphide linkages (Muramoto and Kamiya, 1986).

The proline-rich region is a 48-residue domain where 13 of these residues are proline. This fragment shows homology with a region found in mammalian coagulation Factor XII but is of unknown significance in Factor C (Iwanaga *et al.*, 1992; McMullen and Fujikawa, 1985).

Another domain present in the Factor C protein is the LCCL module, named after the proteins found to contain it: *Limulus* Factor C, vertebrate cochlear protein cochlin (Coch5b2) and Lgl1 (Trexler *et al.*, 2000). As shown for other domains in the protein, this domain is disulphide linked with bonds formed between the most highly conserved cysteine residues Cys1 – Cys4 and Cys5 – Cys6. It is also believed that, when present, disulphide bonds form between Cys2 – Cys3 and Cys7 – Cys8. Secondary structure prediction indicated the presence of two α-helices and six β-strands. The LCCL module is thought to be an autonomous folding domain that plays a role in construction of modular proteins. However, a specific function has not been elucidated for this domain (Robertson *et al.*, 1998; Trexler *et al.*, 2000).

The B chain of Factor C is made up of a serine protease containing a His-Asp-Ser catalytic triad (Blow, 1997). There are over 20 families of serine proteases and Factor C is thought to be a novel serine protease zymogen receptive to α-chymotrypsin (Iwanaga *et al.*, 1992; Rawlings and Barrett, 1994). A conserved Asp is the known site of substrate

binding but zymogen activation is a result of cleavage between a Phe-Ile bond (Tokunaga *et al.*, 1987). Six out of seven cysteine residues form disulphide bonds: Cys1 – Cys2, Cys4 – Cys5 and Cys6 – Cys7. The 7th, Cys3, is thought to link with a Cys residue in chain A. The main structure of the serine protease is made up of two six-stranded β-barrels, with the active site found between the two barrels (Hedstrom, 2002). The sequential activation of the serine protease zymogens results in blood coagulation.

The domain composition and organisation of Factor C is particularly similar to three proteins known as 'selectins' that also play a role in defence systems (Muta *et al.*, 1991). These proteins contain an EGF-like domain, a lectin-like domain and a number of CCP domains and are known as the endothelial leukocyte adhesion molecule (Bevilacqua *et al.*, 1989), the lymph node homing receptor (Siegelman *et al.*, 1989) and the granule membrane protein 140 (Johnston *et al.*, 1989).

Tokunaga *et al.* endeavoured to establish whether there are any immunological or biochemical differences between the *Limulus polyphemus* Factor C and the *Tachypleus tridentatus* Factor C (Tokunaga *et al.*, 1991). Comparisons revealed the two proteins were more or less identical and as purified ttFC was more readily available, it was used for further studies.

Ding *et al.* carried out analysis on Factor C from the *Carcinoscorpius rotundicauda* species of horseshoe crab (crFC) (Ding *et al.*, 1993). crFC was found to be larger than ttFC at 132 kDa and reduction of the chain resulted in two chains of 80 kDa and 52 kDa, representing the heavy and light chains, respectively. Upon activation by LPS, the light chain was shown to undergo similar cleavage to that apparent in ttFC upon exposure to endotoxin.

1.4 Experimental Basis

Characterisation of the key proteins involved in binding endotoxin and a better understanding of the molecular details surrounding LPS is necessary to develop new, highly sensitive biosensors to benefit pharmaceutical and medical device screening as the increase in demand for horseshoe crab blood may lead to an unsustainable burden on the wild population of horseshoe crabs. These new biosensors could be based on direct optical detection rather than relying on the enzymatic activity of the current LAL test. This could

be achieved by incorporating a FRET donor/acceptor pair whose separation is modulated by LPS binding to a Factor C derived conformational switch.

Studies to understand the reason Factor C is so sensitive to minute quantities of LPS have been hindered by the difficulties faced with the production of biologically active recombinant Factor C, along with the fact LPS molecules separate from Factor C after activation. The recognition of the lipid A component of LPS by Factor C is very specific, and details surrounding this type of interaction are lacking. However, attempts have been made to determine the nature and exact site of LPS-binding to Factor C, which inspired the experimental approach taken in this project.

Contradictory results as to the location of the LPS-binding site have been published by Tan *et al.* and Koshiba *et al.* after both groups carried out experiments on different fragments of Factor C (Koshiba *et al.*, 2007; Tan *et al.*, 2000). Tan *et al.* led investigations using the first three CCP domains located between the EGF-like and lectin-like domains. They expressed three fragments from Factor C of the *Carcinoscorpius rotundicauda* species as fusion proteins with enhanced green fluorescent protein (EGFP), using a novel secretory signal. The three fragments expressed were: Sushi123, Sushi1 and Sushi3. Two Factor C derived peptides comprising 34 amino acid regions of the CCPs were also synthesised and designated S1 and S3, alongside another two peptides containing two lysine mutations each: SΔ1 and SΔ3. These peptides each represented approximately half of a CCP module.

The fusion proteins were expressed in Drosophila S2 cells with the selection vector pCoHygro, which was introduced to the cells by a calcium phosphate co-precipitation method (Sambrook *et al.*, 1989). A Western blot was performed to confirm the presence of the fusion proteins, which were purified by anion exchange chromatography. Surface plasmon resonance (SPR) analysis was carried out to elicit protein/peptide interactions with lipid A. The binding constant of polymixin was established for use as a positive control.

Results from the fusion proteins indicated there are multiple binding sites present within the first three CCP modules of Factor C. The peptides highlighted the disadvantages of producing truncated CCP domains and thus a lack of disulphide bonds, as a decreased

affinity for LPS was shown for S1 and S3 (10,000-fold and 100-fold, respectively). Despite this, further experiments were performed using the four peptides.

Circular dichroism (CD) was used to estimate secondary structure contributions from each of the peptides. This revealed that all four peptides largely displayed a random coil structure. Upon addition of LPS, the structures of S1 and SΔ3 became largely α-helical, SΔ1 displayed mostly turns and SΔ3 displayed a high percentage of beta contributions. These findings, for the most part, contradict the previously observed structures of CCP modules, that mostly display β-strands (Norman *et al.*, 1991).

A chromogenic assay was performed on the four peptides. This assay involved using the first steps of the endotoxin reaction for activation of a synthetic substrate enzyme, which ultimately emitted a yellow colour that was monitored by absorbance at 405 nm. The potency of each individual peptide was determined in terms of its 50% endotoxin neutralising concentration (ENC_{50}). A sigmoidal curve was observed for S1, suggesting this peptide binds to LPS and displays positive cooperativity, in comparison to the other three peptides that appear to bind independently. The introduction of lysine mutations into the peptides was based on computational analysis that suggested LPS binding would improve with the presence, at specific sites, of more lysine residues. However, the SΔ1 peptide only displayed a 10-fold increase in LPS binding affinity, and there was no difference evident with SΔ3.

Tan *et al.* suggested that endotoxin and Factor C interact in a "two-prong amplification activation pathway", in which binding of a single LPS molecule to the "LPS-capturing" domain – sushi3 – would encourage positive cooperative binding of the "LPS epitope presenting" domain – sushi1, but also that binding to either one of the two domains would still activate Factor C (Tan *et al.*, 2000).

In conclusion, Tan *et al.* suggested the LPS-sensitivity of Factor C was due to multiple LPS binding sites and the high positive cooperativity of LPS binding. However, flaws of their experimental approach include the neglect of the N-terminal region that includes the EGF-like domain and the Cys-rich region, which could potentially be structurally important. Additionally, the use of truncated peptide sequences did not give a proper representation of the known domain structure of CCPs, and thus it does not seem plausible that accurate results can be drawn from use of these peptides. Experiments were also

lacking negative controls, which should have been used to confirm that it was specifically LPS binding to these domains, and not a different substrate, thus giving false positives.

Koshiba *et al.* expressed Factor C fragments from the *Tachypleus tridentatus* species of horseshoe crab for their investigations into the N-terminal region of Factor C (Koshiba *et al.*, 2007). Six constructs were sub-cloned into the pSecTag2A vector and expressed in HEK293 cells. The constructs each had three tandem copies of the Myc-tag inserted and they represented: the heavy chain (residues 1 – 723); CysEGFCCP123 (residues 1 – 296); LCCL/lectin (residues 300 – 543); CCP4 (residues 551 – 609); CCP5 (residues 669 – 723); and the serine protease domain (residues 738 – 994). Myc-tagged anti-LPS factor (ALF), a known LPS recognising protein, was expressed for use as a positive control (Aketagawa *et al.*, 1986). Purification was achieved by use of agarose beads coupled to the c-Myc polyclonal antibody and separation by SDS-PAGE.

LPS binding was determined by immunoprecipitation experiments and flow cytometric analysis. Results showed binding of the heavy chain fragment and the CysEGFCCP123 construct to LPS. On the other hand, the other four constructs (LCCL/lectin, CCP4, CCP5 and the serine protease) failed to bind LPS. Structure function analysis was performed on a further two constructs that were produced in order to pinpoint a more exact location of LPS binding. These constructs were CysEGF (residues 1 – 116) and CCP123 (residues 117 – 296). Binding tests revealed only the CysEGF construct formed a complex with LPS. The specificity of binding was highlighted by the fact cholesterol and acidic phospholipids could not disrupt the complex. By using a negative control of Gram-positive bacteria, they were able to further confirm Factor C specificity to LPS as no binding took place in the presence of this bacteria.

A feature of the Cys-rich region, which has been identified in a number of LPS-recognising proteins, is a conserved tripeptide motif, consisting of an aromatic residue with a basic residue either side. Two copies of this motif were identified in the Cys-rich region of Factor C: Arg^{36}-Trp^{37}-Arg^{38} and Lys^{55}-Tyr^{56}-Lys^{57}. When the Arg-Trp-Arg motif was mutated to Glu-Trp-Glu, the construct failed to bind to LPS. Furthermore, a substitution of the Trp residue with an Ala also left the construct incapable of binding to LPS. A mutation of Lys-Tyr-Lys had no effect on binding capabilities, therefore establishing that the conserved Arg-Trp-Arg motif plays an important role in LPS-binding.

In conclusion, they proposed the first conserved tripeptide motif found at the N-terminal of Factor C plays a key role in LPS recognition, in the same way that predicted binding of ALF to LPS occurs, where the basic residues are involved with Lipid A's glucosamine (GlcN) I1-phosphate and the aromatic residue associates with a hydrophobic portion (Pristovšek *et al.*, 2005).

Even though Koshiba *et al.* appeared to provide a more specific, detailed, alternative model for LPS binding and thus activation of Factor C, they failed to adequately describe the experiments performed in their structure-function analysis of LPS binding, thus questioning the credibility of the results. In both cases, protein constructs were not checked for correct folding, which could have a significant impact on the outcome of certain, if not all, experiments. Folding is especially important in heavily disulphide bonded proteins, as incorrectly bonded cysteines can result in mis-folding and aggregation, thus leading to improper results. In the experiments carried out by Koshiba *et al.*, they failed to elucidate the extent of the role played, if any, by the second domain in the construct, the EGF-like domain. This domain could be important for proper folding or formation of disulphide bonds, thus playing a key role in LPS binding. No indication was given that these experiments were repeated in any way, casting doubt on the reliability of the results being presented.

Given the conflicts associated with the previous research and the questions that remain unanswered, further characterisation of LPS binding to Factor C is of paramount importance for the development of new endotoxin testing strategies.

1.5 Project Aims

The aims of this project were to identify and understand how Factor C binds LPS; to characterise the conformational changes induced by binding LPS at the molecular level; and to identify why Factor C binding to LPS is so specific to lipid A. To achieve these objectives, full-length synthetic *Limulus polyphemus* Factor C was to be produced in order to understand the overall conformational change that takes place upon LPS binding and fragments of Factor C were produced to reveal the precise location of LPS binding, through structural analysis.

Despite the fact that other Factor C based tests use protein from the *Tachypleus tridentatus* and *Carcinoscorpius rotundicauda* species, the commercial LAL test is made from *Limulus* Factor C. As there are known variances between different LPS binding sites, it is beneficial to study lpFC to identify any differences that may arise between the species in terms of sensitivity or specificity of binding. Additionally, Marine Biotech Limited, the industrial sponsor for this project, use *Limulus polyphemus* for their investigations, and so experiments were carried out using lpFC to coincide with their studies.

As the heavy chain, and thus the N-terminal region of Factor C has been found to be the putative binding site of LPS, the first three CCP domains as well as the EGF-like and Cys-rich fragments were investigated as potential sites of LPS binding. Thus, experimentation endeavoured to express and purify the CCPs as single domains and as multi-domain fragments, along with the Cys-rich domain and the EGF-like domain on their own and together.

1.6 Experimental Approach

The sequence of *Limulus polyphemus* Factor C is not publicly available. For this reason, it was necessary to assemble the lpFC gene from the raw data kindly provided from Washington University, St Louis, in order to synthesise full-length lpFC. Three different eukaryotic expressions systems were chosen to test synthetic production of the full-length protein, to establish which was the most suitable to produce lpFC to fulfil the aim of carrying out a combination of biophysical techniques including circular dichroism, analytical ultracentrifugation and small angle X-ray scattering, to elicit its ligand-free and LPS-bound conformations in solution.

In order to determine the exact location of LPS binding, fragments of the Factor C protein were expressed recombinantly in *E. coli* to allow for their ligand-free and LPS-bound conformations to be determined by high resolution techniques such as nuclear magnetic resonance spectroscopy (NMR) or X-ray crystallography. Lipid binding analyses were performed using circular dichroism and NMR spectroscopy.

NMR is a powerful analytical spectroscopic technique that is used for the study of molecules in solution to obtain structural and dynamic information. Active nuclei are observed as a result of nuclear magnetic spin and transfer of magnetisation allows for the

detection of signals. From this, interactions between nuclei that are near each other can be detected and assignment of the resonances can be made, to piece together the overall structure of the protein under investigation.

2 *Limulus polyphemus* Gene

2.1 Overview

In order to identify and understand how Factor C binds lipopolysaccharide (LPS) and to characterise the conformational changes induced by binding LPS at a molecular level, it was expected that it would be necessary to study the whole, intact protein. This could give insights into why Factor C binding to endotoxin is so specific to lipid A and how Factor C binding to endotoxin changes the structure of the protein, thereby activating it. Ideally, purified native Factor C from amebocytes would be studied in order to determine the ligand free and LPS-bound conformations in solution to see the structural change that occurs from the inactive to the active forms. However, the Factor C protein is not at all abundant in the amebocyte and large quantities of *Limulus* blood would be required in order to purify enough for structural analysis. Additionally, blood and amebocyte lysate is a difficult source to deal with as it clots so easily, therefore additional measures to prevent these risks are required. Eukaryotic expression systems were selected to test synthetic production of the intact protein; yeast (*Pichia pastoris*), insect (*Baculovirus* system) and mammalian (Flp-In™ TREx™ system). Each system comes with its own advantages and limitations and so to determine the most suitable one for expression of lpFC, it was intended to test all three (see Chapter 4).

2.2 Choice of Factor C Sequence

The most widely used amebocyte lysate tests are produced from *Limulus polyphemus* (lp) blood, while the commercially available recombinant Factor C tests use proteins based on the previously determined *Tachypleus tridentatus* (tt) and *Carcinoscorpius rotundicauda* (cr) sequences. Given published differences in LPS binding sites, in order to assess whether differences in sensitivity of current synthetic Factor C based endotoxin reagents exist due to variations in the sequences, it was considered desirable to obtain the *Limulus polyphemus* Factor C sequence. This would allow for a direct comparison of the Factor C genes between the three species.

2.3 Assembly of the *Limulus polyphemus* Factor C Coding Sequence

Neither the amino acid, nor the genomic DNA or cDNA sequences for Factor C from *Limulus polyphemus* were available in any public databases. However, the McDonnell

Genome Institute at Washington University, St Louis (WUSTL) granted pre-release access to their *Limulus polyphemus* genome data (project funded by the National Human Genome Research Institute (NHGRI)). This was mined in an attempt to assemble the *Limulus polyphemus* Factor C coding sequence and gene.

2.3.1 Identification of *Limulus polyphemus* Genomic Sequences Encoding Factor C

Data was received as 'contigs', which are contiguous (or adjacent) sequences of DNA assembled by overlapping fragments of sequence reads from a chromosome, and 'supercontigs', longer stretches of sequence, assembled from correctly oriented contigs joined together systematically to build up the genomes, but separated by gaps. The genomic sequence was built using Roche 454 sequencing technology with 15X coverage, where each nucleotide was 'read' 15 times during the process (Margulies *et al.*, 2005). This technique uses pyrosequencing technology (sequencing by synthesis), and produces several hundred thousand reads up to 1 Kb in length (Mardis, 2008). The McDonnell Genome Institute, WUSTL, also proposed to generate a Bacterial Artificial Chromosome (BAC) library and to use 6X BAC end sequencing, where DNA is cloned into a bacterial cell where it is stable and can be manipulated easily in order to construct the DNA library (Shizuya *et al.*, 1992).

Eukaryotic genes consist of coding regions known as exons that are disrupted by parts of genes that do not code directly for the protein, known as introns. Pre-mRNA is formed during transcription and comprises both exons and introns, however, the introns must be removed to create the mature mRNA. This is achieved by RNA splicing, where small nuclear ribonucleoproteins (snRNPs) attach to the 3' and 5' ends of introns to form a loop. After removal, the exons can then join together. It is important that introns are removed accurately, as remaining intron nucleotides or removal of exon nucleotides can result in a frameshift of the genetic code (Clancy, 2008).

In order to identify which of the numerous contigs and assembled supercontigs were required to build up the full Factor C gene sequence, searchable databases were produced in order to identify regions of similarity between sequences. These were achieved by processing the sequence data in FASTA format using CLC genomics workbench software version 6.0.1 to generate two Basic Local Alignment Search Tool (BLAST) databases.

The supercontigs database comprised 683 sequences ranging from 374,669 bp to 5,251,686 bp and totalling 572,829,000 bp. In comparison, the contigs database consisted of 526,922 sequences ranging from 200 bp to 133,477 bp and totalling 1,685,998,000 bp. The *Limulus polyphemus* genome is estimated to be 270 Mb. However, the information produced from this did not allow for the supercontigs or contigs to be used simply to elucidate the sequence.

Consequently, a tblastn search was carried out using CLC genomics, where the translated nucleotide databases were searched using a known protein query; pre-*Tachypleus tridentatus* factor C (pre-ttFC). This was achieved using the BLAST program tblastn: Protein sequence and translated DNA database, and the following parameters: Standard database genetic code; Matrix: BLOSUM62; Gap cost: Existence 11, Extension 1; Max number of hit sequences: 250. Firstly, the supercontig database was searched against the pre-ttFC amino acid sequence. 139 sequences provided a hit but all had a percentage similarity of less than 30% (Figure 2-1). The likely reason for this lack of hits from the supercontigs database is probably due to not enough supercontigs being pieced together correctly. If this search had identified the Factor C gene, the hits would have been in sequence at adjacent sites on one supercontig representing the exons of the gene and interspersed with intron sequences. The fact that they were scattered across supercontigs and appeared significantly overlapped implies they were unlikely to represent Factor C.

Since no gene candidate could be clearly identified in the supercontig database, the next step was to perform a search of the contig database against the pre-ttFC amino acid sequence. This resulted in 485 hits, with contig size ranging from 51 bp up to 939 bp with varying degrees of similarity, up to 100% similarity in some regions (Figure 2-2). Including overlapping segments, there was a total coverage of 97,134 bp, over 30X the length of the Factor C cDNA. Contigs that were correctly assembled to the ttFC sequence should correspond to fragments of the lp genome and those that were long enough to overlap could be used to find the exons and assemble the entire Factor C gene. The hits were examined manually to select the most closely matching contigs for each region of the protein.

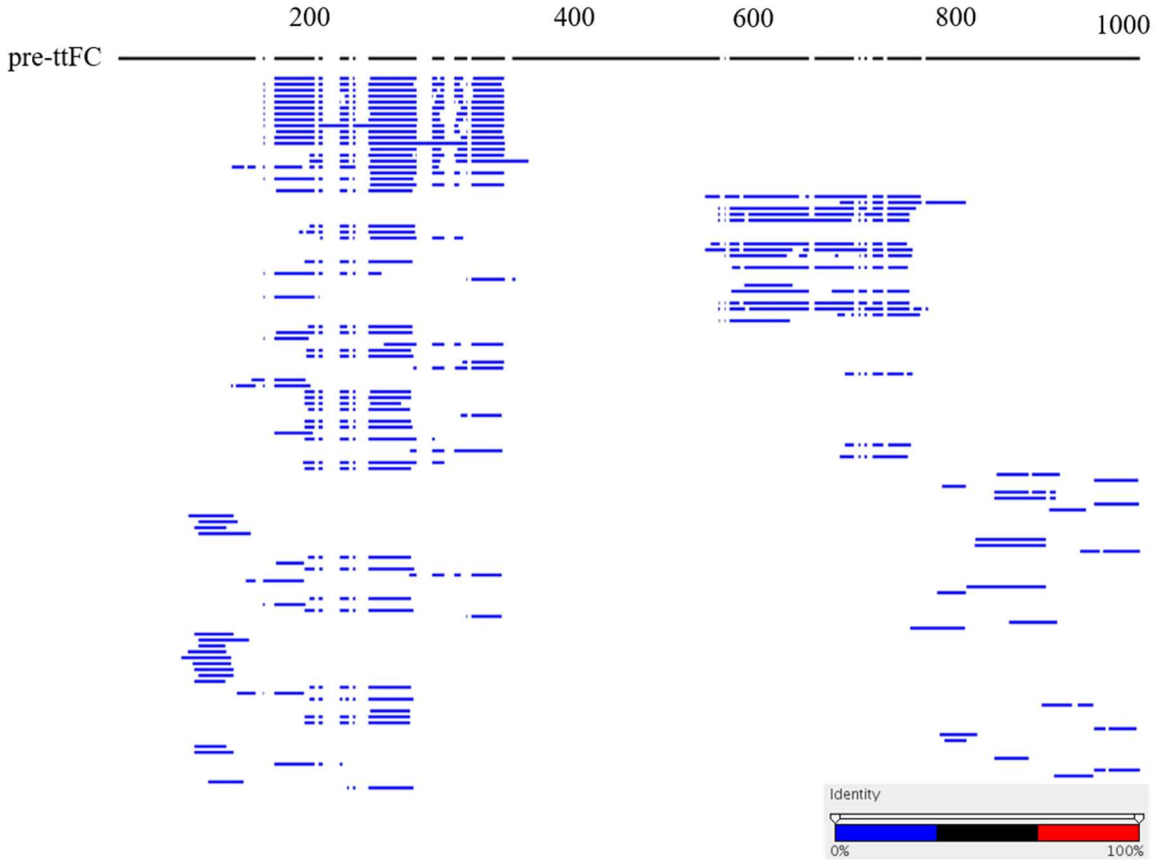

Figure 2-1: Supercontigs BLAST hits against pre-ttFC sequence. A small subset of supercontig hits are represented by each bar and are colour coded according to the quality of the hit. The pre-ttFC sequence is shown by the black bar across the top. Multiple hits for the same region of ttFC were often found within the same supercontig and hits at low similarity scores were found between ttFC and many different supercontigs. A full list of the supercontigs unique identifiers are outlined in Appendix A.

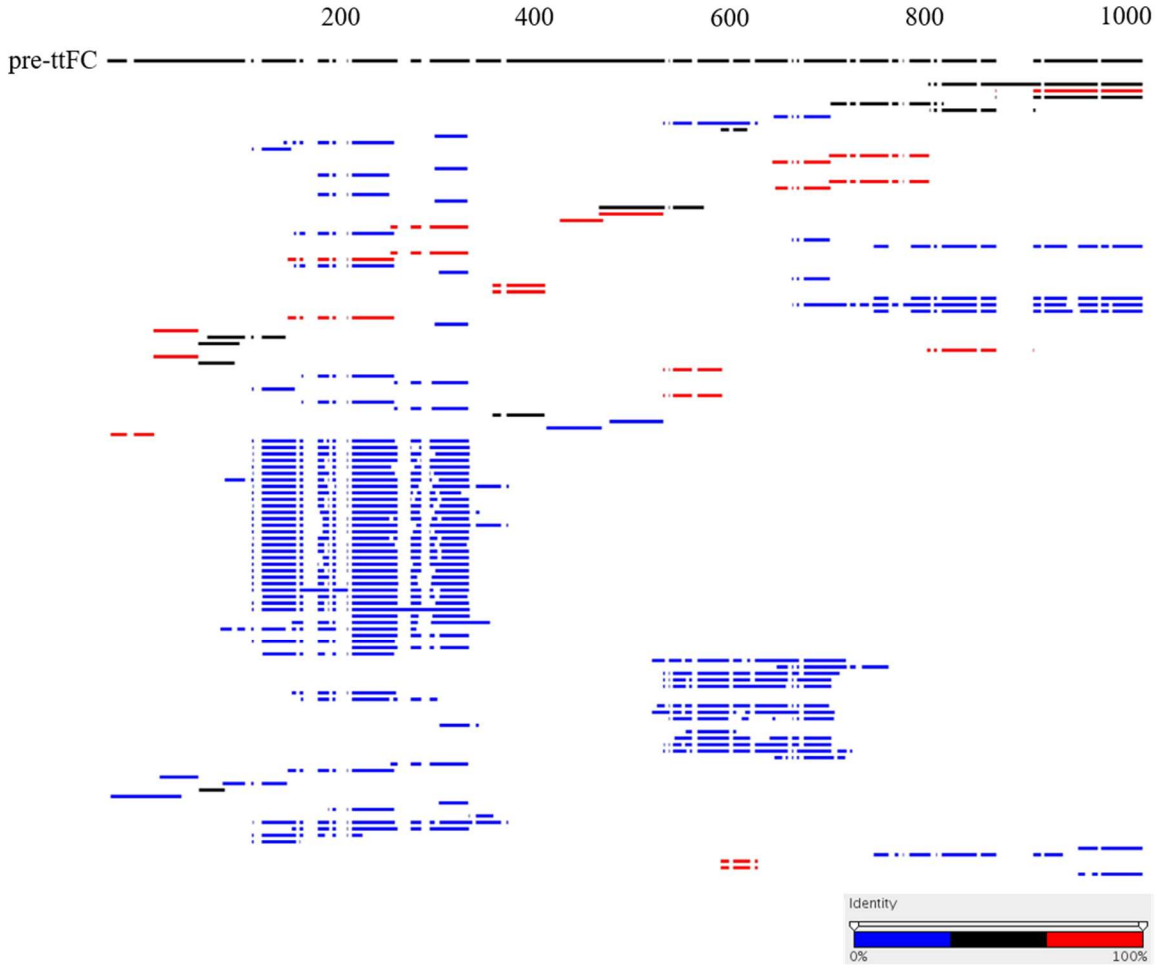

Figure 2-2: Contigs BLAST hits against pre-ttFC sequence. The black bar along the top represents the pre-ttFC sequence, with contig hits shown as bars of varying colour depending on sequence similarity. Each line represents a hit within a specific contig and in this case, there are a number of hits of more than 60% sequence identity. There are often multiple hits to different regions of Factor C in any one contig meaning a contig may appear in more than one line. A full list of the contigs unique identifiers are outlined in Appendix A.

For the assembly of the lpFC cDNA, pairs of the selected contigs encoding adjacent regions of the protein coding sequence presumed to represent exons were compared in an attempt to identify common sequences that would allow the intervening intron sequences to be identified. However, this approach was unsuccessful in assembling the gene.

As a result of failing to produce a sensible construct the contigs identified in the tblastn search were assembled using the ttFC cDNA sequence as a guide in the CLC "assemble sequences to template" tool. The resulting assembly was examined and the regions of each contig contributing to the consensus coding sequence were adjusted by hand to likely splice sites following the consensus splicing rules where the intron sequence typically starts before the GT at the 5' end and after the AG at the 3' end (Figure 2-3).

As highlighted in Figure 2-3 there are a few cases where the contigs only go part way through a module. However, this was to be expected as the right arrangement of the contigs was dependent on the correctness of the assembly, which was likely to contain errors, but also, some modules have internal introns. Furthermore, in some instances, more than one contig covered a domain, which could be due to differences between chromosomes.

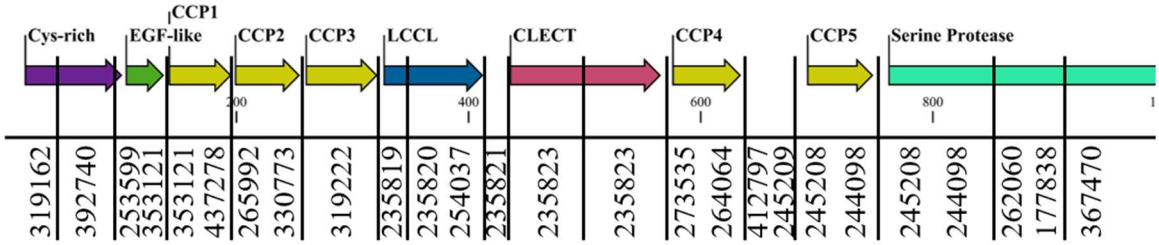

Figure 2-3: Selected contigs for sequence determination. The contigs are shown for each domain of the protein. The vertical black lines give an indication of the contig coverage.

As shown in Figure 2-4, the contigs show a good coverage of the pre-ttFC sequence. However, as to be expected between different species, there are several base pair mismatches. In one instance, two contigs were found with the same sequence that matched the reference sequence and around 20bp either side of the corresponding fragment before diverging. *Limulus polyphemus* is diploid, with a count of 52 chromosomes, and so both chromosomes may have been showing, however, it is more likely that the contigs were assembled wrong thus giving inconsistent results. At a gap in the sequence, the sequence was manually altered to include an extra codon, in line with the ttFC sequence. Additionally, the very N-terminal did not produce any matches and so a targeted search was carried out in order to find this part of the cDNA sequence. The differences that arose could be put down to the sequence data not being good enough, the data being incorrectly assembled or that big differences exist between the species.

After identifying the contigs from the *Limulus* raw data that best matched the ttFC sequence, the sequences were combined in the correct order to obtain the lpFC cDNA sequence. This was then translated using the CLC genomics software to obtain the pre-lpFC protein sequence.

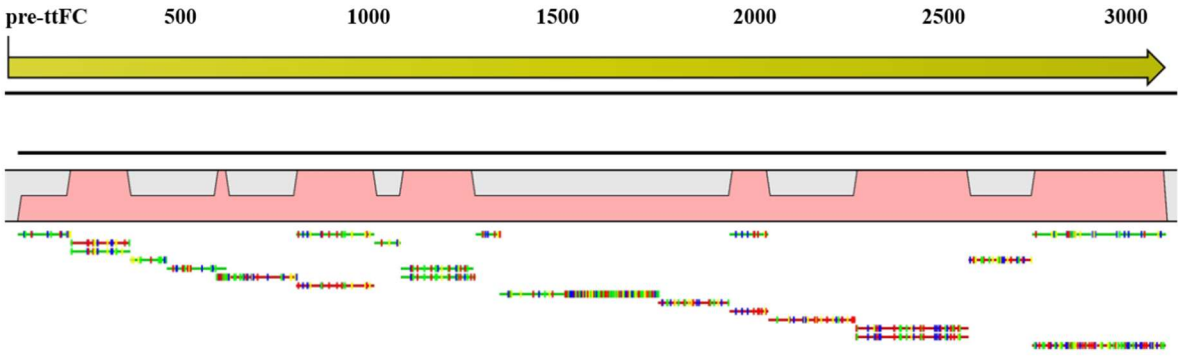

Figure 2-4: pre-ttFC guided lpFC contig assembly. The pre-ttFC open reading frame is shown by the yellow arrow. Coverage is indicated by the pink graph revealing at least one lpFC contig sequence matched the pre-ttFC sequence over the whole sequence. The different colours shown at each contig represent a different base pair mismatch on the lpFC contig (Red = A, Green = T, Blue = C and Yellow = G).

2.3.2 Factor C Sequence Alignment

Having obtained the predicted cDNA sequence for pre-lpFC, and translating this to protein, a protein sequence alignment was created to compare pre-lpFC, pre-ttFC and pre-crFC sequences to identify regions of similarity between the species (Figure 2-5). Identical amino acid sequences would suggest the domains are highly conserved between species and that use of the *Tachypleus* sequence would be sufficient for structural analysis of the protein fragments. It would also suggest a strong likelihood that the LPS binding site may be found in this region as it is thought to be conserved between species. If significant differences were seen between the three species, then the *Limulus* sequence would be required for obtaining accurate data with regards to the native *Limulus* protein.

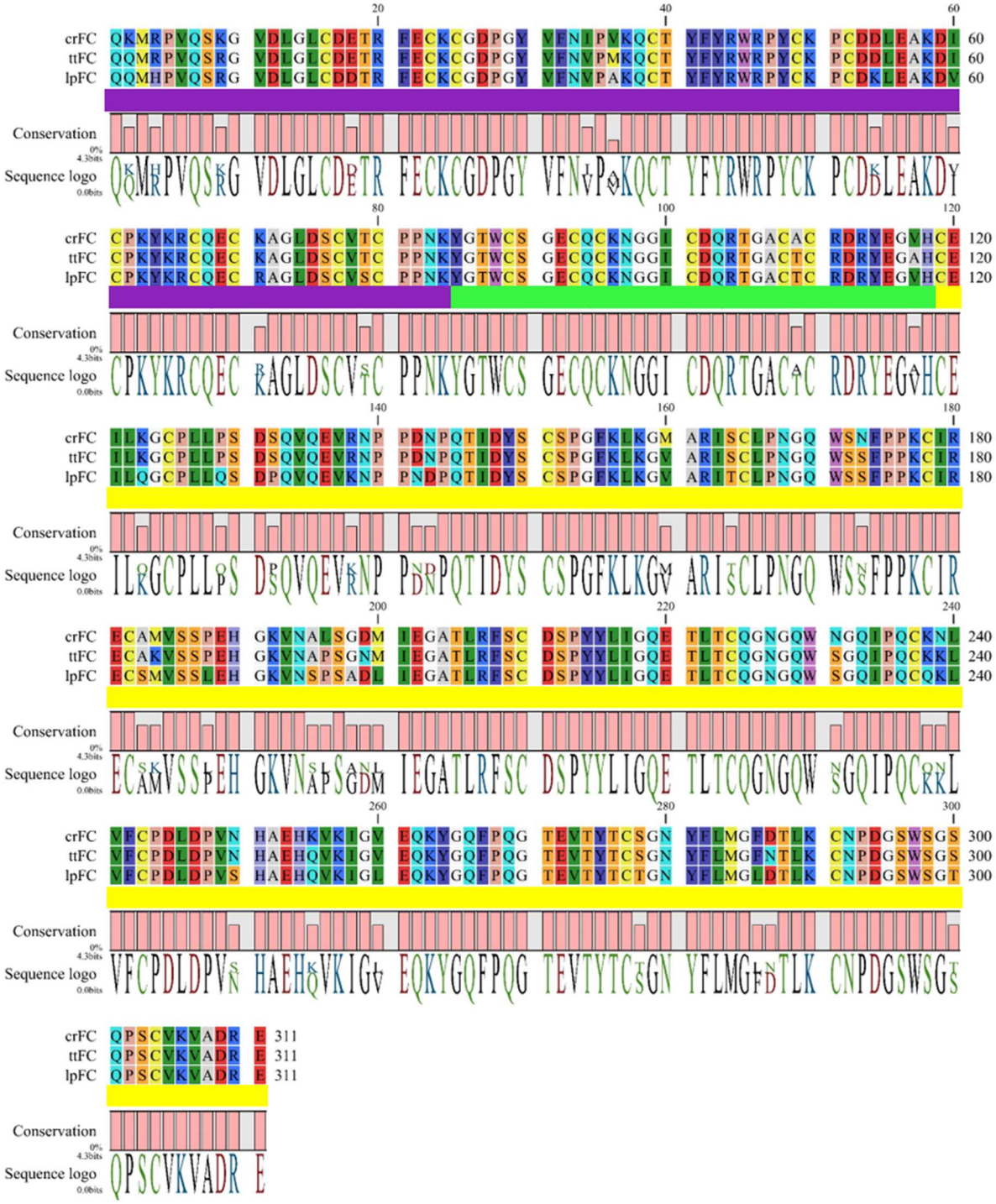

Figure 2-5: Sequence alignment of lpFC, ttFC and crFC N-terminal region. This alignment allows comparison of the Cys-rich, EGF-like and CCP1-3 domains between the three horseshoe crab species. The Cys-rich sequence is shown by the purple bar, the EGF-like by the green bar and the CCPs by the yellow bar. The pink bars indicate the conservation of each residue between the species.

There is a high level of similarity between the N-terminal region of the Factor C gene in all three species of horseshoe crab with the sequence identities for each domain as follows: Cys-rich = 88%; EGF-like = 97%; and CCPs = 86%. However, a few differences do exist between the three species. The majority of these differences are conservative substitutions,

where changes have resulted in amino acids with similar biochemical properties for example, charge, hydrophobicity or size, occupying a particular site. There are also a few non-conservative substitutions, but if we assume that all three proteins bind lipid A in the same way, the non-conservative substitutions suggest that residues at positions 71, 140, 159, 160, 201, 216, 255, 256, 270 and 304 are less likely to be involved in electrostatic interactions with the ligand.

2.4 Design and Assembly of the Synthetic Gene Construct

In order to avoid the production of glycosylated protein in the target eukaryotic expression systems, asparagine (N)-linked glycosylation sites were to be avoided in the final protein sequence. These sites play a role in determining protein structure and shape by post-translationally adding oligosaccharides to asparagine residues, which could potentially lead to structural complications. These sites are not homogenous across any expression system, which can cause conformational or chemical heterogeneity, resulting in considerable differences between each of the samples (Schwarz and Aebi, 2011). Glycosylation is known to strongly influence the folding process, and can also play a role in localisation and binding (Imperiali and O'Connor, 1999). Asn-X_{aa}-Ser/Thr is the consensus sequence for N-linked glycosylation where X_{aa} is not Pro and Thr occurs more often than Ser (Marshall, 1972). Typical practice for removing N-linked glycosylation sites from recombinantly expressed proteins use chemical or enzymatic methods. Release of oligosaccharides can be achieved with the use of mild alkali or mild hydrolysis, however, this often results in protein degradation (Ogata and Lloyd, 1982). Enzymatic methods using the glycosidases PNGase or Endo H are much better for the protein, providing complete removal of oligosaccharides with no protein degradation (Maley *et al.*, 1989). However, the change may have adverse consequences for the stability of the structure.

To identify N-linked glycosylation sites in order to exclude them from the final sequence, the proposed protein sequence for lpFC was entered into the 'Glycosylation site predictor' GlycoEP (www.imtech.res.in/raghava/glycoep/). Both the 'Standard predictor' and 'Advanced predictor' were used, which gave rise to the same results. Five sites were identified at positions 522 and 535 in the CLECT domain, 625 in the CCP4 module, and 768 and 913 in the serine protease domain. Asparagines are likely to be on the surface of proteins, exposed to the aqueous environment. The identified Asparagines (N) were changed to Serine (S). The altered amino acid sequence was again tested using the

glycosylation site predictor, and a further two potential glycosylation sites were identified at positions 766 and 911, also found in the serine protease domain. These were also changed to serine and the resulting sequence now contained no potential N-linked glycosylation sites.

Using the information from the assembled amino acid sequence of *Limulus polyphemus* Factor C, a DNA sequence was generated to allow easy construction of genes for Factor C expression in three different eukaryotic expressions systems; mammalian, yeast and insect.

The predicted horseshoe crab Factor C signal sequence was omitted, and alternative signal sequences were added at the N-terminus for secretion from each system: Ig κ-chain from pSecTag for mammalian expression; Saccharomyces cerevisiae α-mating factor pre-sequence for the yeast, *Pichia pastoris*; and gp67 from *Baculovirus* vector pAcGP67 for insect cell expression. Restriction sites were engineered into the construct to allow for signal sequences to be cut out and to enable re-ligation of the sequence with the appropriate starting sequences. Additional features were added at the N-terminus: a 6xHis-tag was incorporated to allow for purification by nickel affinity chromatography; phiLOV2.1 (Christie *et al.*, 2012) was added as a fluorescent tag; a biotinylation site (Hofmann *et al.*, 1980) was added to be used for detection or purification methods; and a very specific protease cleavage site, human *rhinovirus* 3C (HRV 3C) (Cordingley *et al.*, 1989), was added to enable separation of the expressed Factor C from the tags during purification. A restriction site, XmaI, was incorporated 3' of the 3C site to facilitate subsequent reuse of the vectors to express other proteins, such as the individual domains of Factor C. The designed DNA sequence's total length was 3,943 bp (Figure 2-6).

The predicted lpFC amino acid sequence was run through DNAworks for Drosophila melanogaster codon optimisation (http://helixweb.nih.gov/dnaworks). This also allowed for the exclusion of specific restriction sites that would disrupt the subsequent cloning process, namely; AseI, BglI, BamHI, EcoRI, EcoRV, HindIII, KpnI, NcoI, NdeI, NotI, PstI, SphI, SwaI, XhoI and XmaI. Subsequently, the signal sequences and tags were added to the N-terminal end of the sequence. The full amino acid and DNA sequences for *Limulus polyphemus* Factor C are shown in Appendix B.

The designed DNA sequence was ordered from Genewiz Inc. It was found that the most time and cost effective way to do this was to order two fragments each of less than 3 Kb and so an additional restriction site was required, ideally between domains and roughly in the middle of the sequence. To achieve this, the 'Silent' program from the European Molecular Biology Open Software Suite (EMBOSS) was used. The sequence was analysed in a 'fuzzy' way to recognise all restriction sites up to and including 6 bp in length using the 'silent' function, which identifies cryptic restrictions sites where a single nucleotide change will generate a new site without changing the amino acid sequence of the protein produced. The best solution was to introduce an SpeI site between CCP2 and CCP3, allowing the construct to be ordered as two fragments of size 1631 bp and 2318 bp.

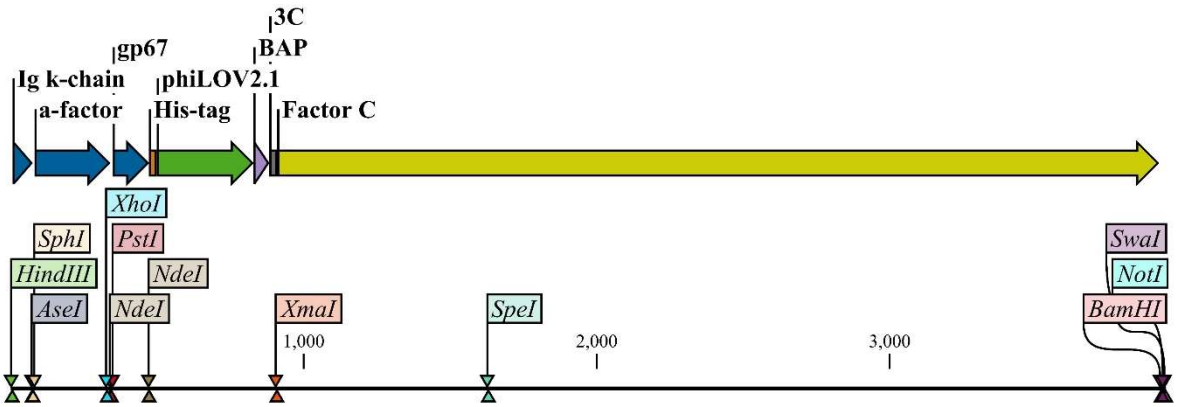

Figure 2-6: Limulus polyphemus Factor C synthetic gene design. Highlighted in blue are the positions of the three signal sequences for the different expression systems: Ig κ-chain; α-factor; and gp67. The various tags for purification and detection are shown at the N-terminal end: His-tag (orange); phiLOV2.1 (green); BAP (purple); and 3C (grey). The restriction sites and positions for use during the molecular cloning process are indicated.

Assembling the lpFC coding sequence and gene from the raw lp genome data allowed a synthetic gene construct to be designed. The resulting construct was to be used to test production of full-length Factor C protein in three eukaryotic expression systems to determine which was the most suitable for protein expression. Determining the sequence for lpFC also allowed comparisons to be made between the different species to identify any differences that may exist between the LPS binding sites. The high-level of similarity between the three N-terminal regions supports the assumption that all three proteins are likely to bind lipid A in the same way.

3 Recombinant Expression in *E. coli*

3.1 Overview

Purification of protein from its natural source, for example animal or plant tissue, is problematic and time-consuming, often resulting in very low yields and protein that is heterogeneously glycosylated. The typical yield of Factor C from horseshoe crab blood is approximately 1.15 mg/ml (Levin and Bang, 1968). In order to overcome the difficulties faced, recombinant protein expression systems have been developed and are now the favoured option for the production of protein to serve a variety of purposes including, but not limited to: structural characterisation, the development of drugs and vaccines, and in industrial processes such as diagnostics. *Escherichia coli* (*E. coli*) is a popular choice for the host organism as it can not only produce high density cell cultures, but molecular biology is quick and easy, and media components are readily available and inexpensive (Rosano and Ceccarelli, 2014). When faced with preparation of samples for nuclear magnetic resonance (NMR) spectroscopy, these features of protein expression are crucial.

Samples produced for analysis by NMR need to be of sufficient concentration (50 – 500 nmoles at 0.1 – 1 mM) and are required to have NMR-active nuclei that interact with the static magnetic field. However, due to the relatively low abundance of naturally active isotopes for specific nuclei, there is a need to incorporate these into the sample through 'labelling'. Active isotopes commonly used for the study of proteins and nucleic acids by NMR include ^{13}C and ^{15}N, but these can be expensive. So, in order to ensure production is economically viable, an adequate yield of recombinantly expressed protein must be attained.

