STRUCTURAL AND FUNCTIONAL STUDIES OF THE TERMINAL PROTEIN COMPLEX OF THE HUMAN COMPLEMENT SYSTEM

Thesis submitted by Carla Clark
for the degree of Doctor of Philosophy
in the school of Chemistry 2011
The University of Edinburgh

ABSTRACT

The membrane attack complex (MAC: C5b1C61C71C81C9n) is the terminal complex of the mammalian complement cascade. Its principal function is to form a self-assembling, potentially cytolytic pore spanning the membrane of a foreign cell. However, inappropriate complement activation and subsequent aberrant MAC action is responsible for tissue damage in several human pathologies.[1] The C-terminal domains or modules that are exclusive to the C6 and C7 components of the MAC, form a putative molecular arm. These modules consist of two complement control protein modules (CCPs) followed by a pair of Factor I-like Modules (FIMs). The C-terminal molecular arms of C6 and C7 are important for linking the complement activation cascade with MAC assembly and thereby ensuring MAC assembly occurs only when needed. They interact with the C345C domain of C5 and C5b. Currently there are no high-resolution structural data for any of the MAC components but efforts are underway to solve the structures of the individual domains. For instance, the structure of the central MAC/perforin domain of C8 has been determined, so has the structure of C5 and the FIM domains of C7. However , the lack of structural information for the MAC is being balanced by an increased understanding of mechanism, and eventually towards the rational design of therapies designed to supress MAC formation.

This report describes the preparation from bacterial and yeast cells of ^{13}C,^{15}N-labelled samples of the C7 CCP-pair and of a triple module consisting of the second CCP followed by the FIM-pair, as well as other constructs from the C6 and C7 C-terminal arms. The NMR samples were used to solve 3D solutions of structures, thus allowing for reconstruction of the four-module C-terminus of C7. Efforts were aimed at obtaining additional structural information for C7. This included SAXS analysis as well as the development of a novel approach using chemical cross-linking followed by tryptic digestion and mass spectrometry-based identification of cross-linked peptides.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr Paul Barlow for the opportunity to work on such an interesting PhD project and for his valued guidance. I would like to express my gratitude f to both Marie Phelan and Janice Bramham , for their constant encouragement and guidance, ever since the initial work on the MAC.

Many thanks to Edinburgh University's School of Chemistry for funding this project.

I gratefully acknowledge old and new members of the biomolecular NMR group members for much fun, laughter, help and advice both in and out of the lab: Dusan, Juraj, Andy, Dave, Marie, Henry, Elin, Yoshi, Dinesh, Isabell, Carina, Christoph, Barbel, Conny, Maria, Nicky, Elisa, Ilias and Fern.

I extend my thanks to Elizabeth Blackburn for invaluable tuition and advice with SAXS data analysis and to John White for his expertise in fermentation, and many interesting discussions. Also a big thanks to Dinesh Soares for his crucial help and expertise in Bioinformatics.

I also wish to thank our collaborators – Dr. Juri Rappsilber and his group and a special thanks to Zhuo Chen and to Dr. Ron Ogata and colleagues for collaboration on this project and for encouraging discussions and ideas.

Finally, I would like to thank my friends and family, in particular Ruth and Lucas, whose support has been immeasurable.

Unless otherwise stated in the text, the work described in this thesis is my own work and has not been submitted for in whole or in part for a degree or any other qualification, neither at The University of Edinburgh nor at any other university.

Carla Clark

CONTENTS

Table of Figures

List of Tables

Abbreviations

2D	two-dimensional
3D	three-dimensional
AOX I	Alcohol oxidase I
AOX II	Alcohol oxidase II
AP	Alternative pathway
BMG	Buffered minimal glycerol
BMM	Buffered minimal methanol
BS2G	Bis[Sulfosuccinimidyl] glutarate
C7-CCF	CCP1-CCP2-FIM1 of protein C7
C7-CFF	CCP2-FIM1-FIM2 of protein C7
CCP	complement control protein
CNS	Crystallography and NMR system
CP	Classical Pathway
DAF	decay accelerating factor
dNTP	2', 3'-dideoxynucleotides
DTT	Dithiothreitol
EDTA	Ethylenediaminetetracetic acid
EGF	Epidermal growth factor-like
FIM	Factor I like module
HSQC	heteronuclear single quantum coherence
IMAC	Immobilized metal-affinity chromatography
LDLRA	Low density lipoprotein receptor class A
MAC	Membrane attack complex
MACPF	MAC of complement and perforin
MASP	Mannan-binding lectin associated serine protease
MCP	membrane cofactor protein
MG	Macroglobulin

MS	Mass Spectrometry
Mw	Molecular weight
MWCO	Molecular weight cut-off
NEB	New England Biolabs
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
OD_{600}	optical density at wavelength=600 nm
PCR	Polymerase chain reaction
PDB	Protein data bank
RCA	Regulator of complement activation
RMSD	Root mean square deviation
SAXS	Small angle x-ray scattering
SOB	Super Optimal Broth
SOC	Super optimal broth with catabolite repression
SDS PAGE	Sodium dodecylsulfate polyacrylamide gel electrophoresis
SPR	Surface plasmon resonance
TOCSY	Total correlation spectroscopy
TSPC	Thrombospondin type-1 repeat (C-terminal)
TSPN	Thrombospondin type-1 repeat (N terminal)
YT	Yeast-tryptone medium
YPD	Yeast-peptone-dextrose medium

INTRODUCTION

1.1 The complement system[1]

1.1.1 An overview

The complement system is both a key molecular component of innate immunity and provides one of the main effector mechanisms of antibody-mediated immunity. Phylogenetic analysis has traced the genetic origins of the complement system to the deuterozome lineage of metazoans that emerged more than 900 million years ago.[2] Many primitive multicellular organisms have relatively simple complement systems while expansion of complement genes by gene duplications has given rise to the more complex complement systems found in higher vertebrates.[3]

The human complement system is composed of over 30 serum and cell-surface proteins. It is required for host defense against invading pathogens, clearance of immune complexes and disposal of apoptotic cell debris and other waste products. It has opsonic, cytolytic and inflammatory role and has been implicated in the pathogenesis of several autoimmune, ischemic and vascular diseases.[1] Furthermore, specific complement components provide an essential link between the innate and acquired immune systems acting as molecular adjuvants, augmenting antigen-specific immune responses.[4] More recently, accumulating evidence suggests that the complement system has many more novel, non-inflammatory roles in the promotion of cell differentiation, tissue remodeling and regeneration, in a variety of biological processes such as animal reproduction and organ regeneration.[5] The multi-functionality of some complement components is governed by their ability to interact with a large number of complement factors, including complement regulators and receptors, as well as non-complement proteins.

Whilst the various biological pathways in which complement participates are currently under investigation, the activation events, the central proteolytic cascade (Fig. 1) and the terminal steps leading to assembly of the membrane-attack complex are well established.

These are key to the role of complement in inflammation and host defense against pathogens. These processes are less understood at the level of atomic resolution of structures. An emerging theme is the exposure of distinct interaction sites on protein molecules occurring as a result of conformational changes induced by limited proteolysis and /or formation of complexes.

1.1.2 The complement cascade [1]

The human complement cascade (Fig. 1) is activated *via* the three canonical activation pathways: the classical, lectin and alternative pathways. Two novel routes to activation - the C2 bypass[6] and extrinsic protease pathway[7] – are also potentially important. All of these pathways entail successive cleavage events, prominent amongst which is conversion of C3 to C3b, which instigates the central amplification cascade, and of C5 to C5b that nucleates assembly of the membrane-attack complex in the so-called terminal pathway.

The classical pathway is initiated by the recognition of immune complexes containing IgG or IgM, and by C-reactive protein, as well as various danger-associated and pathogen-associated molecular patterns (DAMPs and PAMPs).[1] Recognition is achieved *via* the globular domains of C1q, which is part of a heteroligomeric complex C1 that also includes two copies, each of proteins C1r and C1s. The lectin pathway also occurs through the recognition of PAMPs, *via* mannose-binding lectin (MBL) or ficolin proteins in conjunction with MBL-associated serine proteases (MASPs).[1] In both the classical and lectin pathways, target binding activates the proteolytic components (C1r/C1s and MASP-1 or MASP-2, respectively) of the recognition complexes, resulting in the cleavage of C4 to C4b (and C4a) and then in cleavage of C4b-associated C2 to C2a (and C2b) to generate the classical pathway C3 convertase, C4bC2a (Fig. 1). Recently, the C2 bypass pathway was described whereby MASP-2 of the lectin pathway seems to directly

attack and cleave C3 without formation of the corresponding C3 convertase.[6]

Activation *via* the alternative pathway however, occurs constitutively by spontaneous hydrolysis at a slow 'tick over' rate of a thioester bond in C3, forming C3u (also called $C3(H_2O)$). This form of the protein can bind factor B, which is one of numerous proteins within the complement system that contain CCP modules (see below). Binding of factor B to $C3(H_2O)$ renders factor B susceptible to cleavage by factor D. This generates the initiation convertase $C3(H_2O)Bb$, which cleaves C3 to generating the activated C3b molecule and an anaphylatoxin, C3a. C3b is central to the complement system. It's nascently activated thioester interacts with any nearby nucleophile allowing it to covalently attach itself to surfaces. C3b also interacts, non-covalently, with factor B yielding (after factor B cleavage by factor D) the important alternative-pathway C3 convertase, C3bBb.

The C3 convertases of both the classical/lectin and the alternative pathways create a positive-feedback loop by generating more C3b molecules that contribute to formation of further C3bBb (alternative pathway convertase) complexes. Additional molecules of C3b may bind to either C3bBb or C4b2a to generate the C5 convertases (*i.e.* C4b3b3b (CP) or C3b3bBb (AP)) that have a catalytic preference for C5 over C3, generating C5b and the potent anaphylatoxin, C5a. As with the C3a generated as a result of proteolytic cleavage of C3, C5a is a potent inflammatory mediator that triggers a plethora of processes, which culminate in the stimulation of immune cells and the elimination of the pathogen. However, extrinsic proteases such as thrombin and kallikrein have C5 convertase activity and have been reported to cleave C5 to C5b, in the description of a fourth potential complement actvation pathway.[7] C5b is the primary protein in the formation of the terminal complex of complement, the membrane attack complex (MAC), which is the focal point of the work in this report.

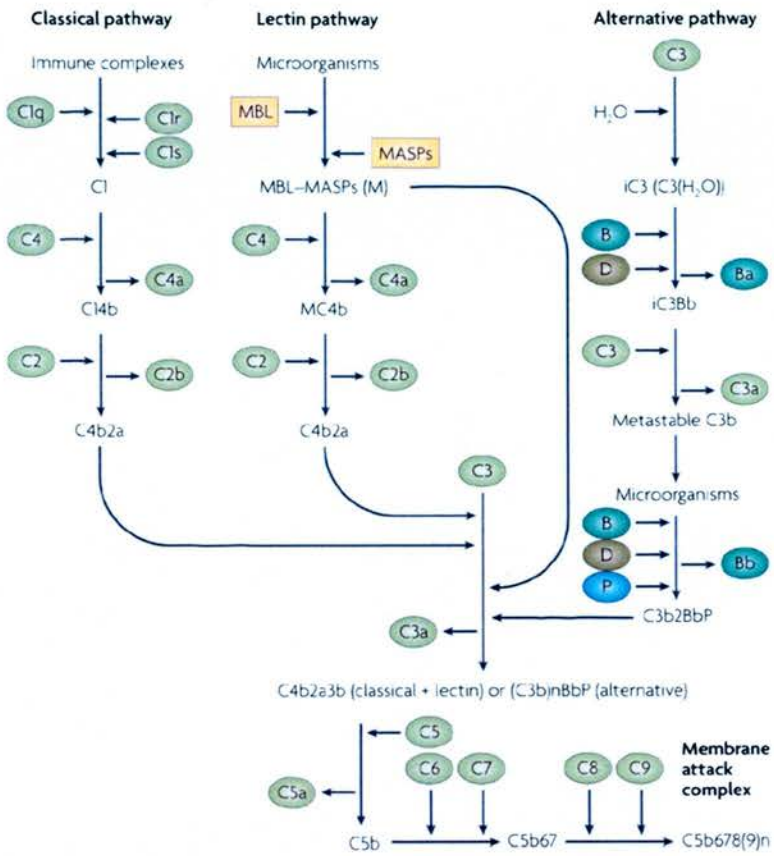


Fig. 1 Activation Pathways of the Complement Cascade[i] All three pathways (lectin, classical and alternative) involve sequential cleavage reactions resulting in the formation of a C3 convertase. The lectin and classical pathway are initiated by molecular recognition of carbohydrates and antigen-antibody complexes on the pathogen surface, respectively. This results in sequential cleavage reactions and the formation of the C4bC2b C3 convertase. The alternative pathway is initiated by spontaneous activation and attachment of C3b to the cell surface, initiating cleavage reactions that result in C3bBb C3 convertase formation. Both converatses cleave C3 forming the activate C3b, which feeds the amplification loop and has a variety of effector actions, including contributing to cleavage of C5 and activation of the terminal pathway.

1.2 The terminal pathway

1.2.1 MAC formation

The MAC is a large (555->1000kDa) heteropolymeric protein complex formed by the sequential addition of four complement proteins (C6, C7, C8 and multiple C9 molecules; Fig. 2) to C5b. The presence of multiple copies of the MAC on a cell mediates one of the principal functions of complement, the lysis of pathogenic cells. Thus the cleavage of C5 to C5b is the final enzymatic event in the complement pathway and subsequent steps involve domain reorganizations rather than cleavage events. C5 is a two-chain plasma protein (190 kDa), which is structurally and genetically related to C3 and C4, belonging to the α2-macroglobulin (α2M) 'superfamily'.[79] As with C3 and C4, C5 cleavage and activation occurs at a specific single site on the α-chain, resulting in a structural transition to the transiently activated C5b* form. Unlike C3 and C4, C5 does not contain a thioester, and therefore upon activation C5b cannot directly bind to cell surfaces. Instead, whilst reportedly[8] remaining attached to C3b within a membrane-tethered C5 convertase complex, the conformational changes that accompany C5 cleavage expose, transiently, a hydrophobic binding surface for C6.[1] If binding does not occur, the C6 binding site exposed in C5b* irreversibly decays to C5b with a half-life of ~2 min at 37 °C.[8] The C5b6 complex, on the other hand, has the capacity to disassociate from the C5 convertase, although it may remain attached.[8] The C5bC6 complex ionically interacts with the membrane surface[9] prior to the attachment of C7 and formation of C5bC6C7. Conformational changes promote release of any C5b-7 complex still attached to the convertase. Simultaneously, the C7 is presumed to undergo a conformational change that induces a hydrophilic to amphiphilic transition, allowing for membrane insertion by part of the C7 molecule. Most evidence suggests that C7 primarily provides the phospholipid-binding site,[10] although C5b has also been proposed to contribute to membrane association.[11] Due to the short half-life of the fluid-phase C5b-7 membrane binding site (10-100ms),[12] and the tendency of the complex to form protein micelles at

physiologic ionic strength and pH[13] attachment of the complex is selectively restricted to nearby membranes that will normally be those of the pathogen. Whilst this is a tight association, the integrity of the lipid bilayer as monitored by XYZ is undisturbed.(REFs) Perturbation of the membrane only occurs following addition of the C8 complex, composed of C8β and its close homologue C8α that is disulfide linked to C8γ, a protein that shares no homology to the other MAC proteins. At this point the C5b-8 complex becomes more deeply buried in the membrane and forms small perturbations of the membrane, causing the cell to become slightly leaky. The mature pore is completed by addition of multiple C9 molecules (as many as 18). These functional pores spanning the cell membrane allow for the passage of ions and small molecules into the cell, ultimately leading to osmotic lysis and cell death.