As described in Chapter 1, it has been determined that the N-terminal region of the Factor C protein is the site of lipopolysaccharide (LPS) binding, but with conflicting claims as to the exact location of the binding site (Koshiba *et al.*, 2007; Tan *et al.*, 2000). This chapter describes how individual domain fragments along with combinations of the domains were produced recombinantly in *E. coli* in order to build a detailed picture of the structure/function relationship of this multi-modular protein and determine the site of LPS binding correctly. The correct conformation for LPS binding may only arise in fragments containing more than one domain.

It was appreciated that protein produced in *E. coli* was likely to have endogenous LPS bound, in which case extra purification steps could be carried out to strip the LPS from the protein, to allow studies of the apo-protein and of complexes with defined LPS-derived ligands such as Lipid A. Extracellular expression would be unfavourable as it would guarantee contact with LPS and could be toxic to the *E. coli*. In addition, it is widely known that intracellular expression of correctly folded disulphide-linked protein can be challenging in *E. coli* (Berkmen, 2012) and so for this reason, an expression vector and an *E. coli* expression strain were chosen to maximise the chance of success, more details of which can be found in section 3.3.1.

3.2 Protein Production

There are three main stages required for recombinant protein production: vector construction; protein expression and protein purification. For the most part, the same procedures were used for all recombinant Factor C fragments expressed. Any differences are described below.

3.3 Vector Construction

DNA encoding proteins can be transferred to different organisms in order to produce recombinant DNA that can then be used for efficient gene expression in the new host.

3.3.1 Expression Plasmid

There are a vast number of expression vectors available for use in protein expression, with various features that must be considered in order to determine the most suitable plasmid for the research being undertaken. pNH-TrxT (Figure 3-1) from the pET expression system produced by the Structural Genomics Consortium (SGC) (Genbank ID: GU269914) for bacterial expression, with a pET28-a backbone, was selected for a variety of reasons (Savitsky *et al.*, 2010). These include the presence of a hexa-histidine (6xHis) tag allowing for simple purification by Ni^{2+} affinity chromatography (Bornhorst and Falke, 2000); a thioredoxin (Trx) fusion partner used to encourage correct protein folding (Kern *et al.*, 2003) and high-level production of soluble protein (LaVallie *et al.*, 1993); a TEV-cleavage site to enable cleavage of the tags from the protein of interest (Carrington and Dougherty, 1987a, b); and primers that can be used for cloning via a ligation independent cloning (LIC) technique (Aslanidis and de Jong, 1990). pNH-TrxT carries a

kanamycin resistance gene for selection. Features of the pET system also present include a strong T7 promoter, lac repressor control and induction of both the chromosomal copy of the T7 polymerase in the host genome and of the promoter in the vector by isopropyl β-thiogalactopyranoside (IPTG) (Studier and Moffatt, 1986). T7 RNA polymerase promoters do not appear in the *E. coli* genome meaning it is extremely selective for its specific promoters (Studier and Moffatt, 1986). This, along with having a transcription rate five times faster than *E. coli* polymerase, makes T7 RNA polymerase a highly effective tool for high-level expression of protein in *E. coli* (Studier *et al.*, 1990).

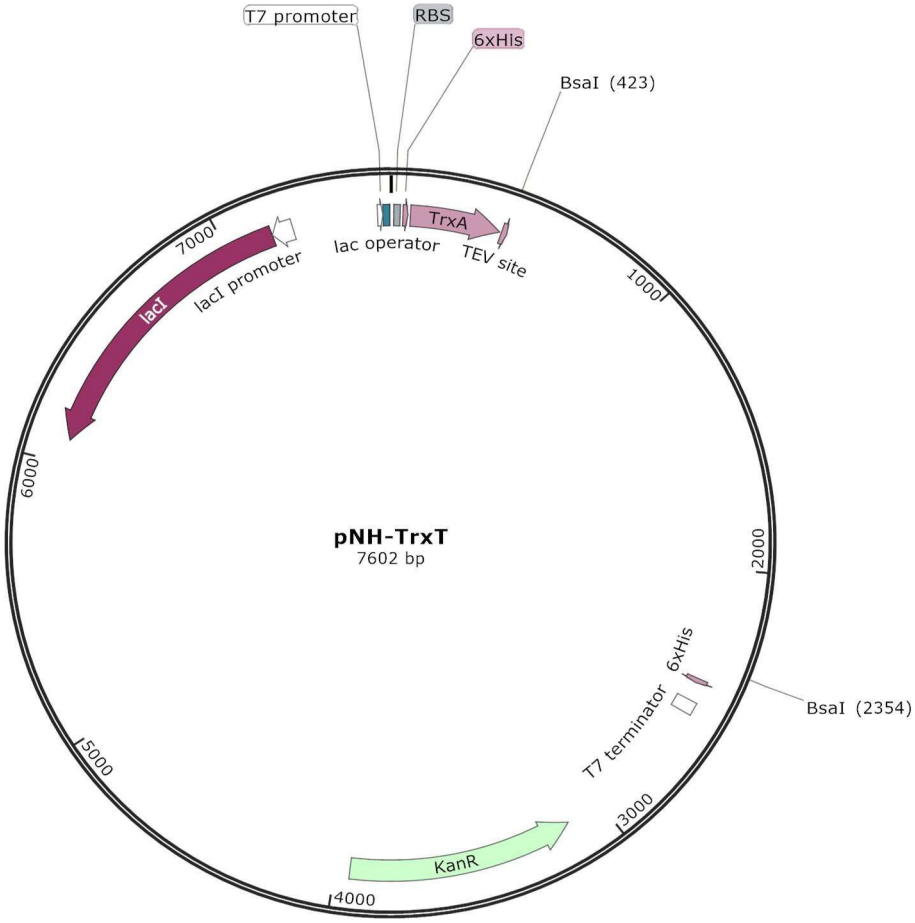

Figure 3-1: pNH-TrxT Vector Map. Created with SnapGene® using the full sequence obtained from Addgene. The two BsaI restriction sites used to linearise the vector are shown at positions 423 and 2354. Other features shown include the E. coli thioredoxin (TrxA), the TEV cleavage site and the 6xHis site.

3.3.2 Ligation Independent Cloning

Developed by Aslanidis and de Jong, ligation independent cloning (LIC) makes use of the T4 DNA polymerase 3' → 5' exonuclease activity to create cohesive ends, 12 nucleotides long, on the vector and insert fragments, allowing for circularisation without the need for ligase, alkaline phosphatase or restriction enzyme treatment of the insert (Aslanidis and de

Jong, 1990). Using this method, constructs from the N-terminal region of Factor C, based on the *Tachypleus tridentatus* sequence, were made for the Cys-rich domain, EGF-like domain and CysEGF domains together, as information on the *Limulus polyphemus* sequence was not yet available. The ttFC protein used was a codon optimised TBIO product, previously produced in the Smith lab. CLC genomics software was used to identify the sequences of the three fragments, which were used to design primers incorporating the LIC overhangs to be used to produce the fragments by the LIC method.

Originally there was a need for primers to be designed for both amplification of the vector and the insert, the former lacking dGMP at their 5' ends and the latter lacking dCMP at their 3' ends. However, the SGC designed their vectors in such a way that simply linearizing the vector and treating the resulting linearized DNA with T4 DNA polymerase and dGTP would be sufficient preparation for the vector for use in the LIC method, and so primers were designed only for the desired inserts (Savitsky *et al.*, 2010). The vector will have noncomplementary tails at each end, which will prevent the formation of circular forms containing only vector.

3.3.2.1 Linearisation of the LIC Vector

pNH-TrxT in DH5α was plated on kanamycin lysogeny broth (LB) agar from a glycerol stock, for selective growth of pNH-TrxT. Individual colonies were picked and used to inoculate 10 ml of LB plus kanamycin in 30 ml universals. Samples were grown overnight at 37°C in the shaking incubator. Plasmid DNA was purified using the Wizard® Plus SV Minipreps DNA purification system (Promega). Initial culture was pelleted at 5,000 × g (Sigma Laboratory Centrifuge 4K15 SciQuip, Rotor 11150) and all Eppendorf centrifugation was carried out at 13,000 × g (Heraeus™ Biofuge Fresco Sorvall benchtop centrifuge). BsaI-HF (NEB) was used for the linearization in a mixture containing 1X NEB Buffer 4, 1X BSA (100 µg/ml), linearized pNH-TrxT vector and BsaI-HF (10 U/µg of DNA). The mixture was incubated for 1 hour at 37°C before heat inactivation for 20 minutes at 65°C. Digested vector was separated from undigested vector using a 0.8% agarose gel (Figure 3-2) made by melting agarose electrophoresis grade in 1X TAE buffer (40 mM Tris-base, 20 mM acetic acid, 1 mM EDTA, pH 8.0) and adding GelGreen™ Nucleic Acid Gel stain (Biotium). Bands were excised on a blue light box (Dark Reader™, Clare Chemical Research) and DNA extraction was carried out using Macherey-Nagel's NucleoSpin® Gel and PCR Clean-up kit.

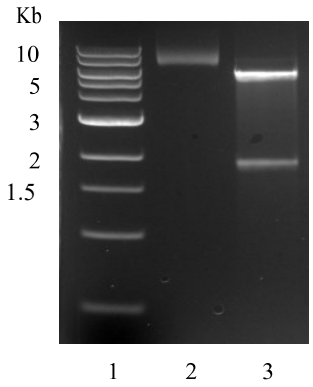

Figure 3-2: Agarose gel analysis of linearised pNH-TrxT vector. Digestion with BsaI resulted in two fragments corresponding to the predicted sizes, 5671 bp and 1931 bp. The larger fragment was excised for treatment with T4 DNA polymerase. 1. Quickload® 1 Kb DNA ladder (NEB), 2. Uncut pNH-TrxT, 3. pNH-TrxT linearised with BsaI-HF.

3.3.2.2 T4 DNA Polymerase Treatment of the Linearised LIC Vector

To generate the long single stranded overhangs required for LIC, 600 ng of BsaI-digested pNH-TrxT vector was treated with 0.06 Units (U) T4 DNA polymerase (NEB) in a reaction containing 1X NEB Buffer 2, 2.5 mM dGTP, 5 mM DTT and 1X BSA (100 µg/ml). The reaction was incubated for 30 minutes at room temperature before addition of EDTA to a final concentration of 10 mM and further incubation for 20 minutes at 75°C to inactivate the polymerase. Due to the competing 3' → 5' exonuclease activity, the polymerase activity and the presence of dGTP, the bases are removed from the 3' ends until the first guanine (G) residue is reached.

3.3.2.3 Polymerase Chain Reaction Amplification of the Inserts

The primers for amplification of the insert fragments included the sequences: TACTTCCAATCC added to the 5' end of the upstream primer and TATCCACCTTTACTG added to the 5' end of the downstream primer (full primer sequences are shown in Appendix C). PCR reactions were set up containing: 1X Pfu polymerase buffer, 0.5 µM Forward primer, 0.5 µM Reverse primer, dNTPs (2.5 mM of each dATP, dCTP, dGTP and dTTP), 1.25 U PfuTurbo DNA polymerase (Agilent Technologies) and 20 ng DNA. The PCR program used for amplification involved 30 cycles of denaturation at 95°C for 30 seconds, annealing at 54°C for 30 seconds, elongation at 72°C for 30 seconds and a final extension of 72°C for 10 minutes.

3.3.2.4 T4 DNA Polymerase Treatment of the PCR Product

The resulting PCR products were analysed using a 2% agarose gel (Figure 3-3) and gel extracted (Macherey-Nagel), then treated with T4 DNA polymerase in the presence of dCTP to create single-stranded ends complementary to those of the vector. 0.2 pmol of

PCR product (calculated using the DNA concentration determined from A_{260} using the assumed e = 50 ng μl^{-1} cm^{-1}, where e = wavelength) was treated with 0.06 U of T4 DNA polymerase in a reaction containing 1X NEB buffer 2, 2.5 mM dCTP, 5 mM DTT and 1X BSA (100 µg/ml).

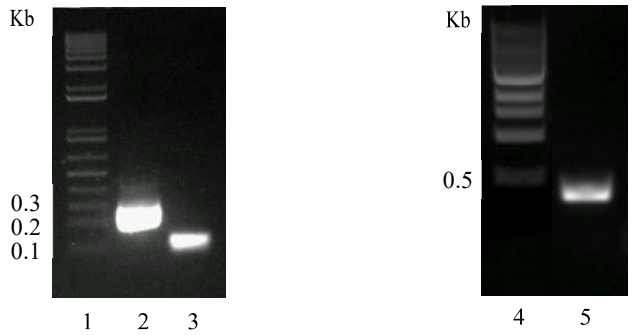

Figure 3-3: Agarose gel analysis of PCR insert products. Cys-rich-, EGF-like- and CysEGF-coding regions amplified in preparation for insertion into linearised pNH-TrxT vector. 1. 1 Kb Plus DNA Ladder (Invitrogen™), 2. Cys-rich PCR product (252 bp), 3. EGF-like PCR product (102 bp), 4. Quickload® 1 Kb DNA ladder (NEB), 5. CysEGF PCR product (348 bp).

3.3.2.5 Annealing of the insert and LIC vector

The complementary ends allowed for non-covalent, intermolecular associations during annealing where vector DNA and insert DNA were incubated in a ratio of 1:5 for 10 minutes at room temperature (RT) before 10 mM EDTA was added and the samples were heated to 75°C before being cooled slowly to room temperature.

3.3.2.6 Transformation into DH5α

The annealed DNA was transformed into Subcloning Efficiency™ DH5α™ Competent Cells (Invitrogen™), a strain providing a transformation efficiency of > 1 x 10^6 cfu/µg plasmid DNA and useful for routine subcloning procedures. Cells were thawed on ice before 1 µl of the annealing mixture was transferred to 20 µl of cells and then placed on ice for 30 minutes. The cells were heat shocked at 42°C for 30 seconds and placed on ice for a further 2 minutes. S.O.C. media (Novagen) was added to the cells and the mixture was incubated at 37°C with shaking for 1 hour. A sample was spread on kanamycin-containing agar and incubated overnight at 37°C. Correctly assembled constructs were determined by sequencing (refer to section 3.3.5).

3.3.3 Thermodynamically Balanced Inside-Out PCR-Based Gene Synthesis

The TBIO method of primer design was developed by Gao *et al.* in order to assemble gene sequences by using smaller fragments to build up the sequence using PCR (Gao *et al.*,

2003). TBIO works to efficiently generate double stranded DNA (dsDNA) from overlapping synthetic single stranded DNA (ssDNA) fragments that cover the entire sequence more efficiently than synthesizing completely complementary dsDNA. The main points are to design the ss fragments so that the overlaps are such that they prime from dsDNA generated in each round of a PCR reaction so that all stages are efficient with the same PCR temperature cycle, and to have the ssDNAs present in relative concentrations such that the inner ds fragment is generated in early cycles and extended in subsequent ones as primers further and further "out" can anneal. An initial fragment of around 0.4 – 0.5 Kb, the 'inside', is produced and gel purified, to be used for bidirectional elongation of the 'outside', using corresponding fragments of 0.4 – 0.5 Kb each time.

DNAworks is a software tool developed by Hoover and Lubkowski to enable automated and optimized design of overlapping oligonucleotides, with a low chance of hairpin formations (Hoover and Lubkowski, 2002). Primer pairs are also constrained to have matching melting temperatures. Overlapping sense strand primers correspond to the N-terminal sequence, whereas the overlapping anti-sense strand primers correspond to the C-terminal sequence.

TBIO was used in an attempt to synthesise a codon optimized fragment for the first three complement control protein domains (CCPs), which lie 3' of the EGF-like domain, with the sequence corresponding to the *Tachypleus tridentatus* protein sequence. The amino acid sequence was submitted to the DNAworks web service using the parameters outlined in Table 3-1. DNAworks was also constrained to exclude the following, frequently used restriction sites from the protein coding region: BamHI, BsaI, EcoRI, HindIII, KpnI, NcoI, NdeI, NotI, PstI, SacI and XhoI. In the highest scoring solution, DNAworks suggested sixteen oligonucleotides with melting temperatures in the range 57-58°C. These sequences were ordered as synthetic oligonucleotides with the addition of restriction sites NcoI and XhoI added at the 5' and 3' oligos in the appropriate frame, respectively, to allow for easy cloning. Correct assembly of the designed oligos was checked using CLC genomics software's 'Assemble sequences' tool. The synthetic oligonucleotides were ordered from Eurogentec.

Parameter	Condition
Standard organism	*E. coli*
Oligo size	55 nucleotides
Annealing temperature	58°C ± 1°C
Oligo concentration	1×10^{-7}
Sodium concentration	5×10^{-2} M
Magnesium concentration	2×10^{-3} M
Codon frequency threshold	10%

Table 3-1: DNAworks parameters for oligonucleotide synthesis. The codon frequency threshold sets a minimum cut-off for the codons to be used for reverse translation of protein sequences into DNA. The codon frequencies are based on the number of times each codon appears in protein regions of the organism's genome.

3.3.3.1 TBIO Experimental Procedure

In order to determine the best conditions for TBIO of the CCPs, PCR reactions were set up for a range of temperatures and with and without additional Mg^{2+}. The combination of oligos used and their concentrations are listed in Table 3-2. The PCR program used for the reactions containing 1X HotStarTaq DNA polymerase buffer, dNTPs (2.5 mM of each), HotStarTaq DNA polymerase (3.75 U, Qiagen), oligos (see Table 3-2) and $MgCl_2$ (if required) included a hot start at 95°C for 15 minutes followed by 30 cycles of denaturation at 94°C for 15 seconds, annealing for 30 seconds and extension at 72°C for 1 minute. The five annealing temperatures tested were: 57°C, 58.1°C, 59.3°C, 59.9°C and 61°C.

Reaction	Oligo combinations and concentrations (nM)							
A				8 200 nM	9 200 nM			
B			7 200 nM	8 120 nM	9 120 nM	10 200 nM		
C		6 200 nM	7 120 nM	8 60 nM	9 60 nM	10 120 nM	11 200 nM	
D	5 200 nM	6 120 nM	7 60 nM	8 40 nM	9 40 nM	10 60 nM	11 120 nM	12 200 nM

Table 3-2: Inner TBIO oligo combinations and concentrations. The CCP coding region of ttFC was built up from combinations of between two oligos and four oligos to determine the optimal reaction conditions and concentrations, and to identify any problematic oligo pairs.

Following PCR using the TBIO oligos, the products were analysed by agarose gel electrophoresis to determine whether the correct constructs had been achieved. The gel revealed an increase in fragments ranging in size from 94 bp for reaction A up to 323 bp for reaction D (results not shown). This suggested that the TBIO reaction had been

successful and therefore, the addition of the next set of 'outside' oligo pairs could take place (Table 3-3). The 'inside' template was cleaned up using the Macherey-Nagel NucleoSpin® Gel and PCR Clean-up kit and used as a template at a concentration of 40 nM for the addition of the next four oligo pairs. It was determined that extra Mg^{2+} was not necessary as agarose gel electrophoresis analysis showed equivalent bands in the presence and absence of the additional Mg^{2+}.

Reaction	**Oligo combinations and concentrations (nM)**								
E				4 200 nM	Inside Template	13 200 nM			
F			3 200 nM	4 120 nM	Inside Template	13 120 nM	14 200 nM		
G		2 200 nM	3 120 nM	4 60 nM	Inside Template	13 60 nM	14 120 nM	15 200 nM	
H	1 200 nM	2 120 nM	3 60 nM	4 40 nM	Inside Template	13 40 nM	14 60 nM	15 120 nM	16 200 nM

Table 3-3: Outer TBIO oligo combinations and concentrations. The CCP coding region of ttFC was built up from combinations of between two oligos and four oligos added to the "inside" template to determine the optimal reaction conditions and concentrations, and to identify any problematic oligo pairs.

TBIO fragments containing the 'outside' oligo pairs were analysed by agarose gel electrophoresis, which revealed fragments of increasing size in reactions E to H (results not shown). These results were very promising. However, clearer, more conclusive results with sharper bands were desired and so conditions were judged to require further optimization.

The repeated "outer" PCR reactions did not produce the same results, so various changes were made to the experimental procedure, including further annealing temperature tests and varying the concentration of inside template used (in the range 2.5 – 20 nM). To check the sequences of the products obtained after the addition of the inner oligonucleotides, the products produced from up to four oligo pairs were TA cloned (3.3.3.2) using the TOPO TA Cloning® kit (Invitrogen™), and transformed into One Shot® TOP10 chemically competent cells (Invitrogen™). Plasmid from selected colonies was prepared for sequencing, which revealed that there were too many errors in the recovered sequences including base pair deletions, missense mutations and frameshift mutations. In order to try to resolve these problems, the assembly was repeated from the start using the proof-reading polymerase KOD Hot Start DNA polymerase (Novagen®)

along with the KOD Hot Start Master Mix in an attempt to minimise the chance of mis-priming events when assembling the construct.

Sequencing results from these additional attempts revealed that base pair deletions were frequently observed in oligo 10. Secondary structure predictions identified oligo 10 as prone to a hairpin loop formation and so the inner PCR reactions were repeated with the addition of dimethyl sulfoxide (DMSO) at concentrations of 5% or 10% to inhibit the formation of intramolecular secondary structures. Oligo 10 was also redesigned maintaining its complementary ends, but with an altered nucleotide sequence, coding for an unchanged polypeptide sequence in between. The CLC genomics software was used to generate suggestions of alternative *E. coli* optimised coding sequences. Several possibilities were generated and the oligo with the lowest secondary structure score was chosen and obtained (see Appendix C for new oligo sequence).

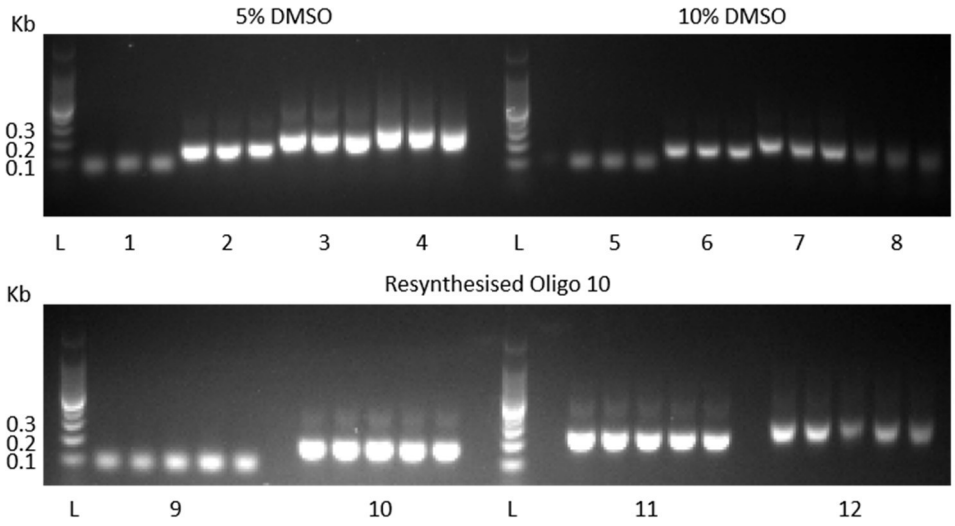

Figure 3-4: ttFC CCPs TBIO PCR products. The upper image shows the products of TBIO PCR reactions including either 5% DMSO (1-4) or 10% DMSO (5-8) for the "inner" set of primers tested at three different temperatures: 54.3°C, 54.9°C and 56.0°C using the original oligo 10. The bottom image shows the products of reactions using the redesigned oligo 10, each tested at five different temperatures: 52.0°C, 53.1°C, 54.3°C, 54.9°C and 56.0°C. L = 1 Kb DNA ladder (Invitrogen™). Reactions labelled: 1, 5 & 9 = 1 primer pair (expected size 94 bp, Table 3-2 reaction A); 2, 6 & 10 = 2 primer pairs (expected size 170 bp, Table 3-2 reaction B); 3, 7 & 11 = 3 primer pairs (expected size 246 bp, Table 3-2reaction C); 4, 8 & 12 = 4 primer pairs (expected size 323 bp, Table 3-2 reaction D).

TBIO synthesis with the redesigned oligo 10 appeared to be more successful than synthesis with the old oligo 10 plus DMSO (Figure 3-4), as judged by agarose gel electrophoresis, since less variability was seen between results. TA cloned plasmids from selected colonies were prepared and sent for sequencing. However, the sequencing results

revealed little success in generating the correct sequence by TBIO, with sequences showing the same errors as before, and so a change in experimental procedure was needed. At this stage, the *Limulus* cDNA sequence had been determined and so the focus switched to producing the CCP fragments from the commercially produced synthetic cDNA using the LIC strategy.

3.3.3.2 TA Cloning

TA cloning relies on the tendency of Taq polymerase to add 3' A-overhangs to dsDNA, and the ability of a fragment of one of the topoisomerases, which is previously attached to the linearised TOPO vector that contains a 3' T overhang at each end, to recombine annealed DNA fragments in a ligase-independent fashion. Thus, PCR amplification products can be treated with Taq and dATP before ligation into a pre-prepared TOPO vector. 2 mM dATP along with 1 X GoTaq® Buffer, 5 U GoTaq® DNA polymerase (Promega) and the PCR product were mixed together and incubated at 72°C for 15 minutes, before being mixed with the pCR™4-TOPO® vector (Thermo Fisher Scientific) in a reaction containing 1 µl vector (10 ng), 0.5-4 µl insert DNA and 1 µl salt solution (1.2 M NaCl, 0.06 M $MgCl_2$). The mixture was incubated at room temperature for 5 minutes before transformation into the TOP10 cells (according to the manufacturer's instructions, Thermo Fisher Scientific), after which the mixture was plated on to kanamycin-containing agar.

3.3.4 IpFC Complement Control Protein Domains

3.3.4.1 Primer design for LIC of CCPs

In order to use LIC to produce the CCP fragments, primers were designed for each construct based on the synthetic *Limulus polyphemus* sequence designed to produce the amino acid sequence determined in Chapter 0. The CCP fragments to be produced were: CCP1, CCP2, CCP3, CCP12, CCP23 and CCP123. For the 5' primers, ~18 bp starting from the 5' end of the desired coding sequence was chosen and synthetic primers designed with this sequence immediately preceded by the vector's LIC sequence in the appropriate frame. For the 3' primers, ~18 bp complementary to the 3' end of the sequence was chosen and synthetic primers designed with this sequence immediately preceded by the vector's LIC sequence in the appropriate frame (3.3.2.3). See Appendix C for full primer sequences.

PCR reactions were set up containing 1 X Pfu DNA polymerase buffer, 0.5 µM LIC forward primer, 0.5 µM LIC reverse primer, dNTPs (2.5 mM of each dATP, dCTP, dGTP and dTTP), 1.25 U PfuTurbo DNA polymerase and Factor C cloned into the mammalian expression vector pcDNA5 as the template (section 4.2.1). The PCR program involved 30 cycles of 94°C for 30 seconds, an annealing temperature of 52°C for 30 seconds and an extension at 72°C for 30 seconds, with a final extension at 72°C for 10 minutes.

The PCR reactions successfully amplified all the desired fragments except for CCP3 alone. Secondary structure prediction for the CCP3 5' primer revealed the probability that it would form a large hairpin. The primer was redesigned such that the final glutamine codon (CAG) was omitted, which reduced the stability of the predicted hairpin, leaving only a weak AT-rich hairpin. The redesigned primer sequence is highlighted in Appendix C. PCR reactions performed with the new primer resulted in amplification of a fragment of the correct size. For all constructs, PCR products of the correct size were generated as judged by agarose gel electrophoresis analysis (Figure 3-5).

The CCP coding fragments were inserted into the pNH-TrxT vector by LIC, following the protocol outlined in section 3.3.2 and their DNA sequences were subsequently checked by DNA sequencing (3.3.5). However, T4 polymerase based LIC was initially unsuccessful, and so a more streamlined version, In-Fusion® cloning (Takara/Clontech), was used instead.

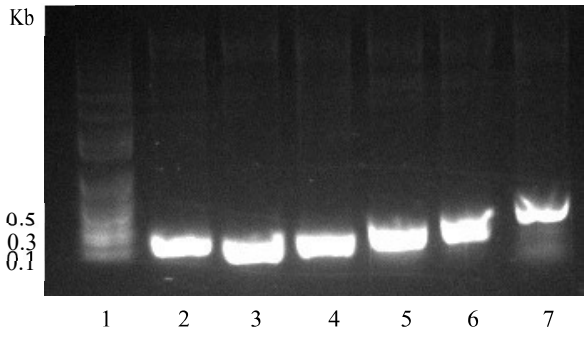

Figure 3-5: Agarose gel electrophoresis analysis of PCR amplified CCP domains. 1. 1 Kb Plus DNA Ladder (Invitrogen™) 2. CCP1 (186 bp), 3. CCP2 (183 bp), 4. CCP3 (221 bp), 5. CCP12 (360 bp), 6. CCP23 (399 bp), 7. CCP123 (576 bp).

3.3.4.2 In-Fusion® HD EcoDry™ Cloning

The In-Fusion® HD EcoDry™ cloning kit (Clontech) relies on an In-Fusion enzyme to efficiently ligate linearised DNA fragments that have at least 15 bp of homologous sequence at their respective ends. This In-Fusion enzyme is believed to be a poxvirus DNA polymerase from vaccinia virus that uses its 3′–5′ exonuclease activity to remove

nucleotides from the 3' end and exposes complementary regions to allow for base pair annealing (Hamilton *et al.*, 2007; Zhu *et al.*, 2007). Solutions containing 20 ng of insert DNA and 50 ng linearised vector DNA were added to an In-Fusion HD EcoDry pellet and mixed well by pipetting. The mixture was incubated for 15 minutes at 37°C, followed by 15 minutes at 50°C before being placed on ice and then transformed into Stellar™ competent cells (according to the manufacturer's instructions, Clontech). Colony PCR (3.3.4.3) was carried out to confirm the presence of the inserts and DNA sequencing (3.3.5) confirmed that the correct sequences had been achieved for all the desired CCP fragments in the pNH-TrxT vector.

3.3.4.3 Colony PCR

Colony PCR allows for colony screening without the need for DNA purification. A dab of colony is transferred to 100 µl of sterile water and boiled for 10 minutes, before PCR is carried out using the appropriate primers for amplification of the DNA the colony is being checked for. Standard PCR reactions and programs can be used and agarose gel electrophoresis is used to analyse the PCR products.

3.3.5 DNA Sequencing

Plasmid DNA was purified using Wizard® Plus SV Minipreps DNA Purification Systems Kit (Promega) and DNA sequencing was performed by DNA Sequencing & Services (MRC PPU, College of Life Sciences, University of Dundee, Scotland, www.dnaseq.co.uk) using Applied Biosystems Big-Dye Ver 3.1 chemistry on an Applied Biosystems model 3730 automated capillary DNA sequencer or Sanger Sequencing Services, Source Bioscience (www.lifesciences.sourcebioscience.com). Sequencing results were analysed using CLC genomics software and compared against the expected sequence to confirm whether the samples contained the correct sequence.

3.4 Protein Expression

E. coli is unable to perform post-translational modifications often necessary for the correct folding and functional properties of proteins (Yin *et al.*, 2007). As disulphide bond formation is an important post-translational modification to achieve the correct structure of the Factor C fragments, SHuffle®, an *E. coli* expression strain designed to produce correctly folded disulphide bonded active proteins was used (Lobstein *et al.*, 2012). This strain is based on the trxB gor suppressor strain, SMG96, that allows cytoplasmic

disulphide bond formation, and incorporates disulphide bond isomerase (DsbC), that proofreads formed bonds and acts as a chaperone for native disulphide bond formation. The cells, SHuffle® T7 Express lysY Competent *E. coli* (New England Biolabs), express a chromosomal copy of T7 RNA polymerase that is under the control of the lacUV5 promoter and induced by IPTG, which initiates transcription of the target gene (Sørensen and Mortensen, 2005). LysY is a variant of T7 lysozyme that lacks amidase activity and therefore this feature makes cells less prone to lysis during induction.

3.4.1 Protein Test Expressions

To investigate protein solubility and establish optimal growth conditions, small scale test expressions were carried out, in which fresh LB cultures were inoculated 1:40 with overnight cultures and 50 μg/ml of kanamycin and grown at 37°C until an optical density at 600 nm (OD_{600nm}) of 0.6-0.8 was reached. At this point, growth temperature and concentration of IPTG were varied to determine the ideal growing conditions for each construct, detailed in 3.7.

3.4.2 Large Scale Protein Expression

Expression conditions were scaled up to produce a larger volume of protein, firstly by expression in two litres (L) of LB to test for solubility and concentration of protein produced before testing unlabelled M9 minimal media (recipe in Appendix D) for the same parameters. In order to produce sufficient isotopically labelled protein for NMR studies, a high-density growth protocol was followed (3.4.2.1).

3.4.2.1 High Density Growth

In order to produce suitable levels of protein to carry out NMR spectroscopy, Marley *et al.* developed a time-efficient and cost-effective method to produce large quantities of isotopically labelled protein (Marley *et al.*, 2001). Initially, high cell numbers are grown in unlabelled rich media, the cells are then recovered and resuspended in isotopically labelled minimal medium, at higher cell densities than they would normally achieve, before protein expression is induced. Thus, isotopically labelled media components are not wasted on producing biomass. For the Factor C fragments, initial growth in 2 x YT media (recipe in Appendix D) was used, which contains double the concentration of yeast extract in comparison with LB medium, at 37°C until an OD_{600nm} = 1.5 – 2. According to Lobstein *et al.*, using glycerol as a carbon source results in poor growth in T7 SHuffle cells and therefore it was important not to use a medium that contained glycerol such as

terrific broth (TB) (Lobstein *et al.*, 2012). The cells were then pelleted (Beckman® Model J2-21 Centrifuge, Rotor JA-10, 9,056 × g) and washed with 1 X M9 media lacking Thiamine, ammonium salts and D-glucose before being pelleted again. This time, the pellets were resuspended into half the original culture volume using M9 minimal medium containing ^{15}N-ammonium chloride ($^{15}NH_4Cl$) for nitrogen labelling or $^{15}NH_4Cl$ and $^{13}C_6$-glucose for carbon labelling to produce a double-labelled sample. The cells were then cultured at 15°C for 1 – 1.5 hours before induction with IPTG (0.1 mM – 0.4 mM depending on construct) and growth overnight at the lowered temperature (Sivashanmugam *et al.*, 2009).

3.5 Protein Purification

In order to isolate the protein of interest from the resulting protein mixture, the cells were first pelleted by centrifugation (Beckman® Model J2-21 Centrifuge, Rotor JA-10, 9,000 rpm), and then resuspended in water and COmplete™, EDTA-free Protease Inhibitor Cocktail (Sigma-Aldrich®), Lysozyme (0.1 mg/ml, Sigma-Aldrich®) and Benzonase® Nuclease-HC (25 U, Novagen®) were added. The protease inhibitor cocktail inhibits the proteolytic activity of a wide range of proteases, lysozyme is used to help with protein extraction by breaking down the bacterial cell wall, and Benzonase® Nuclease-HC degrades nucleic acids.

3.5.1 Cell Lysis

Cell lysis was achieved by sonication (MSE Soniprep 150, Sanyo), where cells are disrupted by high-frequency pressure waves applied via a vibrating probe. A constant low temperature is crucial to prevent the protein from denaturing, and this is maintained by keeping the sample on ice and applying short, 15 second pulses, followed by 15 second pauses (Palmer and Wingfield, 2004).

3.5.2 Centrifugation

High speed centrifugation (40,000 g, Sigma Laboratory centrifuges 3K30 SciQuip, Rotor 12150-H) was used to pellet insoluble cell debris, which was in some cases retained for inclusion bodies protein purification (IBPP, section 3.5.4). The cell lysate containing the protein of interest and other soluble compounds was further purified by Ni^{2+} affinity chromatography (section 3.5.3).

3.5.3 Ni^{2+} Affinity Chromatography

Immobilized metal ion affinity chromatography (IMAC) was first described in 1975 by Porath *et al.*, who recognised the ability of matrix bound metal ions, for example Zn^{2+} and Cu^{2+}, to interact with protein imidazole and thiol groups (Porath *et al.*, 1975). This allowed the development of peptide affinity tags that can be expressed as fusion proteins of the protein of interest in order to purify recombinant proteins (Uhlén *et al.*, 1983). Poly-histidine tags are the most commonly used affinity tags as a strong complex is formed between the histidine imidazole rings and appropriately immobilized Zn^{2+} ions. The cell lysate is passed through the column containing the metal ions and the 6xHis-tagged protein is captured on the matrix. After washing with buffer to remove non-specifically bound proteins, His-tagged protein can be eluted using a buffer containing free imidazole (Bornhorst and Falke, 2000).

For the Factor C protein fragments, soluble cell lysate was filtered (0.8 µm, Minisart® Sartorius stedim biotech) before application onto the pre-charged Ni^{2+}-affinity matrix (Ni-Superflow Resin, Generon). The column was washed with buffers containing 20 mM TrisHCl, 150 mM NaCl and 0.01% NaN_3, pH 7.9 and increasing concentrations of imidazole (binding buffer 5 mM; wash buffer 20 mM; and elution buffer 250 mM). Hydrochloric acid (HCl) was used to lower the pH to 7.9. Sodium azide (NaN_3) was added as an antimicrobial and all buffers were filter sterilised. Flow through from each column wash was collected and analysed by SDS-PAGE (see section 3.7)

3.5.4 Inclusion Bodies Protein Purification

Inclusion bodies formed by *E. coli* contain aggregates usually of inactive, insoluble protein and are often formed during recombinant protein expression (Rudolph and Lilie, 1996). This protein has to be solubilized and refolded correctly in order to produce active protein (Palmer and Wingfield, 2004).

For each of the Factor C constructs that produced inclusion bodies, the inclusion body pellet was resuspended and washed thoroughly three times with a solution of a few grams of methionine per litre of water and 1 mM EDTA. The cell debris was sonicated and centrifuged at 40,000 × g between each wash. Then, 5 volumes of aqueous organic solvent (50:50 acetonitrile-water with 0.2% TFA and some methionine) were added to the pellet and the solution was stirred overnight at room temperature. The soluble and insoluble

fractions from this final stage were separated by centrifugation at 40,000 × g and both the pellet and supernatant were retained for analysis by SDS-PAGE.

3.5.5 Buffer Exchange and Sample Concentration

Protein samples were exchanged into the desired buffer using a 20 ml capacity Vivaspin with a 5,000 kDa molecular weight cut-off (Sartorius Stedim Biotech). This allowed for removal of undesired buffer components, for example imidazole, and concentration of the sample to the required volume for further purification by gel filtration (see section 3.5.7) or analysis by NMR spectroscopy (see section 3.6.2).

3.5.6 TEV-Cleavage

The presence of a TEV-cleavage site within the expression vector allowed for the simple cleavage of the histidine and Trx tags from the protein of interest by TEV-protease. TEV protease is the 27 kDa catalytic domain of Nuclear Inclusion a (NIa) protein, which is encoded by the Tobacco Etch Virus (TEV) and recognises a sequence of the general form E – X_{aa} – X_{aa} – Y – X_{aa} – Q – (G/S) with cleavage occurring between Q and G or Q and S (Carrington and Dougherty, 1987b). The TEV-protease used was expressed in the lab and was His-tagged.

Cleavage was carried out at room temperature overnight in either Tris or Phosphate based buffer at a concentration of 83 μg of TEV-protease per mg of protein. Separation of the protein of interest from the tags and TEV-protease was achieved by Ni^{2+} affinity chromatography (section 3.5.3), gel filtration chromatography (section 3.5.7) or Reversed Phase High-Performance Liquid Chromatography (RP-HPLC, section 3.5.8)

3.5.7 Gel Filtration Chromatography

Gel filtration chromatography was introduced by Lathe and Ruthven and further developed by Porath and Flodin, to be used for the size dependent separation of protein molecules (Lathe and Ruthven, 1956; Porath and Flodin, 1959). This method involves passing a protein sample through a pre-packed Superdex 75 10/300 GL (GE Healthcare Lifesciences) column containing cross-linked agarose and dextran, with a bed volume of 24 ml. The column was attached to an ÄKTA chromatography system controlled with UNICORN™ software version 5.01 (GE Healthcare Lifesciences). This allowed a semi-automated method for sample collection to be run, which, in this case, consisted of a 30

ml equilibration with the appropriate buffer before injection of 0.5 ml of concentrated sample via the injection valve and elution with 1.5 column volumes of buffer. The program was run at 0.5 ml/min to ensure the column's pressure limit (1.8 MPa) was not exceeded and 0.5 ml protein fractions were collected. The UV absorbance at 220 nm and at 280 nm were recorded to identify the volume at which the desired protein was eluted.

3.5.8 Reversed Phase High-Performance Liquid Chromatography

The Reversed Phase High-Performance Liquid Chromatography (RP-HPLC) method typically involves the use of an n-alkyl-silica-based column for elution of proteins in an aqueous mobile phase from immobilized hydrophobic ligands in the stationary phase. This can be achieved by increasing the concentration of a buffer containing an organic solvent, for example the nonpolar solvent acetonitrile (ACN) and an ionic modifier, for example trifluoroacetic acid (TFA) (Aguilar and Hearn, 1996; Aguilar, 2004; Mant and Hodges, 1996).

Samples were prepared by filtering them through a Minisart® 0.2 µm syringe filter (Sartorius Stedim Biotech) and making to 10% ACN, before injection onto the Supelco analytical Discovery® BIO wide pore C8 column (Sigma-Aldrich®), with dimensions 25 cm x 10 mm, a particle size of 10 µm, pore size of 300 Å and base column volume (CV) of 19.635 ml, via the ÄKTA explorer (Amersham Pharmacia Biotech). UNICORN™ software version 5.31 (GE Healthcare Lifesciences) was used to control the method for RP-HPLC, which followed the following protocol: 0.5 CV equilibration with 10% ACN; up to three sample injections of 3 ml each; 0.5 CV column wash with 10% ACN to wash out unbound sample; first gradient step (target 50% ACN in 0.5 CV); second gradient step (target 60% ACN in 2 CV); third gradient step (target 100% ACN in 0.5 CV); 100% ACN for 0.5 CV before returning to 10% ACN for 0.5 CV for re-equilibration. The program parameters included flow rate – 4 ml/min; wavelengths recorded – 280 nm and 220 nm; and the collection of 2 ml fractions by the fraction collector Frac-920 (GE Healthcare Lifesciences), which were analysed by SDS-PAGE (section 3.7).

3.5.9 Lyophilisation

Lyophilisation or freeze-drying, invented in 1906 by Arsène d'Arsonval and Frèdèric Bordasis, is a dehydration process used to remove solvents by sublimation. Samples are completely frozen before being subjected to low pressures. RP-HPLC fractions containing

the protein of interest were divided between 1.5 ml Eppendorf tubes with needle point holes in the lid, deep frozen by submerging them in liquid nitrogen and placed within the chamber of the Heto PowerDry LL1500 freeze dryer (Thermo Scientific), pre-cooled to -110°C, connected to an Edwards XDS5 Pump. Samples were lyophilised overnight and the resulting powdered protein was dissolved in water before addition of stock solutions of appropriate buffer components, in preparation for further analysis by circular dichroism spectroscopy (CD), nuclear magnetic resonance spectroscopy (NMR) or X-ray crystallography.

3.6 Protein Analysis

3.6.1 Circular Dichroism Spectroscopy

Circular dichroism was used to obtain information regarding the structure and stability of the Factor C fragments. Samples were made up in 10 mM Tris, pH 8.0 before being examined by near UV and far UV, more details of which can be found in Chapter 5.

3.6.2 Nuclear Magnetic Resonance Spectroscopy

NMR was carried out to analyse protein folding, ligand binding and to determine the three-dimensional structure of the Factor C fragments. Samples were prepared in NMR Buffer (20 mM Na_2HPO_4, 50 mM NaCl, 0.01% NaN_3, adjusted to pH 7.0 using phosphoric acid) and D_2O was added to make a 5% solution. Details of the NMR experiments used can be found in Chapter 5.

3.6.3 X-Ray Crystallography

Another method that was explored for structure determination of the Factor C fragments was X-ray crystallography. X-ray crystallography requires the production of protein crystals from a protein sample at high concentration. The crystals are then exposed to an X-ray beam to generate diffraction patterns and the diffraction patterns can be analysed to gain information about distribution of electrons within the crystal, and thus the protein's conformation.

Unlabelled samples of the Factor C Cys-rich, EGF-like, CCP1 and CCP2 individual domains were expressed as Trx-fusion proteins and purified to be used for crystallisation trials. Purified samples were buffer exchanged into 10 mM Tris, 100 mM NaCl (made to pH 7.9 using HCl), resulting in concentrations of ~10 mg/ml each. Preliminary screens for

crystallisation conditions were carried out with the help of Dr Alastair Gardiner, in which two 96-well plates were set up for each sample using the JCSG-plus™ and PACT premier™ kits dispensed by a Cartesian robot (Newman *et al.*, 2005; Page *et al.*, 2003). JCSG-plus™ is a sparse matrix screen that tests standard PEG (polyethylene glycol) and salt conditions, in a wide pH range (4.0 – 10.0). The PACT premier™ is a PEG/ion screen exploring pH, cations and anions in varying PEG conditions. Used together, these two screens give a comprehensive investigation of crystallisation conditions.

A total drop size of 1 µl was used, containing 0.5 µl of protein sample and 0.5 µl of reservoir condition. Each plate contained two wells for each condition, one of which a high concentration sample was added, and the other a sample containing half the concentration, after being diluted with Tris buffer. Once trays had been set up, they were stored at 16°C. At the time of setting up crystal trials, it was unknown how the protein would react in low salt concentrations and so the concentration of NaCl was kept at 100 mM. In terms of sample preparation, it would have been optimal for there to be no salt present. This is due to the fact that lots of the conditions being screened included salt, which may have been masked by a high concentration of salt in the sample.

After a few days, liquid-liquid phase separation and/or precipitation was apparent in some of the wells. These features are both regarded as positive as phase separation indicates a metastable transition and precipitation indicates the protein is in a state of supersaturation, both of which can result in crystal growth. After almost four months, crystals appeared in a few of the wells, with the most promising crystals found in the JCSG+ well containing 0.2 M lithium sulphate, 0.1 M sodium acetate, pH 4.5, conc. 50% w/v and precipitant PEG 400. Two types of crystals had been produced, globular and rod-like, with the largest around 100 µm. These crystals were exposed to the X-ray beam on the mardtb “desktop beamline” fitted with the mar345 image plate detector (marXperts) by Dr Aleksander Roszak. The presence of PEG in the buffer means the crystal is cryo-protected and will not freeze in the liquid nitrogen. Unfortunately, the diffraction pattern revealed the presence of salt alone. Further endeavours to produce high enough concentrations of unlabelled samples proved unsuccessful and a decision was made to focus on structural elucidation by NMR.

3.7 Factor C Protein Constructs

3.7.1 *Tachypleus tridentatus* Recombinant Factor C LPS Binding Fragments

The following sections provide details and results regarding the three Factor C fragments that were produced from the *Tachypleus tridentatus* sequence information.

3.7.1.1 Cys-rich

Produced by LIC and expressed at 37°C with 0.1 mM IPTG induction, soluble protein was produced in either unlabelled or ^{15}N-labelled M9 minimal media and purified by Ni^{2+} affinity chromatography (Figure 3-6). Fractions containing soluble protein were combined and the Trx tag was cleaved by overnight cleavage with TEV-protease. Cleaved Cys-rich was separated from the fusion tag by RP-HPLC.

RP-HPLC fractions containing the cleaved protein were lyophilised and the dehydrated samples were re-dissolved in either NMR buffer or Tris buffer in preparation for NMR spectroscopy analysis, CD (results shown in Chapter 5) or X-ray crystallography.

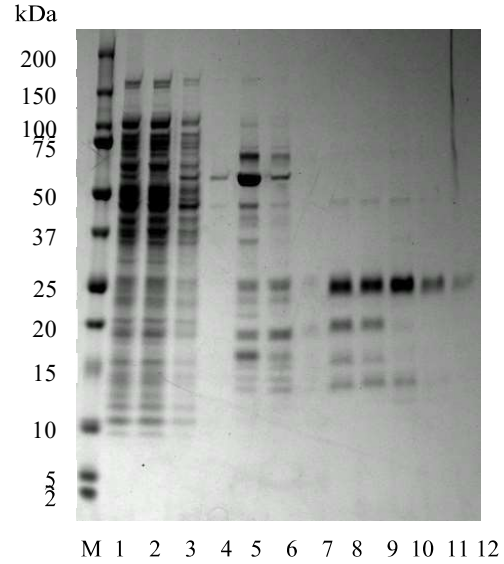

Figure 3-6: pNH-TrxT Cys-rich. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1-2. Flow through 3-4. Binding buffer 5-7. Wash buffer 8-11. Elution buffer 12. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). The strong bands around 23.7 kDa in fractions 8-10 show protein of the correct MW for pNH-TrxT Cys-rich.

Analysis by NMR was carried out on the fusion protein, which showed mainly Trx peaks. Less than 10 peaks were associated with Cys-rich, significantly less than the 80 to be expected. This suggested that the Cys-rich fragment was only partially folded or it was

sampling multiple conformations. To get a clearer picture, the fusion protein was cleaved using TEV-protease and the unseparated mixture was examined by NMR. This showed clearer Trx peaks, which is consistent with the idea that Cys-rich is aggregated, bound to LPS or even interacting with Trx directly. Gel filtration was used to separate the Cys-rich fragment from the Trx. Fractions containing the protein of interest were combined, concentrated and looked at by NMR, but again, only around 10 peaks that could account for the Cys-rich domain were visible, implying the Cys-rich was tumbling slowly and could not be seen.

Reversed-phase HPLC (RP-HPLC) was used to strip the protein sample of any endogenous LPS and to fully separate the protein from the Trx. Fractions containing Cys-rich were lyophilised to remove the solvent, resuspended in H_2O and their NMR spectra acquired. Again, the majority of expected peaks were not observed for this sample, with only around 16 visible. Most of these appeared close to the random coil chemical shift, which suggests these are from a flexible terminus or a loop. Only around 5 were thought to be in more structured parts of the protein. It is likely the protein is in a state of multiple conformations and so experiments were carried out over a range of temperatures from 283 K to 313 K to investigate whether temperature changes would help to stabilise a particular conformation. Slight changes occurred in that sharp peaks got broader, broad peaks got sharper, some peaks appeared and some peaks disappeared, but these were not significant enough to warrant further investigation into effects of temperature differences. Further investigations could look into the effects of varying other conditions such as altering the pH or the addition of LPS to determine any structural changes that may occur.

3.7.1.2 EGF-like

The EGF-like domain was also expressed as a Trx-fusion protein in unlabelled and ^{15}N-labelled M9 media at 37°C with induction by 0.1 mM IPTG. Purification by Ni^{2+} affinity chromatography resulted in a large concentration of the EGF-like fusion protein (Figure 3-7), which allowed for structural investigation by NMR and X-ray crystallography to take place with labelled and unlabelled samples, respectively.

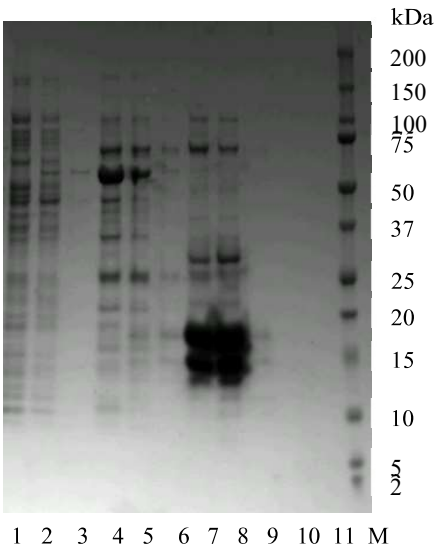

Figure 3-7: pNH-TrxT EGF-like. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). The large band around 17.8 kDa indicates the presence of pNH-TrxT EGF-like, with the lower band around 14 kDa indicating possible proteolytic cleavage of the Trx.

A two dimensional (2D) ^{15}N-HSQC was recorded before and after TEV-cleavage (Figure 3-8). Each peak in the spectrum represents an amide-proton nitrogen pair from one amino acid on the protein backbone. The position of the peaks is affected by the chemical environment around the amino acids and therefore NMR spectra is very informative when the sample is subject to local conformational or chemical changes (for more information see Chapter 6). In Figure 3-8A, the sample has not been cleaved from the Trx. There is a good dispersion of peaks, which suggests the protein is correctly folded. However, upon cleavage (Figure 3-8B), the peaks become disordered, which indicates the protein without the fusion protein is not folded. It is likely that this sample exists in multiple conformations as there were around 40 peaks when there should only be around 30. To try to refold the protein correctly, the pH was lowered by increments of 0.5 from pH 7 to pH 3 using HCl. This did not result in any differences and so the pH was raised back to pH 7 using NaOH and the protein sample was run on gel filtration to ensure complete separation of the EGF-like domain from the cleaved tag.

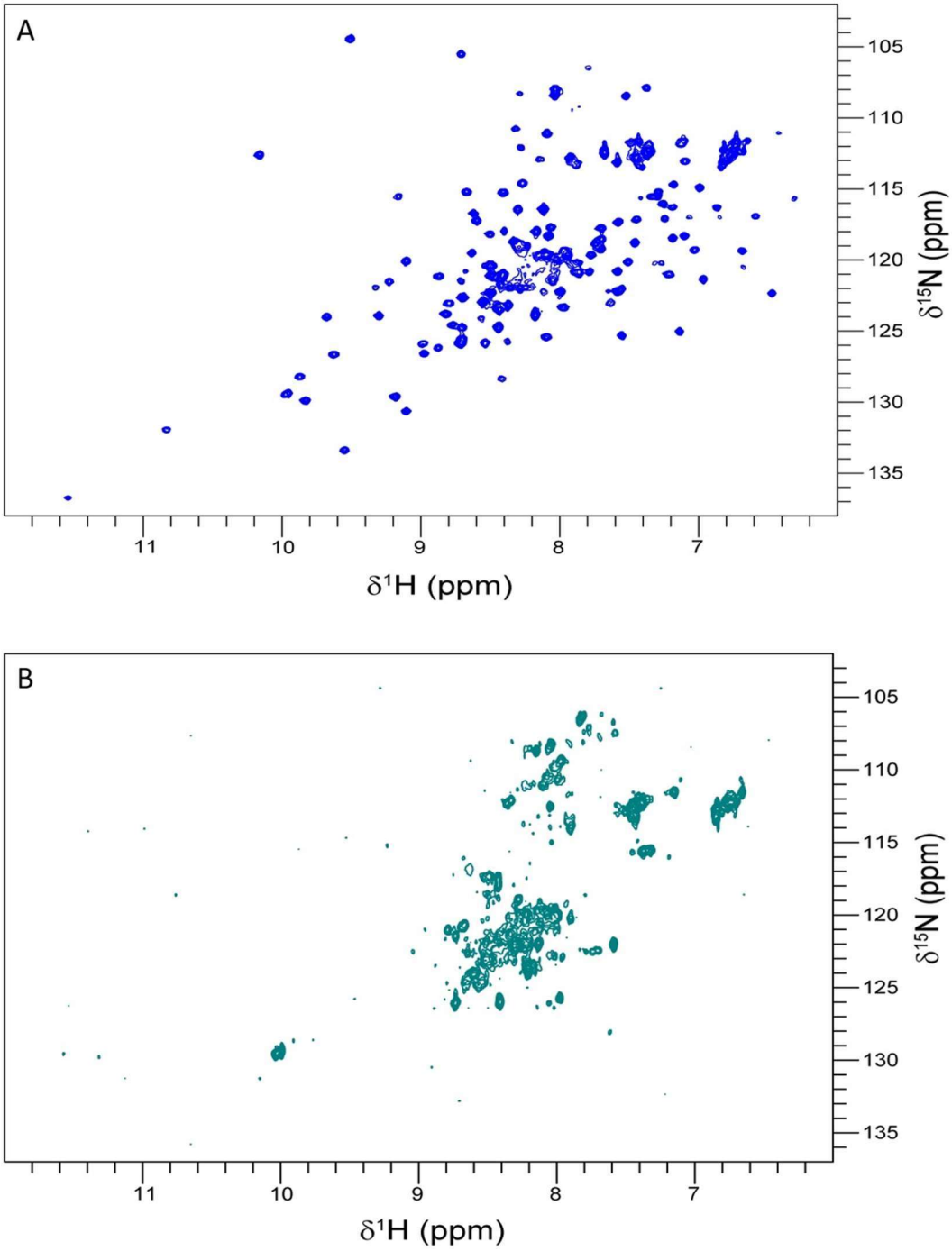

Figure 3-8: 15N-HSQC of the EGF-like domain. A: A spectrum for the EGF-like Trx-fusion protein. B: A spectrum for EGF-like cleaved and separated from the Trx-fusion protein. Proton is along the x-axis and Nitrogen is along the y-axis.

The sample was again examined by NMR where the temperature was lowered in increments from 298 K to 278 K to observe any changes to the structure. Slight differences were observed, including the appearance of peaks at lower temperatures, but these were considered insignificant and so RP-HPLC was used to strip the protein of any endogenous LPS. However, again the NMR did not show very clear peaks and due to time constraints, the focus was changed to a different construct.

3.7.1.3 CysEGF

^{15}N-labelled protein was expressed in minimal media overnight at 15°C, following induction by 0.1 mM IPTG, and purified by Ni^{2+} affinity chromatography (Figure 3-9). The majority of the protein produced was found to be insoluble and so inclusion bodies protein purification was carried out. Soluble protein was buffer exchanged into NMR buffer, concentrated and examined by NMR at 298 K. The spectra of uncleaved protein showed broad peaks, there were no peaks for the Cys or EGF tryptophans and only about 10 peaks of the 100 expected for CysEGF were visible. This suggests the protein may be unfolded, mis-folded, sampling multiple conformations or aggregated. Experiments were run at the lower temperature of 283 K, which resolved a few of the peaks but did not make a significant difference. If this protein was in a monomeric state, either the whole of the protein was not well structured or it was flexible and required other conditions or factors to stabilise its structure.

Gel filtration was carried out using a sample diluted 10-fold, followed by a sample that was undiluted. This was performed to investigate whether correct folding was concentration dependent. It was expected that if the protein was susceptible to concentration dependent aggregation or oligomerisation, the protein would elute at different volumes. However, there was no apparent difference between the two samples. Fractions thought to contain the desired protein, eluted around 10 ml, were combined, concentrated and again looked at by NMR. The 1D-^{1}H experiments revealed protein in the sample, but no peaks were seen in the ^{15}N-HSQC. This suggests that either the sample was not the protein expected or, if it was ^{15}N-labelled, indicates that the protein was misfolded and tumbling slowly. This may be due to it aggregating, it might be bound to LPS micelles or some sort of oligomerisation has taken place. The gel filtration data supported these possibilities as the protein sample is eluted at a volume corresponding to a higher MW protein than expected for CysEGF.

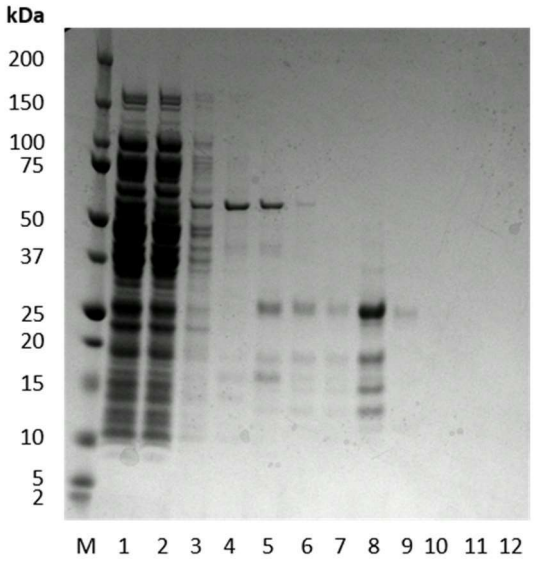

Figure 3-9: pNH-TrxT CysEGF. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through. 2-3. Binding buffer. 4-6. Wash buffer. 7-10. Elution buffer. 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). The strong band in the first elution buffer wash contains protein of the correct MW for CysEGF – 27.2 kDa

SDS-PAGE analysis (data not shown) revealed a laddering of bands, which may be artefacts of gel sample preparation or might actually reflect inter-molecular disulphide bonded forms that are resistant to reduction in the SDS-PAGE sample buffer, revealing that the protein was misfolding in multiple different ways. Additionally, the bands may be a result of proteolytic cleavage. This could be tested by making the samples more reducing for example by adding 10 mM TCEP. If the protein was indeed a misfolded inter-molecular disulphide bonded aggregate, then an alternative route to producing correctly folded protein that could be attempted is treatment with a redox shuffle (e.g. GSH & GSSG) and/or a protein disulphide bond isomerase such as protein disulphide isomerase (PDI) that should catalyse the formation and breakage of disulphide bonds as the protein folds. However, time did not allow to take this construct any further.

3.7.2 *Limulus polyphemus* recombinant Factor C LPS binding fragments

The following sections provide details and results for the complement control proteins (CCPs) that were produced using the *Limulus polyphemus* Factor C sequence described in Chapter 2.

3.7.2.1 CCP1

^{15}N-labelled protein was expressed in minimal media after overnight induction at 15°C with 0.4 mM IPTG, and purified by Ni^{2+} affinity chromatography (Figure 3-10). Soluble protein was buffer exchanged into NMR buffer, concentrated and examined by ^{15}N-HSQC. The data obtained were encouraging with the observation of clearly defined peaks for both the Trx and the CCP1 protein. The presence of a tryptophan not associated with the Trx suggested there was a favourable chance that the protein was properly folded. A T_2 relaxation experiment was set up before cleavage from the Trx. This experiment indicates how the molecules are tumbling in solution. As the Trx and CCP1 are different sizes, they should tumble independently and their ^{15}N signals should relax with different T_2 rates. From the NMR spectra, it was possible to determine which peaks belonged to CCP1 and which belonged to the Trx tag. TEV-cleavage was carried out and a ^{15}N-HSQC was recorded before the CCP1 protein was separated from the Trx tag (Figure 3-11). This showed good chemical shift dispersion indicating that both the Trx and the CCP fragment were folded. The close correspondence between the peak positions indicated that the fusion partners were not interacting, as there is no evidence that a structural change occurred. Clear differences between the two samples are seen in the expanded region of the figure, likely reflecting changes in conformation of linker residues on cleavage.

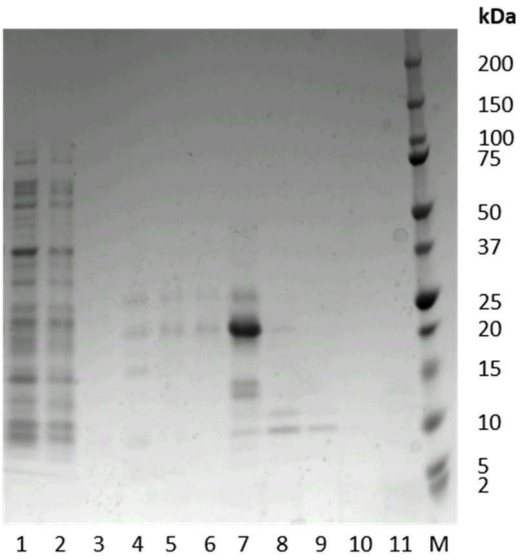

Figure 3-10: pNH-TrxT CCP1. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). The clear band in fraction 7 indicates protein of the correct size (21.1 kDa) for CCP1.

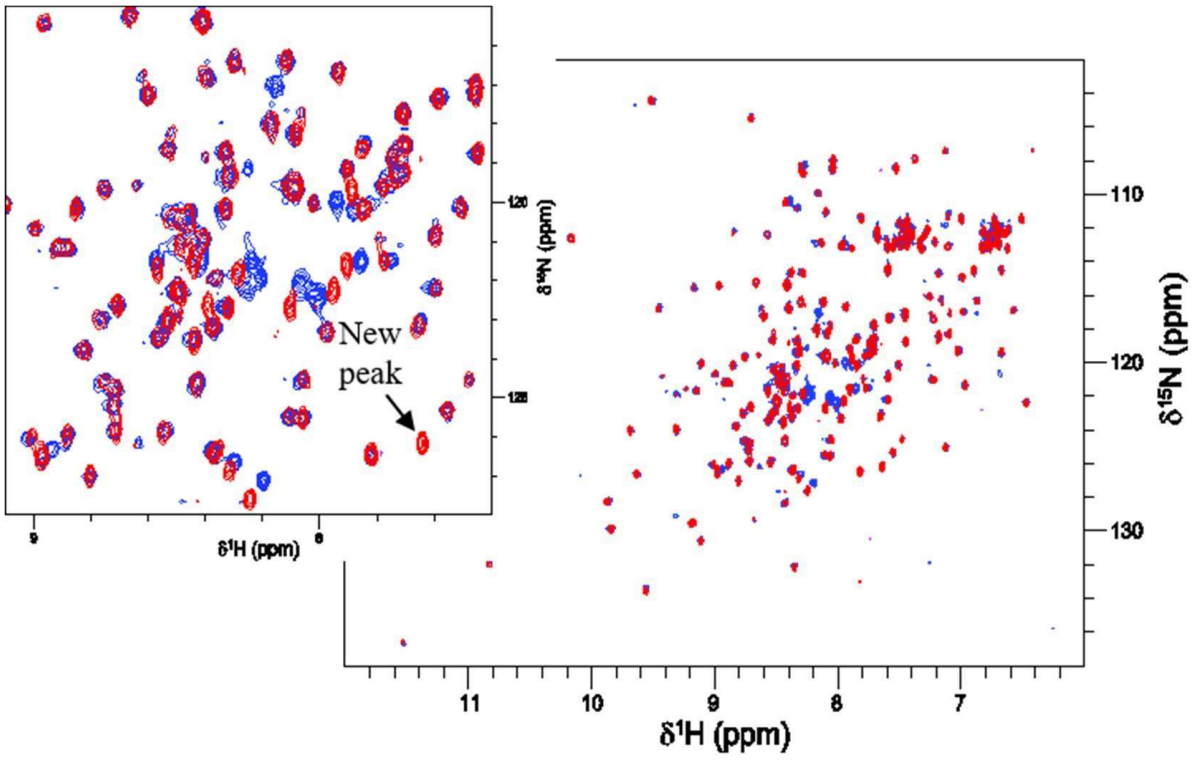

Figure 3-11: 15N-HSQC spectrum of CCP1. The spectrum shows pNH-TrxT CCP1 (blue) and a TEV-cleaved sample of pNH-TrxT CCP1 before gel filtration (red) superimposed. A new peak (indicated in the expanded region) appears in the cleaved sample at chemical shift coordinates expected for a C-terminal residue and a few peaks have shifted likely reflecting slight changes in conformation of linker residues on cleavage.

Gel filtration was carried out to separate the CCP1 from the Trx, however, SDS-PAGE analysis revealed the two proteins were eluted in the same fraction. As Ni^{2+} affinity chromatography also failed to separate the proteins, RP-HPLC was attempted to effectively separate the two from each other in order to perform further experiments for structure determination. This resulted in a severe decrease in protein concentration and further experiments were ineffective and so efforts were directed to a different construct.

3.7.2.2 CCP2

Soluble ^{15}N-labelled protein was expressed in minimal media overnight at 15°C after IPTG induction (0.4 mM), and purified by Ni^{2+} affinity chromatography (Figure 3-12). As with the CCP1 fragment, CCP2 was examined by NMR as a Trx fusion protein and cleaved from the fusion protein but not separated (Figure 3-13). Again, clearly defined peaks were observed for both the Trx and the CCP2 protein and the presence of a tryptophan not related to the Trx suggested folded protein was present in the sample, further indicated by good chemical shift dispersion of the peaks. Slight differences between the samples before and after TEV treatment were observed that likely reflect

changes to the conformation of the linker residues upon cleavage, but cleavage does not appear to affect the overall structure.

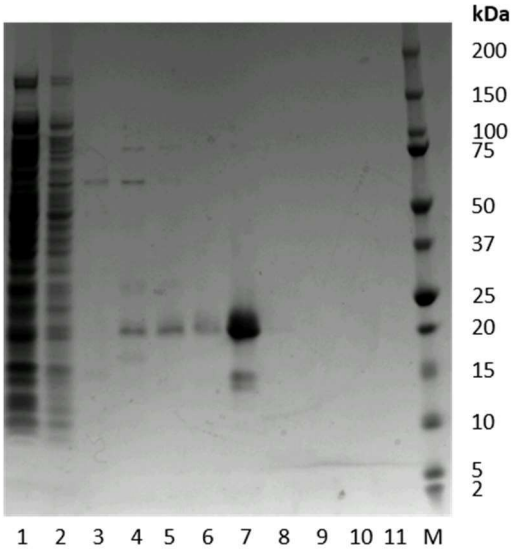

Figure 3-12: pNH-TrxT CCP2. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). The strong band at the start of the elution's reveals protein of the correct MW for CCP2 (20.9 kDa).

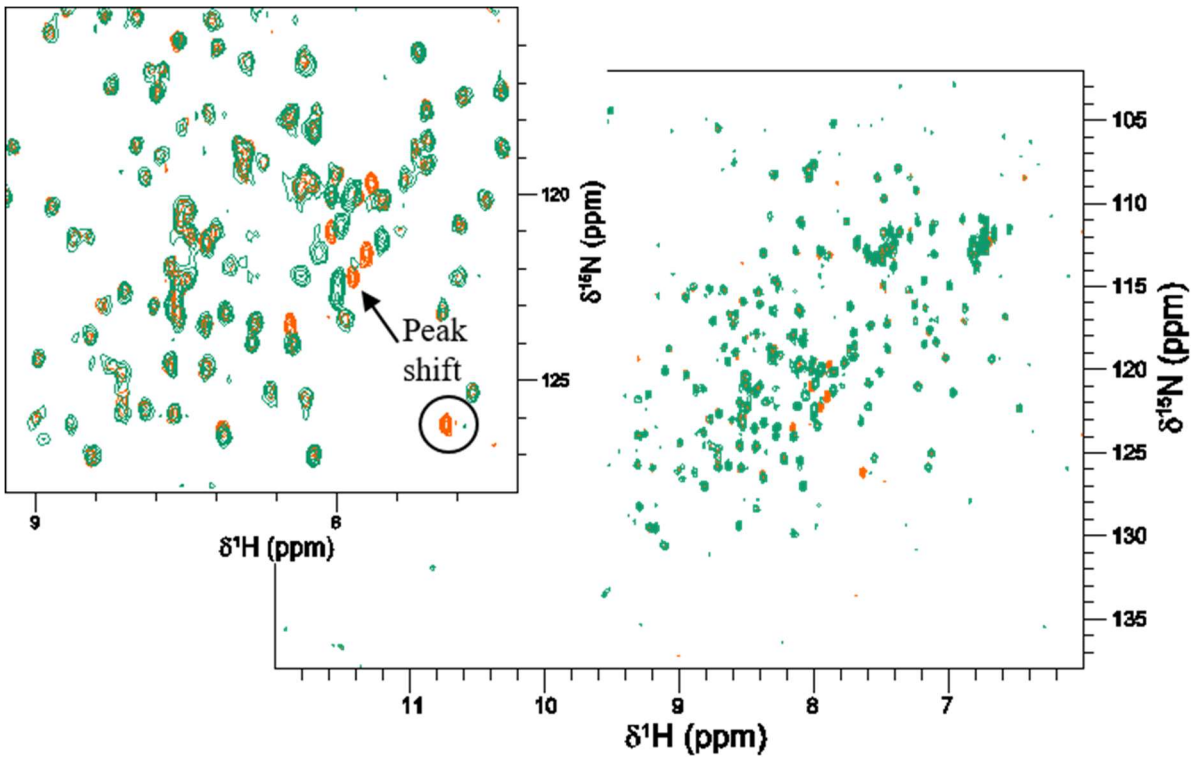

Figure 3-13: 15N-HSQC spectrum of CCP2. The spectrum shows an uncleaved sample of pNH-TrxT CCP2 (teal) and a cleaved sample of pNH-TrxT CCP2 before gel filtration (orange) superimposed. The expanded region reveals that some peaks have shifted positions (indicated) revealing differences between the two samples, suggesting slight conformational changes upon cleavage. A new peak (circled) is evident in the cleaved sample at chemical shift coordinates expected for a C-terminal residue.

Gel filtration, Ni^{2+} affinity chromatography and RP-HPLC were used in an attempt to separate the Trx and CCP2 protein. SDS-PAGE analysis revealed that separation did not occur and that there was a reduction in the protein concentration, meaning further structural determination of this fragment was not feasible. Due to the difficulties faced when separating the cleaved fragments, analysis of this construct was not taken any further.

3.7.2.3 CCP3

^{15}N-labelled protein was expressed in minimal media at 15°C overnight after induction with

0.4 mM IPTG, and purified by Ni^{2+} affinity chromatography (Figure 3-14).

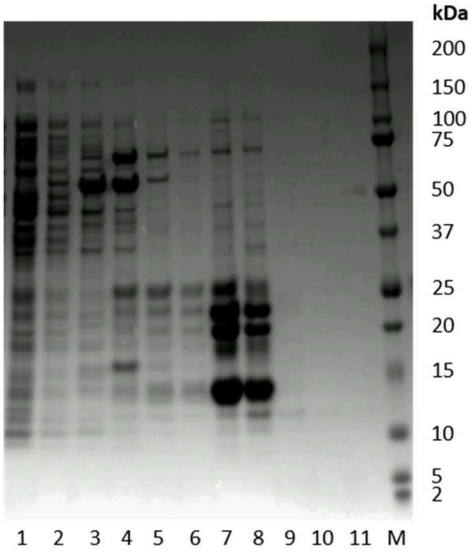

Figure 3-14: pNH-TrxT CCP3. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). The band at 22.2 kDa in fractions 7 and 8 reveal the presence of CCP3 in the samples, with what appears to be truncated Trx at the lower MW.

The soluble protein was buffer exchanged into NMR buffer, concentrated and looked at by NMR. The data from this revealed the peaks were predominantly from the Trx. A few peaks most likely from CCP3 were present but the signals were weak. The sample was cleaved from the fusion protein by TEV-protease and run on gel filtration to separate the proteins from each other. However, SDS-PAGE revealed no separation and so optimisation is required in order to obtain a suitable sample for further analysis.

3.7.2.4 CCP12

^{15}N-labelled protein was expressed overnight at 15°C in minimal media after induction with 0.4 mM IPTG, and purified by Ni^{2+} affinity chromatography (Figure 3-15). Soluble protein was buffer exchanged into NMR buffer, concentrated and examined by NMR. The

spectra showed peaks that were well dispersed, indicating folded protein, but also contained broad peaks at random coil chemical shifts suggestive of unfolded protein. TEV-protease was used to successfully cleave CCP12 from the fusion protein. CCP12 and Trx were separated by Ni^{2+} affinity chromatography. The CCP12 sample was concentrated and examined by NMR, which showed there was correctly folded protein in the sample and no Trx (Figure 3-16). The unfolded protein was also separated from the folded sample. The presence of two tryptophans confirmed the presence of both CCP1 and CCP2. 110 peaks were counted, which corresponded to the correct number expected and so a T_2 relaxation experiment was performed. In this case, the T_2 experiment was important as it gave an indication of how the two domains act in relation to each other. Following this a 3D TOCSY experiment with a mixing time close to 60 ms was carried out. This experiment is a key one in the determination of protein structures and resonance assignment for ^{15}N-only labelled samples. A {^{15}N, ^{13}C} double labelled sample was produced and more information concerning the NMR experiments carried out can be found in Chapter 0.

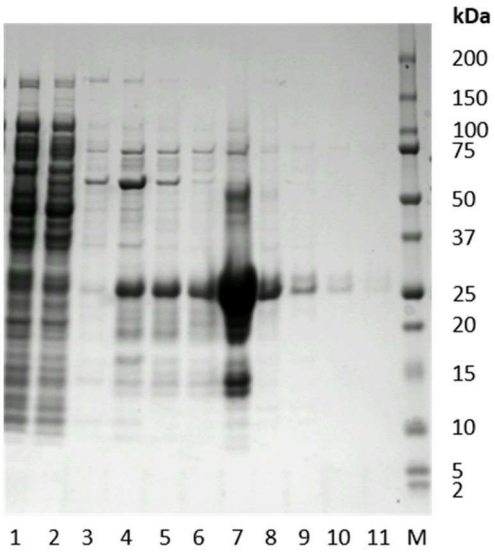

Figure 3-15: pNH-TrxT CCP12. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). Indication of protein at 27.5 kDa confirms the presence of CCP12 in the elution's.

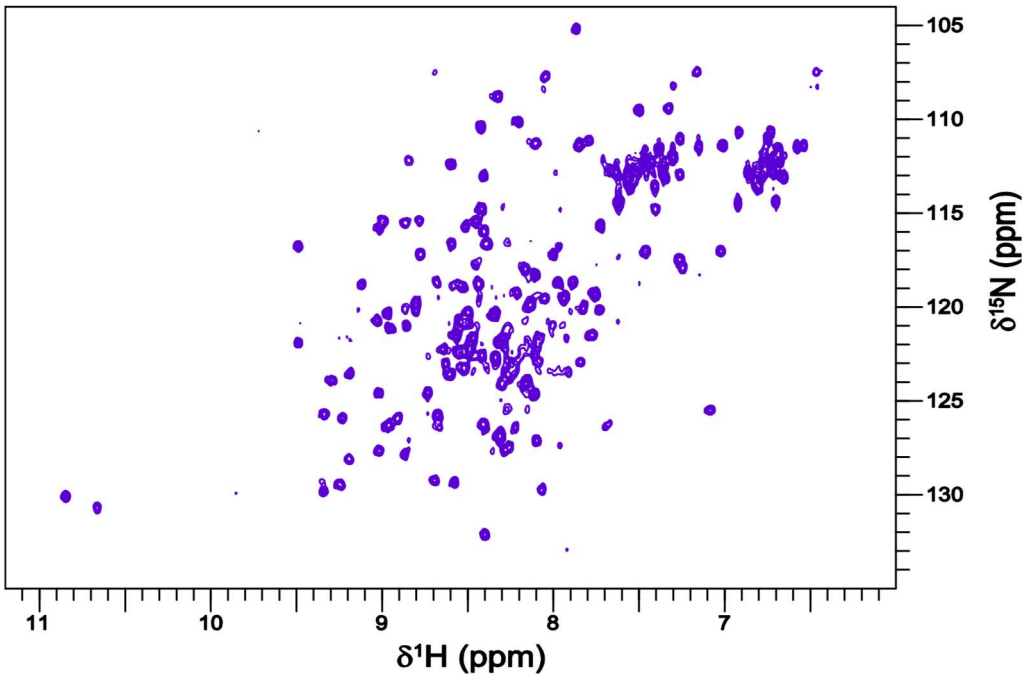

Figure 3-16: 15N-HSQC spectrum of CCP12. The well dispersed peaks indicate folded protein. Proton is along the x-axis and nitrogen is along the y-axis.

To test whether CCP12 contains an LPS binding site, LPS was added to the sample and after incubation at room temperature, NMR experiments were performed. The absence of marked differences suggested that perhaps LPS was already bound and so the sample was run through the RP-HPLC to 'strip' any bound LPS. This resulted in clear chemical shift differences in proton and nitrogen, which suggests a ligand may have been stripped off. However, there are a few other possibilities as to why this may have occurred, for example, oxidation of methionine, and further experimentation is required to confirm any assumptions that can be made.

3.7.2.5 CCP23

^{15}N-labelled protein was expressed at 15°C overnight in minimal media, with the addition of 0.4 mM IPTG for induction. The majority of the fusion protein was found to be insoluble and so was purified from inclusion bodies by RP-HPLC, with soluble protein being purified by Ni^{2+} affinity chromatography (Figure 3-17). Difficulties faced when attempting to express and purify a high enough concentration for structural analysis of this protein resulted in focus being given to other protein constructs that gave better initial results.

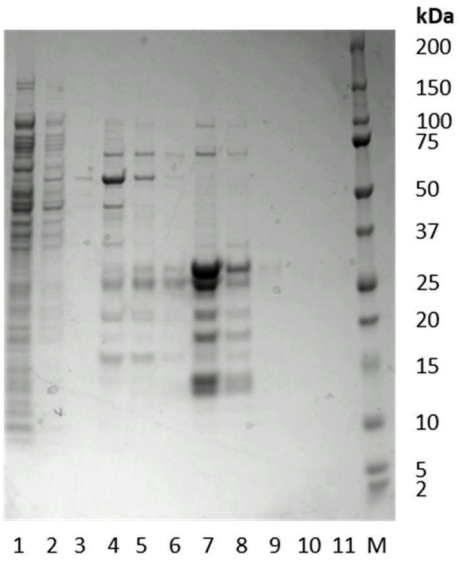

Figure 3-17: pNH-TrxT CCP23. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). CCP23 has a MW of 28.8 kDa, which is shown in fractions 7 and 8 of the purification.

3.7.2.6 CCP123

^{15}N-labelled protein was expressed in minimal media at 15°C overnight after induction with 0.4 mM IPTG, and purified by Ni^{2+} affinity chromatography (Figure 3-18).

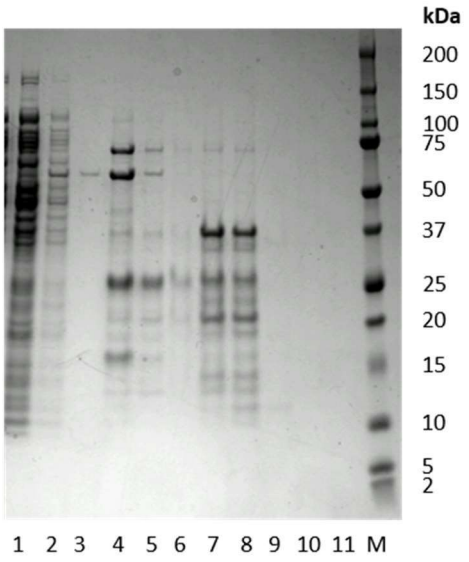

Figure 3-18: pNH-TrxT CCP123. SDS-PAGE analysis of Ni^{2+} affinity chromatography fractions showing purified protein. 1. Flow through 2-3. Binding buffer 4-6. Wash buffer 7-10. Elution buffer 11. Elution buffer with 1 M imidazole. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad). A protein band around 35 kDa in the first two elution's indicates the presence of CCP123 in these fractions.

The fractions containing soluble protein were buffer exchanged, concentrated and examined by NMR, which revealed mainly Trx peaks and evidence of disordered protein. Gel filtration was carried out on the protein sample, which revealed the presence of molecules of several different sizes, confirmed with analysis by SDS-PAGE (Figure 3-19). Some fractions contained high molecular weight species; a strong band appeared around 70 kDa, which could be the chaperone DnaK from *E. coli*. Some of the purer fractions containing protein of the correct molecular weight for CCP123 were combined and examined by NMR. The ^{15}N-HSQC spectra revealed peaks for Trx but only around 10 peaks out of the 170 expected for CCP123. These were at the random coil chemical shift and were broad. This suggests the protein may be quite rigid and that this is why data cannot be produced for the whole protein. Further NMR experiments may provide useful information regarding this sample, for example, running a TROSY experiment, which can pick up signals from things tumbling slower due to their large size. The protein may be aggregated or LPS micelles may be bound and so given more time, RP-HPLC could be used in an attempt to strip this from the protein.

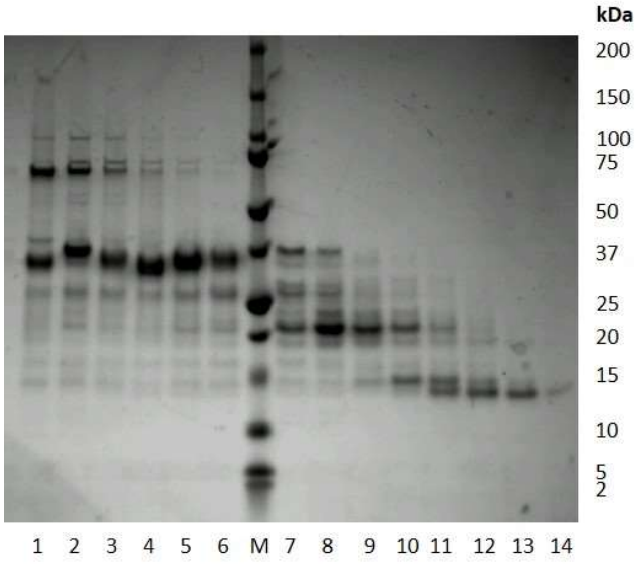

Figure 3-19: CCP123 gel filtration fractions. SDS-PAGE analysis of cleaved CCP123 gel filtration fractions reveals multiple bands for each sample, suggesting the presence of several different molecules. The fractions from lanes 4-6, containing protein of 35 kDa were combined for NMR experiments. Marker = Precision Plus Protein™ Dual Xtra Standards (Bio-rad).

3.8 Summary

In an effort to determine the exact location of LPS binding, nine N-terminal Factor C fragments were successfully cloned into the pNH-TrxT vector using either a ligation independent cloning method or an InFusion® cloning technique. Soluble fusion protein

was produced in *E. coli* and purified for all nine fragments and varying degrees of structural analysis was carried out.

The Cys-rich, EGF-like and CysEGF fragments appeared to be misfolded or sampling multiple conformations. Further experiments are needed to determine the optimal conditions for the analysis of these proteins in order to determine their structures. CCP1 and CCP2 samples produced good spectra indicating folded protein when cleaved from the fusion protein. However, there is a need to optimise separation from the Trx in order to retain a suitable concentration of CCP1 and CCP2 for further structural analysis of the individual fragments by NMR. The overall quality of CCP3 was poor, meaning the sample was not suitable for further analysis. An improvement in expression and purification of this construct is necessary for further studies.

CCP12 produced the best NMR data, which allowed for preliminary structure calculations to be carried out, as detailed in Chapters 0 and 7. For CCP23 and CCP123, it was difficult to express enough protein for structural analysis and further experimentation is required for these constructs.

Although X-ray crystallography was initially unsuccessful, improving the quality and concentration of protein samples may improve its success in future.

4 Recombinant Expression in Alternative Expression Systems

4.1 Overview

For the majority of recombinant protein expression, *E. coli* is the expression system of choice for convenience. This is due, in the most part, to its rapid rate of growth, inexpensiveness and ability to produce large quantities of the target protein (Yin *et al.*, 2007). However, *E. coli* has its limitations and so alternative expression systems including other prokaryotic as well as mammalian, insect and yeast cell systems have been developed for the production of folded proteins that proved unsuccessful in *E. coli* (Tanio *et al.*, 2008). In the case of Factor C, it was anticipated that protein production would be difficult in *E. coli*, especially the expression of full-length protein, and so alternative expression systems were explored: mammalian, insect and yeast.

4.2 Vector Selection

4.2.1 pcDNA™5/FRT/TO for Mammalian Expression

The vector chosen for mammalian expression was pcDNA™5/FRT/TO (Invitrogen™, Figure 4-1). This vector is an attractive choice as it is designed to produce stable transformants as a result of co-transfection into a Flp-In™ T-REx™ host cell with the expression plasmid, pOG44 Flp. This integration allows for the creation of a cell line that can be easily propagated. Unlike many other mammalian expression systems, expression is regulated by the tet operator, and so the level of expression is determined by the amount of tetracycline added.

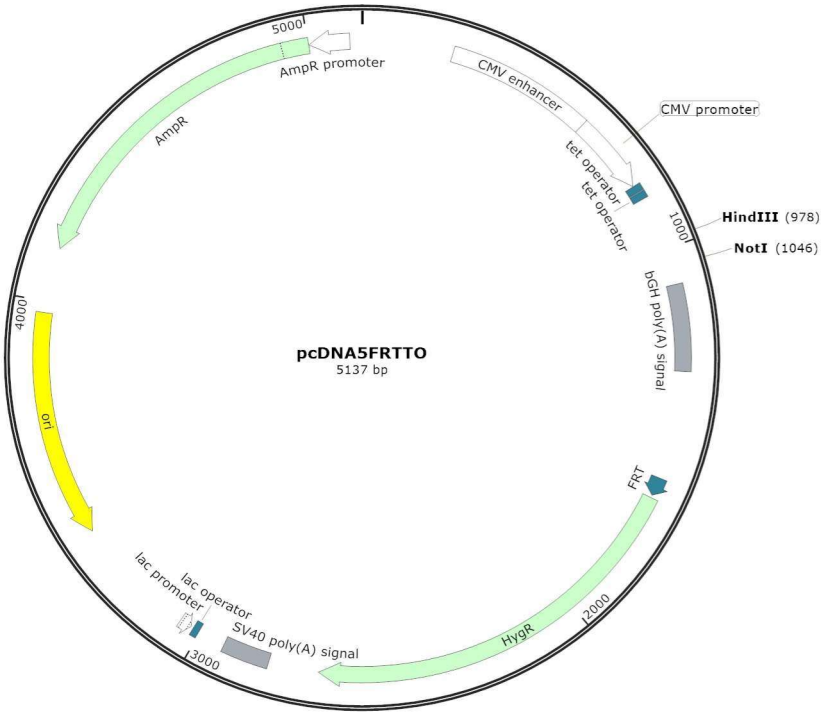

Figure 4-1: pcDNA5/FRT/TO vector map. The HindIII and NotI restriction sites to be used during molecular cloning are shown at positions 978 and 1046, respectively. The tet operator is marked along with other key features of the vector. Created with SnapGene®.

4.2.2 pVL1392 for insect cell expression

The vector of choice for insect expression was pVL1392 (*Baculovirus* transfer vector, BD Biosciences, Figure 4-2), which contains the complete polyhedrin enhancer-promoter sequences for high expression of recombinant protein. The advantages of using an insect expression system include the fact that it is able to carry out complex post-translational modifications and it can produce properly folded proteins. High yields of protein can be produced, and, specifically with regards to Factor C, insect cells have previously been used to produce *Carcinoscorpius rotundicauda* (cr) complement control protein (CCP) fragments, suggesting that protein production is likely to succeed in this system (Tan *et al.*, 2000).

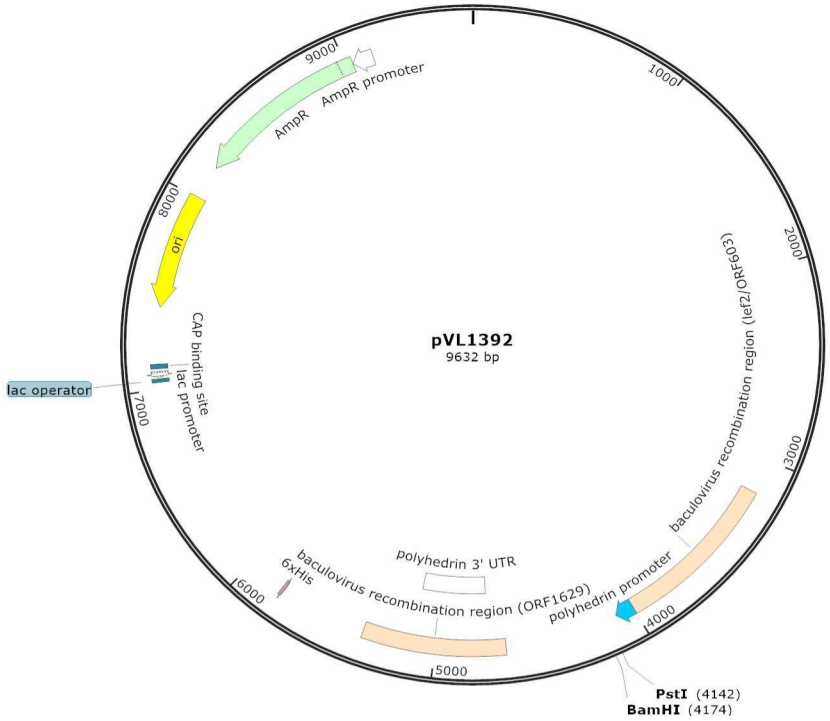

Figure 4-2: pvL1392 vector map. The polyhedron promoter is indicated, as well as the PstI site and BamHI site at locations 4142 and 4174, respectively. Created with SnapGene®.