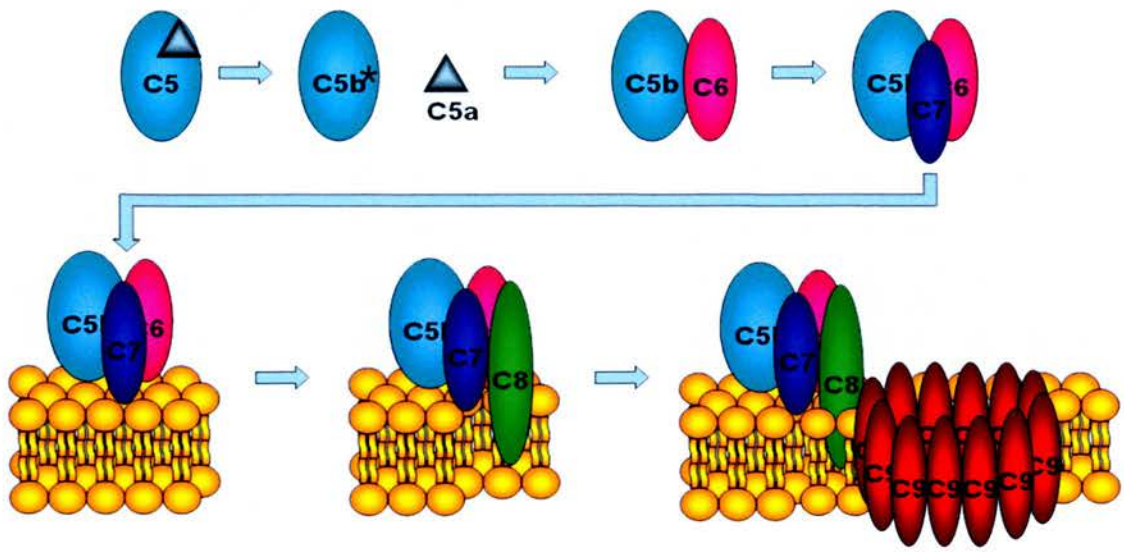


Fig. 2 MAC assembly: C5b*, possibly still bound to C5 convertase, binds C6 and C7 sequentially creating a stable C5b-7 complex that associates with the membrane. Membrane binding allows for addition of C8, which inserts into the membrane, followed by multiple additions of C9, which form a pore, completing MAC assembly.

1.2.2 The MAC in physiology, health and disease

The physiological extent of MAC action, and the ability of MACs to lyse cells, depends on the properties of the target cell. While a single MAC can lyse an erythrocyte,

nucleated cells can endocytose the MAC and repair the damage unless multiple MACs are present.[12] Gram-negative bacteria, with their exposed outer membrane, and enveloped viruses are generally susceptible to complement-mediated lysis. Gram-positive bacteria on the other hand have a thicker peptidoglycan layer protecting the membrane and making them less susceptible to MAC-mediated lysis.[13] Lytic cell death induced by the MAC is thought to be a multi-hit process akin to necrotic cell death; a rapid increase in [Ca^{2+}], followed by loss of mitochondrial membrane polarity, a total loss of adenosine triphosphate (ATP), adenosine diphosphate (ADP) and adenosine monophosphate (AMP) and finally cell death.[14] Furthermore studies on various *Escherichia coli* strains have indicated that the completion of MAC formation (specifically the incorporation of C9) was necessary for lipopolysacharide (LPS) release from the outer membrane. The release of these molecules stimulates activation of the complement system, both through their binding to antibodies and *via* their direct interaction with complement proteins.[15] At sub-lytic levels, however, the MAC can trigger diverse intracellular signalling responses, affecting pathways of the cell cycle, proliferation, differentiation and rescue from cell apoptosis.[16, 17] These sub-lytic functions are thought to contribute to focal tissue repair or cell proliferation at sites of complement activation.[17]

As with complement activation, the terminal or membrane-attack pathway requires strict regulatory control in order to prevent complement-mediated destruction of the host's own tissue. This is achieved through inhibitors present in the fluid-phase and on membranes. Solution-phase inhibition prevents the binding of C5b-7 to membranes and has been attributed to several plasma proteins such as S-protein and clusterin.[1] The accumulation of soluble C5b-7 in haemolytic inhibition assays suggests that this is achieved *via* the elusive C5b-7 membrane-binding site.[18] Host cells also express membrane proteins that protect them from MAC action such as the major MAC inhibitor

CD59, that binds to C8, preventing C9 incorporation and expansion of the pore.[1] Expression of CD59 in cancer-cells is one of the many routes of resistance of tumour cells to complement-mediated lysis, which depend on both extracellular and intracellular factors.[19]

Despite the myriad of proteins involved in complement regulation, inappropriate complement activity can result in MAC-induced tissue damage in several pathologies. Thus prospective drugs inhibiting the MAC could serve a variety of therapeutic uses. In many inflammatory diseases, such as systemic lupus erythematosus, tissue damage has been attributed to host membrane attack.[20] Similarly, in ischemic incidences, such as myocardial infarctions and strokes, the MAC has been associated with tissue necrosis.[21] Furthermore it is suggested that MAC-induced release of growth factors is responsible for several pathologies including choroidal angiogenesis, diabetic retinopathy and nephropathy.[22] Another potential use of MAC inhibition is to increase the success rate of xenotransplantations and organ transplants by preventing rejection of the transplanted organ.[23] Chronic complement activation and MAC formation has been indicated to have a significant role in many neuropathological conditions, including multiple sclerosis, neurodegenerative disorders and various epilepsies.[24] However complement activation and membrane assembly of C5b-9, depending on the pathophysiological context, can play a role in neuroprotection as well as injury.[24] Therefore drugs directing the complement system rather than inhibiting it may provide a better therapeutic rationale.

MAC primary immunodeficiency diseases (PIDs) are the converse of conditions linked with aberrant MAC action. In these conditions about three-dozen mutations, including deletions, single nucleotide polymorphisms and intron-exon boundary mutations lead to complete deficiencies in C5[25], C6[26], C7[27], C8α- γ[28], C8β[29] and C9.[30] Subtotal deficiencies have also been described for C6, C7, C8 and C9, with the protein present in plasma at

approximately 1-5% of the normal concentration levels.[31, 32] Terminal complement deficiencies are associated with recurring meningococcal and gonococcal infections and an increased prevalence of autoimmune diseases such as rheumatoid arthritis, systemic lupus erythematosus, pyoderma gangrenosum and scleroderma. The effective treatment of complement deficiencies would require replacing the missing component of the cascade, either through direct infusion of the protein or through gene therapy. However, neither of these options are currently very feasible and therefore treatment with prophylactic antibiotics and fresh frozen plasma is presently used in emerging therapies aimed at replacement of complement components.[31]

Currently, there are drugs targeting the complement system in the therapeutic pipeline,[33, 34] however, there is only one complement-specific drug approved by the FDA, an antibody targeting C5 and preventing its proteolytic activation for MAC nucleation and release of its pro-inflammatory anaphylatoxin.[35] Presently, the humanised monoclonal antibody Eculizumab (Soliris, Alexion Pharmaceuticals, Cheshire, CT, USA) is marketed as a therapy for paroxysmal nocturnal hemoglobinuria, described by an inability to prevent MAC-mediated lysis of red cells that occurs due to lack on the cells of the GPI-linked regulators CD59 and CD55.[34] However, Eculizumab has potential for treatment of numerous conditions and is currently in clinical trials for treatment of chronic autoimmune indications, including rheumatoid arthritis, atypical haemolytic uremic syndrome, membranous glomerulonephritis, dermatomyositis and lupus.[34] Furthermore, novel C5 antibodies are in development, one of which has shown efficacy in the treatment of myasthenia gravis.[36] Anti-C5 antibodies as well as anti-C8 antibodies (which block C5b-9 formation without interfering with C5 cleavage) were found to reduce tissue damage in rat hearts perfused with human serum highlighting a putative future roll in the prevention of organ rejection in xenotransplantations.[37] Recent results regarding the successful prevention of organ rejection in a hamster-to-rat heart transplant

model using anti-C6 antibodies suggest that the MAC is the only product of complement that needs to be suppressed to permit induction of accommodation.[23]

With the future development of therapeutics specifically targeting the terminal pathway, only terminal complement components will be affected, allowing the beneficial immunoprotective and immunoregulatory effects mediated by 'upstream' complement components to be maintained.[33] Furthermore targeting early MAC formation would prevent C5b-7 docking to the membrane in addition to the prevention of pore-formation.

1.2.3 Early MAC molecular interactions

1.2.3.1 MAC protein domain organization

As drugs directed towards the MAC provide an attractive therapeutic option, structural and functional insights regarding MAC formation, and any putative pre-initiation complexes, would be of invaluable aid to rational drug design. The MAC protein family; C6, C7, C8 (C8α and C8β) and C9, are structurally and genetically related proteins with a similar modular structural organization (Fig. 3) and highly conserved sequences.[38] These proteins share two main distinctive features; the large central domain of the MAC/perforin superfamily (MACPF, ~300-370 aa) and the tandemly arranged, small, disulfide-rich modules located at their N and C termini (~40-80 aa). All MAC proteins have thrombospondin type 1 repeats (TSP) at the N terminus. Exceptionally, C6 has two TSP domains at the N-terminus (TSPN and TSPN2). The TSP-1 domain, or domains, are followed by a low-density lipoprotein receptor class-A domain (LDLRA) that precedes the MACPF. An epidermal growth factor-like domain (EGF) comes after the MACPF and is followed by another TSP (TSPC), which is found in all MAC components except for the smallest protein, C9. Outwith the modular core that is shared by all members of the MAC protein family, human C6 and C7 additionally have a pair of complement control protein modules (CCPs) following EGF and TSPC that precede a pair of factor-I

like modules (FIMs) – these four modules thus form the C terminus of both C6 and C7. An interaction between C6/C7-FIMs-and the C-terminal domain of C5, C345C, is considered to be of particular relevance to the early stages of MAC formation.

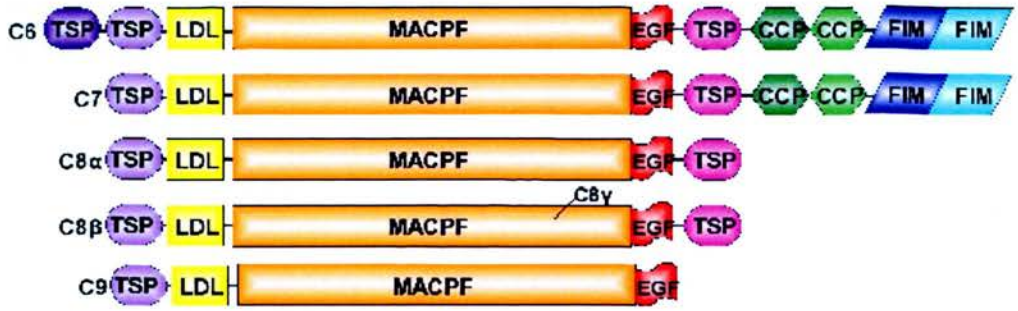


Fig. 3 Domain organization of the MAC protein family: Abbreviations: Thrombospondin type 1 (TSP), low-density lipoprotein receptor class A (LA), MAC/perforin (MACPF), epidermal growth factor-like (EGF), complement-control protein (CCP) and factor I module (FIM).

1.2.3.2 MAC pre-initiation complexes

The three paralogous proteins, C3, C4 and C5 share a similar C-terminal domain called the C345C domain. The C345C domain of C5 appears to function as a binding site for both, the C5 convertase[39] and the pair of C-terminal Factor I-like modules (FIMs) of C6 and C7.[40] In addition to the irreversible binding of C6 to C5b, and then C7 to C5b-6, in MAC formation, it is important to note that solution-phase reversible binding occurs between C6 or C7 and the uncleaved C5. This interaction site was also mapped, using tryptic fragments, to the CCP and FIM module pairs of C6 or C7 and the C345C domain of C5.[41, 42] In blood plasma, such reversible, pre-activation interactions could clearly increase local concentrations of C6 and C7 in the immediate vicinity of the newly activated C5b* molecules (that remain capable of interacting with C6 for only a short period of time), thereby facilitating MAC formation. Although the physiological significance of these reversible binding reactions has not been firmly established, it can be reasoned that they reflect the arrangement of individual components within the stable, fully formed MAC.[43]

1.2.3.3 Lytic MAC formation

Binding studies of truncated C6 proteins showed that the C6-FIMs are an aid to, but not essential for, lytic activity, and (as mentioned above) are involved in binding to C5.[42] Similarly the observations that (i) a C6 variant, devoid of FIMs, is responsible for subtotal C6 deficiency but retains bactericidal activity[44] and (ii) a common form of mouse C6 lacks the FIMs, but is still active,[45] also indicate that the C5-C345C:C6-FIMs interaction is not essential to the stability of the fully assembled MAC. The role of C6-CCPs has not been explored, although binding of truncated C6, devoid of FIMs, to C5 is not fully eradicated unless the CCPs are also absent.[41]

On the other hand, both C5-C345C and C7-FIMs (as recombinantly produced fragments) have been shown to inhibit MAC formation *in vitro* in surface plasmon resonance (SPR) experiments and *in vivo* by inhibition of MAC-mediated hemolysis.[46, 47] Thus the interaction between C7 FIMs and C5-C345C is pivotal for the formation of the MAC. As C7-FIMs and C7 were found to bind C5 and C5-C345C with nearly identical affinities (Table 1)(REF) it was suggested that the FIM-pair are primarily responsible for the reversible binding of C7 to C5, with minor or negligible contributions from the C7-CCPs.

Immobilised protein	Protein in solution	Kd (nM)
C5	C7	0.1
C7	C5	30
C5	C7 FIMs	70
C7 FIMs	C5	40
C5-C345C	C7	3
C7	C5-C345C	30
C5-C345C	C7 FIMs	50

Table 1. List of molecular affinities for C6 and C7 interactions with C5 as derived by SPR.[46, 47]

Interestingly, of 14 gene defects so far described as causing complete C7 deficiency (see section 1.2.2), four are found in exons coding for the C7-CCPs and four in regions coding for C7-FIMs. On the other hand, defects causing C6 deficiency are more evenly

distributed across the C6 gene.[32] This could reflect the biological significance of C7-CCPs and C7-FIMs, and indicates that structural conservation in these modules is necessary for authentic protein-function.[27] Whilst the C7-FIMs appear to have a role in directly binding C5-C345C before and during MAC assembly, the C7-CCPs are likely to have an architectural or mechanical role in the putative domain rearrangements that accompany MAC formation: they could form part of a swinging molecular arm that relocates the C7-FIMs to the C5-C345C during complex formation; subsequently they might participate in secondary interactions with C5b that help to stabilize the self-assembling complex as it accretes additional subunits.

A current model of initial steps in MAC formation[40] (Fig. 4) suggests that the primary C6-binding site is formed by conformational changes elicited by C5 activation exposing a hydrophobic metastable binding region (in C5b*) that must react within minutes with C6. The identification of these sites of interaction, however, is yet to be directly explored. The C5b-6 interaction may be further facilitated by an interaction between C5-C345C and C6-FIMs provided by a putative molecular arm. But if this exists at all, it provides a relatively small amount of the overall binding energy and does not greatly increase the stability of C5b-6; other C5/C5b-binding sites exist within C6. Subsequent binding of C7 to C5b-6 results in replacement of any C6-FIMs:C5-C345C interaction by the higher affinity C7-FIMs:C5-C345C interaction. The displaced C6-FIMs may then re-associate with C6, associate itself with another region of C5, interact with C7, or remain unattached. According to this model, the incorporation of C7 into the complex elicits a conformational change that exposes a membrane-binding site, predominantly formed by but not restricted to C7. This converts the soluble C5b6 complex into a membrane-associating C5b-C7 complex. However, very little is known about the mechanism of these steps or of further steps in the sequential assembly of the MAC on a membrane, despite several decades of effort. It is a spontaneous process that must be directed by

energetically favorable inter-protein domain-domain rearrangements and interactions yielding nascent binding sites, and conformational rearrangements.