4.2.3 Pichia Pink™-HC for Yeast Expression

Pichia Pink™-HC (Invitrogen™, Figure 4-3) was chosen for expression in yeast. There is a limited selection of *Pichia* vectors, but *Pichia pastoris* is known for high-level, large scale expression and secretion of recombinant proteins and has previously been successfully used for many secreted proteins studied by NMR. pPink™-HC was the preferred plasmid for its use of the AOX inducible promoter to drive expression encoding the desired heterologous protein and also for its colour based selection as a result of the presence of the ADE2 gene that catalyses a biosynthesis step of purine nucleotides. The colony colour is determined by the integrant copy number. White colonies contain more copies of the integrated construct and are more desirable than the pink colonies that express very little gene product. This expression system is faster, easier and more cost-effective than other eukaryotic systems and can produce grams per litre under certain circumstances (Clare *et al.*, 1991; Cregg *et al.*, 1987; Sreekrishna *et al.*, 1989).

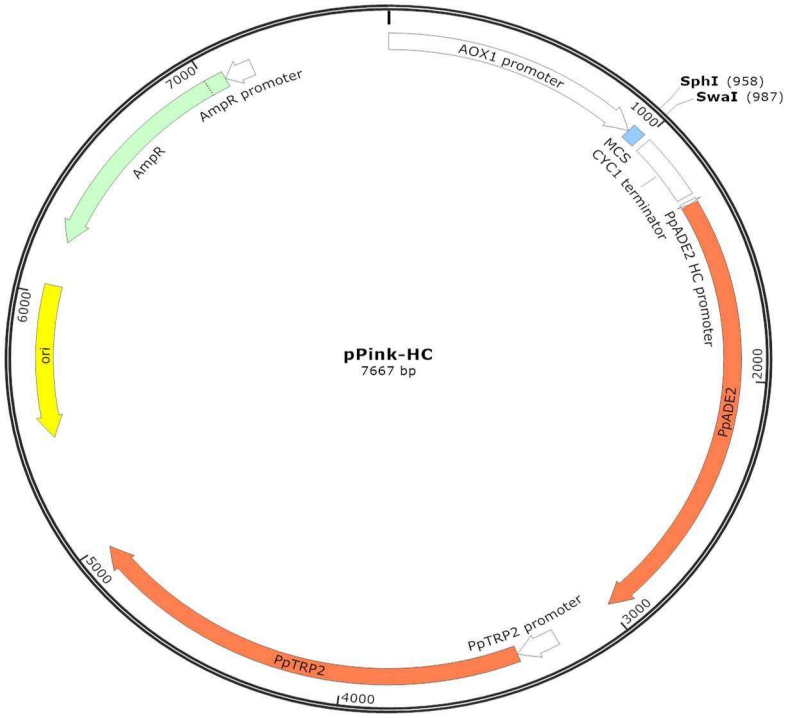

Figure 4-3: Pichia Pink™-HC vector map. The AOX1 promoter is highlighted along with other features of the plasmid including the ADE2 gene promoter. The SphI and SwaI restriction sites for use in molecular cloning are indicated at positions 958 and 987, respectively. Created with SnapGene®.

4.3 Construct Assembly

The lpFC fragments ordered from Genewiz Inc. (Chapter 2), one containing the signal sequences, tags and N-terminal region of lpFC and one containing the C-terminal region of lpFC, were received in powdered form at a concentration of 4 µg each. They were made to a final concentration of 10 µg/ml by reconstituting with TE buffer (10 mM Tris, 1 mM EDTA, pH8), which was used to solubilise the DNA whilst protecting it from degradation. The fragments – lpFC N-terminal plus signal sequences and lpFC C-terminal – were transformed into Subcloning Efficiency™ DH5α™ Competent Cells (Invitrogen™) and overnight cultures were set up in order to purify plasmid DNA using Wizard® Plus SV Minipreps DNA Purification Systems Kit (Promega). The overnight cultures were also used to prepare glycerol stocks for long term storage of the constructs.

4.3.1 Vector Preparation

pcDNA™5/FRT/TO (Invitrogen™) was transformed into Subcloning Efficiency™ DH5α™ Competent Cells and plasmid DNA was purified using the Wizard® Plus SV Minipreps DNA purification system (Promega). DNA was linearised for insertion of the

lpFC gene by a restriction digest using HindIII (2.5 U/µg, NEB) and NotI-HF (5 U/µg, NEB) in a mixture containing 1 X NEB Buffer 2 and 1 X BSA (100 µg/ml) alongside the restriction enzymes and DNA. The digestion was incubated at 37°C for 1 hour before vector dephosphorylation by adding 0.5 U/µg Alkaline Phosphatase, Calf Intestinal (CIP) (NEB), which prevents self-ligation by catalysing the removal of 5' phosphate groups. Linearised DNA fragments were separated by agarose gel electrophoresis (Figure 4-4) and the vector band (5063 bp) was cut out for gel extraction (Macherey-Nagel).

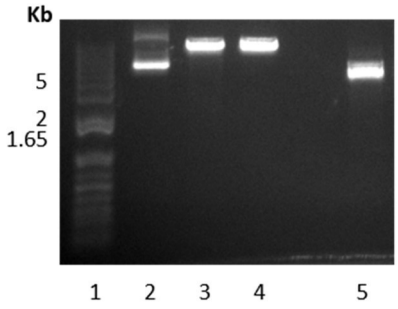

Figure 4-4: Agarose gel electrophoresis of linearised pcDNA5™/FRT/TO. The gel reveals digestion by HindIII and NotI has been successful in linearising the pcDNA5 vector. 1. 1 Kb ladder (Invitrogen), 2. Uncut pcDNA5™/FRT/TO, 3. HindIII single digest, 4. NotI-HF single digest, 5. HindIII and NotI-HF double digest.

4.3.2 Insert Preparation

In order to prepare the two fragments of Factor C to enable them to be ligated together to obtain the full-length sequence, both fragments had to be cut out of the vector they were supplied in (pUC57-kanamycin). The 5' fragment encoding the N-terminal part of the protein including the tags (lpFC-N) was digested using HindIII (5 U/µg, NEB) and SpeI (5 U/µg, NEB) in a mixture containing 1 X NEB Buffer 2 and 1 X BSA (100 µg/ml). The digestion was incubated at 37°C before being analysed by gel electrophoresis to separate the lpFC-N fragment from the pUC57 vector (results not shown). The correct insert band (1630 bp) was cut out for gel extraction (Macherey-Nagel).

The 3' fragment encoding the C-terminal part of the protein (lpFC-C) was cut out from pUC57-kanamycin using SpeI (5 U/µg, NEB) and NotI-HF (5 U/µg, NEB) with NEB CutSmart™ buffer. The mixture was incubated at 37°C for 1 hour before fragments were separated by agarose gel electrophoresis (results not shown) and the insert band (2316 bp) was cut out for gel extraction (Macherey-Nagel).

4.3.3 Vector and Insert Ligation

After successful isolation of the fragments and vector backbone, a three-way ligation was carried out using a 1:3:3 molar ratio (0.025 pmol vector: 0.076 pmol lpFC-N insert: 0.076 pmol lpFC-C insert) in a reaction containing 1 X T4 DNA ligase buffer and T4 DNA ligase (400 U, NEB). The mixture was incubated at room temperature for 10 minutes before being transformed into Subcloning Efficiency™ DH5α™ Competent Cells and plated onto amp-agar plates. Colonies were checked for the correct clones by digestion using HindIII (2.5 U/µg) and NotI-HF (5 U/µg), 1 X NEB buffer 2 and 1 x BSA (100 µg/ml), incubated at 37°C for 1 hour and the correct sequence was confirmed by sequencing using the CMV forward primer and BGH reverse primer (Chapter 3). This construct was the starting point for introduction of lpFC into the vectors for the other expression systems.

4.4 Mammalian

As described above, the lpFC gene was successfully cloned into the pcDNA™5/FRT/TO vector, but with both the insect and the *pichia* signal sequences still present. It had been intended that the insect and *pichia* signal sequences would be cut out from the pcDNA5 lpFC construct using restriction enzymes AseI and NdeI. This would allow for religation of the pcDNA5 lpFC construct, with just the mammalian signal sequence present. However, it was discovered that five AseI and four NdeI sites exist in the pcDNA5 vector, which resulted in the presence of several bands as shown by agarose gel electrophoresis, indicating the construct was being digested into lots of smaller fragments. In an effort to resolve these issues, an alternative cloning strategy was attempted, which involved the use of the pCR2.1 vector. lpFC with all three signal sequences was to be cloned into pCR2.1 and subsequent cloning steps would be taken to isolate the individual signal sequences for ligation with the lpFC into the correct expression vectors.

However, ligation of lpFC into pcR2.1 failed, with sequencing results showing empty vector or sequencing errors. A new strategy was developed to produce a linear version of the plasmid with homologous ends that could be re-joined using InFusion® cloning to produce pcDNA5 lpFC with only the mammalian sequence. The experimental design is illustrated in Figure 4-5 and the primers used f or the amplification are shown in Table 4-1.

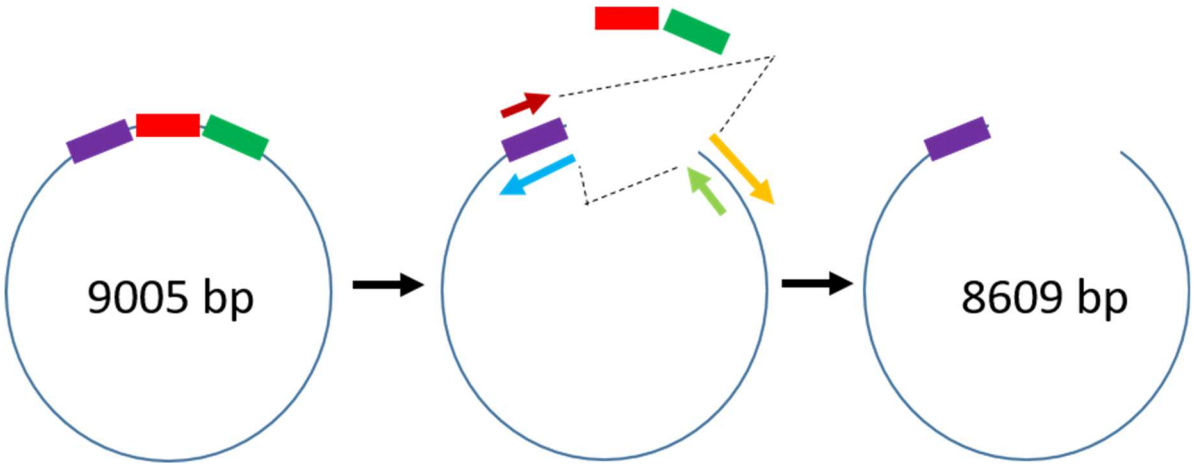

Figure 4-5: Schematic representation of experimental design for pcDNA5 lpFC production. Signal sequences are indicated by the coloured boxes: Ig κ-chain (purple), *Saccharomyces cerevisiae* α-mating factor pre-sequence (red) and gp67 (green). The coloured arrows represent the part of the primer sequence shown in Table 4-1.

5' → 3'	TGGTGACATGCACCATCATC
3' → 5'	GGTGCATGTCACCAGTGGAA

Table 4-1: Primers for the amplification of lpFC with the mammalian signal sequence. The primers are colour coded in line with Figure 4-5 indicating the part of the sequence that corresponds to the primer indicated by the arrows on the figure.

The PCR program used to amplify pcDNA5 with the lpFC gene and no other signal sequences is outlined in Table 4-2. The mixture contained 1 X Pfu DNA polymerase buffer, dNTPs (20 mM), forward primer (10 µM), reverse primer (10 µM), DMSO, DNA (25 ng pcDNA5 lpFC) and Pfu DNA polymerase (2.5 U) after a hot start at 94°C for 5 minutes.

Number of Cycles	**Temperature**	**Length**	**Step**
1	95°C	2 minutes	
30	95°C 53°C 72°C	30 seconds 30 seconds 7 minutes	Denaturation Annealing Extension
1	72°C 4°C	10 minutes Hold	

Table 4-2: PCR program for amplification of pcDNA5/FRT/TO, lpFC and mammalian signal sequence.

DpnI digestion was carried out at 37°C for 1 hour before PCR-clean up using Macherey Nagel's NucleoSpin® Gel and PCR Clean-up kit. InFusion® cloning was used to religate the construct, which was transformed into Stellar™ competent cells and mini-prepped to get purified DNA. DNA sequencing was used to confirm the correct sequence. However, time did not allow this construct to be taken further.

4.5 Insect

4.5.1 Vector preparation

After transformation into Subcloning Efficiency™ DH5α™ Competent Cells and purification of plasmid DNA using Wizard® Plus SV Minipreps DNA purification system (Promega), pVL1392 was linearised with PstI (10 U/µg, NEB) and BamHI-HF (5 U/µg, NEB) for 1 hour at 37°C in a mixture containing the enzymes, vector DNA, 1 X NEB Buffer 4 and 1 X BSA (100 µg/ml). The linearised vector was dephosphorylated as described previously (section 4.3.1). The bands were separated by agarose gel (results not shown) and the linearised vector (9598 bp) was cut out and gel extracted (Macherey-Nagel).

4.5.2 Insert Preparation

The pcDNA™5/FRT/TO lpFC construct described in section 4.3.1 was used to cut out the lpFC synthetic cDNA containing only the insect signal sequence (gp67). Digestion was carried out with Pst-I (5 U/µg) and BamHI-HF (5 U/µg) and samples were analysed by agarose gel electrophoresis (Figure 4-6), which revealed three distinct bands. The correct band (3575 bp) was cut out for gel extraction (Macherey-Nagel), and used for ligation into linearised pVL1392 in a reaction containing 0.076 pmol insert DNA, 0.025 pmol vector DNA, 1 X T4 DNA ligase buffer and 400 U T4 DNA ligase. The ligation mixture was incubated overnight at 16°C before being transformed into Subcloning Efficiency™ DH5α™

Competent Cells and plated onto amp-containing agar. The correct sequence was determined by DNA sequencing.

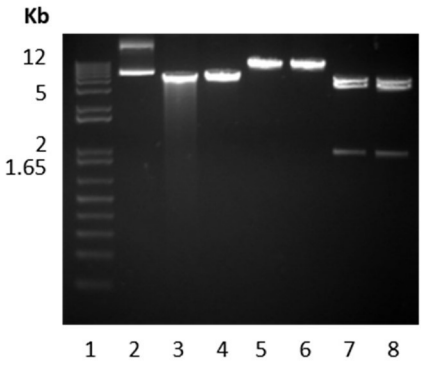

Figure 4-6: Agarose gel electrophoresis of digested lpFC with gp67. In the double digest, the band second from the top (3575 bp) was the correct size for the lpFC with gp67. 1. 1 Kb DNA ladder (Invitrogen), 2. Uncut pcDNA5 lpFC, 3. 5 U PstI single digest, 4. 10 U PstI single digest,

5. 5 U BamHI-HF single digest, 6. 10 U BamHI-HF single digest, 7. 5 U PstI and BamHI-HF double digest, 8. 10 U PstI and BamHI-HF double digest.

4.5.3 Insect Cell Expression

For insect cell expression, the BacMagic™ DNA Kit (Novagen) was used to perform positive selection for recombinant *baculovirus*. The lpFC gene was cloned into the PVL1392 compatible transfer plasmid and co-transfected into insect cells with the BacMagic™ DNA. Homologous recombination within the cells generates recombinant virus DNA and the target gene sequence. Virus is produced from replicating recombinant DNA.

According to the manufacturer's instructions, small scale cell cultures were prepared for transfection, co-transfection was carried out and recombinant virus was amplified for 4 – 5 days. The insect cell culture was analysed by Bradford Assay, SDS-PAGE and Western Blot, following standard procedures for each. A band indicating likely expression of lpFC protein was shown and cells were infected for a larger scale expression. Dr Jan Petersen infected Sf9 cells to express protein. Cells were harvested and again analysed by SDS-PAGE and Western Blot for the presence of protein. Ni^{2+} affinity chromatography (Chapter 3) was used for purification of the protein sample, however this gave rise to uncertainty as to whether or not the protein of interest was being expressed and so a change in medium to BacVector™ insect cell medium (serum-free) occurred. Analysis revealed that the band shown in SDSPAGE and Western Blot was not lpFC, but likely to be a chaperone from the medium, and further experiments with insect cell expression were not taken any further, due to time constraints.

4.6 Yeast

4.6.1 Vector Preparation

pPink™-HC was transformed into Subcloning Efficiency™ DH5α™ Competent Cells and the Wizard® Plus SV Minipreps DNA purification system (Promega) was used to purify plasmid DNA. Linearisation was achieved by sequential restriction digests, first using SwaI (2.5 U/µg, NEB) in a reaction mixture containing 1 X NEB buffer 3.1 and 1 X BSA (100 µg/ml) along with the enzyme and DNA, at 25°C for 1 hour before heat inactivation at 65°C

for 20 minutes and DNA clean-up using Macherey-Nagel's NucleoSpin® Gel and PCR Clean-up kit. The vector was then digested with SphI-HF (5 U/µg, NEB) in a reaction

containing 1 X NEB Buffer 4, 1 X BSA (100 µg/ml), the enzyme and DNA at 37°C for 1 hour before vector dephosphorylation (see section 4.3.1). Linearised DNA was analysed by agarose gel electrophoresis (results not shown), and the resulting band (7638 bp) was extracted for purification by the NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel).

4.6.2 Insert Preparation

As with the mammalian insert preparation, design flaws required a PCR based strategy to be adopted to amplify the *pichia* signal sequence and the lpFC gene as fragments with homologous ends that would allow for the InFusion® cloning of the two fragments into the linearised pPink-HC vector. The design is set out in Figure 4-7 and the primers are listed in Table 4-3.

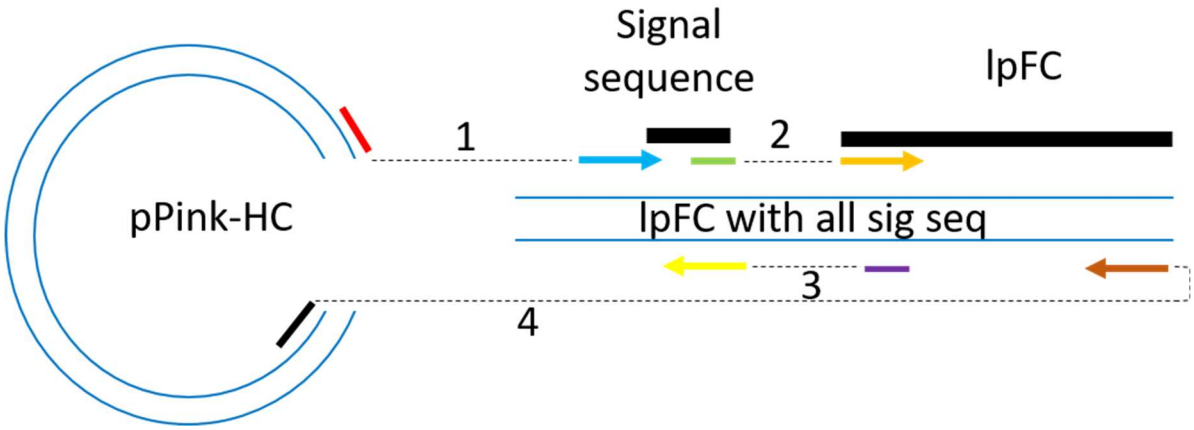

Figure 4-7: Schematic representation of the experimental design for amplification of pPink signal sequence and lpFC. Primers indicated by the broken arrows are numbered and colour coded to relate to the primer sequence highlighted in Table 4-3.

Number	Amplification section	Primer sequence
1	5' → 3' vector and signal sequence	TCCGGACCGGCATGCATGAGATTTCCTTCA
2	5' → 3' signal sequence and lpFC	GAAAAGGCATATGCACCATCAT
3	3' → 5' lpFC and signal sequence	GTGCATATGCCTTTTCTCGAGAGA
4	3' → 5' vector and lpFC	AAAAGGGGCCTGTATTTAAATGGATCCCTAGAT

Table 4-3: Primer sequences for pPink lpFC preparation. The primers are designed to amplify the signal sequence and lpFC independently of each other with homologous ends to each other and to the pPink-HC vector, for 3-way ligation by InFusion® cloning. Primer sequences are colour coded to match the position they relate to on the schematic in Figure 4-7.

4.6.2.1 Signal Sequence

Saccharomyces cerevisiae α-mating factor pre-sequence was amplified from pcDNA5/FRT/TO lpFC in a reaction containing 1 X Pfu DNA polymerase buffer, dNTPs (25 mM each NTP), primer 1 (0.2 µM), primer 3 (0.2 µM) and Pfu DNA polymerase (2.5 U). The PCR program used to amplify the sequence was similar to that outlined in Table 4-2 but with 30 cycles at 95°C for 30 seconds, followed by 53°C for 30 seconds and then 72°C for 30 seconds.

4.6.2.2 lpFC

LpFC was amplified for insertion into the pPink™-HC vector from the pcDNA5/FRT/TO lpFC construct in a reaction containing 1 X Pfu DNA polymerase buffer, dNTPs (25 mM each NTP), primer 2 (0.2 µM), primer 4 (0.2 µM) and Pfu DNA polymerase (2.5 U). The PCR program used to amplify the sequence lpFC fragment followed that shown in Table 4-2 except there was 30 cycles at 95°C for 30 seconds, 50°C for 30 seconds and an extension at 72°C for 3.5 minutes. The amplified fragments were analysed by agarose gel electrophoresis as shown in Figure 4-8.

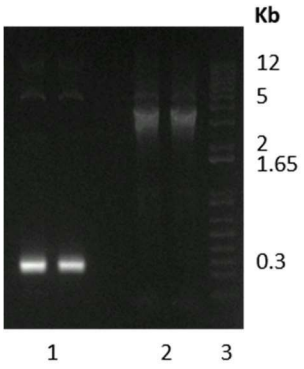

Figure 4-8: pPink PCR products. The bands indicate successful amplification of the α-mating factor pre-sequence and lpFC for cloning into the pPink vector. 1. P. pastoris signal sequence (307 bp), 2. Amplified lpFC (3546 bp), 3. 1 Kb DNA Plus ladder (Invitrogen).

4.6.3 In-Fusion® HD EcoDry™ Cloning

In-Fusion® HD EcoDry™ cloning was used as described in Chapter 3 to clone the α-mating factor pre-sequence (20 ng) and lpFC (20 ng), amplified with the specific primers, into the pPink™-HC vector digested with SwaI and SphI-HF (50 ng). After incubating at 37°C for 15 minutes followed by 50°C for 15 minutes, the mixture was transformed into Stellar competent cells (Clontech) and plated onto amp-containing agar.

Colonies were mini-prepped and the sequence was determined using the AOXI 5' sequencing primer and the CycI 3' sequencing primer. It became clear that during the design and preparation of the sequence for yeast expression, the 'Kozak' sequence (AGATTTCCTTCA), which plays a major role in the initiation of translation, had been omitted. Primers were designed in order to introduce this into the final sequence by InFusion® HD EcoDry™ cloning. The experimental procedure involved linearising the pPink lpFC construct using SphI-HF before In-Fusion® cloning with the two primers (kindly performed by Donald Campbell). DNA sequencing confirmed that the correct sequence including the Kozak sequence had been produced, however, time did not allow for this to be taken any further.

4.7 *Brevibacillus choshinensis*

Brevibacillus choshinensis is a non-pathogenic Gram positive bacterium with low natural levels of secreted proteases, which can be used for heterologous protein expression (D'Urzo *et al.*, 2013). Being Gram-positive, it produces no lipopolysaccharide, which would be a distinct advantage for the production and purification of Factor C fragments (Udaka and Yamagata, 1993). Secretion of proteins into the extracellular medium is efficient and thus the organism is well suited to the production of disulphide bonded proteins such as Factor C (Tanio *et al.*, 2008). *Brevibacillus choshinensis* can be grown on defined growth media and has previously been used to produce isotopically labelled protein for NMR analysis (Tanio *et al.*, 2008). The BIC (*Brevibacillus* in vivo cloning) system provides an efficient method by which secreted proteins can be produced.

Primers were designed for the production of lpFC CysEGF, lpFC CCP12, lpFC CCP23 and lpFC CCP123 in the BIC system. These primers include 15 bp of sequence that is homologous to the ends of the linear vector. The pBIC vectors are provided in linear form and allow for the amplified insert DNA to be inserted straight into the vector of choice, without the need for restriction digests. The full primer sequences are shown in Table 4-4. The fragments were successfully cloned into pBIC-4 and initial test expressions showed promising expression of soluble protein (experiments kindly performed by Donald Campbell).

Amplification section	Primer sequence
5’ lpCysEGF	GATGACGATGACAAACAGCAGATGCACCCAGTGC
3’ lpCysEGF	**CATCCTGTTAAGCTT**TCAGCCTTCATATCGGTCA
5’ CCP1	GATGACGATGACAAAGAGATACTCCAGGGCTGC
3’ CCP2	**CATCCTGTTAAGCTT**TTACTGTTTCTGGCATTGGGGAATTTG
5’ CCP2	GATGACGATGACAAAATCAGGGAATGCAGCATG
3’ CCP3	**CATCCTGTTAAGCTT**TTACTGCTCCCGGTCAGCTAC

Table 4-4: Primer sequences for insertion into the pBic vector. The underlined sequences indicate the 15 bp needed for the 5’ primers. The sequences in bold indicate the 15 bp required for the 3’ primers.

4.8 Summary

Full-length lpFC was successfully cloned into the selected mammalian, insect and yeast expression vectors. Although initial expression tests were performed for insect cell expression, experimental optimisation is needed to ensure the protein being produced is lpFC. Expression of protein in all three eukaryotic expression systems will determine the most suitable system for production of full-length lpFC. For the BIC expression system, the next stage would be to perform a large-scale expression of the constructs that have been produced to determine whether this expression system has advantages over other recombinant expression systems.

5 Structural and Functional Analysis – Circular Dichroism

5.1 Overview

Circular dichroism (CD) is a technique that can be used for the study of proteins to elicit information about structure, stability and interactions with other molecules. CD is the differential absorption of left and right circularly polarised light (CPL) by chiral molecules. CD spectropolarimeters measure the difference between right-handed, clockwise rotations and left-handed, anticlockwise rotations. CD signals are observed in proteins due to the chirality of the peptide bond and contributions from aromatic amino acids or other chiral cofactors present in the protein. The CD effect is most commonly measured by modulation (used for the experiments described below) where signals are received from switching between the L and R components. This occurs from passing the light through a modulator, exposed to an alternating electric field at 50 kHz, that is made up of a piezoelectric quartz crystal and a thin plate of isotropic material, in this case, quartz silica. When measured as a function of wavelength, a CD spectrum is generated and from this several observations can be made (Kelly *et al.*, 2005).

CD signals at specific wavelengths can be used to identify different structural characteristics of the protein being studied. In the far UV region (260 – 190 nm), spectral features give an indication of the types of secondary structures, such as alpha helices, beta sheets and extended coil (or random coil) present within the protein. It is also possible to observe contributions from aromatic amino acids in the far UV region, however, these contributions are often very weak and can be cancelled out by the large contributions associated with alpha helical content. Absorption in the range 260 – 320 nm can report on contributions from aromatic amino acid side chains if they are held rigidly within the protein. Broad signals around 260 nm have been associated with disulphide bond contributions. Specifically, for aromatics, a peak near 290 nm, as well as data between 290 and 305 nm, is usually indicative of contributions from Tryptophan residue(s). Tyrosine is shown by the presence of a peak between 275 and 282 nm and also a shoulder if this hasn’t been concealed by the Tryptophan. Phenylalanine reveals sharp but weak bands between 255 and 270 nm. A near UV (320 – 250 nm) CD spectrum depends on the number and type of any aromatic amino acids present and other features of these including their mobility and location within the protein.

Due to the nature of the experiments carried out and results gained from CD, conformational changes as a result of, for example, ligand binding, can be easily detected. Therefore, CD is an appropriate technique to use in determining structural changes as a result of molecular interactions and therefore binding of LPS to Factor C protein fragments.

As a low-resolution technique, CD is a more straightforward method for obtaining structural information about proteins in comparison to high-resolution techniques such as NMR spectroscopy and X-ray crystallography. The results garnered from CD can give invaluable information about the sample.

5.2 CD Experimentation

CD was carried out for this project primarily as a structural comparison between the recombinant fragments produced. It was also used as a means to determine whether protein domains were correctly folded, to identify the secondary structure contributions from each of the samples and to detect putative conformational changes after the addition of lipopolysaccharide (LPS). All experiments were kindly carried out by Dr Sharon Kelly, using the Jasco J-810 spectropolarimeter and samples were measured in quartz cuvettes of varying path length, as indicated below for each construct along with the experimental parameters used (Table 5-1). For the most accurate data collection, it is imperative that the spectropolarimeter is correctly calibrated and operated and it is also important that the protein concentration has been calculated properly.

Samples must be free from contaminants and any components that give strong signals, including imidazole, that would mask protein signals. There should be no aggregates, which would interfere with light scattering and absorption, affecting the shape and size of the spectrum. Buffer blanks are measured in order to check for any contributions that will need to be subtracted from the resulting spectra and also to ensure the buffer does not give a high absorbance.

Instrument Experimental Parameters	
Data array type	Linear data array * 3
Near UV Start	320 nm
Near UV End	250 nm
Far UV Start	260 nm
Far UV End	190 nm
Data pitch	0.2 nm
Data points	351
Band width	1 nm
Response	2 seconds
Sensitivity	Standard
Scanning speed	10 nm/min
Accumulation	3
Cell length Near UV	0.2 or 0.5 cm
Cell length Far UV	0.1 or .02 cm
Solvent	10 mM Tris, pH 8.0
Temperature	Room temperature

Table 5-1: CD parameters for data acquisition by near and far UV.

5.3 CD Data Analysis

DICHROWEB (Lobley *et al.*, 2002; Whitmore and Wallace, 2004; Whitmore and Wallace, 2008), hosted by Birkbeck College, University of London, provides online analysis for protein spectra from CD experiments. The percentage contribution of several types of secondary structure is calculated including for α-helices, β-sheets and β-turns. Numerous algorithms and reference databases made up of CD spectra from a number of proteins of various structures determined by X-ray crystallography and/or NMR, are used to gain sufficient information to calculate the NRMSD (normalised root mean square deviation), which represents a standard 'goodness-of-fit' parameter. This allows for accurate comparisons to be made between experimental data and previously calculated spectra that have been stored in the reference databases. A perfect fit is indicated by a score of 0, whereas no fit whatsoever gives a score of 1. Several algorithms are available for use, but for the experiments described, analysis was carried out using the CONTIN-LL database and algorithms, which is a variant of Provencher and Glöckner's CONTIN method (Provencher and Glöckner, 1981; Sreerama and Woody, 2000).

5.4 Experimental results for Factor C constructs

For most of the constructs produced, far UV and near UV CD spectra were measured to obtain an estimate of secondary and tertiary structural content using the parameters outlined in Table 5-1. For near UV and far UV CD measurements, protein concentrations of between 0.5 – 1 mg/ml were used. All samples were prepared to 10 mM Tris, pH 8.0. Cells of path length 0.01 cm and 0.02 cm were used for far UV CD spectral measurements. For near UV CD measurements, cells of path length 0.2 – 0.5 cm were used. All spectra were corrected for buffer contribution, protein concentration and cell path length. Contributions from unbound LPS were removed by subtracting spectra for buffer and LPS in isolation from the spectra of protein and LPS together. CCP1 and CCP2 preparations were found to be cloudy following purification, which precluded CD measurement due to excessive light scattering. An overview of the estimates of secondary structure for each Factor C construct is shown in Table 5-2.

Construct	% alpha helix	% beta sheet	% Turns	% Unordered	NRMSD
Cys-rich	24.2	25.2	12.4	38.2	0.051
Cys-rich LPS	25.6	23.4	12.7	38.3	0.042
EGF-like	10.5	33.6	13.5	42.4	0.055
EGF-like LPS	9.2	34.1	13.4	43.3	0.06
Cys-EGF	24.2	25.1	20.9	29.8	0.06
CCP3	18.9	29.7	12.5	38.9	0.107*
CCP3 LPS	18.5	29.8	12.4	39.3	0.114*
CCP12	5.0	41.5	12	41.5	0.057
CCP12 +LPS	5.0	39.0	13	43.0	0.06
CCP23	8.1	35.6	13.4	42.9	0.120*
CCP23 + LPS	8.0	35.6	13.4	43.0	0.123*
CCP123	12.2	28.6	14.9	44.3	0.067
CCP123 + LPS	12.7	28.5	15.7	47.1	0.063

Table 5-2: Secondary structure estimates of Factor C constructs (* denotes poor fit as judged by normalised root mean squared (NRMSD) value).

5.4.1 Cys-rich

The far UV CD spectra of the Cys-rich region in the absence and presence of LPS is shown in Figure 5-1. The spectra were predominantly superimposable indicating that no marked changes in the secondary structural contents in the presence of LPS took place. The spectral data of each were submitted to DICHROWEB, as described above.

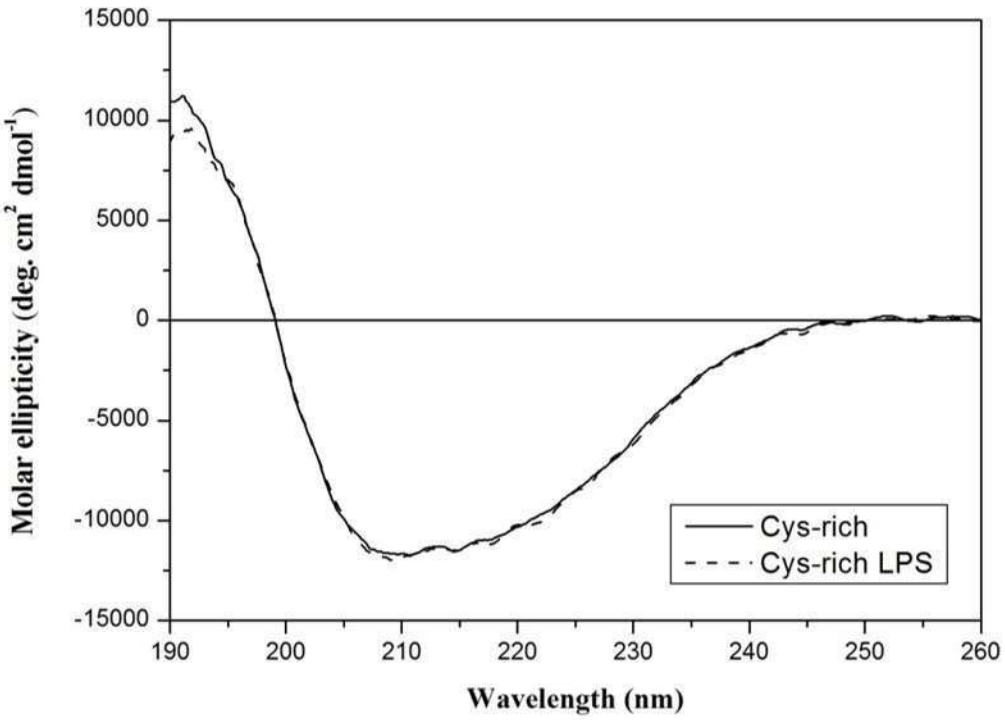

Figure 5-1: Far UV CD spectra of Cys-rich in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

The secondary structure of the Cys-rich construct was estimated to contain 24.2% α-helix and 25.1% beta sheet content. A large proportion of the construct (38.2%) was estimated to be unordered (Table 5-2). Overall, the NRMSD value indicates that the goodness of fit was considered to be reliable. The secondary structural estimates obtained for Cys-rich in the presence of LPS were comparable. Slight differences were observed in the spectral data around 192 nm, which may be due to the slightly higher instrument voltage values required to maintain a constant current in the presence of LPS.

The near UV CD data obtained for the Cys-rich construct in the absence and presence of LPS were of poor quality in terms of their signal to noise ratios and did not exhibit any of the characteristic peaks expected from aromatic residues that are held in a rigid position with the 3D structure of a protein (Figure 5-2). This suggests that any aromatic residues in this construct are probably contained within a flexible region of the structure.

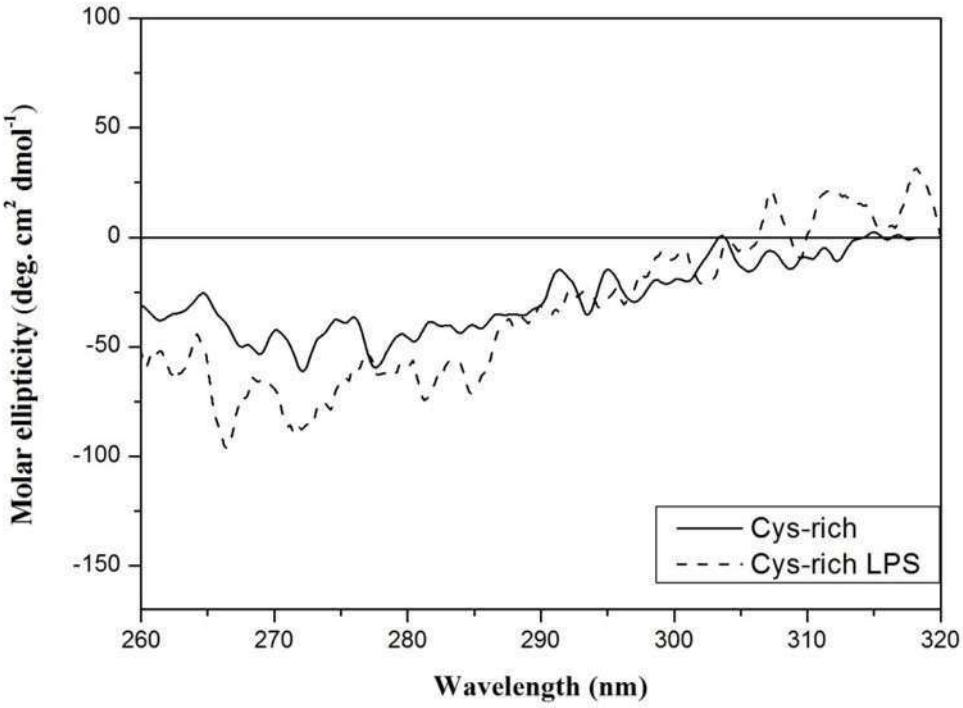

Figure 5-2: Near UV CD spectra of Cys-rich in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

5.4.2 EGF-like

The far and near UV CD spectra of the EGF-like construct in the absence and presence of LPS are shown in Figure 5-3 and Figure 5-4, respectively. LPS did not appear to influence the spectral intensity or shape over the ranges measured, which indicates that there were no significant effects on the secondary or tertiary structure of this construct. The far UV CD spectrum gave a small positive peak around 230 nm, which most likely reflects the contribution from a tryptophan residue. Aromatic contributions in the far UV region are usually masked by the larger contributions from alpha helical structure. The overall shape of the spectrum from the EGF-like construct indicates that there is probably only a small amount of alpha helical structure present in this construct. The large negative minimum around 202 nm suggests a high proportion of unordered structure. The presence of aromatic contributions in the far UV and the presence of characteristic aromatic peaks in the near UV (centred between 280 – 290 nm) indicate that the construct must contain a region with a rigid structure. The secondary structure estimates shown in Table 5-2 indicate that this construct contains around 10% alpha helix, 33.6% beta sheet and 42.4% unordered structure.

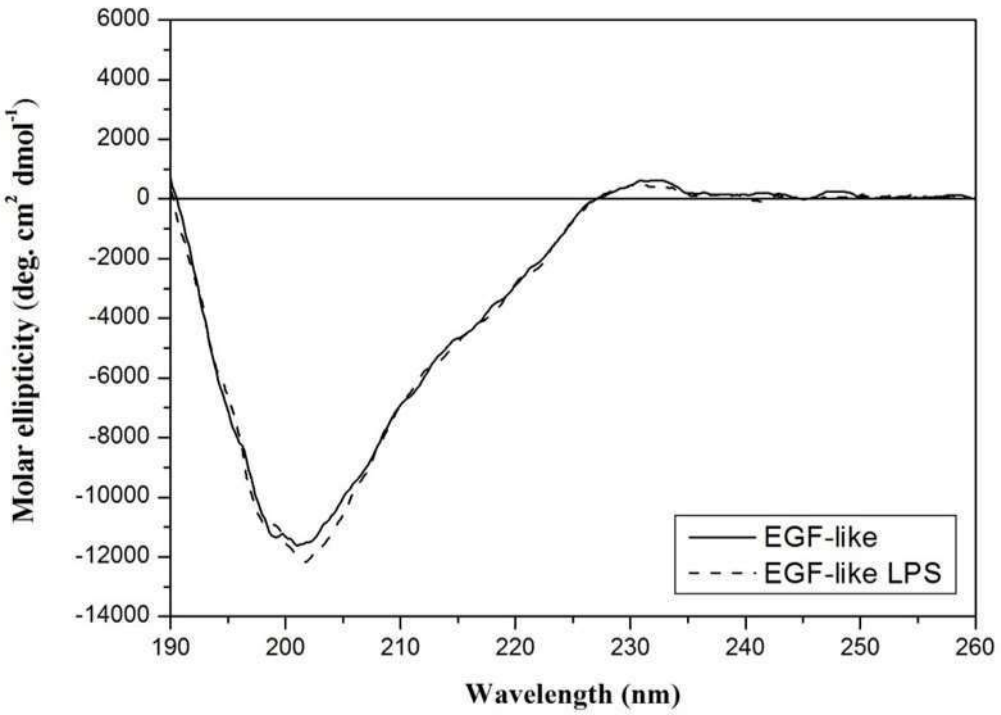

Figure 5-3: Far UV CD spectra of EGF-like in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

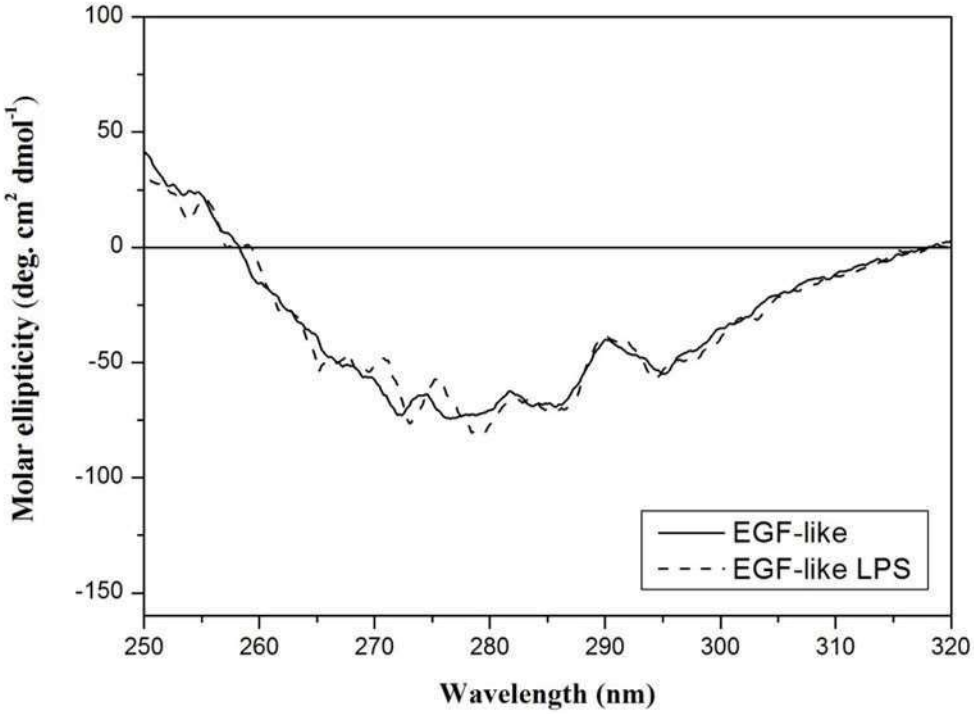

Figure 5-4: Near UV CD spectra of EGF-like in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

5.4.3 CysEGF

Figure 5-5 shows the far UV CD spectra of CysEGF. The far UV CD spectra were recorded in the absence and presence of LPS. The spectra obtained and the secondary structural analysis of the data suggest that the construct undergoes a conformational change upon binding of LPS. From Table 5-2 it can be seen that the construct contains around 22.2% alpha helical structure, 24.4% beta sheet and 31.7% unordered structure. The addition of LPS appears to result in a slight increase in helical content of ~ 2% and a concomitant slight decrease in unordered structure.

There were no significant aromatic contributions evident in the near UV CD region, which indicates that either the tertiary structure surrounding the tryptophan residue is probably flexible or that there was insufficient protein concentration to observe these aromatic contributions. Due to insufficient volume and concentration of this construct it was not possible to measure the effect of the addition of LPS in the near UV CD region.

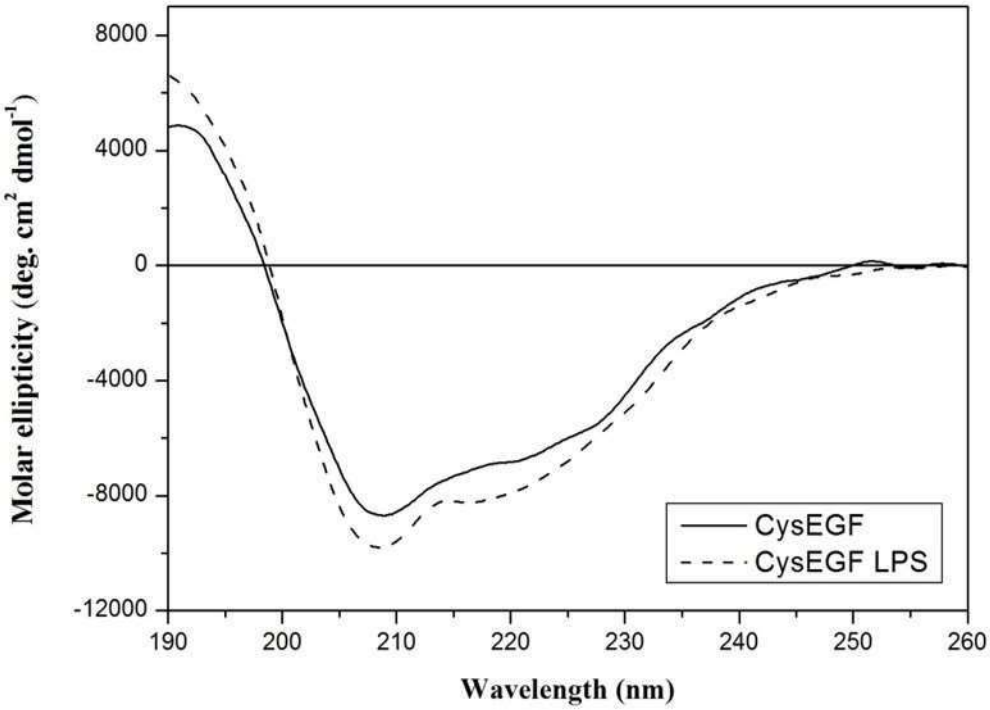

Figure 5-5: Far UV CD spectra of CysEGF in the absence (solid line) and presence (dashed line) of 1 mg/ml LPS

5.4.4 CCP3

It can be seen from Table 5-2 that the CCP3 construct is estimated to contain around 18.9% alpha helical structure, 29.7% beta sheet and 38.9% unordered structure. The presence of LPS did not alter these estimates significantly although the spectral intensities

exhibited a slight reduction between 205 nm and 228 nm (Figure 5-6). Since the amount of sample available for this measurement was limited, it was not possible to obtain the necessary replication to ensure reproducibility in the far UV region. In addition, there was insufficient sample available to obtain a near UV CD spectrum. Given the superimposable spectral shapes obtained between 200 – 205 nm and 230 – 260 nm it is possible that the addition of LPS may have contributed to the minor spectral changes observed in the region associated with alpha helical and beta sheet contributions.

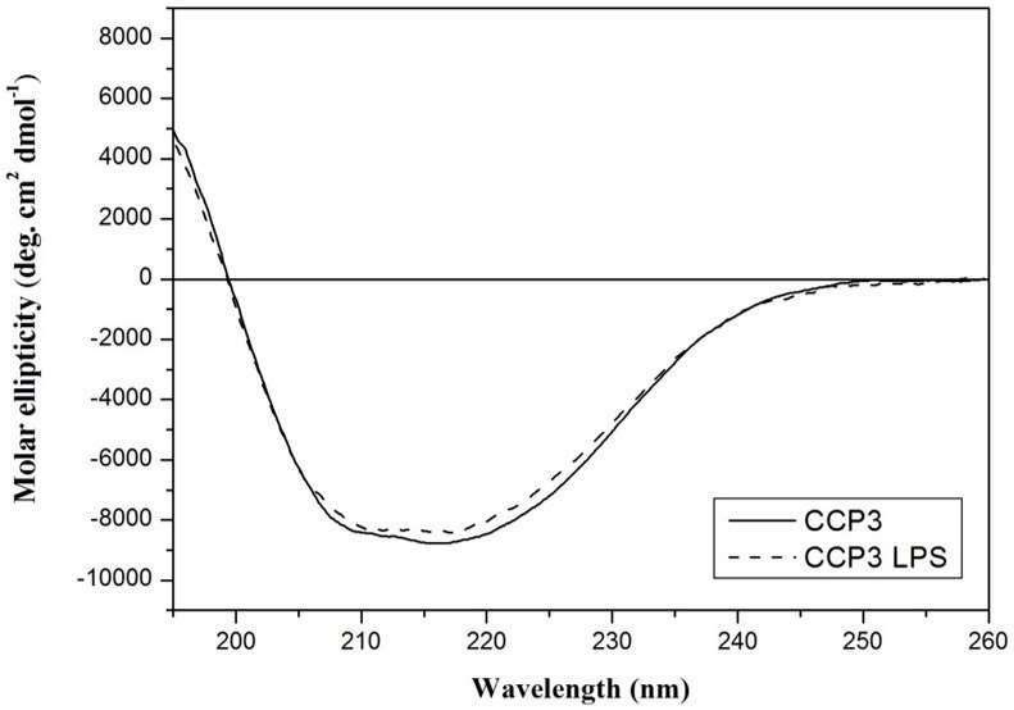

Figure 5-6: Far UV CD spectra of CCP3 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

5.4.5 CCP12

The far and near UV CD spectra of the CCP12 construct in the absence and presence of LPS are shown in Figure 5-7 and Figure 5-8. The far UV CD spectrum exhibits a small positive peak around 235 nm and a large minimum around 195 nm. The lack of minima at 220 nm and 208 nm suggest that there is very little alpha helical content in this construct. The positive peak at 230 nm is indicative of the contribution from the tryptophan residue, which together with the spectral features centred around 285 – 295 nm suggest the tryptophan is held rigidly within the structure of CCP12. Secondary structure estimates were found to give very low helical content (~5%) with the predominant spectral contributions attributed to beta sheet (41%) and unordered structure (41%). When CCP12

was incubated with LPS differences were evident in both the far and near UV CD spectra indicating some secondary and tertiary conformational changes.

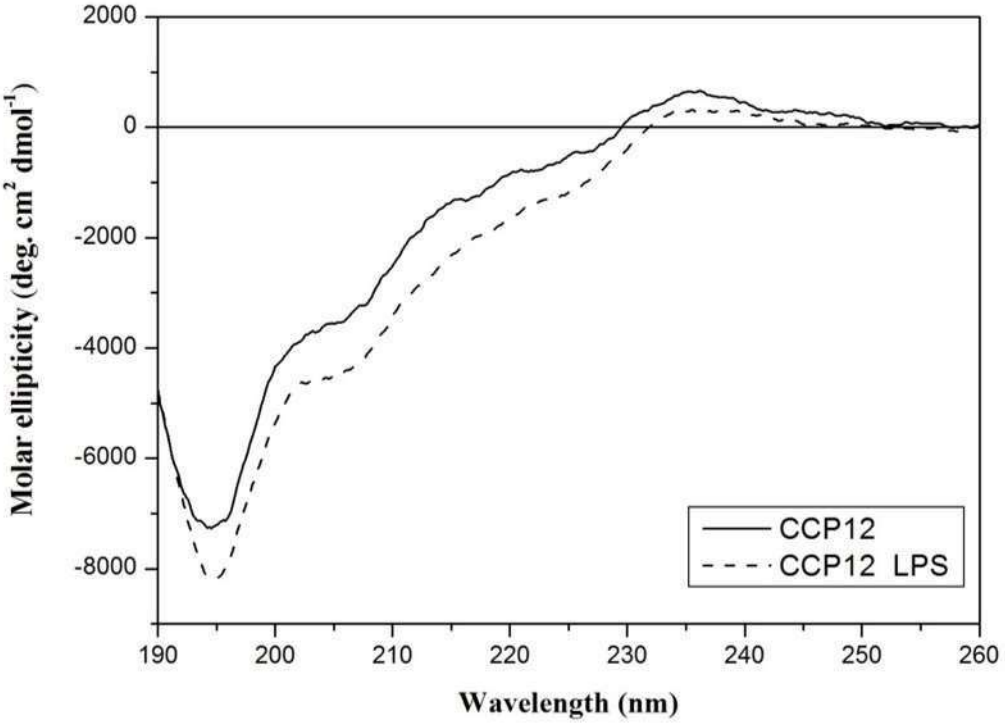

Figure 5-7: Far UV CD spectra of CCP12 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

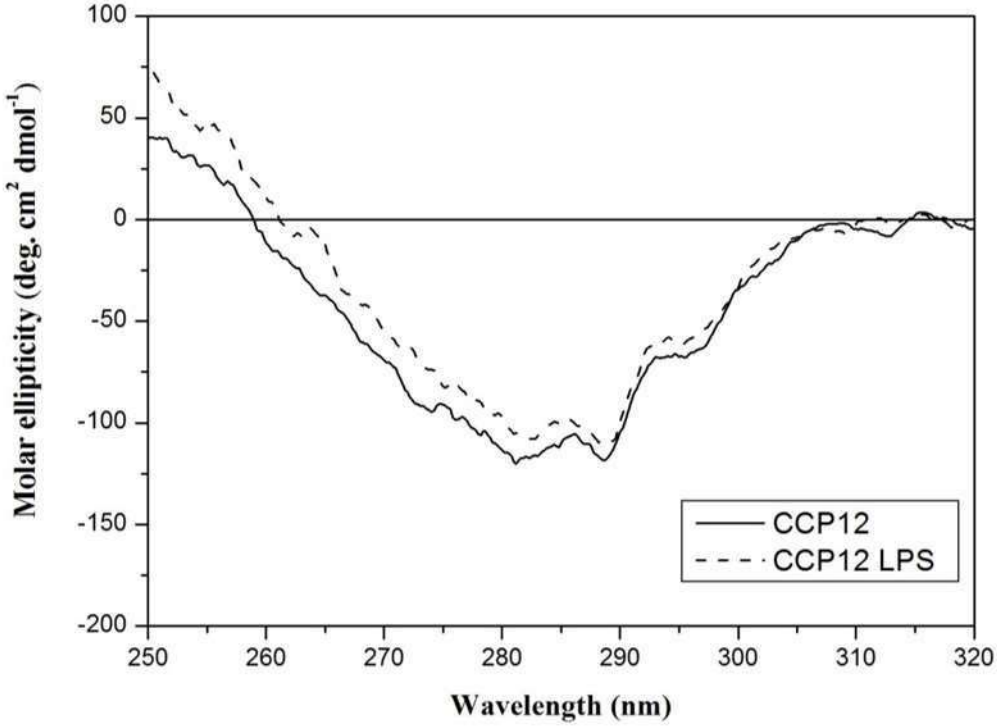

Figure 5-8: Near UV CD spectra of CCP12 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

5.4.6 CCP23

The far and near UV spectra of CCP23 are shown in Figure 5-9 and Figure 5-10. The addition of LPS to the construct did not result in any observable changes in the spectra obtained in the far or near UV regions. The far UV CD spectrum of CCP23 exhibited a minor positive peak centred around 235 nm indicative of aromatic contributions. The shape of the spectrum below 230 nm is unusual with a negative slope down to 210 nm, a small positive shoulder around 205 nm followed by a large minimum centred around 190 nm. The absence of strong negative ellipticities around 220 nm and 208 nm and a large positive peak around 190 nm suggest that the construct contains very little alpha helical structure.

The lack of helical content would account for the observed far UV spectral contributions from aromatic residues held in a rigid asymmetric environment. Spectral features around 235 nm and 205 nm have previously been associated with tryptophan contributions in the enzyme carbonic anhydrase following systematic mutagenesis of the individual tryptophan residues (Freskgard *et al.*, 1994). These results together with the presence of characteristic tryptophan peaks in near UV CD region indicate that the region of the protein that contains the tryptophan residue is rigid and the aromatic residue is held in an asymmetric environment.

The secondary structure estimates obtained for CCP23 had poor NRMSD values, which is probably due to the fact that the observed spectral features do not match well with proteins in the Database. It is clear from the estimates obtained that the construct contains very little alpha helix and that it is composed of mainly beta sheet (35.6%) and unordered structure (42.9%).

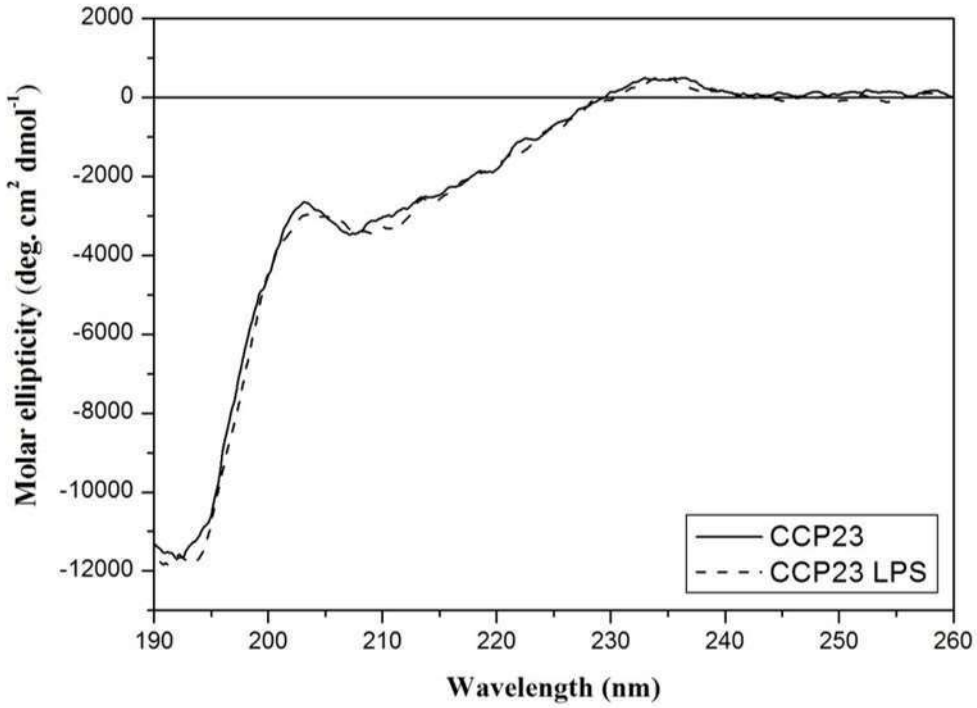

Figure 5-9: Far UV CD spectra of CCP23 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

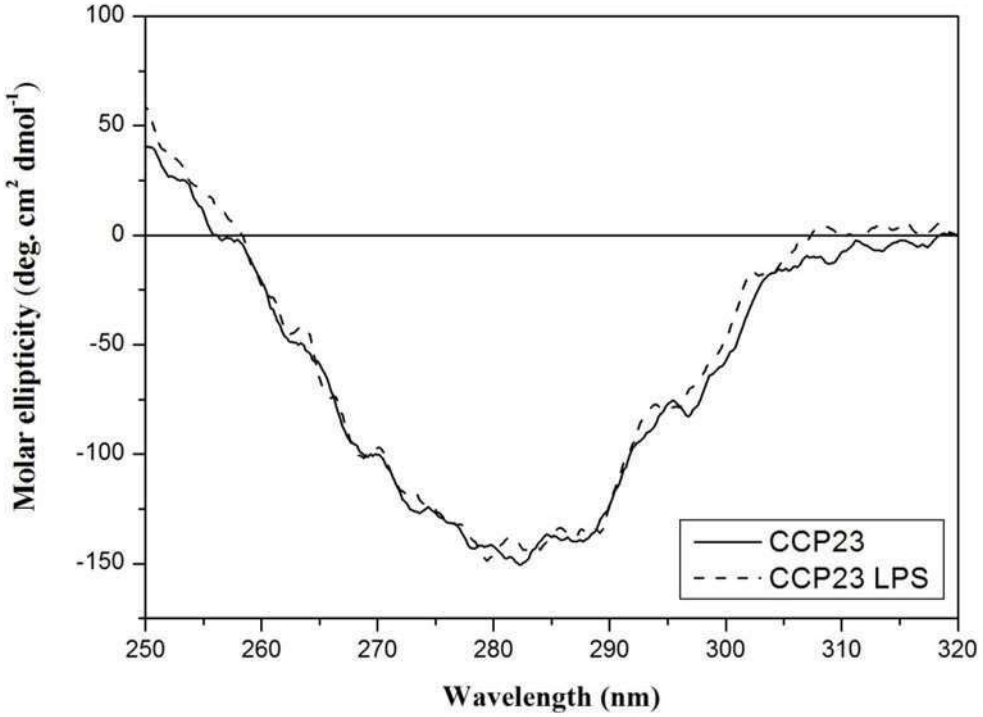

Figure 5-10: Near UV CD spectra of CCP23 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

5.4.7 CCP123

Figure 5-11 and Figure 5-12 show the spectra obtained for the CCP123 construct in the absence and presence of LPS. There were no observable changes in the far UV CD spectrum upon addition of LPS. Secondary structure estimates suggest that the construct contains a low amount of alpha helical structure (~12.2%) and that it predominantly consists of beta sheet and unordered structure (28.6% and 44.3% respectively). Data obtained in the near UV CD region suggest that LPS affects the environment of the aromatic residues. An increase in the intensity of the spectral peaks between 290 – 260 nm can be observed, which could be representative of changes in the local environment of phenylalanine, tyrosine, and/or tryptophan residues.

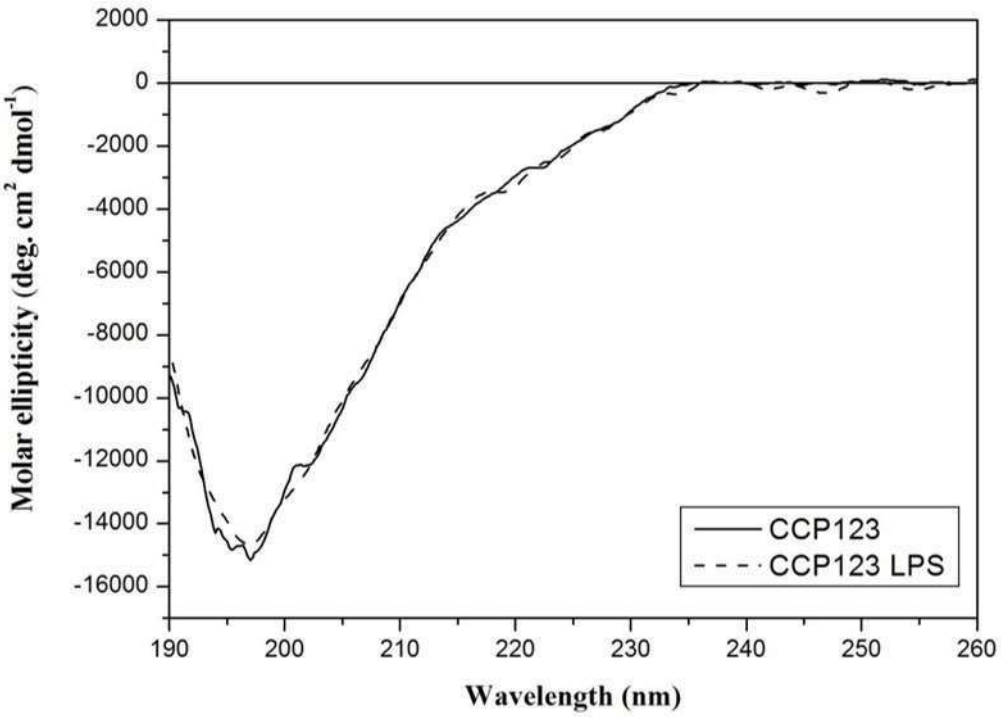

Figure 5-11: Far UV CD spectra of CCP123 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

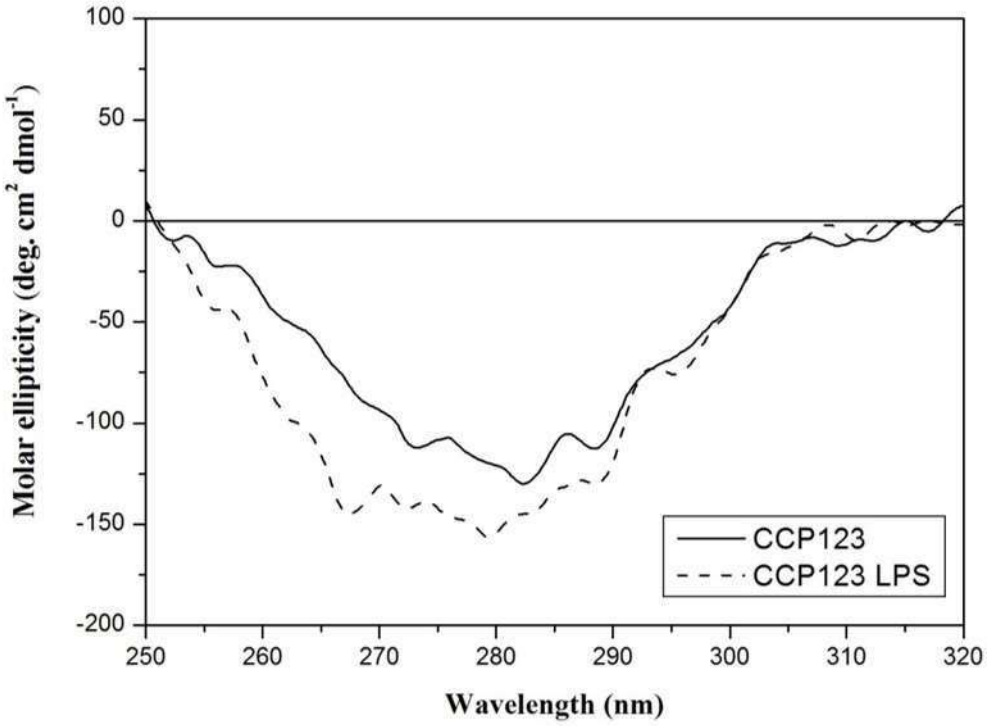

Figure 5-12: Near UV CD spectra of CCP123 in the absence (solid line) and presence (dotted line) of 1 mg/ml LPS.

5.5 Summary

The CD results indicated structural changes in a few of the domains upon exposure to LPS, as set out in Table 5-3. Changes were observed in the CysEGF, CCP3, CCP12 and CCP123 constructs. For the latter three constructs, these findings potentially corroborate the results from Tan *et al.*, who concluded that CCP1 and CCP3 both contain sites for LPS binding (Tan *et al.*, 2000). However, further repeats of these experiments, to ensure the reliability of these results, would be needed before any firm conclusions could be drawn. Alternative LPS and lipid binding assays would also need to be used to obtain quantitative data.

Construct	Far UV	Near UV
Cys-rich	No	No
EGF-like	No	No
CysEGF	Yes	-
CCP3	*	-
CCP12	Yes	Yes
CCP23	No	No
CCP123	No	Yes

Table 5-3: Structural changes observed upon addition of LPS. * indicates inconclusive results due to limited sample available that precluded the performance of replicate experiments to ensure reproducibility. – indicates no data available.

6 NMR

6.1 Overview

Nuclear magnetic resonance spectroscopy (NMR) is an analytical spectroscopic technique that can be used for the study of molecules in solution to obtain structural and dynamic information. NMR depends on nuclear magnetic spin and NMR active nuclei are those that have a non-zero spin quantum number. Those with the spin quantum number of ½ are used most often for NMR studies of organic molecules, in particular ^{1}H, ^{13}C, ^{31}P and ^{15}N. Of these, the proton, ^{1}H and ^{31}P are the most naturally abundant isotopes at close to 100%. In contrast to this, ^{13}C and ^{15}N have natural abundances of only 1.11% and 0.36%, respectively. Since the sensitivity of NMR measurements depends on the abundance of the NMR active nuclei, and sensitivity is usually a limiting factor for protein NMR, enrichment of protein samples for heteronuclear NMR studies is typically required (Cavanagh *et al.*, 1996). Detection of ^{13}C and ^{15}N is less sensitive than detection of ^{1}H as the magnitude of the gyromagnetic ratio (γ), which is a constant property of the nucleus, is smaller in both cases. This can be overcome by the use of NMR experiments that start with the magnetisation originating on ^{1}H, which is then transferred to ^{13}C and/or ^{15}N, before ultimately detecting the signal on ^{1}H.

Interactions between nuclei that are near each other can be detected via either through-space or through-bond effects. Through-bond interactions can be exploited to transfer coherence from one nucleus to another through the bonded network in experiments such as the TOCSY (total correlation spectroscopy) and INEPT (insensitive nuclei enhanced by polarisation transfer) style transfers used in heteronuclear NMR experiments (Figure 6-1). Through-space or dipolar coupling is the direct effect on a nucleus as a result of the magnetic field produced by a different nucleus. Dipolar coupling allows NOE transfer to take place, which can be used to determine intra- or inter-molecular distances. NMR experiments can be one dimensional or multidimensional and the experiments used in this study are described below. The theory of NMR has been well documented in a number of publications (Keeler, 2005; Levitt, 2001).

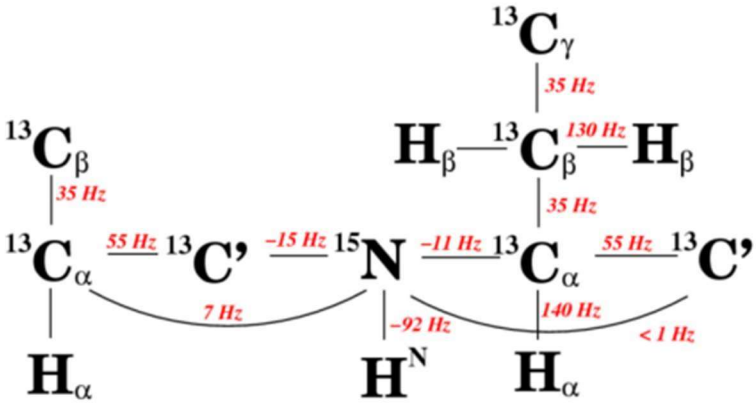

Figure 6-1: ^{1}J - ^{2}J coupling constants. Indicates the size of the coupling constants for the magnetisation transfer in {^{13}C, ^{15}N}- labelled proteins. These give an indication as to how long delays in INEPT style transfers need to be. Figure reproduced from Sattler *et al.* (Sattler *et al.*, 1999).

6.2 NMR Experiments

NMR experiments were carried out on a Bruker AVANCE IIIHD 600 MHz spectrometer equipped with an Ultrashield™ Plus magnet and a TCI triple resonance cryoprobe. Experiments were run using Bruker's Topspin™ v3.2 software using standard Bruker pulse programs, except where noted below.

Experiments described below were carried out using either a ^{15}N-labelled sample or a {^{15}N,^{13}C}-labelled sample of CCP12, produced in M9 media as described in Chapter 3, in an attempt to characterise the structure of this di-domain fragment. All experiments were run at 305 K, unless stated otherwise. Samples were prepared by adding 30 µl deuterium oxide (D_2O) to 570 µl of concentrated protein sample in NMR buffer (20 mM Na_2PO_4, 50 mM NaCl, 0.01% NaN_3, adjusted to pH 7.0 using phosphoric acid). The presence of D_2O in the sample allows for field-frequency locking. Each sample was centrifuged at 13,000 rpm in a microfuge for 5 minutes to remove any particulates before pipetting into a 5 mm NMR tube (Wilmad 535-PP-7).

6.2.1 ^{1}H 1D with solvent suppression

One dimensional (1D) ^{1}H experiments of a low concentration sample in H_2O requires suppression of the signal from the solvent (Hwang and Shaka, 1995). 1D ^{1}H experiments allow signals from non-exchanging hydrogen atoms from the protein to be seen, and helps judge if the sample is in a good condition, free from contamination by small molecules such as imidazole or glycerol. Peaks at specific chemical shifts reveal information as to

the nature of the sample including the presence of aromatic sidechains and backbone NH's, and the dispersion of chemical shifts, in particular amide and methyl signals, gives an indication as to whether the protein sample is folded. One dimensional NMR experiments are sometimes sufficient for the study of small molecules, but fail to provide sufficient information for resonance assignments for larger, more complex biomolecules such as proteins, as the peaks are too poorly resolved and the spectra lack the information required to assign the signals to specific atoms.

6.2.2 Multidimensional Experiments

Multidimensional experiments are crucial for structure determination of small molecules, proteins and nucleic acids. They produce spectra in which the NMR signals are separated into second, third, fourth or more frequency dimensions. These additional dimensions help both by reducing signal overlap, and increasing the information content of each cross-peak observed.

In homonuclear-coupled experiments, both axes of the spectrum display the chemical shift for the same type of nucleus. A cross peak is seen when transfer of magnetisation takes place between two nuclei whose signals have different frequencies. Homonuclear experiments also typically contain diagonal peaks, for which the shifts are identical in both dimensions. Diagonal peaks arise from either magnetisation that has not been transferred, or from transfer between two nuclei with the same chemical shift. In heteronuclear-coupled experiments, transfer of magnetisation occurs between different types of nuclei. The chemical shift for the different nuclei are displayed on different axes. A signal will appear at the intersection of the frequencies, but only if a transfer takes place, and diagonal peaks are not seen.

NMR data acquisition can be lengthy due to the requirements for signal averaging, as addition of scans is needed to improve the signal to noise, and incrementation of time delays that give rise to the indirectly detected frequency dimensions requires multiple repeats of the pulse sequence. In general, it is advantageous to optimise data acquisition. In some experiments, it is desirable to minimise the spectral width (SW). Choosing a narrower SW allows better digital resolution from fewer increments, since the size of the time increments goes as 1/SW. For lpFC CCP12, this was achieved by choosing ^{15}N and ^{13}C sweep widths and offsets that caused aliasing of peaks thus folding signals up field of

the spectral region downfield and vice versa so they appear within the reduced spectral width without overlapping with non-aliased peaks.

6.2.2.1 Heteronuclear Single Quantum Coherence (HSQC)

The HSQC is a very commonly used 2D NMR experiment and reveals all 1-bond correlations between H and N or H and C, depending on the nuclei being used. In a ^{15}N HSQC, magnetisation is transferred by exploiting J-coupling between the ^{15}N nuclei and their covalently attached hydrogen. ^{13}C HSQC experiments obtain a chemical shift correlation map between the directly-bonded ^{1}H and ^{13}C. There is more equilibrium magnetisation on H than C or N, and so the magnetisation starts on H, transfers to C/N and then back to H for detection. The chemical shift evolves before being transferred back for detection on the hydrogen. High sensitivity is achieved due to the detection on proton.

The ^{15}N HSQC experiment mostly gives information about backbone amide groups. However, some side chain groups can also be seen, which can include the Trp side chain NεHε and Asn and Gln side chain groups. Figure 6-2 shows a ^{15}N-HSQC for CCP12.

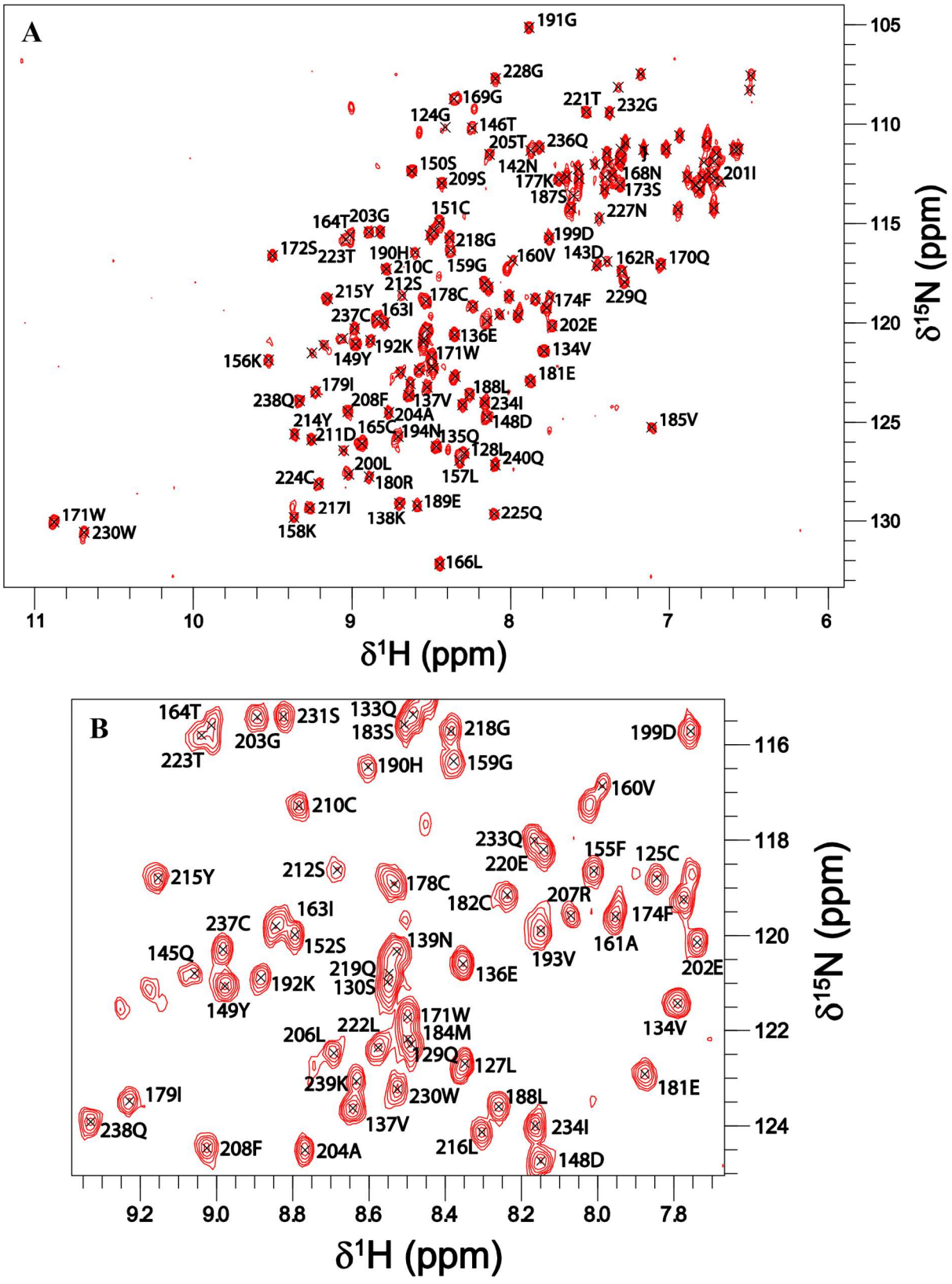

Figure 6-2: CCP12 15N-HSQC. A. Full spectral width spectra displaying all assigned CCP12 cross-peaks, labelled with their residue number and type. B. Zoomed in region of the middle of the spectra for better clarity of the labelled peaks in this area.

6.2.2.2 Triple Resonance Experiments

Triple resonance experiments were used to identify linked sets of the atomic nuclei, in order to obtain backbone ^{1}H, ^{13}C and ^{15}N resonance assignments of the protein. They depend upon heteronuclear $^{1}J/^{2}J$ couplings for magnetisation transfer (Figure 6-1). Each experiment is named after the nuclei involved, organised in the order in which the magnetisation transfer occurs. Examples include: HNCA, CBCA(CO)NH and HN(CA)CO. Nuclei in parentheses are in the magnetisation transfer pathway but their chemical shifts are not detected. Spins that are frequency labelled during acquisition, or indirect dimensions, $^{1}H^{N}$, amide ^{15}N, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, ^{13}CO, $^{1}H^{\beta}$ and $^{13}C^{\beta}$, are represented by HN, N, HA, CA, CO, HB and CB, respectively (Ikura *et al.*, 1990; Kay *et al.*, 1990). Magnetisation transfer pathways for the majority of experiments used in this study are shown in Figure 6-3. In so called “out-and-back” experiments, magnetisation starts on a proton and is then transferred back to this proton for detection. In comparison, in “out-and-stay” experiments, magnetisation is transferred to a different spin where it stays for acquisition. A full list of all the experiments carried out and their acquisition parameters can be found in Table 6-1.

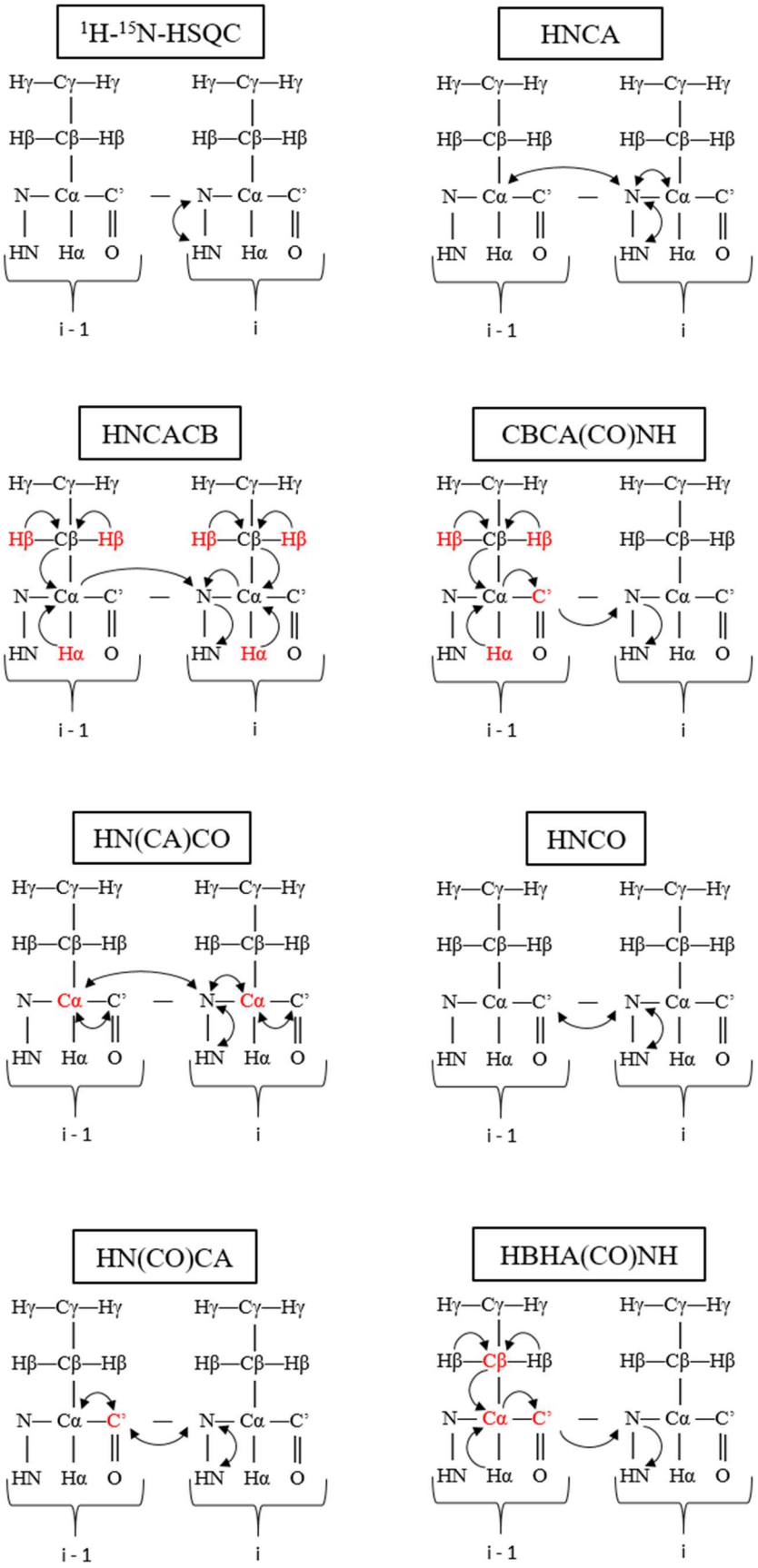

Figure 6-3: Magnetisation transfer pathways. Arrows indicate the transfer of magnetisation between nuclei. Nuclei in red are in the magnetisation transfer pathway but their chemical shifts are not detected.

Experiment	F1 (indirect)			F2 (direct)				F3 (indirect)				NUS	NS	Temp (K)	Pulse Program Reference
	NC	SW	AQ	NC	SW	AQ	QD	NC	SW	AQ	QD				
^{15}N-HSQC	^{15}N	22.5	0.047	^{1}H	16.025	106.496	ST	-	-	-	-	-	8	305	(Mori *et al.*, 1995)
^{13}C-HSQC	^{1}H	14.029	0.059	^{13}C	32.999	12.854	E-Anti	^{1}H	16.025	106.496	ST	-	32	305	(Davis *et al.*, 1992)
HNCA	^{13}C	32	13.256	^{15}N	22.5	24.121	ST	^{1}H	16.025	106.496	ST	37.879	64	305	(Grzesiek and Bax, 1992; Kay *et al.*, 1994; Schleucher *et al.*, 1993)
HNCACB	^{13}C	75	6.716	^{15}N	22.5	24.852	ST	^{1}H	16.025	106.496	ST	20.433	64	305	(Muhandiram and Kay, 1994; Wittekind and Mueller, 1993)
CBCACONH	^{13}C	75	6.716	^{15}N	22.5	24.852	ST	^{1}H	16.025	106.496	ST	20.433	64	305	(Grzesiek and Bax, 1993a; Muhandiram and Kay, 1994)
HNCO	^{13}C	22	24.098	^{15}N	22.5	24.121	ST	^{1}H	16.025	103.496	ST	32.727	16	305	(Grzesiek and Bax, 1992; Kay *et al.*, 1994; Schleucher *et al.*, 1993)
HNCACO	^{13}C	22	24.098	^{15}N	22.5	24.121	ST	^{1}H	16.025	106.496	ST	26.061	72	305	(Clubb *et al.*, 1992; Kay *et al.*, 1994)
HBHACONH	^{1}H	16.025	6.656	^{15}N	22.5	24.121	ST	^{1}H	16.025	106.496	ST	35.606	64	305	(Grzesiek and Bax, 1993a; Muhandiram and Kay, 1994)
HBHANH	^{1}H	16.025	6.656	^{15}N	22.5	24.121	ST	^{1}H	16.025	106.496	ST	35.606	64	305	(Wang *et al.*, 1994)
^{15}N-NOESY-HSQC	^{1}H	13.333	16	^{15}N	22.5	43.125	ST	^{1}H	16.025	106.496	ST	-	8	305	(Sklenar *et al.*, 1993)
^{13}C-NOESY-HSQC	^{1}H	13.333	13.75	^{13}C	32.999	8.034	E-Anti	^{1}H	16.025	106.496	ST	-	16	305	(Davis *et al.*, 1992)
hCCHTOCSY	^{1}H	75	11.312	^{13}C	33	12.854	ST	^{13}C	16.025	106.496	ST	19.922	64	305	(Kay *et al.*, 1993)
HCcHTOCSY	^{1}H	8.333	20	^{13}C	33	6.226	ST	^{1}H	16.025	106.496	ST	25	16	305	(Kay *et al.*, 1993)
T_1 relaxation	^{1}H	26.999	643.21	^{15}N	16.666	102.4	ST	-	-	-	-	-	32	298	(Kay *et al.*, 1992)

Experiment	F1 (indirect)			F2 (direct)				F3 (indirect)				NUS	NS	Temp (K)	Pulse Program Reference
	NC	SW	AQ	NC	SW	AQ	QD	NC	SW	AQ	QD				
T_2 relaxation	^{1}H	27	548.79	^{15}N	16.666	102.4	ST	-	-	-	-	-	48	298	(Kay *et al.*, 1992)
HDEx	^{15}N	39.999	13.157	^{1}H	16.025	106.496	ST	-	-	-	-	-	16	298	(Mori *et al.*, 1995)
^{15}N-NOEref	^{15}N	26.999	107.202	^{1}H	16.666	102.4	ST	-	-	-	-	-	40	298	(Grzesiek and Bax, 1993b)
^{15}N-NOEsat	^{15}N	26.999	107.202	^{1}H	16.666	102.4	ST	-	-	-	-	-	40	298	(Grzesiek and Bax, 1993b)
aroNOESY	^{1}H	13.336	16	^{13}C	74.994	0.044	E-Anti	^{1}H	16.029	106.496	ST	-	1536	305	(Davis *et al.*, 1992)
HBCBCGCDHD	^{1}H	16.025	106.496	^{13}C	40	5.302	ST	-	-	-	-	-	736	305	(Yamazaki *et al.*, 1993)
HBCBCGCDCEHE	^{1}H	16.025	106.496	^{13}C	40	5.302	ST	-	-	-	-	-	736	305	(Yamazaki *et al.*, 1993)

Table 6-1: NMR experiments and their acquisition parameters. NC = Nucleus, SW = Spectral Width (ppm), AQ = Acquisition Time (ms), NUS = Non-Uniform Sampling Amount (%), QD = Quadrature detection mode: ST = States-TPPI, E-Anti = Echo-Antiecho, NS = Number of scans, Temp = Temperature (K). HDEx = Hydrogen Exchange. The aroNOESY was a ^{13}C-edited ^{1}H^{1}H NOESY optimised for aromatics.

6.2.2.3 HNCA, HNCACB and CBCA(CO)NH

The HNCA, HNCACB and CBCA(CO)NH can be used in conjunction for backbone assignment. In the HNCA, magnetisation is passed from ^{1}H to ^{15}N via J-coupling and then via N-Cα J-coupling to the $^{13}Cα$, then back by the reverse pathway to ^{15}N and ^{1}H where detection takes place. A 3D spectrum is produced with chemical shifts for ^{1}H, ^{15}N and $^{13}Cα$. Peaks for both the amide nitrogen's own residue's Cα resonance and for the Cα from the residue before are usually visible, because both N-Cα couplings are similar (Figure 6-1). Two Cα peaks should appear in each NH strip (Figure 6-4), and as coupling to the residues own Cα is stronger these peaks should appear with greater intensity. In the case of lpFC CCP12, the weaker peaks, belonging to the preceding residue were often buried in the noise.