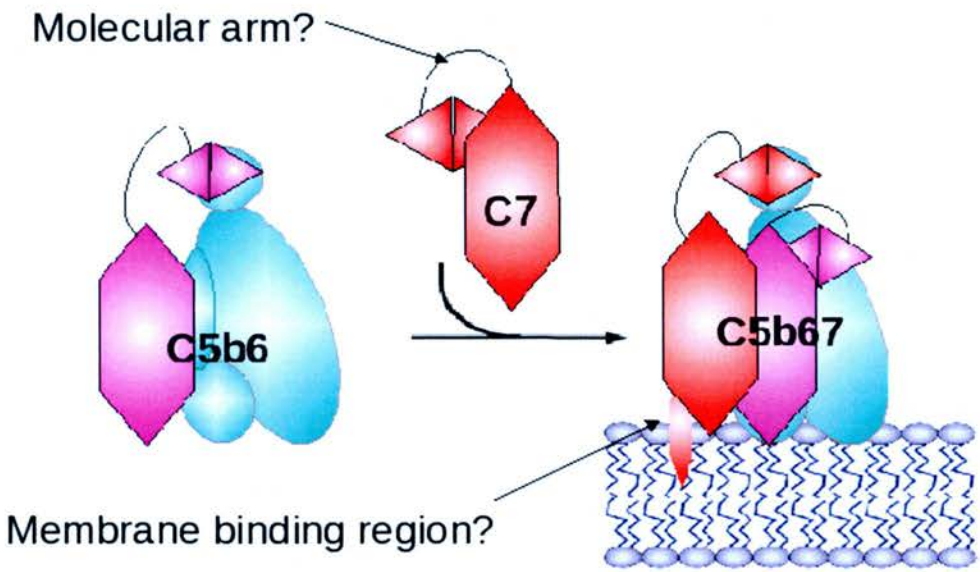


Fig. 4 Schematic of C5b-7 MAC model:[25] First stage: C6 (pink) binds to a metastable binding site on C5b* (cyan), with a possible stabilizing (non-essential) role played by the C6-FIMs:C5-C345C interaction. Second stage: Upon C7 (red) binding any C6-FIMs bound to the C345C domain are displaced by the C7 FIMs due to their higher affinity for C5-C345C. This release of the C-terminal molecular arm is coupled to major conformational rearrangements of C7 that mirror those seen in MACPF-containing relatives such as perforin and cholesterol-dependent cytolysins. This involves conversion of helical clusters into membrane-penetrating beta-hairpins thus allowing C5b6 to target to cell surfaces.

1.3 MAC structure-function relationships

Structural studies on MAC protein modules, module-module interactions , full-length MAC proteins and protein complexes are pivotal to elucidating the mechanisms involved in the formation of the MAC. Probably because MAC proteins have a modular structure, likely exhibiting flexibility at intermodular junctions, they have generally proved refractory to X-ray crystallography analysis when studied as whole molecules.[48] Pioneering negative-stain transmission electron microscopy studies revealed low-resolution structural information on both individual MAC proteins and higher-order MAC complexes (see section 1.3.6), these experiments afford valuable insights of how

this technique is notoriously vulnerable to artifacts arising from sample preparation. They should probably be repeated on the most up-to-date instruments and the data coupled to recently obtained atomic-resolution structures of component domains.

In a very important recent development, two crystal structures of the C8α MACPF domain – one alone,[49] the other with the γ-subunit[50] attached - revealed a fold reminiscent of the bacterial, pore-forming, cholesterol-dependent cytolysins. This immediately suggests models of the later stages of MAC pore-formation (see section 1.3.2). First insights into the C5 structure were provided by NMR, which was used to determine the structure of the C5-C345C domain. Subsequently, the atomic-resolution structure of full-length C5[51] alone, and in a complex with a C5 convertase inhibtor SSL7[52] have also been determined. While this is not a component of the MAC, C5 will likely have many similarities with its activated counterpart C5b (see section 1.3.3), and the C5 to C5b transition might emulate the major conformational changes observed in the case of C3 conversion to C3b. NMR was also employed to provide the first high-resolution structural information on C7, revealing the FIMs[53] structure and highlighting putative regions for interacting with C5-C345C[54] (see section 1.3.4).

Thus, despite much efforts, there are currently no atomic-resolution structures of any full-length MAC proteins, and there is no atomic-level structural information regarding any of the protein-protein interactions and conformational changes involved. A combination of atomic-resolution X-ray/NMR-derived structures and lower-resolution approaches such as electron microscopy, small-angle X-ray scattering and chemical cross-linking (see section 1.4) integrated with functional data, will be necessary for the provision of a comprehensive understanding of the intricate structure-function mechanisms involved.

Domain/Protein	Method	Reference
C5	X-ray Crystallography	Nat. Immunol. (2008) **9**:753-60
C8α-MACPF	X-ray Crystallography	Science (2007) **317**:1552-4
C8α-MACPF-γ	X-ray Crystallography	J.Mol.Biol (2008) **379**:331-42
C7-FIMs	NMR	J.Biol.Chem (2009) **284**:19637-49
C5-C345C	NMR	J. Biol. Chem. (2005) **280**:10636-45

Table 2: Journal Articles of particular relevance to MAC proteins structure.

1.3.1 C8:α, β, γ

Following the attachment of the C5b-7 complex to the target cell membrane, the incorporation of C8 and perturbation of the membrane initiates C9 circular polymerization. With respect to bacterial killing it has been useful conceptually to consider the C5b-8 complex as a receptor to which C9 binds with a conformational change allowing C9 to cross the outer membrane to reach the periplasm.[55, 56] C8 consists of a disulfide-linked C8α,γ heterodimer that is non-covalently associated with C8β. C8α and C8β are paralogous with C6, C7 and C9 (see Fig. 3). C8α has sites for the simultaneous binding of C8β, C8γ and C9.[57] C8β, on the other hand, reciprocally contains a binding site for C8α and an additional site for C5b-7.[58] Binding studies indicate that C8α-MACPF provides the primary binding site for C8β, C8γ and C9 and while the small N- and C-terminal modules stabilize these interactions, they do not confer specificity, *i.e.* interactions involving the small modules are an aid to but not essential for protein-protein interactions.[59-61] In the case of C8β, the TSPN module and the MACPF cooperatively bind C8α-γ[62, 63] and the LDLRA and the MACPF are essential for specificity in binding C5b-7, whilst the C-terminal modules do not confer specificity but stabilize the complex presumably through secondary interactions.[62]

Inclusion of C8γ in the fully formed MAC enhances MAC mediated hemolysis. It is not essential however, and does not result in deeper membrane penetration by C8, but likely maintains C8 in a favourable conformation.[66] It is unrelated to any complement proteins and is a member of the lipocalin protein family. These share a core structure

consisting of an eight-stranded anti-parallel β-barrel that resembles a calyx or cup-like structure. The cup contains a binding pocket used by lipocalin family members to bind small hydrophobic ligands although, to date, the natural ligand/s for C8γ have remained elusive (Fig. 5). The experimentally-derived structures, solved with the C8α indel (C8γ binding sequence of C8α),[64] in the C8α-MACPF-C8γ co-crystal,[65] represent the indel in the upper portion of the ligand-binding pocket of C8γ. This has led to two hypotheses regarding the nature of the C8γ binding pocket. One hypothesis is that the sole role of the binding pocket is to initiate contact between C8α and C8γ, prior to formation of the disulfide link between them. The other possibility is that binding of the putative natural ligand of C8γ is regulated by binding to C8α. According to this model, C8 incorporation into the MAC causes a conformational change that disrupts the C8α indel:C8γ interaction, promoting either the binding or the release of a ligand.

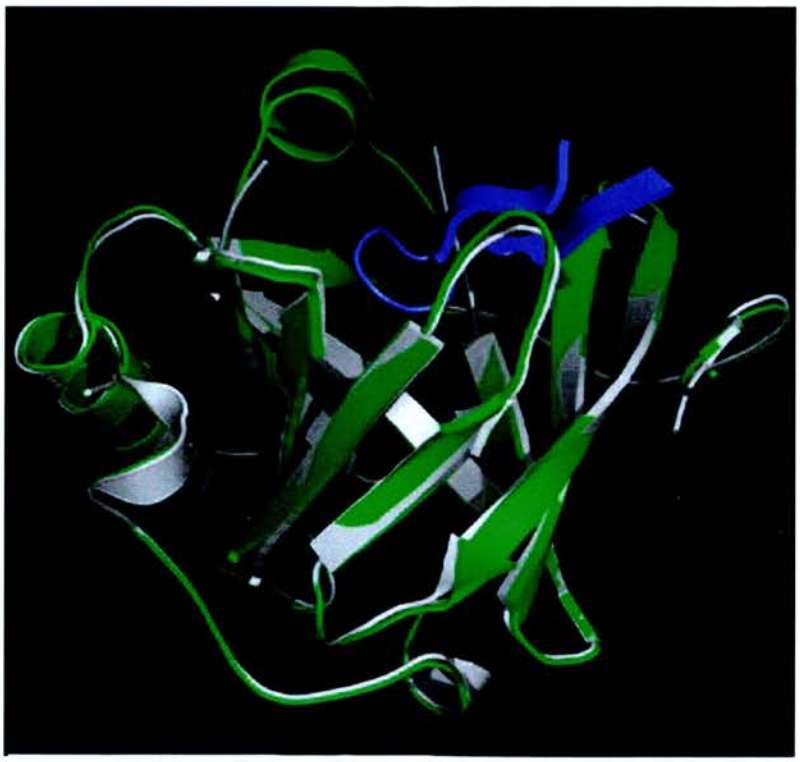


Fig. 5: C8γ and C8α-indel binding. Superposition of the C8γ·indel structure (green) on C8γ (white). The indel is shown in blue. PDB ID: 1IW2 Adapted from paper.[64]

The global architecture of C8 has been probed in a pre-publication release of three-dimensional negative-stain electron microscopy studies.[68] Refinement of the C8 structure, determined by fitting the C8α-MACPF and C8γ structures and a homology model for C8β-MACPF into the density map, converged on a reconstruction whose resolution was 24 Å. According to this low-resolution model (Fig. 6), the C8 quaternary structure consists of a core domain where the two L-shaped MACPF domains (from C8α and C8β) are sandwiched together and a globular protrusion contains the C8γ subunit. Because of the resolution of the reconstruction there were two possible placements of C8β, both plausible given the filling of density space with the modules and their availability for known MAC interactions.[57-63] The orientation in which the MACPFs have

the same orientation in a *cis* configuration was favoured, as the *trans* configuration was in disagreement with the solvent-accessibility of a predicted glycosylation site. This orientation, would result in a closer proximity of the regions of C8α and C8β-MACPF ,thought to be responsible for membrane binding (see next section). Clearly, more detailed structural data are essential for a full molecular understanding of these proteins.

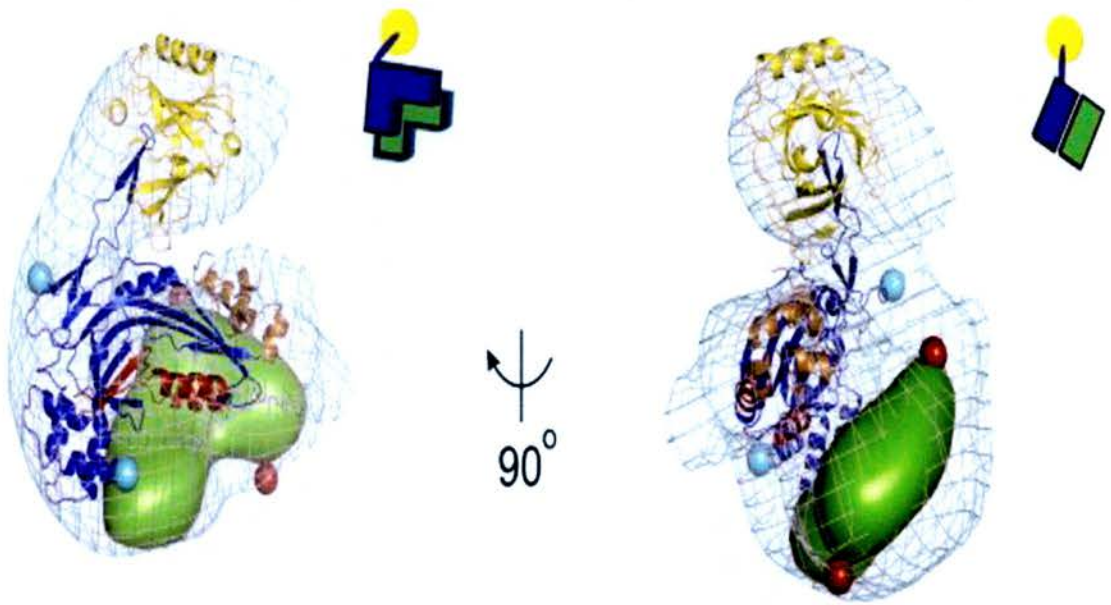


Fig. 6: EM mesh showing relative locations of protein modules in C8 as determined by electron microscopy: The crystal structure (PDB: 2RD7) of C8α-MACPF (blue ribbons) and C8γ (yellow ribbons) were unambiguously fit into the reconstruction (grey mesh). Two possible placements of C8β-MACPF, related by a 180° rotation, interchanging the arms of the "L," were scored. The one shown here corresponds to the lower refinement residual and is rendered as an isosurface filtered to 25 Å (green). The N and C termini of C8α-MACPF are cyan spheres; the N and C termini of C8β-MACPF are brown spheres. Quaternary structure schematics are shown for each view. Adapted from paper.[66]

1.3.2 MACPF and pore formation

Crystal structures of the C8 α-MACPF[49, 65] domain (Fig. 7) revealed a thin L-shaped molecule with a fold similar to that of the bacterial cholesterol-dependent cytolysins (CDCs) with a central kinked four-stranded β-sheet surrounded by α-helices formed by

two structural segments, domains d1 and d3. Absent from MAC proteins, but present in CDCs, the d2 region forms a linker between d1 and d3 and the d4 region responsible for membrane association. Despite an undetectable relationship by sequence analysis, the structural similarity of C8α-MACPF to the CDCs indicates that they share a common mechanism of membrane insertion and pore formation. Insertions between β1–β2 and β3–β4 of the central β-sheet of CDCs form two clusters of α-helices that correspond with two clusters of α-helices (HC1 and HC2) that refold to form transmembrane β-hairpins (TMH1 and TMH2), forming a β-barrel pore , upon membrane insertion.[69] TMH1 is loosely sandwiched between the central β-sheet and the stalk-like β-sheet of d2, while TMH2 is more solvent exposed. The CDC molecules oligomerise, forming a pre-pore on the membrane surface, whereupon transition to the pore state involves a series of conformational changes.[70] Mutational and photo-labeling studies suggest that edge-to-edge stacking of the β-sheet core[68] and movements in d4 upon interaction with the membrane[69] lead to deformation of d2 and the loss of contacts with d3,[72] accompanied by straightening of the bent β-sheet, freeing the helical bundles, allowing d3 to move down on to the membrane surface, and the helical clusters to extend into amphipathic β-hairpins for membrane insertion and subsequent formation of the β-barrel pore.[73] It has been proposed that a similar mechanism exists for pore formation in the MAC whereby the sequential addition of MAC proteins is akin to CDC oligomerization.[74] In this model binding of C8 to C5b-7 arises from C8β-C6/C7 interactions. These potentially include edge-to-edge stacking of MACPF β-sheet cores. These interactions induce unfolding of the putative HC1 and HC2 in C8β, and subsequently those in C8α, to form an extended series of aligned β-hairpins that penetrate the lipid bilayer. These likely create a more energetically favorable environment for deeper membrane insertion by C9.