In the HNCACB experiment, magnetisation starts on ^{1}H is transferred to ^{15}N, $^{13}Cα$ and $^{13}Cβ$, and then back to ^{1}H for detection. Magnetisation is transferred from both $^{13}Cα_{i}$ and $^{13}Cα_{i-1}$ to $^{15}N_{i}$, resulting in two peaks for Cα and two peaks for Cβ in each NH strip. Simultaneous evolution of the carbon signals correlated to an NH occurs meaning they will appear in one dimension, and the chemical shifts for ^{1}H and ^{15}N are evolved in the other two dimensions. The four peaks visible in each strip are $Cα_{i}$, $Cα_{i-1}$, $Cβ_{i}$ and $Cβ_{i+1}$. Those from the NH's own residue appear stronger.

Magnetisation in the CBCA(CO)NH is initially transferred similarly to HNCACB, in that it occurs between $^{1}Hα$ and $^{13}Cα$, and between $^{1}Hβ$ and $^{13}Cβ$, and then from $^{13}Cβ$ to $^{13}Cα$. It then transfers to ^{13}CO, then to ^{15}N and then for detection on ^{1}H. Since CA-CO intra-residue and CO-N inter-residue couplings are strong, only transfer via the directly bonded CACON route is seen and intra-residue cross-peaks are not. Again, simultaneous evolution of the signals results in them appearing in one dimension, however, no chemical shift is evolved on ^{13}CO. Each strip displays $Cα_{i-1}$ and $Cβ_{i-1}$.

6.2.2.4 HNCO and HN(CA)CO

The HNCO is the most sensitive of the triple-resonance experiments, and is used to obtain the CO chemical shifts and assist with backbone resonance assignment. Magnetisation is transferred between ^{1}H and ^{15}N and then selectively via the $^{15}N^{H}$-^{13}CO J-coupling to the ^{13}CO. It is then transferred back via ^{15}N to ^{1}H where the signal is detected. Chemical shifts evolve on all three nuclei. In each NH strip, CO_{i-1} is visible.

The HN(CA)CO produces another 3D spectrum that is used for backbone assignment in combination with the HNCO. Transfer of magnetisation occurs from ^{1}H to ^{15}N, then through N-Cα J-coupling to $^{13}C\alpha$. It is then transferred via $^{13}C\alpha$-^{13}CO J-coupling to the ^{13}CO. It is then transferred back from ^{13}CO to $^{13}C\alpha$, ^{15}N to ^{1}H for detection. Chemical shifts are evolved on ^{1}H, ^{15}N and ^{13}CO but not on $^{13}C\alpha$. The amide nitrogen is coupled to both $C\alpha_i$ and $C\alpha_{i-1}$, meaning both transfers take place, which results in transfer to both CO_i and CO_{i-1}. Two peaks appear in the spectrum for CO_i and CO_{i-1} with the peak that appears in both the HNCO and the HN(CA)CO spectra for a particular NH strip belonging to CO_{i-1}.

6.2.2.5 HBHA(CO)NH and HBHANH

The HBHA(CO)NH is used for the detection of Hα and Hβ resonances. The transfer pathway is the same as that for the CBCA(CO)NH experiment, the only difference being the shifts that are evolved. Magnetisation transfer takes place from $^{1}H\alpha$ and $^{13}C\alpha$, and between $^{1}H\beta$ and $^{13}C\beta$, and then from $^{13}C\beta$ to $^{13}C\alpha$. Next, transfer takes place to ^{13}CO, then to ^{15}N and then for detection on ^{1}H. In this experiment, there is no chemical shift evolution on the carbon atoms, but $^{1}H\alpha$ and $^{1}H\beta$, ^{15}N and ^{1}H evolution takes place. A 3D spectrum is produced with one nitrogen and two proton dimensions. In each NH strip, one Hα peak and one or two Hβ peaks are visible that correspond to the chemical shifts of the preceding residue's resonances. This information cannot be used for sequential assignment on its own, as only the inter-residue correlations are visible, and not the corresponding intra-residue ones. Thus, the HBHANH, is used in conjunction to the HBHA(CO)NH to determine the intra-residue connections.

6.2.2.6 ^{15}N and ^{13}C NOESY-HSQC

In the ^{15}N-NOESY-HSQC, magnetisation is first transferred between all hydrogens using the nuclear Overhauser effect (NOE). From the destination proteins ^{1}H, it is then transferred to bonded ^{15}N nuclei, selecting only amide and NH sidechain resonances, and back to those ^{1}H for detection. Cross-peaks will therefore all be to NH or sidechain NHs. Cross-peaks observed in this spectrum can be interpreted to produce distance restraints for structure calculations. The NOESY mixing time was 120 ms. A longer mixing time gives stronger cross-peaks but risks more spin diffusion, which is a leakage of magnetisation to surrounding nuclei. NOEs from one NH group to all hydrogen atoms that are nearby are shown in each strip.

For the ^{13}C-NOESY-HSQC, again, magnetisation is first transferred using the NOE between all hydrogens. From the destination proteins ^{1}H, it is then transferred to bonded ^{13}C nuclei and back to those ^{1}H for detection. Transfer is either between the aliphatic ^{13}C or the aromatic ^{13}C, and experiments can be optimised for either aliphatics or aromatics. This is dependent on the ^{13}C frequency used in the pulse sequence, whether it is centred on the aliphatic or aromatic carbon region. The NOESY mixing time was 120 ms for the first experiment, which was repeated with a mixing time of 200 ms. The aroNOESY, focusing on the aromatics had a mixing time of 200 ms. NOE cross-peaks from one CH group to the hydrogens close by are shown in each strip and the NOESY centred on the aromatic region is very useful when assigning the aromatic residues as Hβ to H_{aro} connections can be made. NOE cross-peaks were used to derive distance restraints for structure calculations.

6.2.2.7 HCCH-TOCSY

Two versions of an HCCH-TOCSY experiments were recorded: hCCH-TOCSY and HCcH-TOCSY, to be used for side-chain assignment. The magnetisation transfer pathway is the same for both HCCH-TOCSYs. Magnetisation starts on the side-chain hydrogen, is transferred to its attached ^{13}C nucleus then to all the other ^{13}Cs in the same sidechain using a DIPSI2 mixing sequence (^{13}C-TOCSY) (Shaka *et al.*, 1988). Magnetisation is then transferred from the ^{13}C it has arrived on to its attached ^{1}H. This happens simultaneously for all ^{13}Cs in the residue. For the hCCH-TOCSY, the originating H chemical shift is not recorded, but the 'CCH' are, and for the HCcH-TOCSY, all but the 'c' are recorded. Strips are displayed for each carbon frequency in the side-chain.

6.2.2.8 ^{15}N Relaxation Experiments

Relaxation experiments were performed on a ^{15}N-labelled sample at 298 K. ^{15}N heteronuclear NOEs were measured, with and without ^{1}H saturation (NOEref and NOEsat, respectively), by comparing the intensity of signal transferred from amide ^{15}N to ^{1}H.

6.2.2.9 Aromatics

Two experiments are used for the assignment of the aromatic residues' sidechains by connecting their aromatic Hs to their Cβs: hbCBcgcdHD and hbCBcgcdceHE (case rather than brackets indicates which chemical shifts are recorded). In these experiments,

correlations between side-chain $^{13}C\beta$ and ring $^{1}H\delta/\varepsilon$ chemical shifts are identified. Transfer of magnetisation occurs via J-coupling and the paths of magnetisation transfer are as follows:

For $^{1}H\delta$ detection: $H\beta \rightarrow C\beta\ (t_1) \rightarrow C\gamma \rightarrow C\delta \rightarrow H\delta\ (t_2)$
For $^{1}H\varepsilon$ detection: $H\beta \rightarrow C\beta\ (t_1) \rightarrow C\gamma \rightarrow C\delta \rightarrow C\varepsilon \rightarrow H\varepsilon\ (t_2)$

6.2.2.10 Hydrogen Exchange

A hydrogen exchange experiment was carried out where an ^{15}N labelled sample was lyophilised and then dissolved in D_2O. This sample was placed quickly into the magnet and three back to back ^{15}N-HSQCs were recorded. Most of the signals disappeared immediately, and were therefore unobservable in even the first one. Since the rate of exchange of backbone NHs with the solvent depends on how tightly they are hydrogen bonded, the signals that remained were interpreted as indicating the NHs involved in hydrogen bonds.

6.3 Data Processing

Most spectra were processed using Bruker's Topspin™ v3.2 and MDD v1.6 (Orekhov *et al.*, 2003) with the exception of the relaxation data which were processed using AZARA (http://www2.ccpn.ac.uk/azara/, Wayne Boucher).

6.4 Assignment

A key stage in determination of a protein's structure from NMR data is the assignment of resonances observed from NMR spectra to their specific atom in the molecule of interest. This enables the use of the determined distance restraints for calculation of the structure. To use the NMR signals, the atoms giving rise to the individual signals need to be identified. Spectra are used in combination to establish sequential links between backbone amides, which are placed in their sequence context by the assignment of aliphatic sidechain atoms and the atoms of aromatics and longer sidechains. The CcpNmr Analysis software v2.4.1 was used for all spectra analysis prior to structure determination (Vranken *et al.*, 2005) and referenceB (K. Bromek, unpublished) was used to determine the correct referencing for each spectrum (Wishart *et al.*, 1995).

6.4.1 Backbone resonance assignment

All cross peaks evident in the ^{15}N-HSQC were picked and assigned to spin systems, which each represent an individual residue. Triple resonance peaks occurring at the same ^{1}H and ^{15}N chemical shift were assigned to the same spin system and atom type was determined from the specific triple resonance experiment the spectra had been produced from; the HNCA determined the position of the residues own $C\alpha_i$, the HNCACB determined the $C\beta_i$ position and the position of CO_i was established from the HNCO spectrum. Sequential links between backbone amides could be made by identifying the peaks that corresponded in the CBCA(CO)NH for Cα and Cβ and by using the HN(CA)CO to determine the position of the CO in the residue before ($_{i-1}$) (Figure 6-4).

Having assembled the Cα and Cβ chemical shifts, a judgement could be made about which residue type they belonged to. The references provided by Iwadate *et al.* alongside the 'Protein Sequence Assignment' function in CcpNmr analysis, helped with the specific identifications (Iwadate *et al.*, 1999).

Due to the fact that certain residues have atoms with very distinctive chemical shifts, a few residue type assignments can be made easily. Examples of this include serine and threonine residues whose Cβ's have a chemical shift up-field of the Cα's, and glycines that have a Cα but no Cβ's. CcpNmr Analysis provides a 'Link backbone resonances' tool, which suggests residue type based on a comparison of the observed chemical shifts with statistical distributions harvested from BMRB/refDB, to help with amino acid identification (Ulrich *et al.*, 2008; Zhang *et al.*, 2003). Using this tool, the information obtained from the spectra and the amino acid sequence of CCP12, residues could be linked together and assigned. This method is effective until a proline is reached, as this residue does not have an NH and so a break is introduced in the amino acid chain.

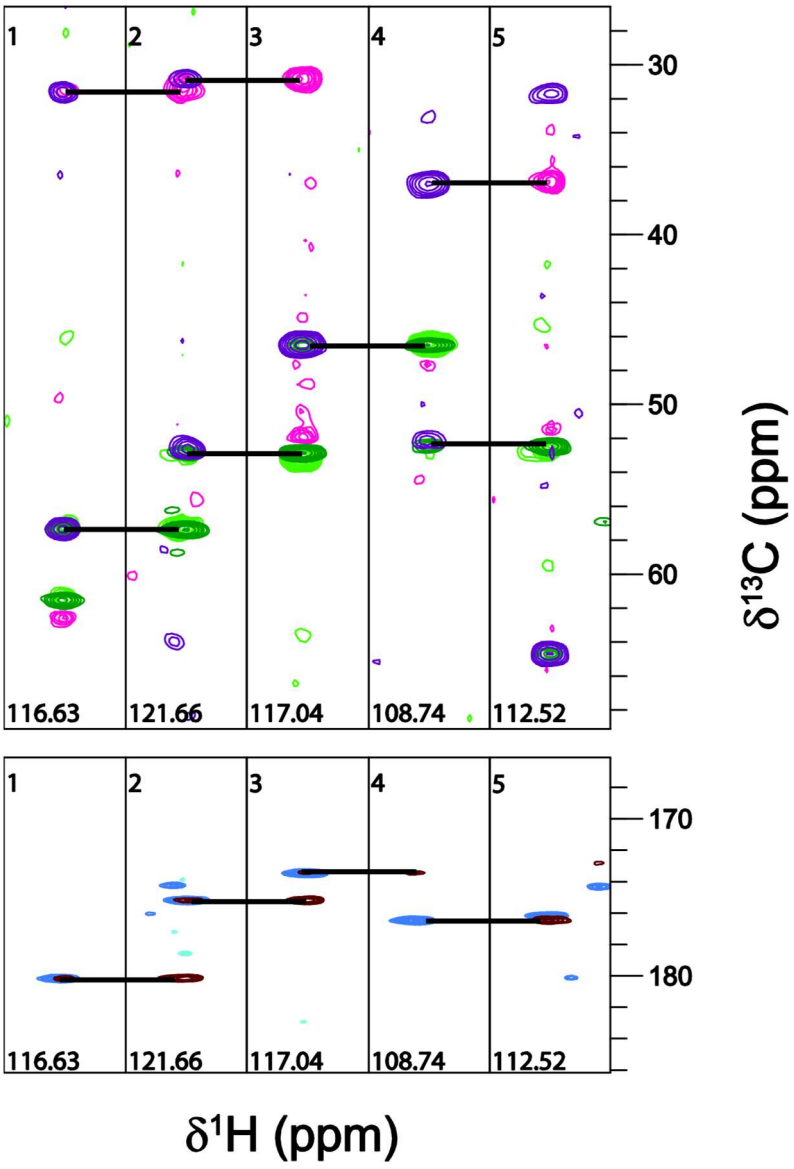

Figure 6-4: Triple resonance assignment. Each vertical strip represents an individual spin system (residue). From left to right: 172Ser, 171Trp, 170Gln, 169Gly and 168Asn. The HNCA experiment gives rise to the green peaks and indicate the chemical shift of the residues own Cα, the pink peaks are from the HNCACB and show the chemical shift of the residues own Cβ, while the purple peaks arise from the CBCACONH and indicate the chemical shift for the preceding residues Cα and Cβ. The residues own CO is indicated by the brown peak from the HNCACO and the preceding residues CO is shown by the blue peak from the HNCO.

To elucidate the position of the Hα's and Hβ's, the HBHANH was used to identify the chemical shifts of a residue's own atoms while the HBHA(CO)NH spectrum helped to confirm a sequential link by revealing Hα and Hβ peaks at the chemical shifts for the residue before.

6.4.2 Sidechain Assignment

Backbone assignments were extended into the sidechain by use of the 3D HCCH-TOCSY experiments. The spectra produced from these experiments allowed for ^{13}C side chain resonances to be correlated with all other ^{13}C and ^{1}H resonances in a spin system. Known Cα and Cβ shifts were used to identify cross-peaks in planes corresponding to the Hα and Hβ shifts of a residue, and additional ^{13}C resonances could then be identified (Figure 6-5). The Hα's of many residues have their chemical shift near 4.7 ppm, which is the approximate shift of the solvent water. This region of the spectrum contains many artefacts due to incomplete solvent suppression and it was therefore often difficult to distinguish and pick peaks. Furthermore, this spectrum often showed regions of overlap, particularly at δ and γ chemical shifts but this could in some cases be resolved by navigating to the chemical shift position in the HCcH-TOCSY. The quality of the data shown by Figure 6-5 highlights the sensitivity problem faced with some experiments, due to the low sample concentration, as peaks are missing and often poorly resolved. The α and β chemical shifts were used to get an indication of where the γ and δ peaks were, so the whole residue could be pieced together. In some cases, signals that were degenerate in the shifts of one nucleus (e.g. ^{1}H) could now be resolved due to correlation with another (e.g. ^{13}C), as they were displayed at different shifts for the different groups.

6.4.3 Proline Assignments

Proline is the most difficult residue to assign, as this residue is lacking an NH. For CCP12, some assignments could be made by using the α and β from the residues that followed prolines. However, there were a number of cases where this could not be achieved due to there being two prolines together in the sequence or because the residues after had not been identified themselves. Thus, in order to complete the proline assignments and identify the cis-peptidyl proline bonds, the $H\alpha_i$ to $H\alpha_{i+1}$ and $H\alpha_i$ to $H\delta_{i+1}$ NOEs were used. This was achieved, for example, by identifying the δs at the α chemical shift for prolines. This allowed for peaks to then be identified at the chemical shift of carbon and from here it was possible to build up the whole picture. Peaks that were at the chemical shift for proline δs but were otherwise unexplained, were investigated and the atoms could, for some, be linked in this way. The peptide bond almost always adopts the trans configuration, which is energetically favoured over the cis configuration. However, the cyclic nature of the proline side-chains mean--s that cis-

peptidyl configurations are more likely to occur than in other residues. In the cis configuration, the omega torsion angle will be close to 0°, in comparison to the trans configuration where it will be 180°. These conformational differences affect the pattern of NOE signals seen, as the conformation of the peptide bond determines whether the NH of the preceding residue is close to the Hδs (trans) or the Hαs (cis) of the proline (Joseph *et al.*, 2012).

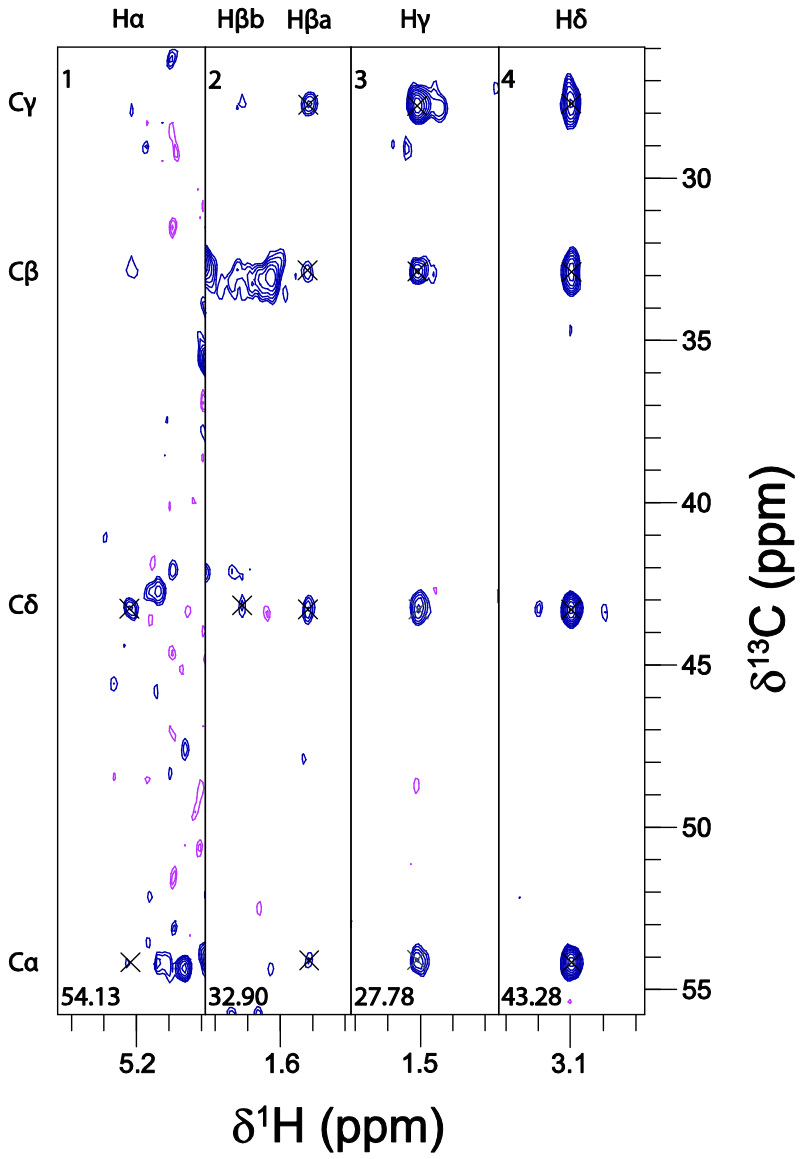

Figure 6-5: hCCHTOCSY spectrum for 207Arg. These strips are displayed vertically as the ^{1}H detected dimension appears along the x-axis and the ^{13}C dimension appears along the y-axis.

6.4.4 Role of NOESY in Assignment

The ^{15}N-NOESY gives strong α_i – NH_{i+1} and NH_i – NH_{i+1} NOEs, which can be used to confirm or establish sequential connectivity. The aromatic side chains were assigned by

using hbCBcgcdHD, hbCBcgcdceHE and the ^{13}C-NOESY experiments (Wang *et al.*, 1994). The ^{13}C-NOESY helps connect the aromatics by linking NOEs such as β – δ.

6.4.5 Assignment Summary

120 of the 121 (99.17 %) residues from CCP12 were at least partially assigned, from the experiments described. A summary of the assignments is shown in Table 6-2 and the chemical shift assignments are summarised in Appendix E.

Resonance Type	Available	Assigned	% Assigned
Overall residues	121	120	99.17
Carbon atoms	563	483	85.79
Hydrogen atoms	719	640	89.01
Nitrogen atoms	158	108	66.46
Backbone amides	108	102	94.44
Backbone non-H	363	338	93.11
Side-chain H	480	407	84.79
Side-chain non-H	358	250	69.83

Table 6-2: CCP12 assignment completeness.

6.5 Summary

NMR experiments were recorded on samples of CCP12 in an attempt to characterise the structure of this di-domain. Several spectra were recorded in order to be able to fully assign the resonances to their specific atoms. A near complete resonance assignment was achieved allowing structure calculations to be carried out for CCP12, as described in Chapter 7.

7 Structure Calculation

NOESY cross-peaks were picked semi-automatically in the 3D NOESY spectra for all assigned resonances. Care was taken to avoid artefacts, noise and overlapped regions. The number of cross peaks was relatively low for a protein of this size, due to a low sample concentration and therefore a low signal to noise. This was also potentially due to anisotropic tumbling due to the elongated nature of the di-domain.

7.1 Structure Calculation by ARIA

ARIA 2.3 (Ambiguous Restraints for Iterative Assignment) was used for structure calculations of lpFC CCP12 (Rieping *et al.*, 2007). This software uses a mathematical way of summing up the various interatomic distances' contributions to individual ambiguous restraints. After each round of structure calculation, the level of ambiguity can be refined based on the structures calculated by excluding the longest interatomic distances (lowest contributors) that contribute to an ambiguous restraint. In this work the standard ARIA protocol was applied but with deviations that include the number of steps in the molecular dynamics stages (Table 7-1), violation thresholds for exclusion of consistently violated restraints, ambiguity thresholds, and the number of structures calculated in the iterations (Table 7-2).

When structures are calculated, many structures will typically satisfy the restraints, but all with slight differences. Structure calculations are repeated from randomised starting states to produce an ensemble of structures from which those that fit the data best and have the lowest restraint energy can be selected. If the structures that have low energy show a high level of structural similarity, this indicates the calculations have converged and the restraints define the structure well.

Step	Temperature (K)	Steps
High temperature	10,000	20,000
Refine	2,000	20,000
Cool 1	1,000	20,000
Cool 2	50	16,000

Table 7-1: Molecular dynamics conditions.

Iteration	Number of Structures	Violation Tolerance (Å)	Partial Assignment Filter
0	20	1000	1.0
1	20	5.0	0.999
2	20	3.0	0.999
3	20	1.0	0.99
4	20	1.0	0.98
5	20	1.0	0.97
6	100	0.5	0.96
7	100	0.3	0.95
8	100	0.3	0.95

Table 7-2: Iterative strategy for CCP12 structure calculations. Strategy of later structure calculations after inclusion of disulphide and hydrogen bond restraints.

7.1.1 Disulphide Bonds

Previous studies of complement control proteins have determined a conserved pattern of disulphide bonds from Cys1 – Cys3 and Cys2 – Cys4 for each module (Reid and Day, 1989). Calculations were first run with NOE restraints, and when it was clear that this pairing of the cysteines was compatible with the NOE only structures, they were imposed as additional distance restraints between the Cys sulphurs.

7.1.2 Hydrogen Bonds

The cross peaks for twenty-six NH groups were visible in the first of the deuterium exchange HSQC spectra. Twenty of these could be assigned unambiguously with the other six unresolved in the HSQC. Preliminary structures calculated with NOE restraints alone were examined using PyMOL to identify the likely hydrogen bonding partners, based on the proximity and relative orientation of potential hydrogen bond acceptors to the NHs identified as slowly exchanging. Slowly exchanging amides and their identified hydrogen bond partners are shown in Table 7-3.

Slowly exchanging NH	Hydrogen bond acceptor
134Val	131Asp
135Gln	150Ser
146Thr	139Asn
148Asp	137Val
149Tyr	161Ala
150Ser	135Gln
155Phe	152Ser
156Lys	179Ile
166Leu	170Gln
170Gln	166Leu
171Trp	126Pro
173Ser	-
177Lys	158Lys
179Ile	156Lys
208Phe	220Glu
209Ser	192Lys
214Tyr	211Asp
215Tyr	238Gln
222Leu	206Leu
230Trp	183Ser

Table 7-3: Hydrogen bond restraints. The HDex spectra identified donor amide groups and well defined structures were analysed to identify the acceptor carbonyl groups.

7.1.3 Dihedral Angle Restraints

Predictions of backbone dihedral angles were carried out using the Dihedral Angles from Global Likelihood Estimates (DANGLE) software in CcpNmr analysis (Cheung *et al.*, 2010). Sequence data and backbone chemical shifts are used to analyse secondary structure and predict possible phi (ϕ) and psi (Ψ) angles. The software uses a stretch of sequence, not just an individual amino acid, to compare the sequence with a database of protein structures, with known chemical shifts. This identifies a number of similar structures that can be analysed and used alongside known angles for specific residues such as prolines (that have a locked ϕ angle) and glycines (that often have a positive ϕ angle), for dihedral angle predictions. The dihedral angle restraints derived from the DANGLE predictions were used to help with convergence of the structure calculations. However, since these restraints are only indirectly based on the experimental data, they were only used in the high temperature and not in the refinement stage of each iteration.

7.1.4 NOE Restraints

The most important restraints used for structure determination in this study were derived from the NOESY experiments. These restraints, derived from the observation of individual cross-peaks in the NOESY spectra, provide information on long range connectivities with defined distances. However, many cross-peaks cannot be assigned unambiguously as representing the distance between a particular pair of atoms. ARIA's algorithm is designed to deal with ambiguous distance restraints, so most peaks were left at least partially unassigned to avoid the structure calculations being biased by incorrectly assigned peaks. ARIA was instructed to use all the cross-peaks from the NOESY spectra to generate distance restraints by matching each cross-peak to the resonances whose chemical shifts lie within a defined tolerance of the peak centre. ARIA initially calibrates the distances according to a normalised 1/distance^6 distribution and subsequently by matrix relaxation refinement that allows for spin diffusion effects (Linge *et al.*, 2003a). After the first few rounds of calculation in which missing and mis-assignments were identified and rectified, the log-normal potential and restraint weighting were also tried (Nilges *et al.*, 2008).

7.1.5 Analysis of Restraints

After each round of structure calculations violated distance restraints that were rejected by ARIA were re-analysed in CcpNmr analysis, to determine the reason for the violation. This allowed for incorrectly assigned peaks to be amended, peaks that had been picked in the noise to be deleted, and at times highlighted unassigned resonances, that could in some cases be assigned. This analysis was followed by subsequent structure calculations until the structures showed good convergence and low experimental energies indicating that the experimental restraints were self-consistent.

7.1.6 Water Refinement

The final stage of the structure calculations is refinement in explicit solvent with a more realistic force field. Earlier in the calculation, for speed, structures are calculated in vacuo with no electrostatic conditions set and with a more simplistic representation of the Van der Waals interactions. The 18 structures with the lowest restraint energy (of 100) from iteration 8 were used for refinement in explicit solvent, according to the pre-defined protocol (Linge *et al.*, 2003b).

7.2 Structure Validation

Measures can be taken to determine the quality of the structures that are calculated by ARIA. This ensures that the structures give a reasonable representation of the protein being studied. Multiple structures are produced as a result of a range of values being used, accounting for distances and angles. Thus, a number of structures are arranged into an ensemble, each fitting the experimental restraints to an approximately equal standard.

7.2.1 Quality of Ensemble Structures

The total number of restraints and the quality of these restraints used in the structure calculation can be used to determine the quality of the structures calculated for the ensemble. The quality of the calculated structures generally improves with the more restraints that are used. 18 water refined structures were used to calculate statistics of the restraints, shown in Table 7-4. Ramachandran plots can be useful to visualise dihedral angles, however, as the structures are still at a preliminary stage, this analysis is not yet justified.

NOE distance restraints	
Ambiguous	477
Unambiguous	725
Intra-residue	396
Sequential (i-j = 1)	197
Short-range (i-j = 2 – 4)	36
Long-range (i-j ≥ 5)	96
Violations > 0.5 Å	0.72
Violations > 0.3 Å	1.33
Violations > 0.1 Å	20.39
Distance restraint RMSD	0.032 Å
Other restraints	
Hydrogen bonds	20
Dihedral angle restraints	206
Disulphide bonds	4
RMSD from the ideal geometry	
Bond length (Å)	0.00345992 ± 0.0001455667
Bond angle (°)	0.460842 ± 0.0219091
Improper angle (°)	1.35534 ± 0.13369

Table 7-4: Statistics of the experimental restraints. Averages were calculated from the 18 water refined structures in the ensemble.

7.3 CCP12 Structure

The UWMN program (Leo Caves, unpublished) calculates an unbiased average structure, then fits all the structures to this and calculates the root-mean-square deviation (RMSD). Results are shown for Cα's and for backbone heavy atoms (N, Cα, CO). In the structure calculations, the CCP1 domain is well defined with a Cα RMSD of 0.828 Å, however, the CCP2 domain is not, with a Cα RMSD of 2.305 Å. This could be because CCP2 has residues with unassigned atoms including Ser186, Ser195, Ser197 and Gly226, that are all lacking their NH, and Pro196 and Ser212 that are lacking their CO. It is clear that the CCP1 module is better defined, with more distinct secondary structure in comparison with CCP2.

Stereo-views of both the CCP1 and CCP2 modules illustrate the convergence seen for each domain and the similarities and differences between them (Figure 7-1). For CCP1, all eighteen structures superimpose well, with little variability. The β-strands are well defined and it represents a typical CCP module structure. CCP2 on the other hand shows more variability, as expected from the higher RMSD. There is poorer definition of the β-strands, and more variation than can be seen for CCP1. As the structures are not completely refined, final conclusions may differ.

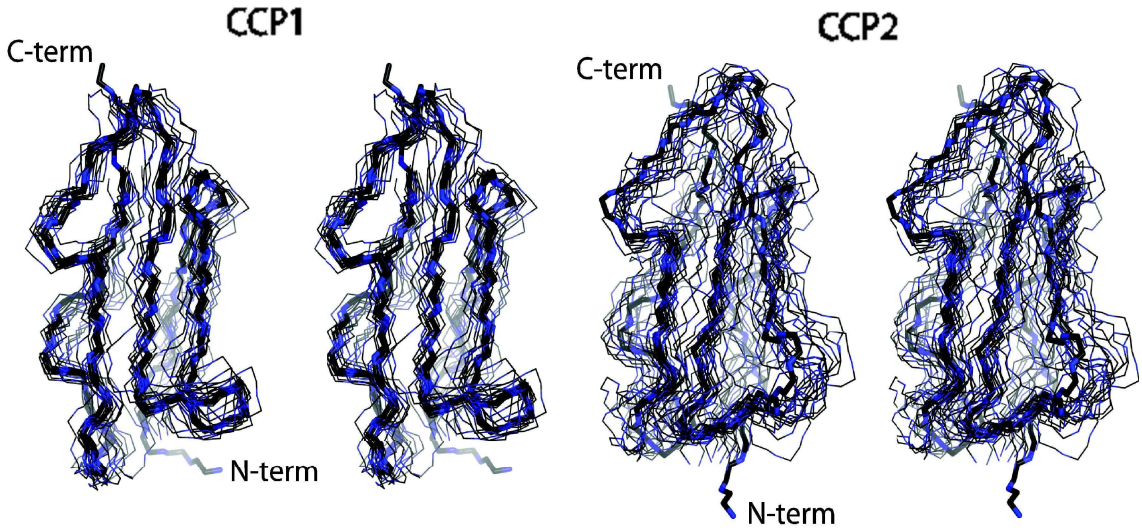

Figure 7-1: Stereo-view of CCP1 and CCP2 domains. The eighteen CCP12 structures in the ensemble are superimposed on the C_αs of each of CCP1 and CCP2. The backbone nitrogens are coloured blue.

Figure 7-2 and Figure 7-3 show the relative orientation between the two modules. The structure is superimposed on the CCP1 module of the structure with the lowest energy.

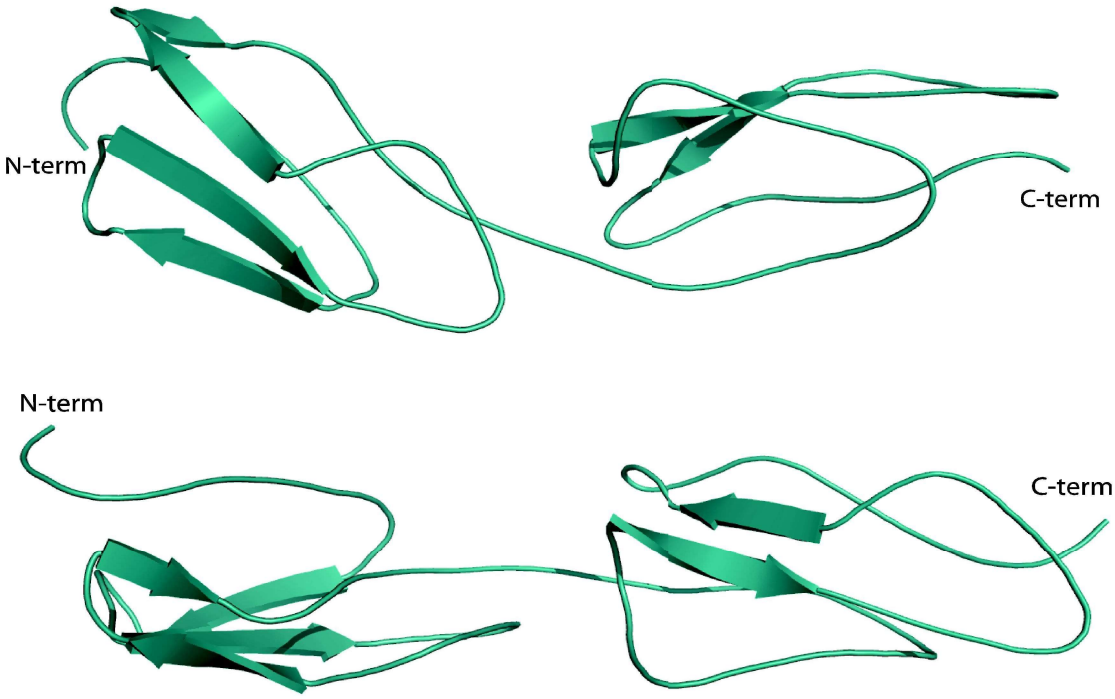

Figure 7-2: Cartoon representation of the representative CCP12 structure. Image shows the relative orientation of the two modules. The bottom image has been rotated 90° on the x-axis.

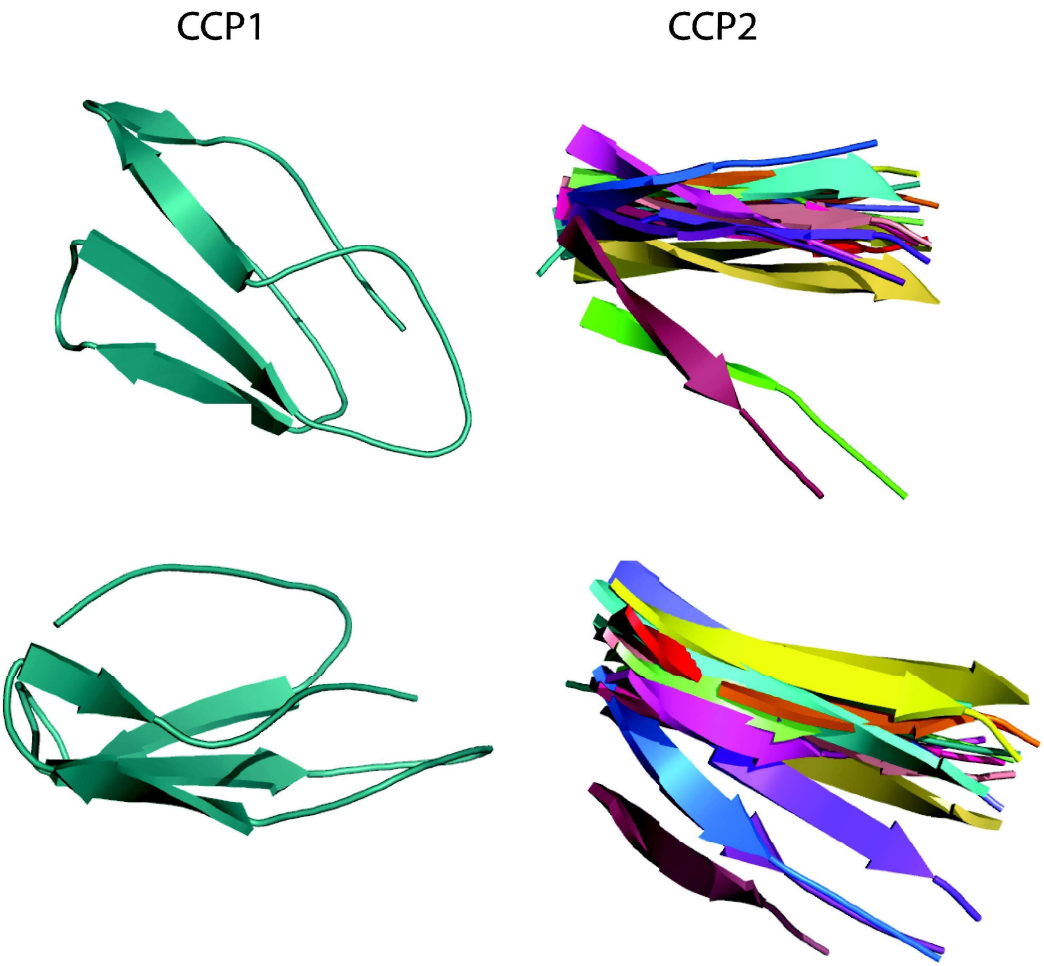

Figure 7-3: Cartoon representation of CCP2 β-strands orientation relative to CCP1. CCP2 shows the backbone cartoon for residues 203 – 210.

In the cartoon representation shown in Figure 7-3 the central β-strand of the β-sheet of CCP2 is shown for all the structures in the ensemble, with all structures superimposed on CCP1, to illustrate the range of orientations seen. One structure of CCP1 is shown as a full cartoon representation. The relative orientation of the two modules is not completely defined by the experimental data used so far, which may reflect real inter-module flexibility.

Electrostatic contact potential at the surface of the protein was analysed (Figure 7-4). At a glance, it is not obvious where a lipid A binding site would be. However, the CCP1 module shows a fairly positive patch close to the inter-module linker. This could suggest a possible binding region for the negatively charged phosphate groups of lipid A. Hydrophobic regions would recognise the acyl chains, and a small pocket adjacent to the positively charged patch gives an indication of where this could occur. The pocket is present in all eighteen structures. However, the pocket does not seem big enough for a

whole acyl chain to fit in. The CCP2 module also displays a positive patch, but this region is not very big.

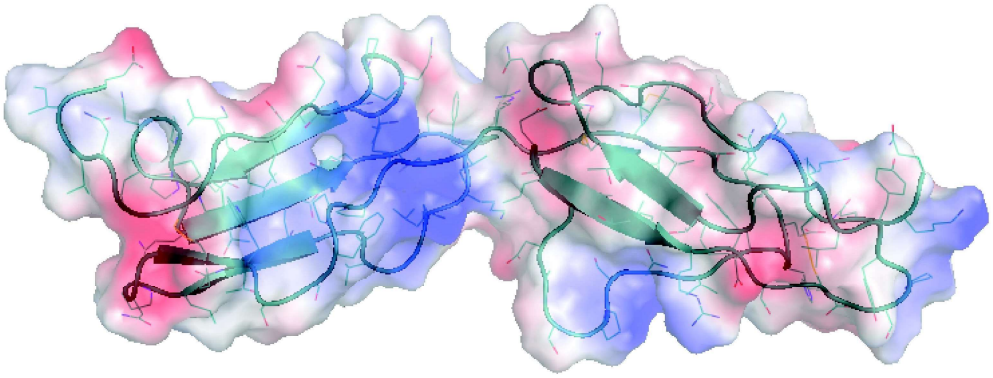

Figure 7-4: CCP12 with surface electrostatics. Positive regions are displayed in blue, negative regions are displayed in red. The CCP12 structure with the lowest energy was used to produce this image.

Figure 7-5 shows the tryptophan residues for both CCP1 and CCP2. In both modules, it can be seen that these residues are buried in the module and stacked against the disulphides, a feature typical to CCPs. This confirms that they are playing a key structural role, and are unlikely to function in lipid binding.

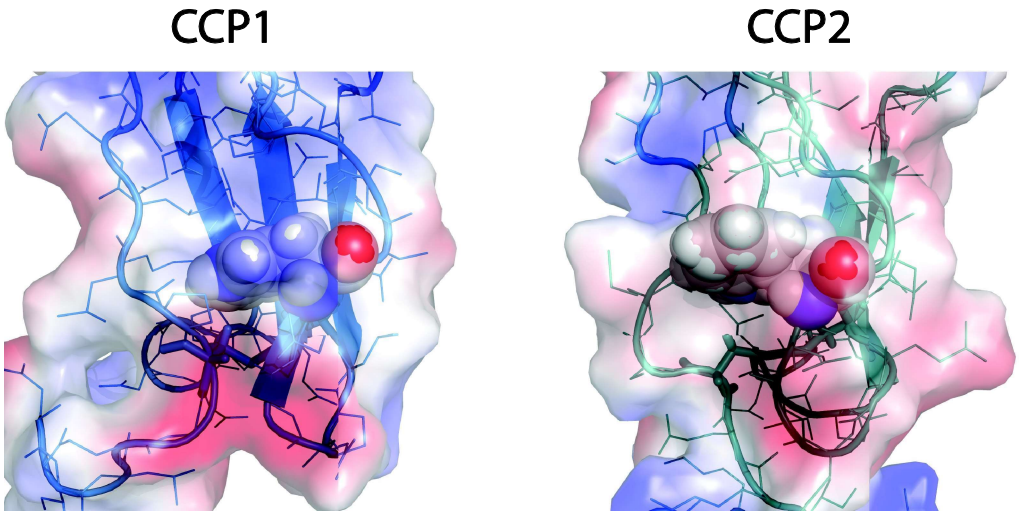

Figure 7-5: CCP1 and CCP2 tryptophans. The tryptophan residues are highlighted by the spheres and disulphide bonds are indicated by the sticks.

7.4 Summary

Overall, it is clear that the structure calculations have not yet produced a well-defined structure for CCP12. The low protein concentration, and thus low signal to noise, resulted in a low number of NOESY cross-peaks, meaning that limited data was available to provide long range distance restraints. However, from the data available, the combinations of residues forming the two disulphide bonds in each module were identified, and 25 out of 26 slowly exchanging amides were assigned to specific hydrogen bonds. Analysis of the consistently violated restraints helped to improve the consistency of the assignment resulting in better convergence of the calculations, and water refinement further enhanced the structures. CCP1 shows similarities to known CCP modules and is better defined in the calculated structures than CCP2, but it is likely that further NMR experiments with higher concentration samples will provide the additional restraints required to bring this domain in line with CCP1 to calculate the overall structure.

8 Discussion

This project aimed to identify and understand how Factor C binds LPS; to characterise the conformational changes induced by binding LPS at the molecular level; and to identify why Factor C binding to LPS is so specific to lipid A. However, throughout this project, difficulties faced with protein expression and purification frustrated attempts to achieve a high enough concentration of cleaved, soluble protein. Thus, although promising results were obtained for a number of Factor C protein fragments, further investigation is required in order to draw firm conclusions.

8.1 Evaluation of the Results Obtained from *E. coli* Expressed Fragments

For each of the *E. coli* expressed Factor C constructs, inferences can be made about their degree of folding and involvement in LPS binding, supported by results from circular dichroism (CD) and nuclear magnetic resonance spectroscopy (NMR). CD analysis of the Cys-rich domain displayed a large proportion of unordered structure, with the presence of a small amount of α-helix and β-sheet. This was supported by the NMR data, that suggested this domain was only partially folded. To determine if correct Cys-rich folding is condition dependent, further experiments are required. This could include altering the pH and observing any changes to the structure by NMR. Disulphide bond mapping would also be valuable for assessing whether a particular fold is consistently adopted by the recombinantly produced protein, and for comparison with the native protein. Cys point mutants could be produced to eliminate the unpaired Cys which may improve the behaviour of the protein. The assumption that the Cys-rich region is a domain may not be true as there is no evidence of similarity at the sequence level to other domains of known 3D structure.

The EGF-like domain appeared highly unordered in results from CD. This was confirmed by the NMR spectra, where peaks that were suggestive of folded EGF-like protein in the fusion protein became largely disordered upon cleavage from Trx. This suggests that the EGF-like domain requires help for correct folding. The CysEGF di-domain fragment was particularly difficult to obtain in soluble form. For the material that was obtained, CD implied a slight conformational change might occur upon the addition of LPS, suggesting that this fragment encompasses a potential LPS-binding

site. If this is an LPS binding fragment, that could account for the problems encountered during expression, as endogenous LPS could have affected the protein's behaviour and solubility. NMR data of a Trx-CysEGF sample shown to contain protein by both SDS-PAGE and UV-Vis spectroscopy did not reveal any cross-peaks. This could be due to the protein being bound to LPS micelles, causing the protein to tumble slowly in solution such that its signals are broadened beyond detection. As the combination of Cys with EGF did not give rise to any clear results, expressing these two domains alongside another domain, for example CCP1, may help with protein folding. A di-domain including the EGF-like domain and CCP1 may also give more information as to the requirements for correct EGF-like domain folding.