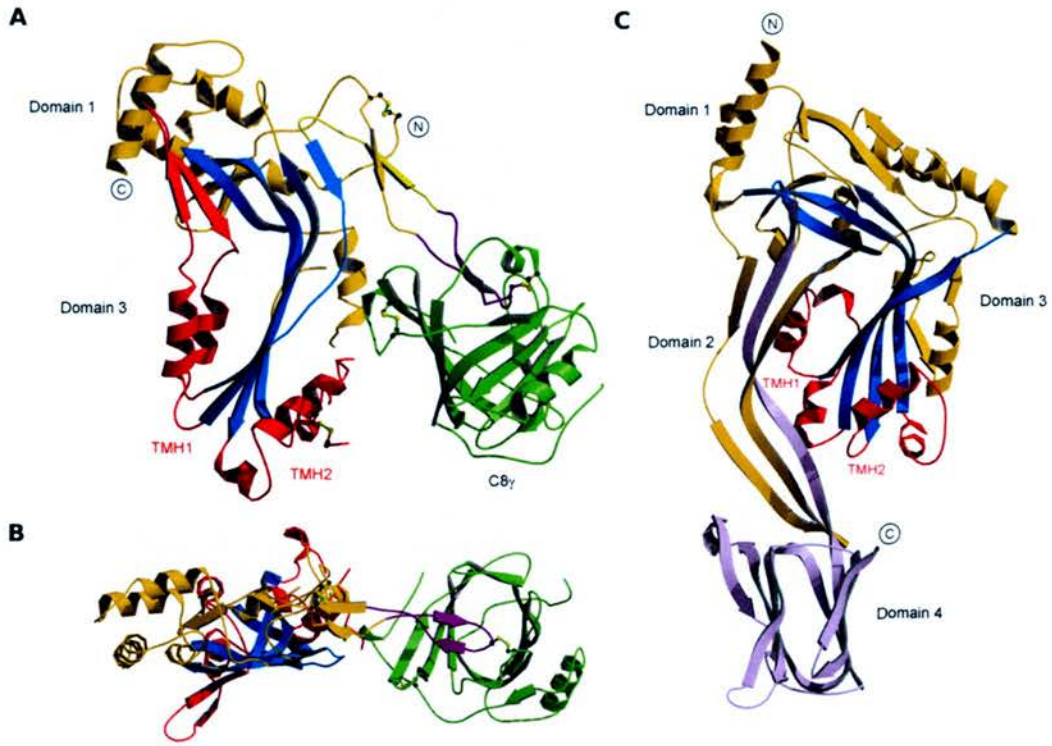


Fig. 7: Structure of C8α-MACPF-γ Vs CDCs. (A) Ribbon representation of αMACPF-γ (PDB ID: 2RD7). C8γ is green, the core β-sheet is blue, the TMH segments are red and, the C8α-indel is purple and the other structural regions are in gold. Regions corresponding to d1 and d3 are labelled accordingly. (B) Top view of C8α-MACPF-γ. (C) Ribbon representation of intermedilysin (ILY) as a representative bacterial CDC (PDB ID: 1S3R). Colours and labels are as in C8α-MACPF-γ. In addition, domain 4 is in pale purple. Adapted from paper.[50]

Electron micrographs of the MAC reveal a torus-like pore with ~100 Å inner-diameter, ~160 Å height, and ~200 Å outer-diameter. Photo-labelling studies indicates that the TSP and LDLRA modules are fully exposed on the exoplasmic side of the polymer, and the EGF is partially embedded in the cylindrical wall of poly(C9) and a small rim at the cytoplasmic base of the pore.[75] An *in silico* representation of the pore, was built, based on the C8α-MACPF structures but using more extended TMH sequences as are found in C9-MACPF. These sequences were modelled as extended beta-hairpins (as in CDCs) and eighteen copies of the molecule were placed in rings, with the β1 to β4 strands of d3

placed on the inside (Fig. 8). The modelled pore structure is consistent with the EM studies.[49]

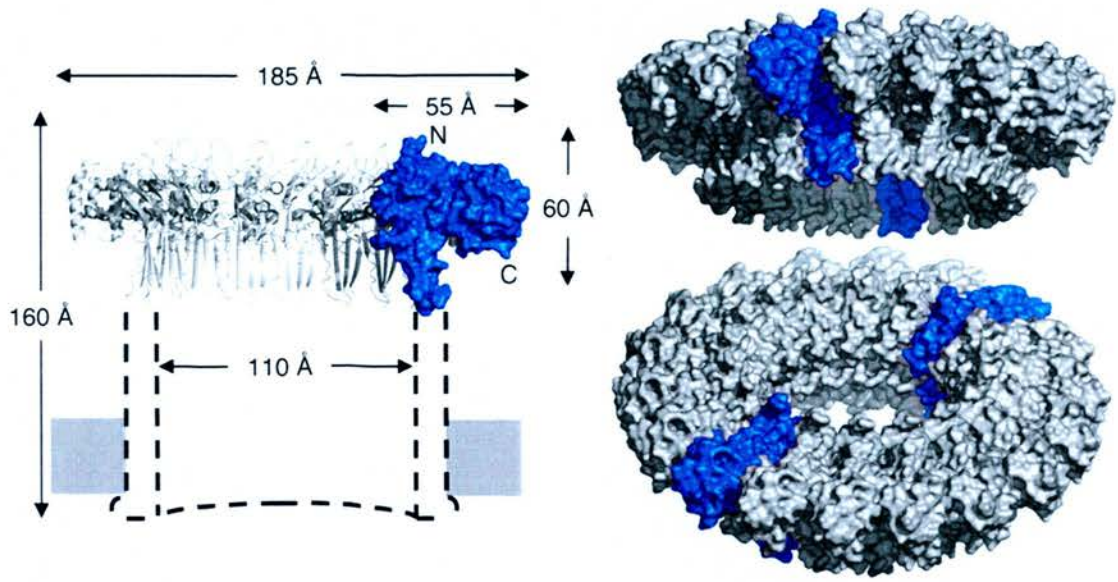


Fig. 8: C9 pore model. Shown is a hypothetical model of the C9 pore derived from a ring of 18 monomers of C8α-MACPF. (Left) Cross section of the pore, with MACPF domain forming the torus. (Right) Two orientations of the torus in surface representation, with two individual monomers highlighted (in blue) for convienience. Adapted from paper.49

While sequence homology (47%) indicates that the C9-MACPF domain is similarly folded to that of C8α; and a hypothetical model of the C9 pore based upon the C8α-MACPF structure is consistent with pore dimensions derived from electron microscopy (see Fig. 8), it does not necessary follow that C9 uses the same structural elements to enter a membrane as similarly folded CDC proteins. Disruption of TMH1 in C9 by cleavage[76] and by glycosylation,[77] in conjunction with photo-labelling studies,[78] indicate that the transmembrane region proposed from the release of the C8α-MACPF structure may be more important for formation of circular poly(C9) as opposed to forming the membrane pore itself. A recent study that probes C9 topology by glycosylation mapping, anti-peptide antibody binding, and disulfide modification suggests that there are two modes of membrane interaction in C9. C5b-9 complexes containing less than four C9

molecules are anchored to the membrane in a monotopic fashion utilizing the conserved helices, while in fully formed poly(C9) other structural elements may refold and becomes staves of a transmembrane barrel.[75] This would reflect the ability of C9 to bind lipids without MAC incorporation.[78]

The recent solving of the X-ray crystal structure of the murine perforin monomer, in combination with a cryoEM reconstruction of the entire perforin pore, indicates that there is remarkable flexibility in the mechanism of action of the conserved MACPF/CDC fold. It also provides new insights into how related immune defence molecules, such as complement proteins, assemble into pores.[79] Perforin, like the MAC proteins, contain a MACPF domain followed by an EGF-like domain. The EGF-like domain is a short (~30 aa) disulfide-rich, flexible domain that, in the perforin protomer structure,[79] is in close association with the HC1 region and the C-terminal end of the MACPF. The EGF-like domain and the C-terminus of the MACPF domain together form a shelf upon which the bulk of the MACPF domain "sits". The HC1 region of the MACPF domain is loosely held between this shelf and the central beta sheet core. Earlier studies of C9 had indicated that its EGF-like domain similarly folds up against the equivalent putative β-hairpin-forming region.[81]

Unexpectedly, the cryoEM evidence suggests that the perforin MACPF domain within the pore-form is in the opposite orientation to the MACPF domains within pores formed by the CDCs. It appears that the perforin molecule does not undergo a major collapse in the pore-form as has been observed in, for example, pneumolysin pores.[73] Instead it is postulated that perforin, and by extension MAC proteins C8 and C9, achieve membrane insertion without a major buckling of the molecule but *via* extended CH1 and CH2 sequences. When these regions unfurl to form loops, they are therefore long enough to permit passage over the shelf and to thereby reach and insert into the membrane, without

buckling the molecule. B-factor analysis, in agreement with EM-based evidence, indicates that the shelf region in perforin is flexible and is less ordered[79] in comparison to the domain 2 region of CDCs which similarly associates with the TMHs. An attractive idea is that MACPF:EGF shelf interactions are lost upon inclusion of monomers into the MAC promoting extension of TMHs, mirroring the deformation of d2 and subsequent extension of TMHs implicated in CDC pore formation.

Comparison of the putative MACPF TMH regions in C8 and C9 with C6 and C7 (Fig. 9) one can see that the CH1 segment is shorter and largely hydrophilic suggesting that membrane penetration as a β-hairpin structure would be limited. The disulphide loop region in CH2 is shorter in C6 and C7, however, C7's sequence is still relatively hydrophobic in nature, while the sequence in C6 has many charged residues. This is consistent with C7 providing the principal membrane binding site, although contributions from C5b and C6 have not been ruled out. The much shorter TMH regions proposed for C7 could be possible membrane binding regions as deep penetration of the membrane is not required, however other regions of the protein and the C5b-7 complex may also be associated with the membrane. It is clear that any structural information regarding the domain arrangements and re-arrangements necessary for the sequential binding of MAC proteins will be the key to elucidating the underlying mechanisms of MAC formation and function.

Fig. 9: Comparison of TMH sequences in MAC proteins. Secondary structure is shown as a cartoon above the sequnces. PFO refers to perforin and ILY to the CDC intermedylsin. Sequences are aligned with positively charged residues in red, negatively charged residues in green and neutral/hydrophobic residues in yellow. Figure from paper.[50]

1.3.3 C5b

Full-length C5 is the 188-kDa precursor of the proteolytically activated 185-kDa C5b molecule. Therefore, (in the absence of a C5b structure) insights into structure-function relationships in C5b can be gained from the X-ray crystallography structure of its inactive counterpart.[51, 52] Members of the α2M superfamily, such as C5, have been suggested to share a common architecture prior to cleavage of the bait/anaphylotoxin regions, with the development of features unique to each protein.[82] As with all α2M family members C5 has a stable core of eight macroglobulin domains (MG), MG 1-5 and part of MG 6 are formed by the β chain, followed by a linker (LNK) domain (Fig. 10A). The α-chain begins with the cleavable anaphylatoxin domain (ANA), followed by the second portion of MG 6 and MG 7. A second insert incorporates the CUB domain and the thioester-containing domain (TED), or the structurally equivalent C5d domain in C5 (that does not contain an internal Cys-Gln thioester bond as in C3), with the C-terminus ending in the final MG domain, MG8. C3, C4 and C5 additionally contain an

anchor connecting MG8 to the C345C domain. Together these domains form two superdomains: the MG core or superhelix formed by the α and β-chains, and the CUB-TED/C5d-MG8 superdomain anchored to the MG core *via* MG8. While the MG core is structurally stable in a right-handed superhelical formation, the CUB-TED/C5d-MG8 superdomain undergoes substantial conformational changes following proteolytic cleavage.[83]

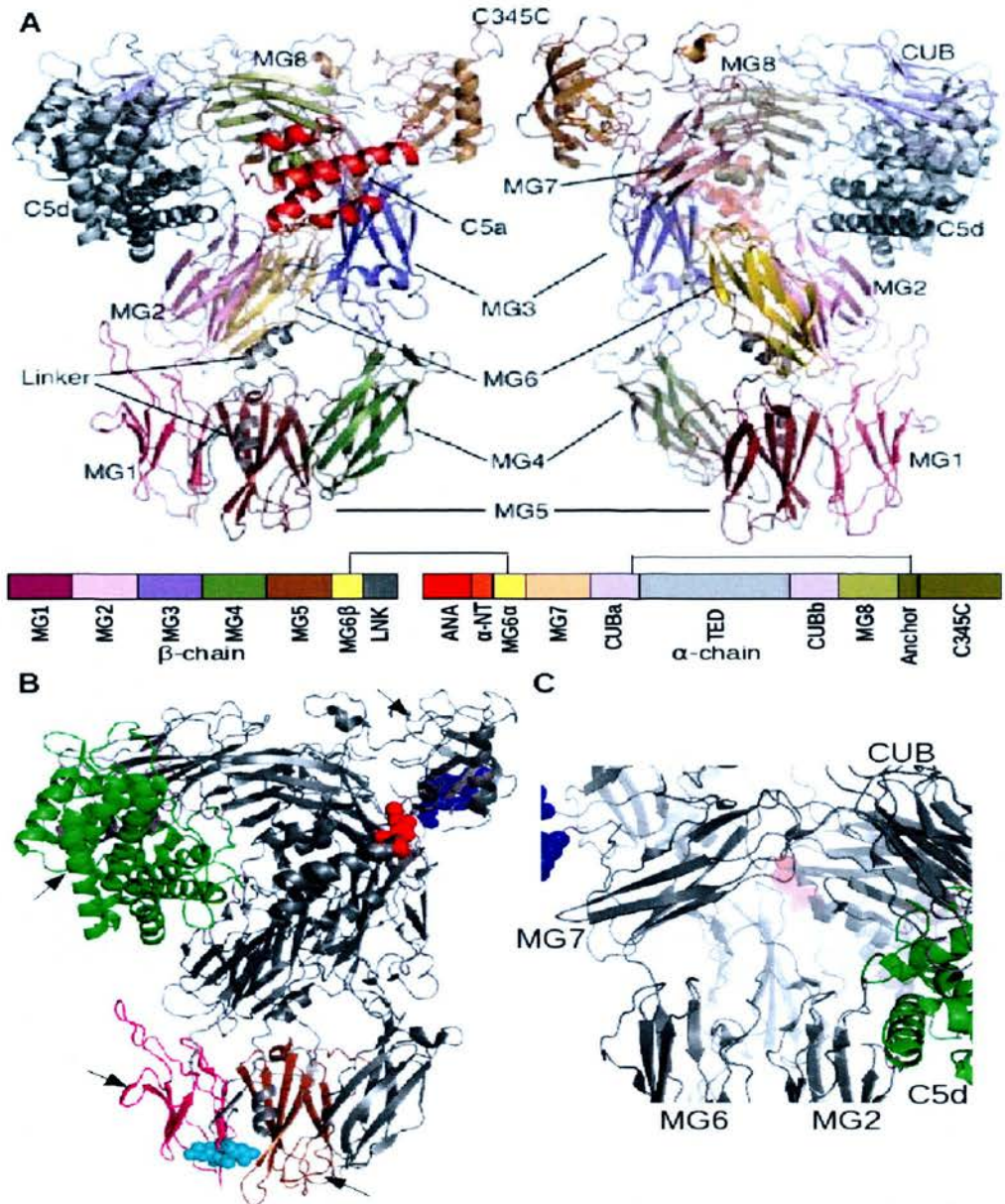


Fig. 10: C5 structure. (A) Ribbon representation. (B) C5 crystal structure (PDB ID:2A73) colour coded as in A. (B) Potential binding sites of C5 for C3bBbC3b and C4bC2C3b: the convertase cleavage site Arg751-Leu752 (red, spheres) sandwiched between the MG3 and MG8 domains; residues 1628–1633 of the flexibly attached C345C domain (blue, spheres); the C5d domain (green, ribbon); A region at the bottom of the molecule formed by MG1 (pink, ribbon) and MG5 (brown, ribbon) and residues of the linker (cyan, spheres). Potential binding sites for C6 and C7: the C345C, C5d and MG1 and MG6 are marked with arrows (C) A close up of the cavity on the opposite face of the convertase cleavage site created by MG2, MG6, MG7, CUB and C5b, colouring is as in (B). (A) is adapted from paper,[51] (B) and (C) were produced in Pymol.