As a result of precipitation in the CCP1 and CCP2 single domain samples, CD analysis was not possible due to excessive light scattering. However, ^{15}N-HSQC NMR spectra of the two domains individually before precipitation showed a good dispersion of peaks implying the presence of correctly folded protein. Low sample concentrations and the difficulty faced when attempting to separate the domains from the fusion protein meant studies with these fragments were not taken any further. If the CCP12 domain was validated as a site for LPS binding, it would be useful to optimise protein expression and purification of the individual domains, to localise the binding site more precisely. This would also confirm whether the presence of a single domain is sufficient for LPS-binding, or if the di-domain is required for correct function.

For the more tractable CCP12 di-domain, the results from CD and NMR analysis were in agreement, with both showing a high β- strand content and very low helical contributions, in line with what is already known about the structure of CCP modules. CCP12 displayed well-dispersed peaks in NMR spectra, indicating good folding of the fragment, and allowing the structure to be calculated for this fragment as described in Chapter 7. As expected for CCP domains, the conserved tryptophan residues' side chains were shown to be buried within the structure. More work to further refine the structure is required, but the initial structural studies are of sufficient quality to begin to draw conclusions. According to the CD data, a structural change occurs upon addition of LPS, implying the presence of an LPS binding site/s within this di-domain. However, the lack of sample at a high enough concentration meant these experiments could not be

reproduced. These results are very promising, and give rise to further scope for additional structure/function analyses.

Expression of soluble CCP23 proved rather unsuccessful and thus, no NMR data was acquired for this di-domain. CD suggested the presence of rigidly held aromatics, but a poor NRMSD of the fits of the solution from DICHROWEB to the far UV CD spectrum suggests correct folding had not been achieved or it does not represent any structures present in the database. CCP123 appeared to produce correctly folded protein, as the CD spectra showed a high contribution from β-sheets, and low α-helical content, typical of CCP modules. However, ^{15}N HSQC NMR spectra contained only a small number of peaks that could be attributed to CCP123, mostly in the random coil chemical shift region. This is consistent with the tri-domain fragment tumbling slowly in solution, perhaps due to the elongated conformation one would predict for a relatively rigid rod of CCP domains joined end to end. NMR experiments designed to detect signals from slowly tumbling molecules may enable this fragment to be studied, although ^{15}N TROSY experiments were attempted unsuccessfully. Previous NMR studies of triple CCP domain fragments have often required the protein to be deuterated to improve its relaxation properties (Smith *et al.*, 2002).

In the UV spectra of a number of fragments, a higher A_{260} than A_{280} value was observed, suggestive of nucleic acid contamination. However, LPS also gives a higher A_{260} than A_{280}, implying endogenous LPS may be bound to these fragments. Although RP-HPLC was used in most cases, this technique may not be sufficiently denaturing to strip tightly bound LPS molecules from the protein fragments.

8.2 Comparison of CCP12 with other LPS Binding Proteins

Comparisons between CCP12 and known LPS-binding proteins, namely recombinant anti-LPS Factor (rALF) (Yang *et al.*, 2009) and the ferric hydroxamate uptake receptor (FhuA) (Ferguson *et al.*, 2000) were made in an attempt to identify regions of similarity between the three, in an effort to locate the LPS-binding site within the CCP12 modules. Figure 8-1 displays the three structures for comparison. Lipid A contains phosphate groups that are negatively charged, meaning the positively charged residues of CCP12 and other LPS binding proteins are likely to bind it. It is clear that rALF and FhuA have much more positively charged patches than CCP12, suggesting that, if this

is indeed the site for LPS binding, the molecular mechanisms of recognition between the three are different.

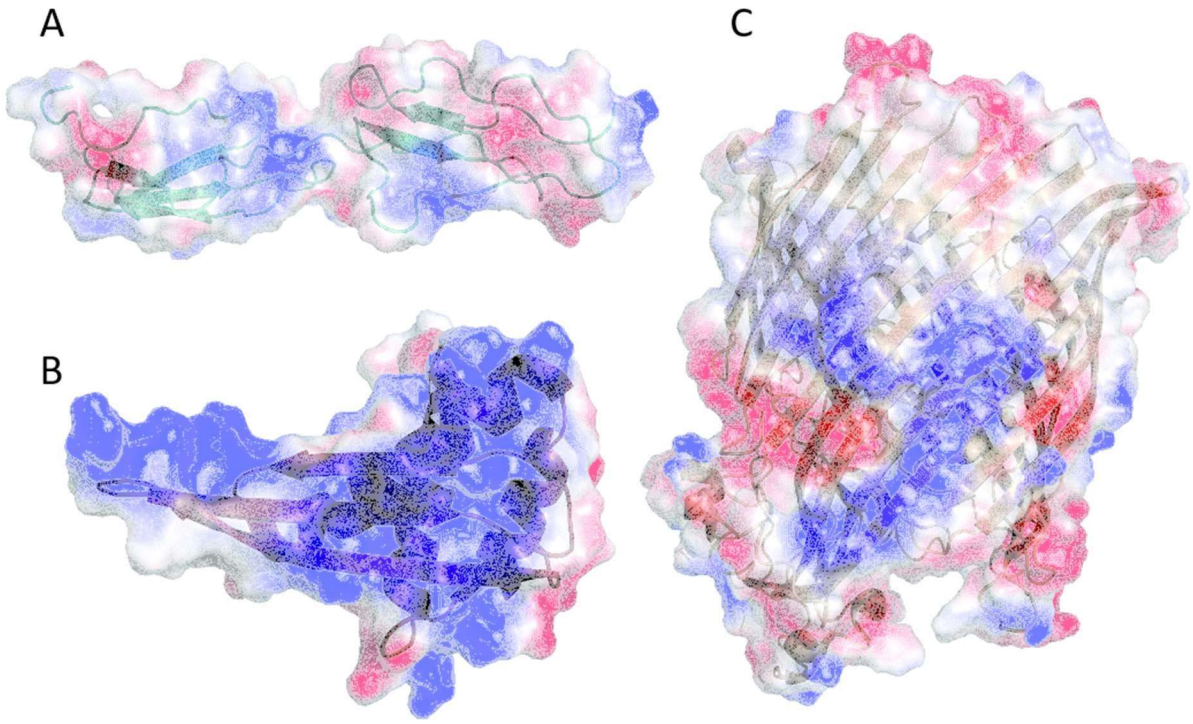

Figure 8-1: Surface comparison of CCP12, rALF and FhuA. A. CCP12. B. rALF (PDB: 2JOB). C. FhuA (PDB: 1QFG). The blue patches on the surface of the structures indicate positive regions and the red patches signify negative regions. Very large areas of the rALF and FhuA are positively charged in comparison to CCP12. Although negative patches are evident, they are much smaller in size, suggesting that if this is the LPS-binding site, that the proteins have different mechanisms of recognition. Pictures made in PyMOL using protein contact potential tool and coloured at matching electrostatic potentials. Images are not to scale.

Koshiba *et al.* identified a tripeptide motif in Factor C (Arg-Trp-Arg) thought to play an important role in LPS-binding (Koshiba *et al.*, 2007). They demonstrated that this feature is conserved amongst other LPS binding proteins, including human LPS-binding protein and *Limulus* anti-LPS factor (Figure 8-2). CCP12 was examined to establish whether or not this motif was present. While CCP12 does not exhibit a conserved tripeptide motif of the form basic-aromatic-basic, there is a surface exposed Phe next to a positive patch that could potentially function in the same way.

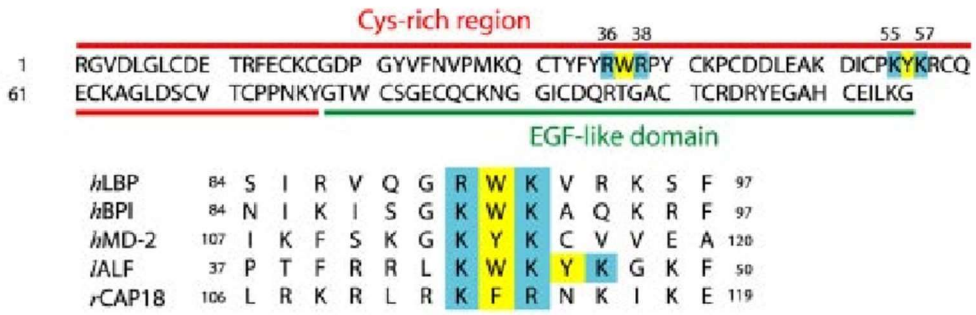

Figure 8-2: The inferred conserved tripeptide-motif for LPS-binding. Comparisons are made between hLBP, human LPS binding protein; hBPI, human bactericidal permeability-increasing protein; hMD-2, human MD-2' lALF, Limulus anti-LPS factor; and rCAP18, rabbit cationic antimicrobial protein, identifying a conserved basic-aromatic-basic motif. Figure reproduced from Koshiba *et al.*. (Koshiba *et al.*, 2007).

Based on their conclusion that CCP1 from *Carcinoscorpius rotundicauda* was sufficient for LPS binding, Tan *et al.* designed a 34 amino acid peptide (S1) from residues of the N-terminal sushi domain region in Factor C, and claimed it was a functional LPS-binding peptide (Tan *et al.*, 2000). For comparison, the amino acids in CCP12 corresponding to those that made up this peptide are highlighted on the CCP12 structure (residues 154 – 187, Figure 8-3). As illustrated in the figure, the peptide comprises just the two C-terminal extended strands of CCP1 and the intervening loop, and then extends through the linker and a few residues into CCP2. It seems unlikely that this peptide would adopt its native conformation in isolation since it represents less than half of a CCP domain and would expose several normally buried hydrophobic residues in this state. It also contains two cysteine residues that would usually form disulphide bonds with other Cys residues in the two domains. In the absence of their normal partners and under oxidizing conditions, they would very likely form non-native intra- or inter-molecular disulphides. Thus, it seems questionable that the peptide would adopt the same 3D structure as it does in CCP1, making it an unlikely candidate for an LPS binding site.

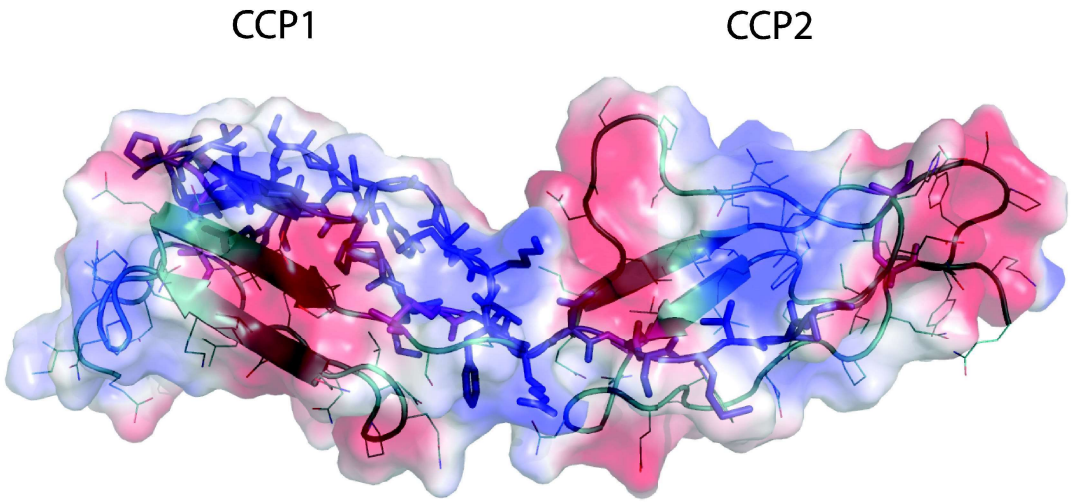

Figure 8-3: CCP12 displaying the S1 peptide. Indicated by sticks, the peptide is shown to extend from CCP1 into CCP2 (Tan *et al.*, 2000).

8.3 Overall Conclusions

Overall, no clear conclusions can yet be drawn as to the mechanism of Factor C binding to LPS, but, as shown by the results, there is plenty of scope to investigate this further.

8.3.1 Recombinant Expression in *E. coli*

This project determined that production of Factor C fragments in *E. coli* is possible and that at least some adopt well defined 3D conformations. However, it was not determined to what extent the presence of endogenous LPS affected the fragments produced. Optimised recombinant protein expression in *E. coli* would allow an easy means to produce samples from which a more thorough understanding of the way in which Factor C binds to LPS could be gained.

8.3.2 Recombinant Expression in Alternative Expression Systems

Identifying the *L. polyphemus* gene sequence and consequently cloning full-length Factor C into the different eukaryotic expression systems has paved the way for future experiments. Optimising expression of Factor C in mammalian, insect and yeast expression systems and identifying which is the most suitable for full-length protein expression will enable attempts to be made to characterise the overall structure of Factor C, and thus the structural change that occurs as a result of LPS binding This will result in a better understanding of the mechanism of Factor C binding to LPS.

Further investigations into the use of *Brevibacillus choshinensis* for protein expression, to determine the advantages this expression system has over other recombinant expression systems, will help to identify the role this system could play in the search to understand the endotoxin sensing nature of Factor C.

8.3.3 Lipid Binding Investigations

Structural changes were observed by CD in a few of the domains upon addition of LPS, indicating where further investigations should focus as discussed in Chapter 5. Increasing the concentration of purified, cleaved protein will enable more experiments to be carried out, improving reliability of the results observed. Employing other biophysical techniques such as isothermal titration calorimetry (ITC) could be useful for measuring the binding affinity between LPS and fragments of or full-length Factor C protein, which would assist in identifying the lipid binding domain.

8.3.4 Towards the structure of Factor C

From a CCP12 sample of relatively low concentration, valuable information has been garnered in the search to understand the mechanism of LPS binding to Factor C. Although the overall findings from the project only produced a preliminary NMR-based structure for one fragment, initial investigations into several other Factor C fragments produced results that suggest the acquisition of further structural data will be possible.

9 References

Aguilar, M.-I., and M. T. W. Hearn, 1996, High-resolution reversed-phase high-performance liquid chromatography of peptides and proteins, Methods in Enzymology, v. Volume 270, Academic Press, p. 3-26.

Aguilar, M. I., 2004, Reversed-phase high-performance liquid chromatography: Methods Mol Biol, v. 251, p. 9-22.

Aketagawa, J., T. Miyata, S. Ohtsubo, T. Nakamura, T. Morita, H. Hayashida, S. Iwanaga, T. Takao, and Y. Shimonishi, 1986, Primary structure of *limulus* anticoagulant anti-lipopolysaccharide factor: Journal of Biological Chemistry, v. 261, p. 7357-7365.

Appella, E., I. T. Weber, and F. Blasi, 1988, Structure and function of epidermal growth factor-like regions in proteins: FEBS Letters, v. 231, p. 1-4.

Ariki, S., K. Koori, T. Osaki, K. Motoyama, K.-i. Inamori, and S.-i. Kawabata, 2004, A serine protease zymogen functions as a pattern-recognition receptor for lipopolysaccharides: Proceedings of the National Academy of Sciences of the United States of America, v. 101, p. 953-958.

Armstrong, P., and M. Conrad, 2008, Blood Collection from the American Horseshoe Crab, Limulus Polyphemus: Journal of Visualized Experiments : JoVE, p. 958.

Aslanidis, C., and P. J. de Jong, 1990, Ligation-independent cloning of PCR products (LIC-PCR): Nucleic Acids Res, v. 18, p. 6069-74.

Bang, F. B., 1956, A bacterial disease of Limulus polyphemus: Bull Johns Hopkins Hosp, v. 98, p. 325-51.

Bazan, J. F., 1993, Emerging families of cytokines and receptors: Current Biology, v. 3, p. 603-606.

Berkmen, M., 2012, Production of disulfide-bonded proteins in Escherichia coli: Protein Expression and Purification, v. 82, p. 240-251.

Berkson, J., and C. N. Shuster, 1999, The Horseshoe Crab: The Battle for a True Multiple-use Resource: Fisheries, v. 24, p. 6-10.

Bevilacqua, M. P., S. Stengelin, M. A. Gimbrone, and B. Seed, 1989, Endothelial leukocyte adhesion molecule 1: an inducible receptor for neutrophils related to complement regulatory proteins and lectins: Science, v. 243, p. 1160.

Blow, D. M., 1997, The tortuous story of Asp…His…Ser: Structural analysis of α-chymotrypsin: Trends in Biochemical Sciences, v. 22, p. 405-408.

Bornhorst, J. A., and J. J. Falke, 2000, Purification of proteins using polyhistidine affinity tags: Methods Enzymol, v. 326, p. 245-54.

Botton, M. L., and R. E. Loveland, 1989, Reproductive risk: high mortality associated with spawning by horseshoe crabs (*Limulus polyphemus*) in Delaware Bay, USA: Marine Biology, v. 101, p. 143-151.

Brockmann, H. J., 1990, Mating behavior of horseshoe crabs, *Limulus polyphemus*: Behaviour, v. 114, p. 206-220.

Carrington, J. C., and W. G. Dougherty, 1987a, Processing of the tobacco etch virus 49K protease requires autoproteolysis: Virology, v. 160, p. 355-362.

Carrington, J. C., and W. G. Dougherty, 1987b, Small Nuclear Inclusion Protein Encoded by a Plant Potyvirus Genome Is a Protease: Journal of Virology, v. 61, p. 2540-2548.

Cavanagh, J., W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton, 1996, Protein NMR Spectroscopy: Principles and Practice, Academic Press.

Ceramil, A., and B. Beutler, 1988, The role of cachectin/TNF in endotoxic shock and cachexia: Immunology Today, v. 9, p. 28-31.

Cheung, M.-S., M. L. Maguire, T. J. Stevens, and R. W. Broadhurst, 2010, DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure: Journal of magnetic resonance, v. 202, p. 223-233.

Chow, J. C., D. W. Young, D. T. Golenbock, W. J. Christ, and F. Gusovsky, 1999, Toll-like receptor-4 mediates lipopolysaccharide-induced signal transduction: Journal of Biological Chemistry, v. 274, p. 10689-10692.

Christie, J. M., K. Hitomi, A. S. Arvai, K. A. Hartfield, M. Mettlen, A. J. Pratt, J. A. Tainer, and E. D. Getzoff, 2012, Structural Tuning of the Fluorescent Protein iLOV for Improved Photostability: Journal of Biological Chemistry, v. 287, p. 22295-22304.

Clancy, S., 2008, RNA Splicing: Introns, Exons and Spliceosome: Nature Education, v. 1, p. 1.

Clare, J. J., F. B. Rayment, S. P. Ballantine, K. Sreekrishna, and M. A. Romanos, 1991, High-level expression of tetanus toxin fragment C in Pichia pastoris strains containing multiple tandem integrations of the gene: Bio/technology (Nature Publishing Company), v. 9, p. 455-460.

Clubb, R., V. Thanabal, and G. Wagner, 1992, A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C′ chemical shifts in 15N13C-labelled proteins: Journal of Magnetic Resonance (1969), v. 97, p. 213-217.

Cordingley, M. G., R. B. Register, P. L. Callahan, V. M. Garsky, and R. J. Colonno, 1989, Cleavage of small peptides in vitro by human rhinovirus 14 3C protease expressed in Escherichia coli: Journal of Virology, v. 63, p. 5037-5045.

Cregg, J. M., J. F. Tschopp, C. Stillman, R. Siegel, M. Akong, W. S. Craig, R. G. Buckholz, K. R. Madden, P. A. Kellaris, and G. R. Davis, 1987, High–Level Expression and Efficient Assembly of Hepatitis B Surface Antigen in the Methylotrophic Yeast, Pichia Pastoris: Nature Biotechnology, v. 5, p. 479-485.

Davis, A. L., J. Keeler, E. D. Laue, and D. Moskau, 1992, Experiments for recording pure-absorption heteronuclear correlation spectra using pulsed field gradients: Journal of Magnetic Resonance (1969), v. 98, p. 207-216.

Day, A. J., J. Ripoche, A. C. Willis, and R. B. Sim, 1987, Structure and polymorphism of human factor H: Complement, v. 4, p. 147-148.

Ding, J. L., M. A. A. Navas, and B. Ho, 1993, Two forms of Factor C from the amoebocytes of *Carcinoscorpius rotundicauda*: purification and characterisation: Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology, v. 1202, p. 149-156.

Drickamer, K., 1988, Two distinct classes of carbohydrate-recognition domains in animal lectins: Journal of Biological Chemistry, v. 263, p. 9557-9560.

D'Urzo, N., M. Martinelli, C. Nenci, C. Brettoni, J. L. Telford, and D. Maione, 2013, High-level intracellular expression of heterologous proteins in Brevibacillus choshinensis SP3 under the control of a xylose inducible promoter: Microbial Cell Factories, v. 12, p. 12.

Ferguson, A. D., W. Welte, E. Hofmann, B. Lindner, O. Holst, J. W. Coulton, and K. Diederichs, 2000, A conserved structural motif for lipopolysaccharide recognition by procaryotic and eucaryotic proteins: Structure, v. 8, p. 585-592.

Freskgard, P.-O., L.-G. Martensson, P. Jonasson, B.-H. Jonsson, and U. Carlsson, 1994, Assignment of the contribution of the tryptophan residues to the circular dichroism spectrum of human carbonic anhydrase II: Biochemistry, v. 33, p. 14281-14288.

Gaboriaud, C., V. Rossi, I. Bally, G. J. Arlaud, and J. C. Fontecilla-Camps, 2000, Crystal structure of the catalytic domain of human complement C1s: a serine protease with a handle: The EMBO Journal, v. 19, p. 1755.

Gao, X., P. Yo, A. Keith, T. J. Ragan, and T. K. Harris, 2003, Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences: Nucleic Acids Research, v. 31, p. e143-e143.

Grzesiek, S., and A. Bax, 1992, Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein: Journal of Magnetic Resonance (1969), v. 96, p. 432-440.

Grzesiek, S., and A. Bax, 1993a, Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins: Journal of Biomolecular NMR, v. 3, p. 185-204.

Grzesiek, S., and A. Bax, 1993b, The importance of not saturating water in protein NMR. Application to sensitivity enhancement and NOE measurements: Journal of the American Chemical Society, v. 115, p. 12593-12594.

Hamilton, M. D., A. A. Nuara, D. B. Gammon, R. M. Buller, and D. H. Evans, 2007, Duplex strand joining reactions catalyzed by vaccinia virus DNA polymerase: Nucleic acids research, v. 35, p. 143-151.

Hedstrom, L., 2002, Serine Protease Mechanism and Specificity: Chemical Reviews, v. 102, p. 4501-4524.

Hofmann, K., S. W. Wood, C. C. Brinton, J. A. Montibeller, and F. M. Finn, 1980, Iminobiotin affinity columns and their application to retrieval of streptavidin: Proc Natl Acad Sci U S A, v. 77, p. 4666-8.

Hoover, D. M., and J. Lubkowski, 2002, DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis: Nucleic Acids Research, v. 30, p. e43-e43.

Hwang, T.-L., and A. J. Shaka, 1995, Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients: Journal of Magnetic Resonance, Series A, v. 112, p. 275-279.

Ikura, M., L. E. Kay, and A. Bax, 1990, A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin: Biochemistry, v. 29, p. 4659-4667.

Imperiali, B., and S. E. O'Connor, 1999, Effect of N-linked glycosylation on glycopeptide and glycoprotein structure: Current Opinion in Chemical Biology, v. 3, p. 643-649.

Iwadate, M., T. Asakura, and M. P. Williamson, 1999, Cα and Cβ carbon-13 chemical shifts in proteins from an empirical database: Journal of biomolecular NMR, v. 13, p. 199-211.

Iwanaga, S., 2002, The molecular basis of innate immunity in the horseshoe crab: Current Opinion in Immunology, v. 14, p. 87-95.

Iwanaga, S., S.-i. Kawabata, and T. Muta, 1998, New Types of Clotting Factors and Defense Molecules Found in Horseshoe Crab Hemolymph: Their Structures and Functions: Journal of Biochemistry, v. 123, p. 1-15.

Iwanaga, S., T. Miyata, F. Tokunaga, and T. Muta, 1992, Molecular mechanism of hemolymph clotting system in *Limulus*: Thrombosis research, v. 68, p. 1-32.

Johnston, G. I., R. G. Cook, and R. P. McEver, 1989, Cloning of GMP-140, a granule membrane protein of platelets and endothelium: sequence similarity to proteins involved in cell adhesion and inflammation: Cell, v. 56, p. 1033-1044.

Joseph, A. P., N. Srinivasan, and A. G. De Brevern, 2012, Cis-trans peptide variations in structurally similar proteins: Amino Acids, v. 43, p. 1369-1381.

Kakinuma, A., T. Asano, H. Torii, and Y. Sugino, 1981, Gelation of limulus amoebocyte lysate by an antitumor (1→3)-β-D-glucan: Biochemical and Biophysical Research Communications, v. 101, p. 434-439.

Karima, R., S. Matsumoto, H. Higashi, and K. Matsushima, 1999, The molecular pathogenesis of endotoxic shock and organ failure: Molecular Medicine Today, v. 5, p. 123-132.

Kawabata, S.-i., T. Osaki, and S. Iwanaga, 2003, Innate immunity in the horseshoe crab, Innate Immunity, Springer, p. 109-125.

Kay, L. E., M. Ikura, R. Tschudin, and A. Bax, 1990, Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins: Journal of Magnetic Resonance (1969), v. 89, p. 496-514.

Kay, L. E., L. K. Nicholson, F. Delaglio, A. Bax, and D. Torchia, 1992, Pulse sequences for removal of the effects of cross correlation between dipolar and chemical-shift anisotropy relaxation mechanisms on the measurement of heteronuclear T1 and T2 values in proteins: Journal of Magnetic Resonance (1969), v. 97, p. 359-375.

Kay, L. E., G. Y. Xu, A. U. Singer, D. R. Muhandiram, and J. D. Formankay, 1993, A Gradient-Enhanced HCCH-TOCSY Experiment for Recording Side-Chain 1H and 13C Correlations in H2O Samples of Proteins: Journal of Magnetic Resonance, Series B, v. 101, p. 333-337.

Kay, L. E., G. Y. Xu, and T. Yamazaki, 1994, Enhanced-Sensitivity Triple-Resonance Spectroscopy with Minimal H2O Saturation: Journal of Magnetic Resonance, Series A, v. 109, p. 129-133.

Keeler, J., 2005, Understanding NMR Spectroscopy, Wiley.

Kelly, S. M., T. J. Jess, and N. C. Price, 2005, How to study proteins by circular dichroism: Biochim Biophys Acta, v. 1751, p. 119-39.

Kern, R., A. Malki, A. Holmgren, and G. Richarme, 2003, Chaperone properties of Escherichia coli thioredoxin and thioredoxin reductase: Biochemical Journal, v. 371, p. 965-972.

Kilpatrick, D. C., 2002, Animal lectins: a historical introduction and overview: Biochimica et biophysica acta, v. 1572, p. 187-197.

Koshiba, T., T. Hashii, and S. Kawabata, 2007, A structural perspective on the interaction between lipopolysaccharide and factor C, a receptor involved in recognition of Gram-negative bacteria: J Biol Chem, v. 282, p. 3962-7.

Kreutz, M., U. Ackermann, S. Hauschildt, S. W. Krause, D. Riedel, W. Bessler, and R. Andreesen, 1997, A comparative analysis of cytokine production and tolerance induction by bacterial lipopeptides, lipopolysaccharides and Staphyloccocus aureus in human monocytes: Immunology, v. 92, p. 396-401.

Lathe, G. H., and C. R. J. Ruthven, 1956, The separation of substances and estimation of their relative molecular sizes by the use of columns of starch in water: Biochemical Journal, v. 62, p. 665-674.

LaVallie, E. R., E. A. DiBlasio, S. Kovacic, K. L. Grant, P. F. Schendel, and J. M. McCoy, 1993, A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm: Biotechnology (N Y), v. 11, p. 187-93.

Leschen, A. S., and S. J. Correia, 2010, Mortality in female horseshoe crabs (Limulus polyphemus) from biomedical bleeding and handling: implications for fisheries management: Marine and Freshwater Behaviour and Physiology, v. 43, p. 135-147.

Levin, J., and F. B. Bang, 1968, Clottable protein in Limulus; its localization and kinetics of its coagulation by endotoxin: Thromb Diath Haemorrh, v. 19, p. 186-97.

Levitt, M. H., 2001, Spin dynamics: basics of nuclear magnetic resonance, John Wiley & Sons.

Linge, J. P., M. Habeck, W. Rieping, and M. Nilges, 2003a, ARIA: automated NOE assignment and NMR structure calculation: Bioinformatics, v. 19, p. 315-316.

Linge, J. P., M. A. Williams, C. A. E. M. Spronk, A. M. J. J. Bonvin, and M. Nilges, 2003b, Refinement of protein structures in explicit solvent: Proteins: Structure, Function, and Bioinformatics, v. 50, p. 496-506.

Liu, J. S., and C. L. Passaglia, 2009, Using the Horseshoe Crab, Limulus Polyphemus, in Vision Research: Journal of Visualized Experiments : JoVE, p. 1384.

Lobley, A., L. Whitmore, and B. A. Wallace, 2002, DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra: Bioinformatics, v. 18, p. 211-212.

Lobstein, J., C. A. Emrich, C. Jeans, M. Faulkner, P. Riggs, and M. Berkmen, 2012, SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm: Microb Cell Fact, v. 11, p. 56.

Lozier, J., N. Takahashi, and F. W. Putnam, 1984, Complete amino acid sequence of human plasma beta 2-glycoprotein I: Proceedings of the National Academy of Sciences, v. 81, p. 3640-3644.

Maley, F., R. B. Trimble, A. L. Tarentino, and T. H. Plummer, 1989, Characterization of glycoproteins and their associated oligosaccharides through the use of endoglycosidases: Analytical Biochemistry, v. 180, p. 195-204.

Mallett, S., and A. N. Barclay, 1991, A new superfamily of cell surface proteins related to the nerve growth factor receptor: Immunology today, v. 12, p. 220-223.

Mant, C. T., and R. S. Hodges, 1996, Analysis of peptides by high-performance liquid chromatography, Methods in Enzymology, v. Volume 271, Academic Press, p. 3-50.

Mardis, E. R., 2008, Next-Generation DNA Sequencing Methods: Annual Review of Genomics and Human Genetics, v. 9, p. 387-402.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, 2005, Genome sequencing in microfabricated high-density picolitre reactors: Nature, v. 437, p. 376-380.

Marley, J., M. Lu, and C. Bracken, 2001, A method for efficient isotopic labeling of recombinant proteins: Journal of Biomolecular NMR, v. 20, p. 71-75.

Marshall, R. D., 1972, Glycoproteins: Annual Review of Biochemistry, v. 41, p. 673-702.

McMullen, B. A., and K. Fujikawa, 1985, Amino acid sequence of the heavy chain of human alpha-factor XIIa (activated Hageman factor): Journal of Biological Chemistry, v. 260, p. 5328-5341.

Mole, J. E., J. K. Anderson, E. A. Davison, and D. E. Woods, 1984, Complete primary structure for the zymogen of human complement factor B: Journal of Biological Chemistry, v. 259, p. 3407-3412.

Mori, S., C. Abeygunawardana, M. O. Johnson, and P. C. M. Vanzijl, 1995, Improved Sensitivity of HSQC Spectra of Exchanging Protons at Short Interscan Delays Using a New Fast HSQC (FHSQC) Detection Scheme That Avoids Water Saturation: Journal of Magnetic Resonance, Series B, v. 108, p. 94-98.

Muhandiram, D. R., and L. E. Kay, 1994, Gradient-Enhanced Triple-Resonance Three-Dimensional NMR Experiments with Improved Sensitivity: Journal of Magnetic Resonance, Series B, v. 103, p. 203-216.

Muramoto, K., and H. Kamiya, 1986, The amino-acid sequence of a lectin of the acorn barnacle Megabalanus rosa: Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, v. 874, p. 285-295.

Muta, T., T. Miyata, Y. Misumi, F. Tokunaga, T. Nakamura, Y. Toh, Y. Ikehara, and S. Iwanaga, 1991, Limulus factor C. An endotoxin-sensitive serine protease zymogen with a mosaic structure of complement-like, epidermal growth factor-like, and lectin-like domains: Journal of Biological Chemistry, v. 266, p. 6554-6561.

Nakamura, T., T. Morita, and S. Iwanaga, 1985, Intracellular Proclotting Enzyme in Limulus (Tachypleus tridentatus) Hemocytes: Its Purification and Properties: Journal of Biochemistry, v. 97, p. 1561-1574.

Nakamura, T., T. Morita, and S. Iwanaga, 1986, Lipopolysaccharide-sensitive serine-protease zymogen (factor C) found in Limulus hemocytes. Isolation and characterization: Eur J Biochem, v. 154, p. 511-21.

Nakamura, T., F. Tokunaga, T. Morita, and S. Iwanaga, 1988a, Interaction between lipopolysaccharide and intracellular serine protease zymogen, factor C, from horseshoe crab (Tachypleus tridentatus) hemocytes: J Biochem, v. 103, p. 370-4.

Nakamura, T., F. Tokunaga, T. Morita, S. Iwanaga, S. Kusumoto, T. Shiba, T. Kobayashi, and K. Inoue, 1988b, Intracellular serine-protease zymogen, factor C, from horseshoe crab hemocytes. Its activation by synthetic lipid A analogues and acidic phospholipids: Eur J Biochem, v. 176, p. 89-94.

Newman, J., D. Egan, T. S. Walter, R. Meged, I. Berry, M. Ben Jelloul, J. L. Sussman, D. I. Stuart, and A. Perrakis, 2005, Towards rationalization of crystallization screening for small-to medium-sized academic laboratories: the PACT/JCSG+ strategy: Acta Crystallographica Section D: Biological Crystallography, v. 61, p. 1426-1431.

Nilges, M., A. Bernard, B. Bardiaux, T. Malliavin, M. Habeck, and W. Rieping, 2008, Accurate NMR structures through minimization of an extended hybrid energy: Structure, v. 16, p. 1305-1312.

Norman, D. G., P. N. Barlow, M. Baron, A. J. Day, R. B. Sim, and I. D. Campbell, 1991, Three-dimensional structure of a complement control protein module in solution: Journal of Molecular Biology, v. 219, p. 717-725.

Novitsky, T. J., 1984, Discovery to commercialization - the blood of the horseshoe-crab.: Oceanus, v. 27, p. 13-18.

Ogata, S.-I., and K. O. Lloyd, 1982, Mild alkaline borohydride treatment of glycoproteins—a method for liberating both N-and O-linked carbohydrate chains: Analytical biochemistry, v. 119, p. 351-359.

Okun, S. B., 2012, Mating in the moonlight: the battle to save the American horseshoe crab: Ocean & Coastal LJ, v. 18, p. 195.

Orekhov, V. Y., I. Ibraghimov, and M. Billeter, 2003, Optimizing resolution in multidimensional NMR by three-way decomposition: Journal of Biomolecular NMR, v. 27, p. 165-173.

Page, R., S. K. Grzechnik, J. M. Canaves, G. Spraggon, A. Kreusch, P. Kuhn, R. C. Stevens, and S. A. Lesley, 2003, Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome: Acta Crystallographica Section D: Biological Crystallography, v. 59, p. 1028-1037.

Palmer, I., and P. T. Wingfield, 2004, Preparation and Extraction of Insoluble (Inclusion-Body) Proteins from Escherichia coli: Current protocols in protein science / editorial board, John E. Coligan ... [*et al.*], v. CHAPTER, p. Unit-6.3.

Park, B. S., and J.-O. Lee, 2013, Recognition of lipopolysaccharide pattern by TLR4 complexes: Exp Mol Med, v. 45, p. e66.

Porath, J., J. A. N. Carlsson, I. Olsson, and G. Belfrage, 1975, Metal chelate affinity chromatography, a new approach to protein fractionation: Nature, v. 258, p. 598-599.

Porath, J., and P. E. R. Flodin, 1959, Gel Filtration: A Method for Desalting and Group Separation: Nature, v. 183, p. 1657-1659.

Pristovšek, P., K. Fehér, L. Szilágyi, and J. Kidrič, 2005, Structure of a synthetic fragment of the LALF protein when bound to lipopolysaccharide: Journal of medicinal chemistry, v. 48, p. 1666-1670.

Provencher, S. W., and J. Glöckner, 1981, Estimation of globular protein secondary structure from circular dichroism: Biochemistry, v. 20, p. 33-37.

Raetz, C. R. H., and C. Whitfield, 2002, Lipopolysaccharide Endotoxins: Annual Review of Biochemistry, v. 71, p. 635-700.

Rawlings, N. D., and A. J. Barrett, 1994, [2] Families of serine peptidases, Methods in Enzymology, v. Volume 244, Academic Press, p. 19-61.

Reddy, K. N. N., and G. Markus, 1972, Mechanism of activation of human plasminogen by streptokinase Presence of active center in streptokinase-plasminogen complex: Journal of Biological Chemistry, v. 247, p. 1683-1691.

Reid, K. B. M., and A. J. Day, 1989, Structure-function relationships of the complement components: Immunology Today, v. 10, p. 177-180.

Rieping, W., M. Habeck, B. Bardiaux, A. Bernard, T. E. Malliavin, and M. Nilges, 2007, ARIA2: Automated NOE assignment and data integration in NMR structure calculation: Bioinformatics, v. 23, p. 381-382.

Robertson, N. G., L. Lu, S. Heller, S. N. Merchant, R. D. Eavey, M. McKenna, J. B. Nadol, R. T. Miyamoto, F. H. Linthicum, J. F. Lubianca Neto, A. J. Hudspeth, C. E. Seidman, C. C. Morton, and J. G. Seidman, 1998, Mutations in a novel cochlear gene cause DFNA9, a human nonsyndromic deafness with vestibular dysfunction: Nat Genet, v. 20, p. 299-303.

Rosano, G. L., and E. A. Ceccarelli, 2014, Recombinant protein expression in Escherichia coli: advances and challenges: Front Microbiol, v. 5, p. 172.

Rudloe, A., 1983, The effect of heavy bleeding on mortality of the horseshoe crab, Limulus polyphemus, in the natural environment: Journal of Invertebrate Pathology, v. 42, p. 167-176.

Rudolph, R., and H. Lilie, 1996, In vitro folding of inclusion body proteins: The FASEB Journal, v. 10, p. 49-56.

Sambrook, J., E. F. Fritsch, and T. Maniatis, 1989, Molecular cloning, v. 2, Cold spring harbor laboratory press New York.

Sattler, M., J. Schleucher, and C. Griesinger, 1999, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients: Progress in Nuclear Magnetic Resonance Spectroscopy, v. 34, p. 93-158.

Savage, C. R., J. H. Hash, and S. Cohen, 1973, Epidermal growth factor location of disulfide bonds: Journal of Biological Chemistry, v. 248, p. 7669-7672.

Savitsky, P., J. Bray, C. D. Cooper, B. D. Marsden, P. Mahajan, N. A. Burgess-Brown, and O. Gileadi, 2010, High-throughput production of human proteins for crystallization: the SGC experience: J Struct Biol, v. 172, p. 3-13.

Schleucher, J., M. Sattler, and C. Griesinger, 1993, Coherence Selection by Gradients without Signal Attenuation: Application to the Three-Dimensional HNCO Experiment: Angewandte Chemie International Edition in English, v. 32, p. 1489-1491.

Schwarz, F., and M. Aebi, 2011, Mechanisms and principles of N-linked protein glycosylation: Current Opinion in Structural Biology, v. 21, p. 576-582.

Shaka, A. J., C. J. Lee, and A. Pines, 1988, Iterative schemes for bilinear operators; application to spin decoupling: Journal of Magnetic Resonance (1969), v. 77, p. 274-293.

Shizuya, H., B. Birren, U. J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon, 1992, Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector: Proceedings of the National Academy of Sciences of the United States of America, v. 89, p. 8794-8797.

Siegelman, M. H., M. Van De Rijn, and I. L. Weissman, 1989, Mouse lymph node homing receptor cDNA clone encodes a glycoprotein revealing tandem interaction domains: Science, v. 243, p. 1165.

Sivashanmugam, A., V. Murray, C. Cui, Y. Zhang, J. Wang, and Q. Li, 2009, Practical protocols for production of very high yields of recombinant proteins using Escherichia coli: Protein Science : A Publication of the Protein Society, v. 18, p. 936-948.

Sklenar, V., M. Piotto, R. Leppik, and V. Saudek, 1993, Gradient-Tailored Water Suppression for 1H-15N HSQC Experiments Optimized to Retain Full Sensitivity: Journal of Magnetic Resonance, Series A, v. 102, p. 241-245.

Smith, B. O., R. L. Mallin, M. Krych-Goldberg, X. Wang, R. E. Hauhart, K. Bromek, D. Uhrin, J. P. Atkinson, and P. N. Barlow, 2002, Structure of the C3b Binding Site of CR1 (CD35), the Immune Adherence Receptor: Cell, v. 108, p. 769-780.

Sreekrishna, K., L. Nelles, R. Potenz, J. Cruze, P. Mazzaferro, W. Fish, M. Fuke, K. Holden, and D. Phelps, 1989, High-level expression, purification, and characterization of recombinant human tumor necrosis factor synthesized in the methylotrophic yeast Pichia pastoris: Biochemistry, v. 28, p. 4117-4125.

Sreerama, N., and R. W. Woody, 2000, Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set: Analytical Biochemistry, v. 287, p. 252-260.

Stuart, L. M., and R. A. B. Ezekowitz, 2005, Phagocytosis: Elegant Complexity: Immunity, v. 22, p. 539-550.

Studier, F. W., and B. A. Moffatt, 1986, Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes: J Mol Biol, v. 189, p. 113-30.

Studier, W. F., A. H. Rosenberg, J. J. Dunn, and J. W. Dubendorff, 1990, Use of T7 RNA polymerase to direct expression of cloned genes, Methods in Enzymology, v. Volume 185, Academic Press, p. 60-89.

Sørensen, H. P., and K. K. Mortensen, 2005, Advanced genetic strategies for recombinant protein expression in Escherichia coli: J Biotechnol, v. 115, p. 113-28.

Tai, J. Y., and T. Y. Liu, 1977, Studies on Limulus amoebocyte lysate. Isolation of pro-clotting enzyme: Journal of Biological Chemistry, v. 252, p. 2178-2181.

Tai, J. Y., R. C. Seid, R. D. Huhn, and T. Y. Liu, 1977, Studies on Limulus amoebocyte lysate II. Purification of the coagulogen and the mechanism of clotting: Journal of Biological Chemistry, v. 252, p. 4773-4776.

Tan, N. S., M. L. Ng, Y. H. Yau, P. K. Chong, B. Ho, and J. L. Ding, 2000, Definition of endotoxin binding sites in horseshoe crab factor C recombinant sushi proteins and neutralization of endotoxin by sushi peptides: FASEB J, v. 14, p. 1801-13.

Tanio, M., T. Tanaka, and T. Kohno, 2008, 15N isotope labeling of a protein secreted by Brevibacillus choshinensis for NMR study: Analytical Biochemistry, v. 373, p. 164-166.

Toh, Y., A. Mizutani, F. Tokunaga, T. Muta, and S. Iwanaga, 1991, Morphology of the granular hemocytes of the Japanese horseshoe crabTachypleus tridentatus and

immunocytochemical localization of clotting factors and antimicrobial substances: Cell and Tissue Research, v. 266, p. 137-147.

Tokunaga, F., T. Miyata, T. Nakamura, T. Morita, K. Kuma, and S. Iwanaga, 1987, Lipopolysaccharide-sensitive serine-protease zymogen (factor C) of horseshoe crab hemocytes. Identification and alignment of proteolytic fragments produced during the activation show that it is a novel type of serine protease: Eur J Biochem, v. 167, p. 405-16.

Tokunaga, F., H. Nakajima, and S. Iwanaga, 1991, Further Studies on Lipopolysaccharide-Sensitive Serine Protease Zymogen (Factor C): Its Isolation from Limulus polyphemus Hemocytes and Identification as an Intracellular Zymogen Activated by α-Chymotrypsin, Not by Trypsin: Journal of Biochemistry, v. 109, p. 150-157.

Trent, M. S., C. M. Stead, A. X. Tran, and J. V. Hankins, 2006, Invited review: Diversity of endotoxin and its impact on pathogenesis: Journal of Endotoxin Research, v. 12, p. 205-223.

Trexler, M., L. Bányai, and L. Patthy, 2000, The LCCL module: Eur J Biochem, v. 267, p. 5751-7.

Udaka, S., and H. Yamagata, 1993, High-level secretion of heterologous proteins Bacillus brevis, Methods in Enzymology, v. Volume 217, Academic Press, p. 23-33.

Uhlén, M., B. Nilsson, B. Guss, M. Lindberg, S. Gatenbeck, and L. Philipson, 1983, Gene fusion vectors based on the gene for staphylococcal protein A: Gene, v. 23, p. 369-378.

Ulrich, E. L., H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, and Z. Miller, 2008, BioMagResBank: Nucleic acids research, v. 36, p. D402-D408.

Van Amersfoort, E. S., T. J. C. Van Berkel, and J. Kuiper, 2003, Receptors, Mediators, and Mechanisms Involved in Bacterial Sepsis and Septic Shock: Clinical Microbiology Reviews, v. 16, p. 379-414.

Vranken, W. F., W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue, 2005, The CCPN data model for NMR spectroscopy: Development of a software pipeline: Proteins: Structure, Function, and Bioinformatics, v. 59, p. 687-696.