The structure of C5[51] and C5 complexed with a convertase inhibitor protein, SSL7 from *Staphylococcus aureus*,[52] in combination with binding studies[85] and observations from conservation analysis,[51] have suggested up to five regions of C5 that might be involved in binding to the C5 convertase enzymes, covering extensive regions of the molecule (See Fig. 10B, C). Following enzymatic cleavage on a complement-activating cell surface the C5b* molecule remains attached to the convertase enzyme, awaiting C6 binding. It has been proposed that the C5b* conformation will be akin to that of an intermediate form of C3b (C3b*) where the CUB and TED domains are completely detached from the β-chain, prior to attachment to the target cell-surface.[83] Recent structural studies on C3u (also called C3(H_2O)), the spontaneously activated form of C3 from the alternative pathway, structurally and functionally akin to C3b, suggest that the TED domain is significantly extended away from the MG1-MG8 domains in solution.[85] An attractive option is that the conformational change induced by C5 activation propels the C5d domain towards the C345C domain and away from the MG core for binding C6. This is in agreement with proteolytic stripping[86] and photolabelling[87, 89] experiments that have indicated that the α-chain of C5b is proximal to the membrane facing the MAC pore, while the β-chain is further from the membrane, peripherally positioned within the MAC. Binding regions within C5b for C6/C7, other than the C345C domain (see next section), have not been identified. Other regions of C5 must be involved in the interaction, however, as the flexibly attached C345C domain would not be sufficient to discriminate against MAC formation with native C5.[51]

The extensive convertase binding sites on C5b and the sites for C6/C7 binding likely share close proximity or some overlap. This would account for early MAC complex release from the convertase upon binding of C6, and in particular of C7, to C5b. The proposal that C5d[51] is a putative C6 binding partner therefore seems highly plausible and would parallel the importance of the C3b equivalent, *i.e.* the TED domain, in covalent

attachment to target-cell membranes. Moreover, it has also been proposed that the transient nature of C5b* is conferred by a further movement of the C5 CUB-C5d-MG8 superdomain to adopt a more C3b-like conformation (see section 1.4.2).[51] This would result in the C5d domain being closely associated with the MG core and therefore shielded from interaction with C6. Similarly, occupation of the convertase binding site by SSL7 is likely to reflect the binding of this bacterial protein to activated C5b, an interaction that results in the inhibition of hemolysis and soluble MAC formation.[84] Thus according to this binding model, occupation of the SSL7-binding region within C5/C5b (primarily provided by MG1 and MG5 with minor contributions from MG2 and MG6) competes with functionally critical interactions between C5/C5b and C6/C7, by directly or indirectly masking C6/C7-binding sites.

1.3.4 C7-FIMs:C5-C345C

Even the best characterized of the early MAC assembly interactions - that between C5-C345C and the C7-FIMs - where three-dimensional structures are available for both domains, is poorly understood. The structure of C5-C345C consists of an oligosaccharide/oligonucleotide-binding fold that is similar to the fold of the netrin-like module family. In this structure, two helices (α-subdomain) pack against a five-stranded β-barrel (β-subdomain, strands A-E, Fig. 11). While mutagenesis studies[87] indicate the DE loop of the C5-C345C domain in convertase binding, no mutations have been discovered that affect its interaction with FIMs. Unpublished (Dr Ron Ogata, Torrey Pines Institute, Florida) mutagenesis studies have excluded *AB*, *BC*, and *CD* loops, and the Cys-Ser-Ser-Cys bulge in the loop connecting the sub-domains, leaving the unexplored α-subdomain to be provisionally identified as a likely candidate.[53] Comparisons made with C3 and C4, *i.e.* structural homologues with no affinity for C6 or C7-FIMs, highlighted characteristics of the α-subdomain that are unique to C5; this further implicates this subdomain in C6/C7-FIMs binding. The highlighted residues

consist of a pair of exposed hydrophobic sidechains, Phe[1654] and Leu[1655], within α-helix-2, adjacent to a strikingly electronegative patch that extends over both helices. C7-FIM1 and C6-FIM2 were implicated as the domains that dominate FIMs binding to C5b in the context of a C5bC6C7 complex. This is due to the high pIs of C7-FIM1 (9.5) and C6-FIM2 (9.2) indicating a compatibility with the electronegative patch of C5-C345Cs α-subdomain.[90] Moreover, an absence of detectible binding of lone C7-FIM2 with C5-C345C[53] may reflect the importance of C7-FIM1 in the interaction, but does not exclude a co-operative binding mode of the FIMs modules.

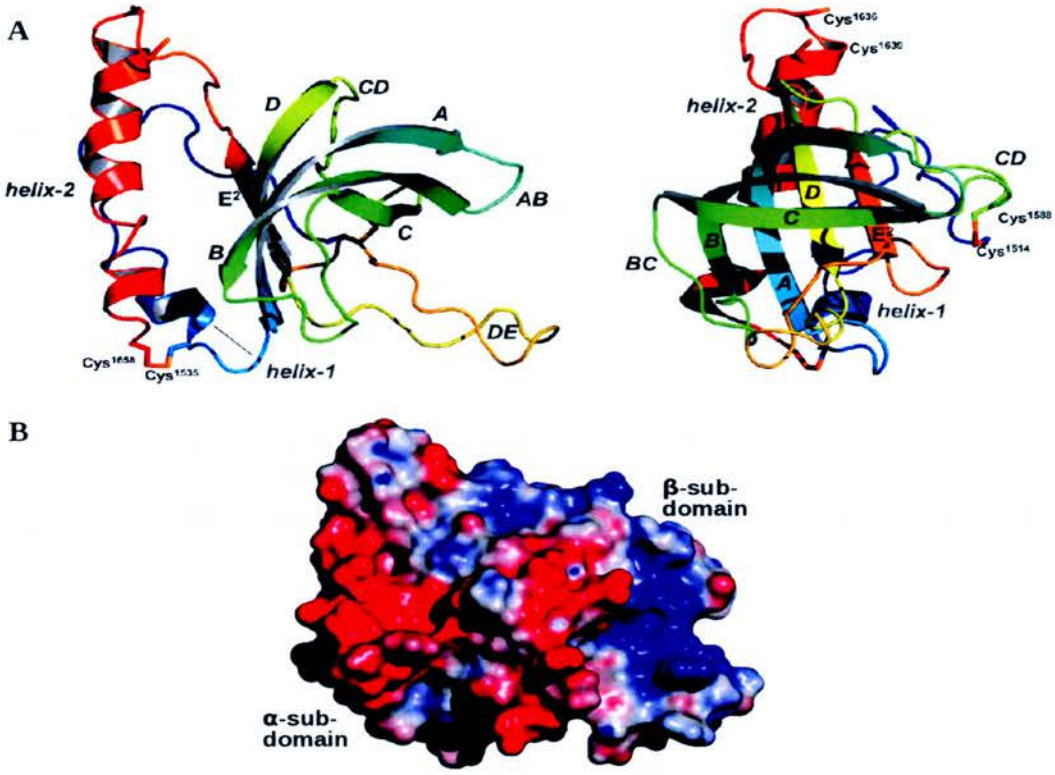


Fig. 11: C5-C345C structure. (A) Two orthogonal views of the structure of C5-C345C in a ribbon representation. Strands and loops of the β-subdomain and helices of the α-subdomain are annotated. Color scheme progresses sequentially from blue (N-terminus) to red (C-terminus). (B) Electrostatic surface representation, red is negative charge and blue is positive charge. A range of -4/+4kT is used. Adapted from paper.[54]

The NMR solution structure of C7-FIMs[53] reveals two type-1 follistatin domains (FD) that have an intimate interface in a homodimer-like, pseudosymmetrical arrangement (Fig. 12). Each FIM is further subdivided into two subdomains. An N-terminal FOLN subdomain, which is a disulphide bonded section that includes a β-hairpin, connected *via* a 3_{10} helix to a C-terminal KAZAL domain, which consists of an α-helix packed against a triple stranded antiparallel β-sheet. The N-terminal domains constitute the 'body' of the homodimer with a 60° angle between the 'wings' formed by the KAZAL domains. While the linker region Ala^{767}-Ala^{772} between the domains appears to be flexible, the two domains are immobile with respect to one another on an NMR time-scale (ps-ns and ms timescales). The intimate interface between the modules is supported by both electrostatic and hydrophobic interactions (Fig.12), including module-bridging hydrogen bonds and three salt bridges (Arg^{704}–Glu^{800}, Lys^{716}–Glu^{800}, and Asp^{726}–Arg^{824}), as well as aromatic-to-cysteine and aromatic-to-aromatic interactions. The C7-FIMs structure highlights a region in FIM1, worthy of mutagenesis studies; on the FIM1 face, opposite to that of the interface with FIM2, an electropositive patch is interrupted by hydrophobic residues and could provide a complementary binding surface for the previously mentioned electronegative and hydrophobic patches on C5-C345C (Fig. 12). The FIMs lie at the C terminus of C7, which is consistent with them having at least one solvent-exposed face. Due to the length of the linker between the preceding CCP domain of C7 and FIMs, it is neither possible to predict which face of the FIMs this is likely to be nor which face will presumably be less accessible due to the bulk of the parent protein.

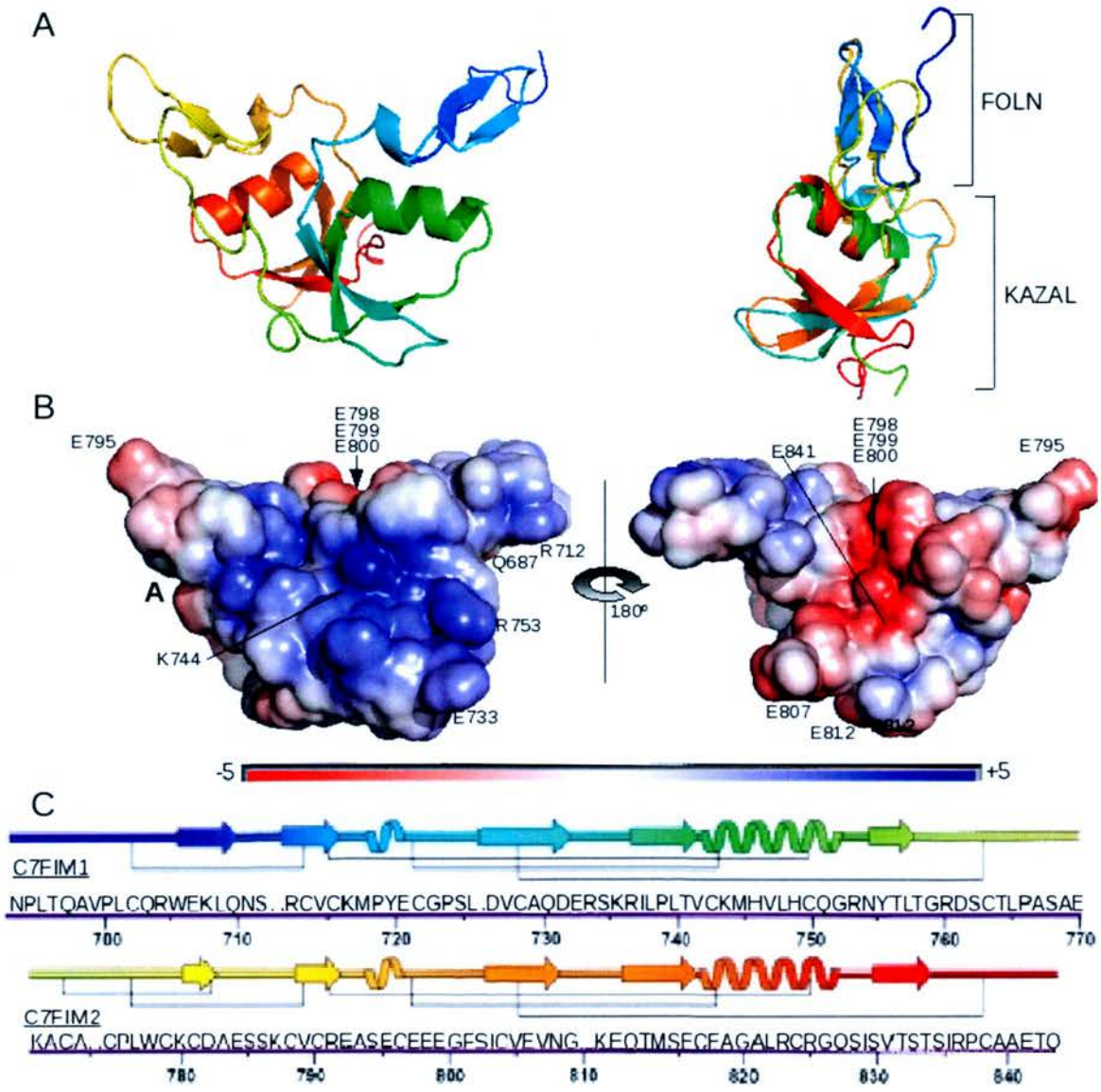


Fig. 12: C7-FIMs structure. (A) Ribbon representation of C7-FIMs NMR structure on the left (PDB ID: 2WCY) with putative with FIM1 and FIM2 overlayed on the right. (B) Surface electrostatic representations. (C) Cartoon showing the secondary structure and sequence of C7FIM1 and C7FIM2. Adapted from paper.[53]

A second binding mode is postulated based on comparison of C7-FIMS with other proteins that contain multiple FDs. The close association of the two C7-FIM modules is

in marked contrast with the structures of other FD-containing structures all of which have been solved in complex with their respective ligands.[91, 92] These adopt an open, elongated arrangement, displaying large tilt angles between FDs and a lack of symmetry. It has been proposed that FD pairs undergo a closed to-open transition upon ligand binding. Hence, by extension, it is conceivable that the C7-FIMs (and C6-FIMs) open up during MAC self-assembly, with the enthalpic cost of disruption of the intermodular interface compensated for by other protein-protein interactions within the complex.[53] Although it is currently postulated that C7/C6-FIMs are in the closed conformation in protein monomers, it is also possible that FIMs-MACPF/module interactions within the protein itself maintain the FIM pair in an open configuration.

1.3.5 CCPs

CCP modules are found extensively throughout the proteins of the complement system, and particularly within the RCA family. The common splice variants of the RCA family each contain between 4 and 30 tandemly arranged CCP modules. CCP modules are approximately 60 amino acid residues in length and are characterised by a consensus sequence that includes four invariant cysteines (disulfide linked Cys-I–Cys-III and Cys-II–Cys-IV), an almost invariant tryptophan and highly conserved prolines, glycines and hydrophobic residues.[94] While there are no experimentally derived structures for the two CCP modules of C6 or of C7, there are now over 40 structures of CCP modules, derived by NMR or X-ray crystallography, within the Protein DataBase (PDB).[93] These structures revealed a characteristic fold; a compact hydrophobic core wrapped in a β-sheet framework of up to eight β-strands, held together by the two strictly conserved disulphide bridges; there is generally a so-called "hypervariable loop" between strands B and C that is a region of poor sequence conservation and varied length, and often a β-bulge (again of variable composition) between strands E and F.[91] Despite sharing a broadly similar structure, CCP modules within complement proteins are functionally

non-equivalent, and indeed carry out a wide range of different functions. These include protein-protein recognition, protein-carbohydrate recognition, and spacing or positioning roles. This presumably reflects the versatility of a structural scaffold that has been adapted by evolution to suit a variety of purposes. The abundance of CCP structures has prompted homology-based, large-scale modeling of the CCP module family,[94] an exercise that highlighted the importance of the very diverse surface electrostatic surface properties displayed by even closely related CCP modules in their various functions.

The sequences of the C7-CCPs and C6-CCPs are fairly typical of CCP modules permitting the production of homology-based models that have been deposited in the CCP module database[94] (Fig. 13, models available for C7-CCP1, C7-CCP2 and C6-CCP1). While the modeled modules display the characteristic CCP module structure, analysis of surface electrostatics indicates that the surface of C7-CCP1 is highly electronegative. On the other hand, the C7-CCP2 model has an extensive electropositive surface patch that dominates one face together with its N-terminal end; its C-terminal end features an electronegative surface region. C6-CCP1 on the other hand is largely neutral but has a prevalence of positive charge and negative charge at its N and C-terminus respectively. The differences between the C7-CCP modules are reflected in their respective pI's of 4.8 and 7.9.[53] A similar difference is observed for C6-CCPs with respective pI's of 4.4 and 7.6.