Wang, A. C., P. J. Lodi, J. Qin, G. W. Vuister, A. M. Gronenborn, and G. M. Clore, 1994, An efficient triple-resonance experiment for proton-directed sequential backbone assignment of medium-sized proteins: Journal of Magnetic Resonance, Series B, v. 105, p. 196-198.

Wang, X., and P. J. Quinn, 2010, Lipopolysaccharide: Biosynthetic pathway and structure modification: Progress in lipid research, v. 49, p. 97-107.

Whitmore, L., and B. A. Wallace, 2004, DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data: Nucleic Acids Research, v. 32, p. W668-W673.

Whitmore, L., and B. A. Wallace, 2008, Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases: Biopolymers, v. 89, p. 392-400.

Wishart, D. S., C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, and B. D. Sykes, 1995, 1H, 13C and 15N chemical shift referencing in biomolecular NMR: Journal of Biomolecular NMR, v. 6, p. 135-140.

Wittekind, M., and L. Mueller, 1993, HNCACB, a High-Sensitivity 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins: Journal of Magnetic Resonance, Series B, v. 101, p. 201-205.

Yamazaki, T., J. D. Forman-Kay, and L. E. Kay, 1993, Two-dimensional NMR experiments for correlating $^{13}C\ \beta$ and $^{1}H\ \delta/\varepsilon$ chemical shifts of aromatic residues in ^{13}C-labeled proteins via scalar couplings: Journal of the American Chemical Society, v. 115, p. 11054-11055.

Yang, Y., H. Boze, P. Chemardin, A. Padilla, G. Moulin, A. Tassanakajon, M. Pugnière, F. Roquet, D. Destoumieux-Garzón, Y. Gueguen, E. Bachère, and A. Aumelas, 2009, NMR structure of rALF-Pm3, an anti-lipopolysaccharide factor from shrimp: model of the possible lipid A-binding site: Biopolymers, v. 91, p. 207-20.

Yin, J., G. Li, X. Ren, and G. Herrler, 2007, Select what you need: A comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes: Journal of Biotechnology, v. 127, p. 335-347.

Young, N. S., J. Levin, and R. A. Prendergast, 1972, An invertebrate coagulation system activated by endotoxin: evidence for enzymatic mediation: Journal of Clinical Investigation, v. 51, p. 1790-1797.

Zhang, H., S. Neal, and D. S. Wishart, 2003, RefDB: a database of uniformly referenced protein chemical shifts: Journal of biomolecular NMR, v. 25, p. 173-195.

Zhu, B., G. Cai, E. O. Hall, and G. J. Freeman, 2007, In-FusionTM assembly: seamless engineering of multidomain fusion proteins, modular vectors, and mutations: Biotechniques, v. 43, p. 354-359.

10 Appendices

Appendix A: Supercontig and Contig identifiers

Supercontigs in the order they appear in Figure 2-1:

Unique Identifier	Number of appearances
No_name-221	42
No_name-1	4
No_name-74	6
No_name-139	3
No_name-34	11
No_name-6	3
No_name-207	4
No_name-249	4
No_name-30	3
No_name-31	5
No_name-222	2
No_name-247	4
No_name-90	2
No_name-333	9
No_name-404	3
No_name-414	1
No_name-617	1
No_name-531	2
No_name-582	3
No_name-44	2
No_name-501	1
No_name-36	1
No_name-343	2
No_name-542	1
No_name-669	1
No_name-340	1

Contigs in the order they appear in Figure 2-2:

Unique Identifier	Number of appearances
gnI|BL_ORD_ID|244097	1
gnI|BL_ORD_ID|367470	1
gnI|BL_ORD_ID|177838	9
gnI|BL_ORD_ID|245208	4
gnI|BL_ORD_ID|244098	4
gnI|BL_ORD_ID|235822	1
gnI|BL_ORD_ID|235823	2
gnI|BL_ORD_ID|319222	3
gnI|BL_ORD_ID|107887	1
gnI|BL_ORD_ID|265992	5
gnI|BL_ORD_ID|235820	1
gnI|BL_ORD_ID|254037	1
gnI|BL_ORD_ID|107355	1
gnI|BL_ORD_ID|146844	1
gnI|BL_ORD_ID|65600	1
gnI|BL_ORD_ID|330773	2
gnI|BL_ORD_ID|392740	1
gnI|BL_ORD_ID|353121	2
gnI|BL_ORD_ID|262060	1
gnI|BL_ORD_ID|253599	2
gnI|BL_ORD_ID|273536	4
gnI|BL_ORD_ID|264064	3
gnI|BL_ORD_ID|177839	3
gnI|BL_ORD_ID|319162	1
gnI|BL_ORD_ID|32197	40

Appendix B: Amino Acid and DNA Sequences of *Limulus polyphemus* Factor C

Amino acid sequence:

MVLASFLVSGLVLGLLAQQMHPVQSRGVDLGLCDDTRFECKCGDPGYVFNVPAKQCTYFYRW
RPYCKPCDKLEAKDVCPKYKRCQECRAGLDSCVSCPPNKYGTWCSGECQCKNGGICDQRTGAC
TCRDRYEGVHCEILQGCPLLQSDPQVQEVKNPPNDPQTIDYSCSPGFKLKGVARITCLPNGQWSS
FPPKCIRECSMVSSLEHGKVNSPSADLIEGATLRFSCDSPYYLIGQETLTCQGNGQWSGQIPQCQK
LVFCPDLDPVSHAEHQVKIGLEQKYGQFPQGTEVTYTCTGNYFLMGLDTLKCNPDGSWSGTQPS
CVKVADREVNCDSKAVDFLDDVGEAVRIHCPAGCSLTAGTVWGTAIYHELSSVCRAAIHAGKV
PNSGGAVHVVNNGPYSDFLASDLNGIKSEELKSLAQSFRFDYVSSSTAGRKSGCPDGWFEIEENC
VYVTSKQRAWERAQGVCTNMAARLAVLDKDVIPSSLTETLRGKGLATTWIGLHRLDADNHFIW
ELMDRSSVALSDSLTFWAPGEPGSETNCVYLDIQDQLQPVWKTKSCFQPSSFVCMMDLSDKNK
AKCKDPGPLENGHAKLHGQSIDGFYAGSSVRYSCEVLHYLSGTETVSCTSSGTWSAPKPRCIKVI
TCQTPPVPSYGSVDIKPPSRTNSISRIGSPFLRLPRLPLPLARAAKPPPKARSSPPSTVDLASKVKLP
EGHYRVGSRASYTCESRYYELLGSQGRRCNSNGKWSGRPASCIPVCGRSDSPRSPFIVSGSSTEIG
QWPWQAGISRWLADHNMWFLQCGGALLNEKWIITAAHCVTYSATAEIIDPSQFKFYLGKYYRD
DSKDDDYVQVREALEIHVNPNYDPGNLNFDIALIQLKTSIALTTRVQPICLPTDLTTRENLKEGTL
AVVTGWGLSESNTYSEMIQQAVLPVVAASTCEQGYQDSGSPLTVTENMFCAGYKQGRYDACS
GDSGGPLVFADDSRTDRRWVLEGIVSWGSPNG CGKPNQYGGFTKVNVFLSWIRQFI*

cDNA sequence:

ATGGTACTCGCTAGCTTTCTCGTAAGCGGACTGGTCTTGGGTCTGTTGGCTCAGCAGATGCA
CCCAGTGCAGAGTCGGGGAGTAGACCTCGGCCTGTGCGACGACACCCGCTTCGAGTGTAAA
TGCGGCGATCCTGGCTACGTGTTCAATGTTCCGGCAAAGCAATGTACCTACTTCTACCGCTG
GCGGCCTTATTGCAAACCATGTGATAAACTGGAGGCGAAGGATGTGTGCCCCAAGTATAAG
CGCTGCCAAGAATGCCGGGCAGGCCTCGATTCCTGTGTAAGCTGCCCACCAAACAAATACG
GCACTTGGTGCAGTGGCGAATGCCAATGCAAAAACGGCGGTATCTGCGACCAAAGGACCGG
TGCGTGCACCTGTCGTGACCGATATGAAGGCGTTCATTGTGAGATACTCCAGGGCTGCCCGT
TGCTGCAAAGCGACCCTCAAGTACAAGAGGTTAAGAACCCGCCCAACGACCCCCAGACAAT
AGATTATAGCTGCTCCCCTGGTTTCAAACTGAAGGGCGTAGCACGCATAACATGCCTCCCGA
ATGGCCAATGGTCGTCGTTTCCCCCGAAGTGCATCAGGGAATGCAGCATGGTGAGCAGCCT
GGAGCACGGCAAGGTTAACAGCCCCAGCGCCGATCTGATCGAGGGCGCTACCTTGCGCTTC
AGCTGCGATTCCCCGTATTACCTGATCGGACAGGAGACCTTGACGTGCCAGGGCAACGGCC
AGTGGAGTGGTCAAATTCCCCAATGCCAGAAACTCGTTTTCTGCCCAGACTTGGACCCGGTA
AGCCATGCCGAGCACCAGGTAAAGATAGGCCTGGAACAGAAGTACGGTCAGTTCCCCCAGG
GAACGGAAGTGACCTACACCTGCACCGGAAACTATTTTCTCATGGGCTTGGATACTCTCAAA
TGTAACCCAGACGGCTCCTGGTCCGGAACTCAGCCCAGCTGTGTGAAAGTAGCTGACCGGG

AGGTCAACTGTGATTCGAAAGCCGTAGATTTCCTCGACGACGTGGGCGAAGCTGTTCGCATC
CACTGCCCTGCGGGTTGCTCCCTCACGGCGGGAACAGTTTGGGGTACAGCCATATACCACGA
GCTGAGCTCGGTGTGTCGCGCAGCCATTCATGCCGGTAAAGTCCCGAACTCGGGCGGTGCTG
TGCACGTTGTGAACAATGGCCCCTACAGCGACTTTTTGGCATCGGACTTGAACGGCATCAAA
AGCGAGGAATTGAAGAGCCTGGCTCAATCGTTCCGATTCGATTATGTCTCCAGCAGCACTGC
TGGCCGGAAGTCCGGATGCCCGGATGGATGGTTCGAAATAGAGGAAAACTGCGTGTACGTC
ACGAGCAAGCAGCGGGCATGGGAACGCGCCCAGGGTGTATGTACTAACATGGCCGCACGTC
TGGCTGTGCTCGATAAAGACGTGATCCCGAGCAGTCTGACAGAGACCCTGCGTGGTAAGGG
CCTGGCGACCACCTGGATTGGCCTCCATCGATTGGATGCCGACAATCATTTCATTTGGGAGC
TCATGGATAGGAGCAGTGTGGCACTCTCCGACTCGTTGACATTTTGGGCACCGGGTGAACCG
GGAAGCGAAACCAATTGCGTCTACCTCGACATACAGGATCAGCTCCAACCGGTCTGGAAGA
CCAAAAGCTGCTTTCAGCCTTCGAGCTTCGTTTGCATGATGGATTTGAGCGATAAGAACAAA
GCCAAATGCAAGGACCCCGGACCGCTCGAAAACGGACACGCAAAGTTGCACGGCCAAAGC
ATCGACGGTTTTTACGCTGGAAGCTCCGTGCGCTACTCGTGCGAGGTCCTCCACTATTTGTCC
GGCACGGAAACCGTGAGCTGCACCTCCTCCGGTACATGGTCCGCACCGAAACCGCGCTGTA
TCAAGGTGATAACCTGCCAGACACCACCCGTCCCAAGCTACGGCAGCGTGGACATTAAACC
TCCGTCCCGCACTAACTCCATCAGCCGCATCGGTTCCCCTTTCCTGCGCCTCCCTAGGCTGCC
TCTGCCCCTGGCGCGTGCCGCCAAACCACCCCCTAAGGCTCGCAGCAGCCCCCCAAGCACA
GTCGATTTGGCGTCGAAAGTGAAACTGCCCGAGGGACACTATCGCGTCGGCTCCCGGGCTTC
GTATACCTGTGAGTCCCGGTACTACGAATTGCTGGGAAGTCAAGGCAGGCGCTGCAACTCC
AACGGAAAGTGGAGCGGCCGACCCGCTAGTTGCATCCCAGTGTGCGGTCGCTCCGATAGTC
CACGATCCCCCTTCATCGTGAGCGGCTCGTCCACTGAAATTGGACAGTGGCCTTGGCAAGCT
GGCATCTCCCGATGGCTCGCGGATCATAATATGTGGTTCCTGCAATGTGGCGGAGCCCTCCT
GAATGAGAAGTGGATCATCACTGCCGCGCACTGCGTTACATACTCGGCCACCGCGGAGATT
ATCGACCCGAGCCAATTCAAGTTCTACCTGGGTAAGTACTACCGCGATGATAGCAAAGACG
ATGATTACGTACAAGTACGCGAGGCGTTGGAAATCCACGTTAACCCAAATTACGACCCAGG
CAACCTCAACTTCGATATTGCTCTGATTCAGCTCAAAACTAGTATCGCCCTGACCACACGGG
TACAACCCATCTGCCTGCCGACTGACCTGACTACGCGGGAGAATCTCAAAGAGGGAACCCT
CGCGGTTGTGACCGGCTGGGGACTCTCGGAGTCCAACACTTATAGCGAGATGATTCAACAG
GCTGTTTTGCCAGTAGTGGCAGCTAGCACATGCGAGCAGGGATACCAAGATTCGGGTAGTC
CTCTCACCGTAACTGAGAATATGTTTTGTGCCGGCTACAAACAGGGCCGCTATGATGCCTGC
TCCGGTGACAGCGGTGGACCACTGGTGTTTGCGGACGACAGCCGTACAGATCGTCGCTGGG
TATTGGAAGGCATTGTGTCGTGGGGTAGCCCCAACGGCTGCGGCAAACCCAACCAGTACGG
CGGCTTTACCAAGGTAAACGTATTCCTGAGTTGGATCAGGCAGTTCATC

Appendix C: Primer Sequences

Tachypleus tridentatus Cys-rich, EGF-like and CysEGF amplification primers for ligation independent cloning.

Construct	**Primer Sequences**	**Melting temperature (T_m)**
Cys-rich 5’	TACTTCCAATCCCAGCAGATGCGTCCGGTTC	60.6°C
Cys-rich 3’	TATCCACCTTTACTGTCATTTGTTCGGTGGGCAGGTAAC	61.5°C
EGF-like 5’	TACTTCCAATCCAACAAATACGGTACCTGGTG	56.8°C
EGF-like 3’	TATCCACCTTTACTGTCAACCTTCGTAACGGTCAC	59.4°C

LIC complementary overhangs are underlined.

Limulus polyphemus complement control protein primers for ligation independent cloning.

Construct	Primer Sequence	Melting temperature (T_m)
CCP1 5’	TACTTCCAATCCATGGAGATACTCCAGGGCTGC	60.6°C
CCP1 3’	TATCCACCTTTACTGTTCCCTGATGCACTTCGG	59.3°C
CCP2 5’	TACTTCCAATCCATGATCAGGGAATGCAGCATG	58.1°C
CCP2 3’	TATCCACCTTTACTGTTTCTGGCATTGGGGAATTTG	58.2°C
CCP3 5’	TACTTCCAATCCATGCAGAAACTCGTTTTCTGC	56.8°C
CCP3 3’	TATCCACCTTTACTGCTCCCGGTCAGCTAC	59.3°C
CCP3 5’ v2	TACTTCCAATCCATGAAACTCGTTTTCTGC	53.8°C

LIC complementary overhangs are underlined.

Oligonucleotides for Thermodynamically Balanced Inside-Out PCR

Number	**Oligonucleotide sequences**
1	CCATGGAAATTTTGAAGGGGTGCCCG
2	TTGAAGGGGTGCCCGCTGCTGCCGAGCGACAGCCAGGTGCAGGAAGTGCGTAATC
3	TGCAGGAAGTGCGTAATCCGCCGGATAACCCGCAGACCATTGATTATAGCTGTAG
4	AGACCATTGATTATAGCTGTAGCCCGGGTTTCAAATTGAAAGGTGTGGCGCGTAT
5	AGGTGTGGCGCGTATTAGCTGCCTTCCGAATGGCCAGTGGAGTAGCTTTCCGCCG
6	GAGTAGCTTTCCGCCGAAGTGCATTCGTGAGTGCGCGAAAGTGAGCAGCCCTGAA
7	AAGTGAGCAGCCCTGAACATGGCAAAGTTAATGCGCCGTCTGGGAACATGATTGA
8	CGTCTGGGAACATGATTGAAGGCGCGACACTGCGTTTTTCGTGCGATAGCCCGTA
9	TTGCCCTGGCAGGTCAGTGTCTCCTGTCCAATCAGATAGTACGGGCTATCGCACG
10	CAGTTTCTTGCACTGCGGAATCTGGCCGGACCACTGTCCATTGCCCTGGCAGGTC
10v2	CAGTTTCTTGCACTGCGGGATTTGGCCTGACCACTGCCCATTGCCCTGGCAGGTC
11	TCTGCATGATTCACGGGGTCCAGATCCGGGCAAAAAACCAGTTTCTTGCACTGCG
12	GCCCATATTTCTGTTCCACGCCAATTTTCACCTGATGTTCTGCATGATTCACGGG
13	AGCAGGTGTAGGTAACTTCCGTACCCTGGGGAAACTGCCCATATTTCTGTTCCAC
14	TCAGGGTATTAAAGCCCATCAGAAAATAGTTGCCCGAGCAGGTGTAGGTAACTTC
15	CTGGCTGCCTGACCAGCTGCCGTCCGGATTGCATTTCAGGGTATTAAAGCCCATC
16	CTCGAGCTATTCACGATCCGCCACTTTCACGCAGCTCGGCTGGCTGCCTGACCA

Appendix D: Buffer Recipes

M9 Minimal Media

To make 1 L of 5X M9 stock solution:

Na_2HPO_4	34.08 g
KH_2PO_4	15 g
NaCl	2.5 g

Make to ~800 ml with dH_2O
Adjust pH to 7.0
Make to 1 L with dH_2O
Autoclave

For 1 L 1X M9 minimal media:

dH_2O	800 ml
5X M9 stock solution	200 ml
$MgSO_4$ (1 M)	1 ml
$CaCl_2$ (1 M)	50 µl
Thiamine (50 mg/ml)	400 µl
D-glucose	1.5 g
$(NH_4)_2SO_4$	0.5 g
Antibiotic	50 µg/ml

For ^{15}N-labelled media, $(NH_4)_2SO_4$ is replaced with $^{15}NH_4Cl$

For ^{13}C-labelled media, D-glucose is replaced with $^{13}C_6$-glucose

2xYT media

For 1L:

Tryptone	16 g
Yeast Extract	10 g
NaCl	5 g

Appendix E: CCP12 chemical shift assignments

	H	N	C	C_a	
121 Ile			175.61	61.10	H^a 4.14, H^b 1.79, C^b 38.65
122 Leu	8.36	127.54	176.60	55.14	H^a 4.35, H^{b2} 1.55, H^{b3} 1.55, H^g 1.53, H^{da*} 0.83, H^{db*} 0.86, C^b 42.28, C^g 26.98, C^{da} 23.76, C^{db} 24.78
123 Gln	8.32	121.72	175.75	55.49	H^a 4.30, H^{ba} 1.95, C^b 29.98
124 Gly	8.42	110.27	173.47	45.35	H^{a2} 3.83, H^{a3} 3.83
125 Cys	7.90	118.69	172.54	57.50	H^a 4.62, H^{ba} 1.50, H^{bb} 1.57, C^b 38.92
126 Pro			176.63	63.25	H^a 4.39, H^{ba} 1.95, H^{bb} 2.39, H^{ga} 2.08, H^{gb} 2.19, H^{da} 3.33, H^{db} 3.90, C^b 32.74, C^g 27.84, C^d 50.68
127 Leu	8.34	122.65	177.09	55.88	H^a 4.33, H^{ba} 1.59, H^{bb} 1.63, H^g 1.65, H^{da*} 0.92, H^{db*} 0.96, C^b 41.99, C^g 27.22, C^{da} 24.51, C^{db} 24.71
128 Leu	8.29	126.57	177.02	53.47	H^a 4.54, H^{ba} 1.23, H^{bb} 1.43, H^g 1.48, H^{da*} 0.09, H^{db*} 0.68, C^b 43.49, C^g 26.54, C^{da} 25.19, C^{db} 22.84
129 Gln	8.49	122.29	176.06	55.64	H^a 4.29, H^{b2} 1.97, H^{b3} 1.97, C^b 29.50
130 Ser	8.55	121.08	173.30	58.43	H^a 4.45, H^{ba} 3.64, H^{bb} 3.82, C^b 64.32
131 Asp	8.94	126.07	175.64	51.39	H^a 4.89, H^{ba} 2.52, H^{bb} 2.77, C^b 42.84
132 Pro			177.47	64.56	H^a 4.32, H^{ba} 1.88, H^{bb} 2.27, H^{g2} 1.98, H^{g3} 1.98, H^{d2} 3.78, H^{d3} 3.78, C^b 32.12, C^g 27.35, C^d 50.92
133 Gln	8.48	115.36	174.81	55.96	H^a 4.07, H^{b2} 1.74, H^{b3} 1.74, H^{g2} 1.99, H^{g3} 1.99, C^b 29.02, C^g 34.55
134 Val	7.79	121.42	175.70	61.58	H^a 4.39, H^b 2.10, H^{ga*} 0.81, H^{gb*} 0.93, C^b 33.24, C^{ga} 21.41, C^{gb} 22.37
135 Gln	8.47	126.21	175.19	55.65	H^a 4.33, H^{ba} 1.85, H^{bb} 1.95, H^{g2} 2.18, H^{g3} 2.18, C^b 28.60
136 Glu	8.35	120.57	176.23	54.09	H^a 4.96, H^{ba} 1.58, H^{bb} 1.66, H^{ga} 2.02, H^{gb} 2.07, C^b 34.12
137 Val	8.65	123.67	175.29	62.43	H^a 4.08, H^b 1.86, H^{g1*} 0.80, H^{g2*} 0.80, C^b 33.87, C^{g1} 20.94, C^{g2} 20.94
138 Lys	8.70	129.09	174.86	55.94	H^a 4.51, H^{ba} 1.63, H^{bb} 1.77, H^{ga} 1.14, H^{gb} 1.30, H^{d2} 1.66, H^{d3} 1.66, H^{e2} 2.90, H^{e3} 2.90, C^b 34.98, C^g 26.42, C^d 29.65, C^e 42.10
139 Asn	8.54	120.30	174.28	50.41	H^a 5.01, H^{ba} 2.57, H^{bb} 2.63, C^b 42.57
140 Pro				30.29	H^a 5.17, H^{ba} 2.15, H^{bb} 2.59, H^{ga} 1.93, H^{gb} 2.00, H^{da} 3.69, H^{db} 3.94, C^b 32.93, C^g 26.00, C^d 50.80
141 Pro			177.20	65.36	H^a 4.03, H^{ba} 1.87, H^{bb} 2.37, H^{g2} 2.16, H^{g3} 2.16, H^{da} 3.76, H^{db} 3.83, C^b 31.94, C^g 27.91, C^d 50.08
142 Asn	7.88	111.29	174.46	53.15	H^a 4.55, H^{ba} 1.86, H^{bb} 2.70, C^b 39.03
143 Asp	7.46	117.05	171.90	53.35	H^a 4.67, H^{ba} 2.60, C^b 39.78
144 Pro			176.51	63.88	H^a 4.42, H^{ba} 1.95, H^{bb} 2.17, H^{ga} 1.86, H^{gb} 2.26, H^{da} 3.40, H^{db} 3.86, C^b 32.54, C^g 27.99, C^d 50.64
145 Gln	9.06	120.76	176.30	56.15	H^a 4.41, H^{ba} 2.14, H^{bb} 2.38, H^{g2} 2.48, H^{g3} 2.48, C^b 30.71, C^g 34.31
146 Thr	8.25	110.26	173.32	60.20	H^a 5.47, H^b 4.02, H^{g2*} 1.24, C^b 73.07, C^{g2} 22.09
147 Ile	8.55	120.98	173.62	61.52	H^a 4.50, H^b 1.16, H^{g1a} 0.56, H^{g1b} 1.26, H^{g2*} -0.09, H^{d1*} 0.21, C^b 40.06, C^{g1} 27.49, C^{g2} 17.43, C^{d1} 13.37
148 Asp	8.16	124.73	175.60	53.29	H^a 5.44, H^{ba} 2.29, H^{bb} 2.58, C^b 44.42
149 Tyr	8.98	121.05	175.58	57.41	H^a 5.54, H^{ba} 2.48, H^{bb} 2.68, H^{d*} 6.79, H^{e*} 6.49, C^b 43.67, C^{d*} 133.55, C^{e*} 117.56
150 Ser	8.63	112.34	171.80	57.75	H^a 4.48, H^{b2} 3.89, H^{b3} 3.89, C^b 65.45
151 Cys	8.45	114.92	175.37	52.32	H^a 5.34, H^{ba} 2.43, H^{bb} 3.23, C^b 42.73
152 Ser	8.80	119.98	170.74	58.61	H^a 4.57, H^{ba} 3.65, H^{bb} 3.96, C^b 62.26
153 Pro			177.58	64.66	H^a 4.42, C^b 31.76
154 Gly	8.85	112.12	173.30	44.85	H^{aa} 3.60, H^{ab} 4.19
155 Phe	8.01	118.66	174.09	56.22	H^a 5.01, H^{ba} 2.54, H^{bb} 3.31, H^{d*} 6.84, H^{e*} 7.24, C^b 42.40, C^{d*} 131.49, C^{e*} 132.27
156 Lys	9.53	121.89	174.88	53.97	H^a 4.54, H^{b2} 1.61, H^{b3} 1.61, H^{g2} 1.26, H^{g3} 1.26, H^{d2} 1.57, H^{d3} 1.57, H^{e2} 2.93, H^{e3} 2.93, C^b 35.73, C^g 24.69, C^d 28.91, C^e 42.10
157 Leu	8.33	127.00	176.57	56.25	H^a 4.51, H^{ba} 1.51, H^{bb} 1.58, H^g 1.25, H^{da*} 0.58, H^{db*} 0.67, C^b 42.91, C^g 27.48, C^{da} 25.61, C^{db} 25.67
158 Lys	9.37	129.81	174.64	55.02	H^a 4.60, H^{ba} 1.44, H^{bb} 1.71, H^{ga} 1.22, H^{gb} 1.35, C^b 33.69, C^g 24.51

	H	N	C	C_a	
159 Gly	8.38	116.39	172.17	43.23	$H^{a\,a}$ 3.61, $H^{a\,b}$ 4.60
160 Val	8.02	117.22	174.07	61.33	H^{a} 4.20, H^{b} 2.16, $H^{g\,a*}$ 0.88, $H^{g\,b*}$ 1.00, C^{b} 32.91, $C^{g\,a}$ 19.97, $C^{g\,b}$ 21.84
161 Ala	7.96	119.62	176.65	53.17	H^{a} 4.30, H^{b*} 1.48, C^{b} 19.75
162 Arg	7.30	117.47	175.89	54.66	H^{a} 5.68, $H^{b\,a}$ 1.50, $H^{b\,b}$ 2.06, $H^{g\,a}$ 1.31, $H^{g\,b}$ 1.46, $H^{d\,a}$ 3.01, $H^{d\,b}$ 3.12, C^{b} 33.88, C^{g} 27.24, C^{d} 43.55
163 Ile	8.85	119.77	173.96	60.16	H^{a} 4.67, H^{b} 1.99, $H^{g\,1a}$ 1.14, $H^{g\,1b}$ 1.30, $H^{g\,2*}$ 0.71, $H^{d\,1*}$ 0.15, C^{b} 41.66, $C^{g\,1}$ 25.16, $C^{g\,2}$ 18.27, $C^{d\,1}$ 13.86
164 Thr	9.01	115.43	173.19	60.73	H^{a} 5.36, H^{b} 3.97, $H^{g\,2*}$ 1.19, C^{b} 72.02, $C^{g\,2}$ 21.61
165 Cys	8.94	125.99	173.63	54.40	H^{a} 4.12, $H^{b\,a}$ 1.91, $H^{b\,b}$ 2.47, C^{b} 36.16
166 Leu	8.45	132.04	176.93	54.19	H^{a} 4.47, $H^{b\,a}$ 1.37, $H^{b\,b}$ 2.36, H^{g} 1.78, $H^{d\,a*}$ 0.81, $H^{d\,b*}$ 0.83, C^{b} 39.84, C^{g} 27.00, $C^{d\,a}$ 25.56, $C^{d\,b}$ 21.87
167 Pro			176.18	64.63	H^{a} 4.32, $H^{b\,a}$ 1.87, $H^{b\,b}$ 2.36, $H^{g\,a}$ 1.88, $H^{g\,b}$ 1.99, $H^{d\,a}$ 3.69, $H^{d\,b}$ 3.78, C^{b} 31.79, C^{g} 27.42, C^{d} 50.71
168 Asn	7.40	112.53	176.46	52.25	H^{a} 4.51, $H^{b\,a}$ 2.68, $H^{b\,b}$ 3.13, C^{b} 36.98
169 Gly	8.37	108.74	173.47	46.48	$H^{a\,a}$ 3.51, $H^{a\,b}$ 3.87
170 Gln	7.06	117.04	175.19	52.77	H^{a} 4.45, $H^{b\,a}$ 1.82, $H^{b\,b}$ 2.03, $H^{g\,2}$ 2.15, $H^{g\,3}$ 2.15, C^{b} 30.82, C^{g} 33.28
171 Trp	8.49	121.66	180.16	57.37	$N^{e\,1}$ 130.09, H^{a} 4.99, $H^{b\,a}$ 3.03, $H^{b\,b}$ 3.37, $H^{d\,1}$ 7.28, $H^{e\,1}$ 10.88, $H^{e\,3}$ 7.16, $H^{z\,2}$ 7.41, $H^{z\,3}$ 6.59, $H^{h\,2}$ 6.71, C^{b} 31.55, $C^{d\,1}$ 125.90, $C^{e\,3}$ 120.21, $C^{z\,2}$ 114.62, $C^{z\,3}$ 121.69, $C^{h\,2}$ 124.56
172 Ser	9.51	116.63	180.19	61.48	H^{a} 4.02, $H^{b\,a}$ 4.12, $H^{b\,b}$ 4.25, C^{b} 62.67
173 Ser	7.32	113.02	172.79	56.90	H^{a} 4.40, $H^{b\,a}$ 3.61, $H^{b\,b}$ 3.92, C^{b} 64.87
174 Phe	7.78	119.21	173.40	55.78	H^{a} 4.85, $H^{b\,a}$ 2.85, $H^{b\,b}$ 3.16, C^{b} 38.17
175 Pro				61.35	H^{a} 3.92, $H^{b\,a}$ 1.57, $H^{b\,b}$ 2.01, $H^{g\,a}$ 1.80, $H^{g\,b}$ 1.96, $H^{d\,a}$ 3.54, $H^{d\,b}$ 3.67, C^{b} 30.56, C^{g} 28.02, C^{d} 50.44
176 Pro			173.66	62.37	H^{a} 4.62, $H^{b\,a}$ 1.55, $H^{b\,b}$ 1.93, $H^{g\,a}$ 1.76, $H^{d\,a}$ 1.65, $H^{d\,b}$ 2.82, C^{b} 33.63, C^{g} 27.82, C^{d} 49.67
177 Lys	7.70	112.77	175.40	53.61	H^{a} 4.48, $H^{b\,a}$ 1.48, $H^{b\,b}$ 1.73, $H^{g\,2}$ 1.38, $H^{g\,3}$ 1.38, $H^{d\,2}$ 1.59, $H^{d\,3}$ 1.59, $H^{e\,2}$ 2.88, $H^{e\,3}$ 2.88, C^{b} 36.17, C^{g} 24.72, C^{d} 28.73, C^{e} 42.25
178 Cys	8.54	118.91	174.05	55.13	H^{a} 5.25, $H^{b\,a}$ 2.47, $H^{b\,b}$ 2.59, C^{b} 42.71
179 Ile	9.23	123.48	174.07	59.02	H^{a} 4.59, H^{b} 1.87, $H^{g\,1a}$ 1.24, $H^{g\,1b}$ 1.38, $H^{g\,2*}$ 1.03, $H^{d\,1*}$ 0.78, C^{b} 40.02, $C^{g\,1}$ 27.05, $C^{g\,2}$ 19.26, $C^{d\,1}$ 12.60
180 Arg	8.89	127.76	174.82	56.79	H^{a} 3.56, $H^{b\,2}$ 1.39, $H^{b\,3}$ 1.39, $H^{g\,2}$ 1.15, $H^{g\,3}$ 1.15, $H^{d\,2}$ 2.43, $H^{d\,3}$ 2.43, C^{b} 30.87, C^{g} 26.59, C^{d} 42.83
181 Glu	7.88	122.90	177.38	54.98	H^{a} 4.72, $H^{b\,a}$ 1.61, $H^{b\,b}$ 1.82, $H^{g\,a}$ 1.93, $H^{g\,b}$ 2.01, C^{b} 32.59, C^{g} 37.03
182 Cys	8.24	119.16	173.36	57.13	H^{a} 4.35, $H^{b\,a}$ 1.57, $H^{b\,b}$ 2.12, C^{b} 44.36
183 Ser	8.51	115.60	174.27	58.70	H^{a} 4.32, $H^{b\,a}$ 3.90, $H^{b\,b}$ 3.96, C^{b} 64.05
184 Met	8.50	122.12	176.16	56.24	H^{a} 4.26, $H^{b\,a}$ 1.98, H^{e*} 2.05, C^{e} 16.97
185 Val	7.11	125.27	175.09	61.99	H^{a} 4.18, H^{b} 1.85, $H^{g\,a*}$ 0.68, $H^{g\,b*}$ 0.87, C^{b} 33.09, $C^{g\,a}$ 20.05, $C^{g\,b}$ 21.08
186 Ser			173.88	59.38	H^{a} 4.32, $H^{b\,2}$ 3.82, $H^{b\,3}$ 3.82, C^{b} 64.41
187 Ser	7.59	113.52	171.37	57.77	H^{a} 4.35, $H^{b\,a}$ 3.67, $H^{b\,b}$ 3.73, C^{b} 64.80
188 Leu	8.26	123.61	175.42	53.87	H^{a} 4.31, $H^{b\,a}$ 1.05, $H^{b\,b}$ 1.41, H^{g} 1.21, $H^{d\,a*}$ 0.67, $H^{d\,b*}$ 0.76, C^{b} 45.08, C^{g} 27.03, $C^{d\,a}$ 24.46, $C^{d\,b}$ 27.19
189 Glu	8.59	129.19	177.15	58.44	H^{a} 3.71, $H^{b\,a}$ 1.56, $H^{b\,b}$ 1.73, $H^{g\,a}$ 1.56, $H^{g\,b}$ 1.83, C^{b} 28.97, C^{g} 35.89
190 His	8.61	116.48	173.03	57.04	H^{a} 3.92, $H^{b\,a}$ 2.29, $H^{b\,b}$ 2.65, C^{b} 25.91
191 Gly	7.89	105.11	172.83	46.48	$H^{a\,a}$ 4.22, $H^{a\,b}$ 4.43
192 Lys	8.89	120.90	173.72	55.00	H^{a} 4.64, $H^{b\,a}$ 1.59, $H^{b\,b}$ 1.61, $H^{g\,a}$ 1.22, $H^{g\,b}$ 1.27, $H^{d\,a}$ 1.55, $H^{d\,b}$ 1.63, $H^{e\,2}$ 2.93, $H^{e\,3}$ 2.93, C^{b} 36.93, C^{g} 24.73, C^{d} 29.03, C^{e} 42.20
193 Val	8.15	119.93	174.67	60.25	H^{a} 4.25, H^{b} 1.64, $H^{g\,a*}$ 0.43, $H^{g\,b*}$ 0.51, C^{b} 33.89, $C^{g\,a}$ 20.54, $C^{g\,b}$ 21.60
194 Asn	8.71	125.69	173.39	52.49	H^{a} 4.81, $H^{b\,a}$ 2.49, $H^{b\,b}$ 2.54, C^{b} 41.35
195 Ser				54.38	H^{a} 5.05, $H^{b\,a}$ 3.66, $H^{b\,b}$ 3.74, C^{b} 64.89

	H	N	C	Cα	
196 Pro				63.47	H^{a} 4.51, H^{ba} 2.02, H^{bb} 2.28, H^{g2} 1.97, H^{g3} 1.97, H^{da} 3.64, H^{db} 3.96, C^{g} 26.93, C^{d} 50.76
197 Ser			173.72	57.46	H^{a} 3.89, C^{b} 63.82
198 Ala	8.27	125.33	176.82	53.07	H^{a} 4.24, H^{b*} 1.38, C^{b} 19.54
199 Asp	7.76	115.70	175.37	53.36	H^{a} 4.65, H^{ba} 2.36, H^{bb} 2.44, C^{b} 42.46
200 Leu	9.02	127.58	175.35	53.30	H^{a} 4.30, H^{ba} 1.29, H^{bb} 2.03, H^{g} 1.74, H^{da*} 0.90, H^{db*} 1.11, C^{b} 39.11, C^{g} 26.35, C^{da} 23.20, C^{db} 26.24
201 Ile	6.72	112.97	175.52	58.73	H^{a} 4.79, H^{b} 1.87, H^{g1a} 0.80, H^{g1b} 1.15, H^{g2*} 0.67, H^{d1*} 0.80, C^{b} 41.14, C^{g1} 24.80, C^{g2} 18.60, C^{d1} 13.93
202 Glu	7.74	120.14	176.46	58.07	H^{a} 3.38, H^{ba} 1.67, H^{g2} 1.97, H^{g3} 1.97, C^{b} 30.02, C^{g} 35.29
203 Gly	8.90	115.42	176.49	44.59	H^{aa} 3.44, H^{ab} 4.45
204 Ala	8.77	124.50	176.83	52.99	H^{a} 4.46, H^{b*} 1.66, C^{b} 20.08
205 Thr	8.14	111.49	173.92	60.27	H^{a} 5.70, H^{b} 3.96, H^{g2*} 1.18, C^{b} 72.82, C^{g2} 21.74
206 Leu	8.70	122.45	174.46	53.29	H^{a} 4.59, H^{ba} 0.41, H^{bb} 0.91, H^{g} -0.17, H^{da*} -0.13, H^{db*} 0.48, C^{b} 46.00, C^{g} 25.52, C^{da} 25.59, C^{db} 23.08
207 Arg	8.08	119.58	175.12	54.13	H^{a} 5.22, H^{ba} 1.51, H^{bb} 1.72, H^{g2} 1.51, H^{g3} 1.51, H^{d2} 3.10, H^{d3} 3.10, C^{b} 32.90, C^{g} 27.78, C^{d} 43.27
208 Phe	9.03	124.45	174.62	57.13	H^{a} 5.10, H^{ba} 2.71, H^{bb} 2.75, H^{d*} 7.08, H^{e*} 6.80, H^{z} 6.64, C^{b} 43.03, C^{d*} 132.71, C^{e*} 130.59, C^{z} 128.54
209 Ser	8.44	112.96	172.34	57.30	H^{a} 4.58, H^{ba} 3.75, H^{bb} 3.81, C^{b} 65.65
210 Cys	8.79	117.28	174.16	54.24	H^{a} 5.13, H^{ba} 2.32, H^{bb} 2.74, C^{b} 42.72
211 Asp	9.26	125.85	175.79	53.76	H^{a} 4.57, H^{ba} 2.53, H^{bb} 2.65, C^{b} 40.70
212 Ser	8.68	118.59		57.49	H^{a} 4.64, H^{ba} 3.82, H^{bb} 3.94, C^{b} 63.00
213 Pro			174.88	64.03	H^{a} 4.86, H^{ba} 1.88, H^{bb} 2.37, H^{ga} 1.06, H^{gb} 1.88, H^{da} 3.31, H^{db} 3.54, C^{b} 32.18, C^{g} 24.15, C^{d} 49.58
214 Tyr	9.37	125.59	174.33	58.88	H^{a} 4.33, H^{ba} 2.47, H^{bb} 3.43, H^{d*} 6.71, H^{e*} 6.43, C^{b} 39.29, C^{d*} 132.91, C^{e*} 117.41
215 Tyr	9.16	118.80	173.57	55.96	H^{a} 4.73, H^{b2} 2.90, H^{b3} 2.90, H^{d*} 7.00, H^{e*} 6.67, C^{b} 41.33, C^{d*} 133.81, C^{e*} 117.93
216 Leu	8.31	124.15	176.56	56.32	H^{a} 4.55, H^{ba} 1.33, H^{bb} 1.71, H^{g} 1.14, H^{da*} 0.57, H^{db*} 0.61, C^{b} 43.48, C^{g} 27.66, C^{da} 25.85, C^{db} 25.98
217 Ile	9.27	129.33	175.22	60.01	H^{a} 4.33, H^{b} 1.84, H^{g1a} 1.28, H^{g1b} 1.37, H^{g2*} 0.87, H^{d1*} 0.74, C^{b} 38.03, C^{g1} 26.71, C^{g2} 17.07, C^{d1} 11.72
218 Gly	8.39	115.70	175.18	42.92	H^{aa} 3.57, H^{ab} 4.52
219 Gln	8.55	120.75	174.26	55.85	H^{a} 4.17, H^{ba} 1.83, H^{bb} 2.15, H^{g2} 2.41, H^{g3} 2.41, C^{b} 30.26, C^{g} 34.62
220 Glu	8.14	118.23	175.89	58.18	H^{a} 4.14, H^{ba} 2.09, H^{bb} 2.16, H^{ga} 2.41, H^{gb} 2.47, C^{b} 30.75, C^{g} 36.55
221 Thr	7.53	109.38	173.73	59.72	H^{a} 5.58, H^{b} 3.97, H^{g2*} 1.13, C^{b} 71.71, C^{g2} 21.61
222 Leu	8.58	122.37	175.53	53.61	H^{a} 4.69, H^{ba} 1.49, H^{bb} 1.70, H^{g} 1.55, H^{da*} 0.58, H^{db*} 0.74, C^{b} 47.95, C^{g} 27.26, C^{da} 25.36, C^{db} 22.73
223 Thr	9.05	115.83	174.43	61.45	H^{a} 5.48, H^{b} 3.83, H^{g2*} 1.13, C^{b} 71.41, C^{g2} 21.44
224 Cys	9.21	128.09	174.46	55.00	H^{a} 3.83, H^{ba} 1.92, H^{bb} 2.38, C^{b} 37.28
225 Gln	8.11	129.65	177.09	55.14	H^{a} 4.54, H^{ba} 2.13, H^{bb} 2.54, H^{ga} 2.11, H^{gb} 2.51, C^{b} 31.02, C^{g} 35.14
226 Gly			173.31	46.56	H^{a2} 3.79, H^{a3} 3.79
227 Asn	7.44	114.77	176.46	51.92	H^{a} 4.54, H^{ba} 2.72, H^{bb} 3.16, C^{b} 37.09
228 Gly	8.10	107.71	173.06	46.17	H^{aa} 3.34, H^{ab} 3.69
229 Gln	7.29	117.93	175.91	53.03	H^{a} 4.37, H^{ba} 1.79, H^{bb} 2.12, H^{ga} 2.06, H^{gb} 2.18, C^{b} 30.15, C^{g} 33.04
230 Trp	8.53	123.22	179.01	57.53	N^{e1} 130.58, H^{a} 4.90, H^{ba} 3.25, H^{bb} 3.36, H^{d1} 7.34, H^{e1} 10.69, H^{e3} 7.08, H^{z2} 7.31, C^{b} 31.66, C^{d1} 126.86, C^{e3} 119.74, C^{z2} 115.31
231 Ser	8.83	115.39	179.05	60.14	H^{a} 4.19, H^{ba} 4.04, H^{bb} 4.38, C^{b} 63.31
232 Gly	7.38	109.41	171.34	44.68	H^{aa} 3.90, H^{ab} 4.19
233 Gln	8.17	118.01	175.45	54.61	H^{a} 4.30, H^{ba} 1.80, H^{bb} 1.96, H^{g2} 2.37, H^{g3} 2.37, C^{b} 30.02, C^{g} 33.91
234 Ile	8.17	124.02	174.74	59.25	H^{a} 3.55, H^{b} 1.63, H^{g1a} 1.15, H^{g1b} 1.44, H^{g2*} 0.81, H^{d1*} 0.71, C^{b} 37.32, C^{g1} 27.93, C^{g2} 16.75, C^{d1} 12.46

	H	N	C	C_a	
235 Pro			172.76	62.84	H^a 4.54, H^{ba} 1.52, H^{bb} 1.92, H^{ga} 1.60, H^{da} 1.94, H^{db} 2.87, C^b 33.32, C^d 49.86
236 Gln	7.82	111.15	174.16	53.97	H^a 4.54, H^{ba} 1.80, H^{bb} 1.95, H^{ga} 2.16, H^{gb} 2.26, C^b 33.47, C^g 33.89
237 Cys	8.98	120.30	174.66	53.50	H^a 5.45, H^{ba} 2.66, H^{bb} 2.87, C^b 40.65
238 Gln	9.33	123.92	174.10	54.33	H^a 4.83, H^{ba} 1.64, H^{bb} 1.91, H^{g2} 2.25, H^{g3} 2.25, C^b 33.05, C^g 33.74
239 Lys	8.63	123.03	175.55	56.20	H^a 4.19, H^{ba} 1.25, H^{bb} 1.31, H^{ga} 0.85, H^{gb} 0.98, H^{da} 1.13, H^{db} 1.26, H^{e2} 2.71, H^{e3} 2.71, C^b 33.68, C^g 24.33, C^d 29.21, C^e 41.88
240 Gln	8.10	127.13	180.59	57.82	H^a 4.11, H^{ba} 1.83, H^{bb} 1.98, H^{g2} 2.21, H^{g3} 2.21, C^b 30.32, C^g 34.51