In light of the model structures for both C7-CCP modules, an electrostatically favourable close association between these CCP modules is concievable, however, with a four-residue long linker between the modules bending will be resisted (see below) and any such association is likely to be sterically restricted. Moreover, the negative patch at the C-terminal end of C7-CCP2 could form an electrostatic interaction with the positive patch on FIM1.[53] The potential for extensive interactions between the CCPs within C7 and the FIMs raises the possibility that the C7 FIMs (and indeed the C7-CCPs) could in

fact adopt a different – presumably more open - orientation (with respect to one another) when they are in the context of the longer protein. These putative surface charges could also play a direct role in intra/inter-protein interactions. Or they may have indirect but vital architectural roles in putative module rearrangements that accompany formation of the preactivation complex(es) of C5 with C6 and/or C7, or in formation of the successive post-activation complexes.

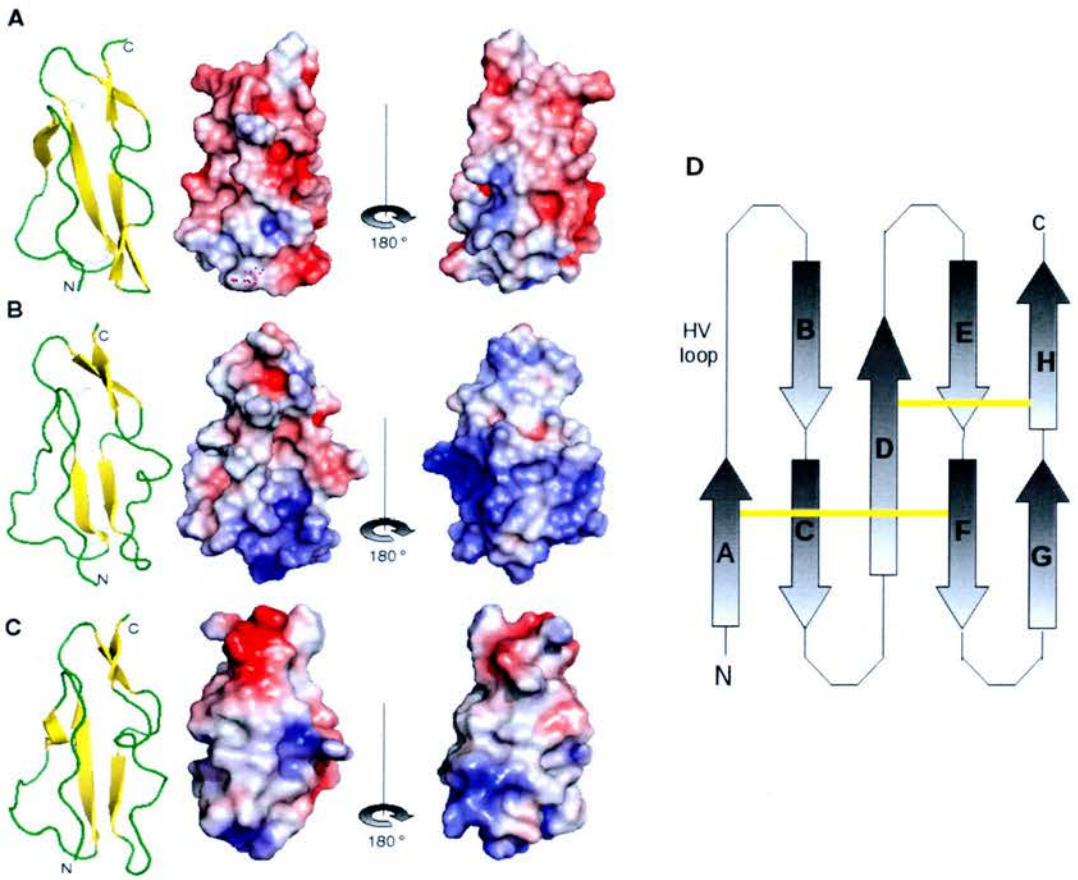


Fig. 13: CCP model structures of the MAC. (A) Model structure of C7-CCP1 showing secondary structure in standard view (left) and surface electrostatic properties displayed in standard view (middle) and with 180° view of the "back" of the module. (B) The same views as in A but for C7-CCP2. (C) The same views as in A but for C6-CCP2. Models are from the CCP module model database.[94] (D) Representative schematic of CCP module.

The relatively short linking sequence of four residues ($Q^{627}KIA^{630}$) between CCP1 and CCP2 suggests a degree of intermodular rigidity on the grounds that the hydrophobic portion of the linker will likely be buried; this would reduce the length of the potentially flexible part of the linker. Similarly, surface-exposed hydrophobic residues at the C-terminal end of CCP1 and the N-terminal end of CCP2 suggest that the two domains could be intimately associated with a fixed intermodular angle. On the other hand, values of intermodular tilt and twist between CCPs appear to vary in a way that can not easily be predicted from the sequence or even from knowledge of the structures of the contributing modules.[93]

Further downstream, there is 13-aa linking sequence between the final cysteine residue of CCP2 and first cysteine of FIM1, *i.e.* $V^{689}QKENPLTQAVPK^{701}$; the length and predominantly polar nature of these residues hints at a high degree of solvent accessibility and flexibility. However, these residues could also be sandwiched between CCP2 and FIM1, possibly shielding surface-exposed hydrophobic residues of FIMs and thereby contributing to a more opened up conformation of the two C-terminal FIMs (see previous section).

Thus an experimentally derived three-dimensional structure of the C-terminus of C7, consisting of both CCPs and FIMs, is needed to understand how these modules fit and work together. Structural insights may shed light on the putative swinging arms of C6 and C7 and their potential role as safety catches on their MACPF domains, ensuring they do not become activated inappropriately. The current work thus set out to test various predictions of this hypothesis.

1.3.6 C6, C7 and early MAC formation

Although the structure of the C7-FIM pair has been solved, little is known so far regarding the overall architecture of the protein monomers and even less is understood about the C5b-6 and C5b-7 complexes, with most knowledge arising from early electron-micrograph studies.[95, 96] For both C6 and C7 molecules, two structural regions were identified: a larger globular domain and smaller filamentous appendage (Fig. 14). Taking molecular dimensions of C8α-MACPF into account (each lobe is 60-70Å long) the globular regions are likely dominated by the MACPF (Fig. 14), with the appendage accredited to a molecular arm containing the C-terminal modules. In the EM images, the molecular arm of C6 folds back on itself, making contact with the globular region. Whilst for C7 the molecular arm was fully extended, away from the globular region. Although these conformations may hold true for the molecules in a physiological setting, the negative staining methods used dehydrate the molecule and can result in exposure of hydrophobic regions and molecular distortions. The extremity of the extension of C7's molecular arm may be a bi-product of the staining method used. However, the C7 EM images show a highly pliant structure and it was suggested that its molecular arm unfurls in binding C5b-6; with the EM images representative of the arm-open conformation. Conversely, the N-terminal modules may constitute the molecular arm depicted in EM images.

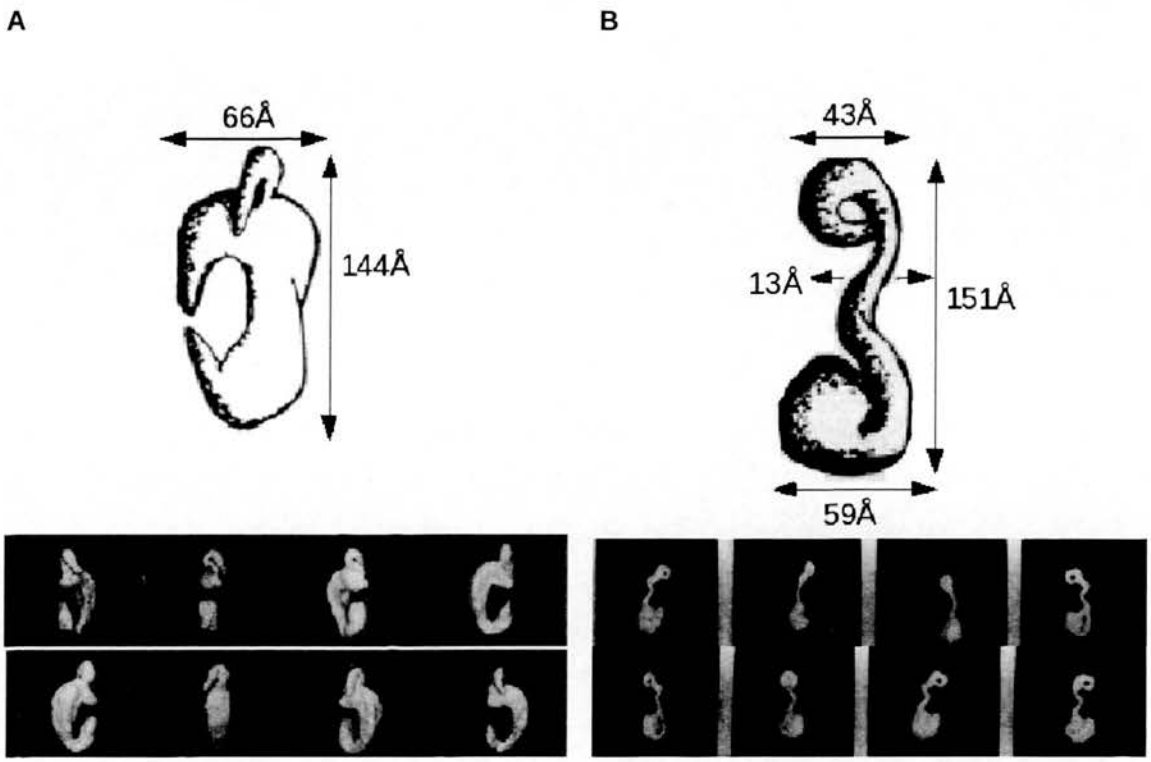


Fig. 14: EM Images. (A) Representative image of C6 (B) Representative EM image of C7. Adapted from paper.[96]

While the perforin structure has implicated a close association between the EGF and the helical clusters within the MACPF,[77] the relative location and flexibility of the other modules with respect to the MACPF domain (and one another) requires further investigation. Whether these modules harbour binding sites that confer specificity for other MAC proteins (as in C8β), or if they only stabilise the primary MACPF interactions, is yet to be fully explored. A potential role for C6-TSPC in binding C5 has been suggested.[97] Moreover, the mechanisms by which MAC proteins reveal or cooperatively create their protein binding sites are currently unknown. The predominant theory is the'modular fusion hypothesis.[96] The hypothesis is that the modules of each

subunit bind to corresponding modules of the other MAC subunits sequentially, exhibiting cooperativity. As a result, the inter-subunit interactions cause the proteins to unfold, exposing binding sites for protein adjuncts and/or hydrophobic sidechains for membrane attachment. The observation that MAC complexes aggregate in the absence of membranes – and in the case of C5b-7, form rosettes – supports the exposure of hydrophobic side-chains.[95] However, no studies have provided evidence as to the mechanism by which these side-chains are exposed.

The CDC/Perforin mechanisms of oligomerisation and pore formation used to illustrate the later stages of MAC formation, in combination with the molecular fusion hypothesis, can be further extrapolated to the formation of the early MAC (Fig. 15). Following activation of C5 and movement of the alpha-chain exposing a metastable binding site for C6, C6 binds facilitated by the C5-C345C domain were possibly provided by a molecular arm. The C7-FIMs displaces the C6-FIMs:C5-C345C interaction resulting in a fixed positioning of the C7 molecular arm, which may be accompanied by edge to edge stacking between C6 and C7 MACPF domains as in CDCs/Perforin, possibly in addition to modular interactions. Nevertheless, formation of the C5b-7 complex results in an amphiphilic to hydrophobic transition for membrane insertion.

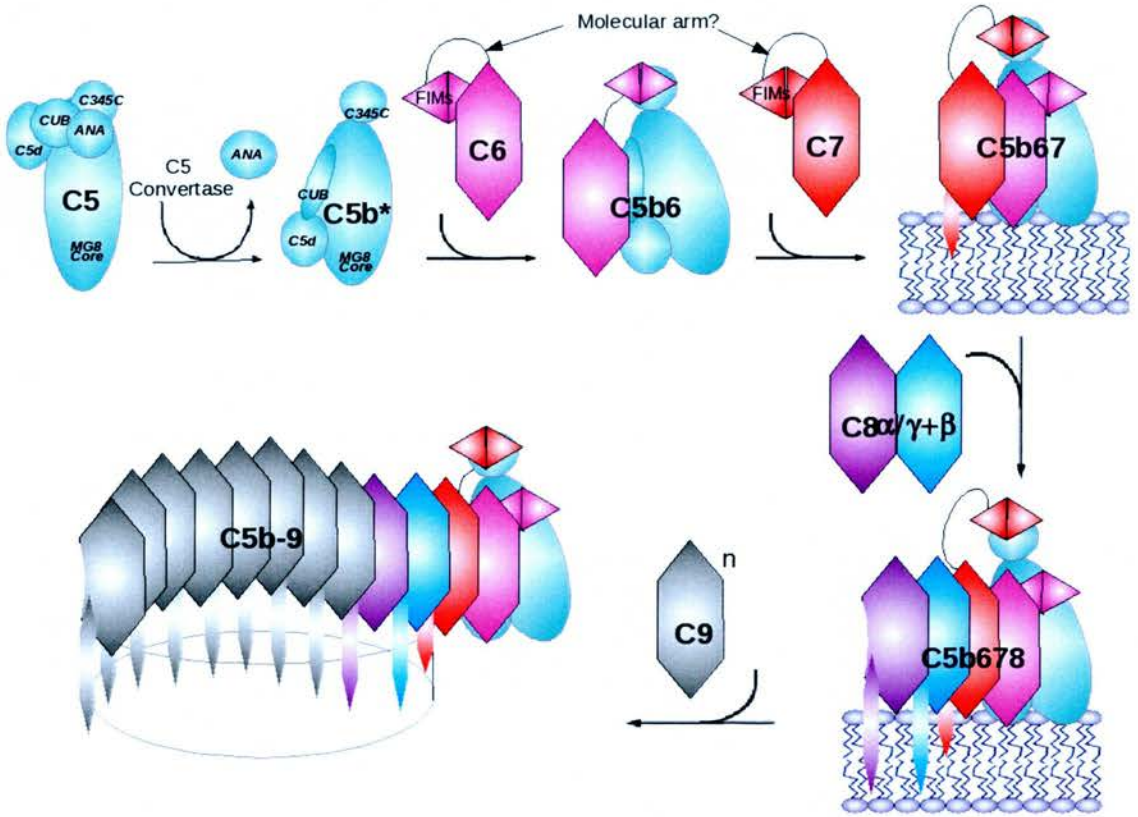


Fig. 15: MAC formation model: C5/C5b is shown in teal, C6 in magenta, C7 in red, C8α-γ in pruple, C8β in blue and C9 in grey. Domains of C5 (structure known) are labelled as previously described. Briefly, the image shows the activation of C5 to the metastable C5b* and binds C6 including a C5-C345C:C6-FIMs interaction. This is later displaced by C7-FIMs, by C7 incorporation into the complex and subsequent membrane binding. This allows for the addition of C8 and multiple C9 molecules to form the MAC pore.

1.4. Chemical cross-linking

1.4.1 Analysis of 3D structures and protein-protein interactions

The use of chemical cross-linkers in combination with mass spectrometry (XLMS) to infer structural information about protein conformations and protein-protein interactions was introduced at the end of the 1990s.[98-105] Cross-linking reagents can act as molecular rulers for the estimation of spatial relationships in protein structure-function studies. There are two principal applications of chemical cross-linking of proteins. It can be used in 3D structural analysis to estimate the spatial relationships within a protein by

evaluation of intra-molecular cross-links. This can give an indication of the proximity of domains to one another, which can be applied to conformational changes. Alternatively, it can be applied to protein-protein interaction analysis to estimate the spatial relationship between two (or more) binding partners by evaluation of inter-molecular cross-links.[99, 101, 105] This can identify potential binding sites and highlight domains or protein regions worthy of further analysis. Furthermore cross-linking can provide spatial information on flexible regions of proteins or transient conformations, which are commonly absent from X-ray crystallography structures, and allows structural analysis of proteins in solution of higher-order multiple-protein complexes that are not amenable to NMR studies.

A typical cross linking study is a multistep process: the cross linking reaction, the optional purification of a single cross-linked species (monomer, dimer etc.), the cleavage of the cross-linked proteins into peptides by enzymatic digestion, the optional separation of the peptides, and the mass spectrometry based detection, analysis, and identification of the cross-linked peptides. In recently developed innovative strategies, cross-linking reactions are conducted in living cells by directly incorporating reactive groups into the protein, using the cell's own biosynthetic machinery.[100]

Since proof of principle experiments conducted over a decade a go,[98, 99] the application of XLMS has expanded through developments in methodology, instrumentation, and bioinformatics.[100-105] The Rappsilber group developed a method[99] that was used to analyse by cross-linking the largest complex to date by extending the 13-subunit, 530-kDa RNA polymerase II (Pol II) X-ray structure to a 15-subunit, 670-kDa complex of Pol II with the initiation factor TFIIF.[106] Following cross-linking, proteolysis, and LC-MS/MS of the resultant peptides, an algorithm automatically finds and validates cross-linked peptides using peptide MS/MS fragmentation spectra. The use of a 1:1 mixture of stable isotope-labelled and non-labelled cross-linkers results in the appearance of

doublets in the mass spectra of cross-linked peptides with 1:1 signal ratio, separated by a mass corresponding to the difference between the heavy and light isomers. This reduces the false positive rates of the process. A combination of a standard database search tool (Mascot)[107] and a purpose built cross-link database (XDB) containing all possible combinations of cross-linked peptides is used to generate the list of cross-linked peptides. Each cross-link identified is assigned a score. The confidence score is calculated by database searches conducted under conditions that yield only false results by searching an XBD with reversed sequences and searching the XBD using a false mass for the cross-linker. The development of this methodology has expanded the complexity of the protein complexes studied by cross-linking.

The PolI II complex X-ray structure and a model of its binding partner, the initiation factor TFIIF, were validated by cross-linking and analysis of the Poll II:TFIIF complex identified the sites of interaction between the proteins.[106] Using the same technique to understand the global fold of protein monomers in the absence of atomic resolution structures of the full-length protein thus seems achievable, given the provision of structures of its individual domains.[108, 109] Ultimately, for a single protein, distance constraints derived from cross-linking sites could lead to structure elucidation. However, with the sparsity of the identifiable cross-links produced for a single protein with current approaches, and the lack of software programs that calculate structure models solely from cross-linking restraints, this degree of structure elucidation is currently unattainable .[103]

Cross-linking, in conjunction with mass spectrometry is therefore a very promising tool to yield structural information on proteins and protein complexes that are difficult to address using standard structural methods. Given that MAC formation is driven by conformational changes and domain-domain interactions and that protein complexity

increases with the sequential addition of each monomer, the MAC would provide a good model system for cross-linking/MS studies. Cross-linking could provide valuable structural information regarding the MAC, as the large and predicted flexible nature of the MAC proteins makes standard structural analysis by X-ray crystallography and NMR highly challenging. However, the general application of cross-linking to produce a model of a protein in the absence of atomic structures is yet to be achieved and strategies and techniques are currently in development. Therefore atomic resolution structures are required to develop and confirm reliable techniques and the resultant data.

1.4.2 C3/C3b

$C3^{ref}$ has the same domain arrangements as the C5 structure[51, 52] As with the transition of C5 to C5b, conformational changes induced by proteolytic cleavage allow for binding to new partners and additionally expose the reactive thioester in the TED domain for binding to target surfaces. Crystal structures of both C3 and C3b are available,[412-414] which makes the transition from C3 to C3b an ideal candidate to test using the Rappsilber group's XLMS method. Moreover, there has been much controversy regarding the C3b structure and the orientation of its domains (Fig. 15). Whilst there is a general consensus that the MG core is a structurally stable platform, relativity unchanged between C3 and C3b, the location of the TED and CUB domains has been the subject of dispute. In C3 the TED domain is closely associated with the CUB and MG8, which shields the thioester. In two C3b structures (Gros *et al*, PDB: 2IO7;[111] Wiesmann *et al*, PDB:2ICF)[112] the activated thioester is fully exposed for covalent attachment to target surfaces and is more than 85 Å away from the buried site in native C3, with the TED contacting MG1. This has come about through a reorientation of the CUB domain that has retained all of its secondary structure. A third, nominally higher resolution, C3b structure (Murthy *et al*, PDB:2HRO)[113] reveals a marked loss of secondary structure in the critical CUB domain and a different position and orientation of the TED . It

subsequently emerged that closer scrutiny of the PDB entry for 2HR0 reveals physically implausible features that are claimed to undermine the validity of the deviant C3b mode.[113] Only when the experimental diffraction images are made available can the model be either verified or falsified.[114] Nonethess, the arguments against 2HR0 and the underlying diffraction data were rebutted,[115] functional implications of the structure have been supported and paper continues to receive citation. The controversy regarding the structure of C3b and the orientation of its domains, could be resolved with the aid of XLMS, as this would reveal the true location of the TED in solution and could also shed light on the compactness of the CUB domain. Moreover, if the cross-linking technique can be successfully applied to C3 and C3b, it implies that a similar approach will yield useful structural information for components of the MAC.

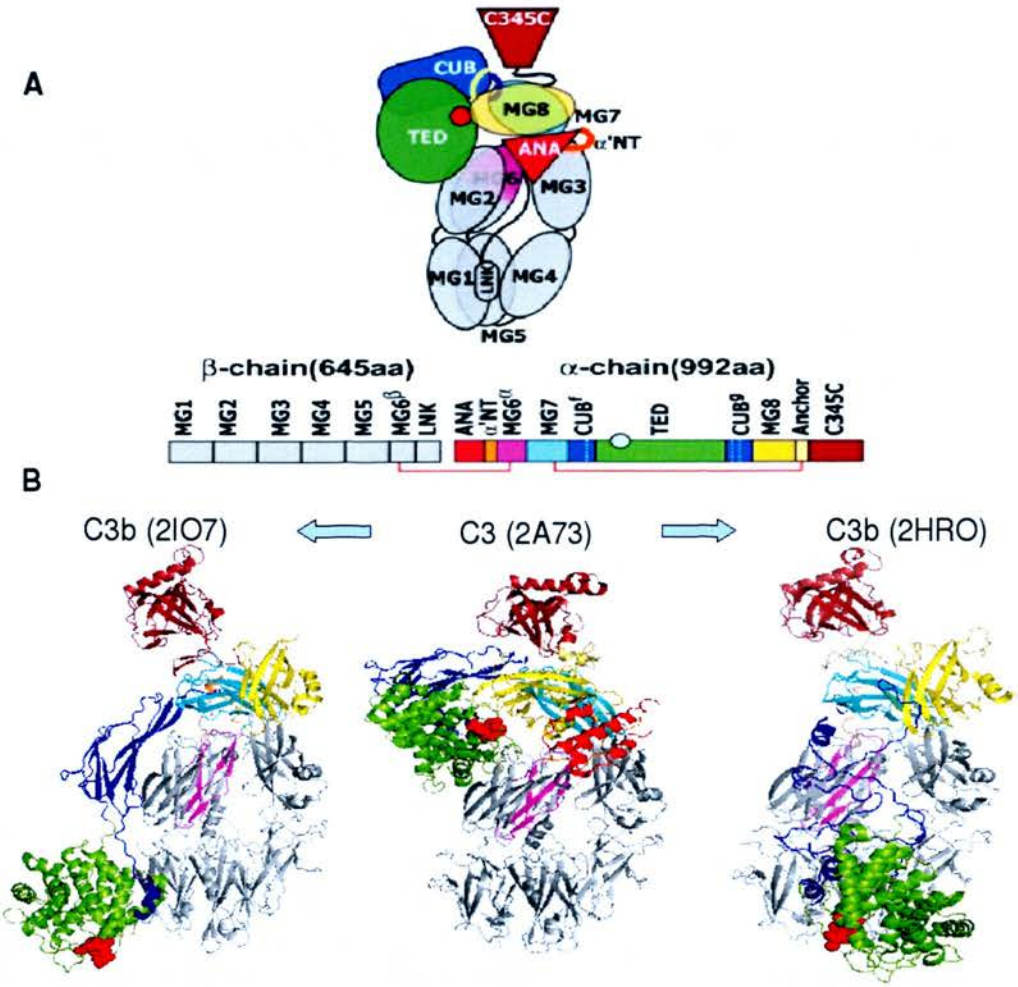


Fig. 16: C3 structure and structural transitions. (A) A schematic representation of C3 domains, sequence and organization.[ref] (B) Cartoon representation of crystal structures of native C3 and the two proposed structures for C3b. Diagrams were generated in Pymol.[34] The thioester of the TED domain is shown space-filled in red. Abbreviations and colour code: MG, macroglobulin domain (1-6 grey, 7 cyan, 8 yellow); LNK, linker domain (grey); ANA, anaphylatoxin domain (red); CUB domain (blue); anchor (pale yellow); C345C domain (burnt red). Adapted from paper.[110]

1.4.4 C7 Architecture

The application of XLMS to an isolated MAC protein, for which the structures of individual domains have been established, should help to establish the arrangement of domains and thus reveal the architecture of the protein. In the case of the MAC we envisage a mechanism driven by domain re-arrangements (as is the case in the C3 to C3

transition) and thus structural information at this level could be highly valuable. While features of the domain arrangements in C5b (from the C5, C3 and C3b structures), in C8 (from the C8 EM reconstruction) and in C9 and poly-C9 (EM and photo-labelling) have emerged, the domain arrangements within C6 and C7 remain unexplored.

In the specific case of C7, XLMS might give an indication as to the locations of the N-terminal and C-terminal modules with respect to the central MACPF. In particular, XLMS might shed light on how these presumably autoregulatory modules are organized relative to the helical clusters (HCs) of the MACPF, which are hypothesized to undergo conformational rearrangements forming transmembrane hairpins. If HC1 and HC2 are predominantly accessible in C7 then the parallels with perforin and the CDCs are obvious and extrapolations with regard to mechanisms of activation are valid. If on the other hand the HCs of the MACPF are shielded by the other modules in C7, this would support our "safety catch" hypothesis (in which the C-terminal modules maintain C7 in an inactive form and they must be released in order for C7 to undergo its activating conformational transition). Moreover, cross-linking studies could potentially reveal the conformation of the FIMs - are they open or closed – in addition to their location in the context of the whole molecule.

The XLMS technique may capture flexible molecule in a conformation or range of conformations adopted within the time-scale of the cross-linking reactions. Therefore it is valuable to combine these data with other data, such as NMR-based protein dynamics, studies to build up an understanding of the flexibility between modules, and particularly the nature of the C-terminal modules and their relation to the MACPF. Taken together these findings may confirm or refute a CDC/Perforin membrane-binding mode and current hypotheses regarding the swinging arm of C7. They should expand the structural understanding of the mechanisms involved in MAC assembly. They would also represent

an early example of the use of XLMS to define the tertiary structure of a protein for which no high-resolution structure is available.

1.5 Project Aims

The aim of this PhD project is to employ various biophysical techniques including NMR, and mass spectrometry in conjunction with cross-linking to study the structure and function of the terminal components of the complement pathway. Applying this highly structural approach to C7 of the MAC will provide a more detailed account of the early steps in MAC assembly.

In order to study the domains within the C5b-7 proteins—particularly C5-C345C and the CCPs and FIMs of C6 and C7, recombinant expression from a system capable of yielding at least milligram amounts of highly purified, soluble, fully folded protein is required for further characterization and binding studies. Protein fragment studied include C6-CCPs, C7-CCPs, C7-CCP2-FIM1-FIM2 (CFF), C7-CCP1-CCP2-FIM1 (CFF), C7-CCP2-FIMs (CF), C7-FIMs, C5-C345C and C3-C345C.

With the completion of the C7-FIMs solution structure, a NMR derived structure of the pair of CCP modules preceding the FIMs will add significantly to the limited information on the MAC proteins. Moreover NMR studies of a CFF module triplet will be used to link the structures of the FIMs and CCPs, giving a more complete picture of the C7 C-terminal arm. Combining insights from NMR studies with intra-molecular cross-linking data will provide information regarding the domain arrangements within C7 and will be discussed in terms of the implications for MAC self-assembly.

METHODS

2.1 DNA manipulation, recloning and expression vectors

2.1.1 Estimation of DNA concentrations

A UV spectrometer (Eppendorf BioSpectrometer, Eppendorf, Hamburg, Germany) was used to estimate DNA concentration as nucleic acids have an absorption maximum at 260 nm. The degree of sample purity was evaluated by calculating the $A_{[260]}/A_{[280]}$ ratio to assess protein contamination and the $A_{[260]}/A_{[230]}$ ratios to assess organic compounds. Values of 1.8 for $A_{[260]}/A_{[280]}$ and 2.0-2.2 $A_{[260]}/A_{[230]}$ signify pure DNA samples.

2.1.2 Agarose gel electrophoresis

1%w/v agarose gels were made by heating agarose in Tris-acetate-EDTA (TAE) until dissolved, followed by addition of ethidium bromide (50 ng/ml). The solution was then poured into molds to set. DNA samples were combined with 6x loading buffer (bromophenol blue (0.25%w/v), xylene xyanol, sucrose (40%v/v)) and loaded onto the gel. The gel was run at constant voltage (200 V) for approximately 30 minutes.

2.1.3 pET15b *Escherichia coli* expression vector

The pET15b vector (Invitrogen, for vector map see Appendix A), is a 5.7 kb bacterial expression vector which possesses a 6xHis-tag at the N-terminus, a subsequent thrombin cleavage site for removal of the tag and multiple cloning sites for target gene insertion. Ampicillin resistance is conferred *via* expression of the gene from the TEM I promoter, while the strong bacteriophage T7 promoter is used for high-level expression of the gene of interest via T7 polymerase provided by the host cell. 5' of the promoter is a *lac* operator for selective expression in the presence of IPTG to prevent repression by the *lac* repressor encoded by the *lacI* gene.

2.1.4 Transformation of *E.coli*

2.1.4.1 Preparation of electrocompetent *E.coli*

A vial of Origami™ B(DE3)pLysS Competent Cells (expression strain, Invitrogen) or

One Shot® TOP10 cells (storage, Invitrogen) were thawed on ice for inoculation of a 5 ml yeast-tryptone (YT, Appendix B) starter culture (overnight/37°C/200 rpm). This was used to inoculate a larger 500 ml YT culture (~2 hours/ 37°C/200 rpm) until an OD_{600} value of 0.3-0.4 was reached. All subsequent steps were performed on ice and using ice-cold media that was either autoclaved or sterile filtered (0.2 µm MWCO) to maintain cell health and prevent contamination. The cells were washed with 500 ml ice-cold autoclaved H_2O and twice with 50 ml 10%v/v glycerol by resuspension following centrifugation (10 min/1000 xg/4°C). Lastly, the cells were resuspended in 1 ml of GYT medium (Appendix B) and diluted if necessary to achieve the desired concentration of 2.5 x 10^{11} cells/ml (100x dilution OD_{600}= ~3.75). 40 µl aliquots (~1x10^{10} cells) were flash frozen and stored at -80°C.

2.1.4.2 Plasmid DNA extraction

E.coli colonies containing the pET15b constructs were grown in 5-10 ml of YT (overnight/37°C/200 rpm) containing carbenicillin (50 µg/ml) for selection of the vector. Carbenicillin was substituted for ampicillin due to its longer half-life. The plasmid DNA was extracted using a *QIAGEN® Plasmid Mini Prep kit* as specified by the manufacturer, yielding 2-10 µg of DNA per construct.

2.1.4.3 E.coli transormation by electroporation

10-50 ng of purified pET15b construct DNA was combined with 40 µl of freshly thawed electrocompetent *E.coli* cells for transformation by electroporation. The mixture was transferred to an electroporation cuvette and incubated on ice for 5 minutes. The cells were then pulsed with a charging voltage of 2500 V, capacitance 25 µF and resistance 200 mΩ, 1 ml of room temperature Super Optimal Broth (SOB) with subsidiary glucose (SOC) was added immediately and then transferred to a 15 ml Falcon™ tube for shaking incubation (1hr/37°C/200 rpm) to allow expression of the antibiotic resistance gene. 10-

250 µl aliquots were spread on YT agar plates with the carbenicillin (50 µg/ml) for vector selection and incubated at 37°C overnight. For Origami™ B cells kanamycin(15 µg/ml), tetracycline (12.5 µg/ml) and chloramphenicol (34 µg/ml) were additionaly included for selection of the *gor* mutation *trxB* mutation and pLysS plasmid respectively (See 2.2.5.1 Overview of Origami™ B pLysS host strain)

2.1.5 Reclone into *eukaryotic* expression vectors

2.1.5.1 pGAPZαB and pPICZαB Pichia pastoris expression vectors

pGAPZαB and pPICZαB vectors (Invitrogen, see Appendix A) are 2.9 kb and 3.6 kb respectively and are used for recombinant protein expression by *P.pastoris.* These vectors are used if the protein is normally secreted, glycosylated or directed to intracellular organelles. This is ideal for all target proteins due to the presence of disulphide bonds which result in the packaging of protein into insoluble inclusion bodies when expressed in *E.coli.* While pGAPZαB is designed for high-level, constitutive expression via the glyceraldehyde-3-phosphate dehydrogenase (GAP) promoter, the pPICZαB vector is designed for tightly regulated, methanol-inducible expression for the gene of interest via the AOX I promoter. While higher yields are expected for expression from pGAPZαB, the inducible expression provided by pPICZαB is ideal for NMR labelling strategies. Both vectors also contain the native *S.cerevisiae* α-factor secretion signal that constitutes efficient secretion of most proteins by *P.pastoris.*[116] The vectors also contain a TEF I and EM7 promoters to drive expression of the *Sh ble* gene in *P.pastoris* and *E.coli* respectively. Furthermore the vectors contain a pUC origin which permits replication and maintenance of the plasmid in *E.coli*. Lastly, they contain SacI, PmeI and BstXI restriction sites for vector linearization for efficient integration into the *P. pastoris* genome.

2.1.5.2 Cloning strategy

To reclone the C6 CCP, C7 CCP, C7 FIMs, C7 FIM(1), C7 FIM(2), C7 CFF, C7 CCF, C7

CF, C5 C345C and C3 C345C target genes from the pET15b vector, a standard strategy was employed (Fig. 17) whereby the PCR amplified target gene is sub-cloned into the high copy TOPO vector to obtain high yields of the target gene for insertion into pGAPZαB. A shorter cloning strategy was attempted for recloning into pPICZαB, for direct ligation of the target-gene PCR products without TOPO subcloning (Fig. 17).

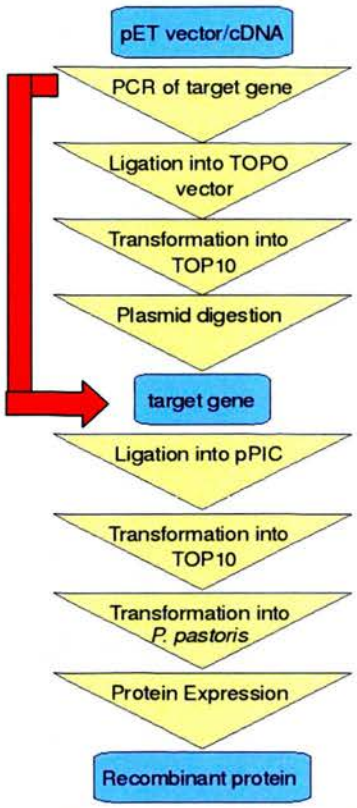


Fig. 17: Cloning Strategy Flowchart. The recloning stategy for pGAPzαB clones is depicted. Technical steps are shown in yellow and progress landmarks shown in blue. The red arrow highlights the steps omitted in the pPICzαB cloning strategy.

2.1.5.3 Amplification of target genes

The polymerase chain reaction (PCR) was used to amplify the C6-CCP, C7-CCP, C7-FIMs, C7-FIM(1), C7-FIM(2), C7-CFF, C7-CCF, C7-CF, C5-C345C and C3-C345C coding sequences from the pET15b vector using complementary primers (Table. 3). 1 µl of forward and reverse primers (10 µM) were mixed with 200-250 ng of template DNA,

1 ul triphosphates (dNTPS, 2.5 mM, Invitrogen) and 0.75 µl of DMSO was used to prevent primer dimerisation. 1 U of Pfu polymerase (Fermentas®) was used for amplification for insertion into pGAP to produce blunt-end PCR products for TOPO®-subcloning; while 1U of Taq polymerase (Invitrogen) was used for amplification for insertion into pPICZαB as Taq has independent terminal transferase activity which results in the addition of 1> adenosine nucleotides at the 3' end of the extension product. This extends the short 3-4 nucleotide overhang from the 3' restriction enzyme site ensuring successful digestion for direct insertion into pPIC vector. The final reaction volume was made up to 20 µl with ddH_2O prior to PCR using the program in Table. 4.

Oligo. Name	DNA Sequence	Restriction site
F C3-C345C	GTATCT**CTCGAG**AAAAGAGAGGCTGAAGCTGCTGAGGAGAATTGC	XhoI
R C3-C345C	CTAG **TCTAGA**CTATCAGTTGGGGCACCCAAACACAACAT	XbaI
F C5-C345C	GTATCT**CTCGAG**AAAAGAGAGGCTGAAGCTGCTGATTGTGGG CA	XhoI
R C5-C345C	CTAG **TCTAGA**CTATTAGCATCCATTTAAAAAGATATCTTC	XbaI
F C6-CCP	GAAGC **TGCA**GCTTCCGGGTGTCCTCAG	PstI
R C6-CCP	CTAG**TCTAGA**CTATTTTTCACAGGTGAGAGAGTTTG	XbaI
F C7-CCP	GAAGC**TGCA**GAATTCTGTCCATCACCTCCT	PstI
R C7-CCP	CTAG**TCTAGA**CTATTGTACACAGCGGGCATTCT	XbaI
F C7-CCP1	CTAG**TCTAGA**CTATTTGAGACTCCACCCAGTTTG	XhoI
R C7-CCP1	CTAG**TCTAGA**CTATAGCTCACACACGTTTTG	PstI
F C7-FIM1	GTATCT**CTCGAG**AAAAGAGAGGCTGAAGCTAAAGAAAATCCGT	XhoI
R C7-FIM1	CTAG**TCTAGA** CTATGAGGCAGGCAGAGTACAGC	XbaI
F C7-FIM2	CTATCT**CTCGAG**AAAAGAGAGGCTGAAGCT GCTGAGAAAGCTTG	XhoI

Table. 3: Primer oligonucleotides. The oligonucleotide naming system indicate whether the primer is for forward (F) or reverse (R) priming directions. Restriction enzyme sites are shown in bold.

Step	Temperature (°C)	Time (min)	Repetitions
Initialisation	95	1	1
Denaturation	95	0.5	14
Touch-down annealing steps	T_m	0.5	
Elongation	72	1.0/kb DNA	
Denaturation	95	0.5	19 to 24x
Annealing	T_m-7	0.5	
Elongation	72	1.0/kb DNA	
Final Elongation	72	5	/

Table. 4: PCR cycling program. Annealing temperatures were determined using the melting temperatures of the primers as indicated by manufacturers. In cases where the forward and reverse primers melting temperatures did not match, the lower melting temperature was used.

2.1.5.4 TOPO cloning reaction

For cloning into pGAP the blunt-end PCR products were sub-cloned into the plasmid vector, pCR®4Blunt-TOPO® using the Zero Blunt®TOPO® cloning kit (Invitrogen). Following protocol 2 µl of PCR product, 1 µl of salt solution and 1 µl of pCR®4Blunt-TOPO® were mixed and incubated at room temperature for 10 mins. The reaction was then placed on ice and subsequently used for transformation into One Shot®Top10 chemically competent cells.

2.1.5.5 Transformation of plasmids into E.coli chemically competent cells

Approximately 5 ng of the ligation mixture was incubated with 50 µl of One Shot® TOP10 Competent Cells for 30 mins on ice, heat shocked for 30sec in a pre-warmed 42°C water bath followed by the immediate addition of 250 µl of room temperature SOC and incubation on ice for 5 mins. The cell culture was incubated for 1 hour at 37°C, 200 rpm, and 50 µl of cells were plated on low salt LB agar plates containing zeocin™ (100 µg/ml) and grown overnight at 37°C.

2.1.5.6 Plasmid DNA extraction

TOP10 colonies containing the TOPO vectors with inserts and the pPIC and pGAP

vectors were grown overnight at 30°C in 250 ml of low salt Lysogeny Broth (LB)[47] containing kanamycin (TOPO, 50 µg/ml) and Zeocin™ (pPIC/pGAP, 100 µg/ml) for selection. The plasmid DNA was extracted using a *QIAGEN Plasmid Maxi Prep kit* as specified by the manufacturer, yielding 100-500 µg of DNA per construct.

2.1.5.7 Restriction enzyme digestion

Double digestion of the TOPO® constructs (pGAP) and taq-PCR products (pPIC) and the corresponding digestion of the empty pPIC/pGAP vectors was performed overnight at 37°C and the enzymes inactivated by incubation at 62°C. The restriction enzymes used for each construct are shown in Table. 3. DNA gel electrophoresis was employed to ensure successful digestion and to separate the digested fragments. The insert and vector bands were excised accordingly for solubilization and DNA extraction using the QIAquick® Gel Extraction kit (QIAGEN).

2.1.5.8 Ligation reactions

A 3-fold molar excess of insert was used to ligate into 50 ng of the pPIC/pGAP vector. The T4 Ligase Quick Ligation Kit (New England Biolabs) was used whereby insert and vector are mixed with 1 µl of 10x Quick Ligation Buffer (NEB) and 1 µl of T4 DNA ligase (400 U/µl, NEB) to a final volume of 20 µl. The reaction mixture was mixed thoroughly and then incubated at room temperature for 5 mins, followed by transformation into TOP10 chemically competent E.coli (see 2.1.4.3 *E.coli* transformation by electroporation).

2.1.5.9 Screening of E.coli colonies

TOP10 colonies potentially containing pPIC/pGAP and insert were screened via PCR. Small samples of colony cells were dissolved in 20 µl of ddH_2O and heated at 95°C for 3 mins to lyse the bacterial cells and solubilise the vector DNA. 2 µl of this was mixed

with 1.5 ul each of 10 μM forward and reverse primers (Table. 3) and 5 μl of PCR Master Mix® (Promega), which contains Taq DNA polymerase (50 U/ml), dNTPs (400 μM each) and $MgCl_2$ (3 mM). The PCR was then carried out using the program outlined in Table. 5 and the presence of the insert established by visualisation of the correct sized band following agarose gel electrophoresis.

Step	Temperature (°C)	Time (min)	Repetitions
Initialisation	95	1	1
Denaturation	95	0.5	
Annealing	50	0.5	
Elongation	60	1.0/kb DNA	30
Hold	4	hold	/

Table. 5: Mastermix screening PCR program

2.1.5.10 Sequencing of plasmid DNA

To confirm correct insertion of the genes of interest into the *P. pastoris* expression vectors, the DNA was sequenced using the chain termination method. α-factor primers (Addgene) directed against the α-factor signal sequence adjacent to the inserts were added to 200-300 ng of plasmid DNA along with 4 μl of ABI prism BigDye® terminator mix (Applied Biosystems, Foster City, CA). The BigDye® contains a mixture of all four standard deoxynucleotides (dATP, dGTP, dCTP and dTTP), all four fluorescently labelled dideoxynucleotides (ddATP, ddGTP, ddCTP and ccTTP) and AmpliTaq DNA Polymerase. Amplification of the target sequence using BigDye® and the PCR program outlined in Table. 6 results in a mixture of fragment lengths due to random incorporation of dideoxynucleotides which terminates chain elongation due to the absense of a 3'-hydroxyl group. The mixture was then separated by polyacrylamide gel electrophoresis, where a pattern of fluorescent bands corresponding to the terminating ddNTPs for each fragment can be read off using the ABI 3730 instrument carried out by the School of Biological Sciences Sequencing Service (SBSSS), Edinburgh.

Step	Temperature (°C)	Time (min)	Repetitions
Initialisation	95	0.5	1
Denaturation	96	0.5	
Annealing	50	4	
Elongation	60	1.0/kb DNA	24
Hold	4	hold	/

Table. 6: PCR program for DNA sequencing

2.1.6 Transformation of *P.pastoris*

2.1.6.1 Preparation of DNA

As above the TOP10™ colonies containing the pPIC/pGAP constructs were grown in low salt LB and had DNA extracted using a QIAGEN Plasmid Maxi Prep kit. The purified DNA was linearized using the restriction enzyme SacI for incorporation into *P. pastoris*. Complete linearization was assessed by agarose gel electrophoresis (see 3.5.1)

and when complete,the enzyme was inactivated by incubating at 62°C for 20 mins. To purify DNA from the DNA-enzyme mixture, samples were subject to phenol-chloroform extraction; whereby an equal volume of phenol: chloroform: isoamyl alcohol (25:24:1, saturated with 10 mM Tris, pH 8.0, 1 mM EDTA) was added to the DNA sample followed by vortexing and centrifugation (1 min/10,000 rpm, 4°C). The resultant bi-phasic mixture contains the DNA in the aqueous phase, which was removed and the above steps repeated until no protein was visible in the organic phase. Finally an equal volume of chloroform was added to the DNA sample and the centrifugation steps repeated to remove any residual phenol from the sample.

Ethanol precipitation was then used to concentrate the DNA. Firstly the salt concentration was adjusted to 0.3 M, using 3 M sodium acetate, pH 5.2, providing Na^+ ions. The solution is mixed with 2.5 volumes of ice-cold ethanol (100%), vortexed and placed on ice for 1 hour. Adding ethanol to the solution disrupts screening of charges by water and the Na^+ ions and the DNA's phosphate backbone form stable ionic bonds,

precipitating the DNA out of solution. The precipitate is then collected by centrifugation for 30 mins/13,000 rpm/4°C and removal of the supernatant. To remove residual salts from the DNA 750 µl of 70%v/v ethanol was added and then centrifuged for 5 mins/13,000 rpm/4°C. The resultant DNA pellet was air-dried and resuspended in the desired volume of ddH_2O.

2.1.6.2 Preparation electrocompetent P. pastoris

A 5 ml Yeast Peptone Dextrose (YPD, Appendix B) overnight culture of *P. pastoris* (KM71H strain) was used to inoculate 500 ml of YPD medium and grown at 30°C until an OD_{600} of 1.3-1.5 was achieved. The cells were centrifuged (5 mins/1500 xg/4°C) and resuspended in 500 ml of ice-cold ddH_2O. Centrifugation was repeated and the cells resuspended in 250 ml of ice-cold ddH_2O. After another round of centrifugation the cells were resuspended in 20 ml of ice-cold 1 M sorbitol, which was reduced to 1 ml following a final round of centrifugation.

2.1.6.3 P. pastoris transformation by electroporation

5-10 µg of purified and linearized plasmid DNA was combined with 80 µl of freshly prepared electrocompetent cells for transformation by electroporation. This was transferred to an electroporation cuvette and incubated on ice for 5 mins. The cells were then pulsed for 5 seconds with a charging voltage of 1500 V, capacitance 25 µF and resistance 200 mΩ. 1 ml of ice-cold 1 M sorbitol was added immediately and then transferred to a 15 ml Falcon™ tube for incubation without shaking for 3 hours at 30°C. 150 µl and 300 µl aliquots were spread on 100 µg/ml and 300 µg/ml YPDS, Zeocin™ plates and incubated at 30°C for 3 days.

2.2 Protein production, manipulation and purification

2.2.1 Estimation of protein concentrations

A UV spectrometer (Eppendorf BioSpectrometer, Eppendorf, Hamburg, Germany) was

used to estimate protein concentrations. UV absorbance was detected at 280 nm due to the primary absorbance by tryptophan, tyrosine and cysteine with their molar absorption coefficients decreasing in that order. The ratios $A_{[280]}/A_{[260]}$, and $A_{[280]}/A_{[320]}$ were used to assess the purity of the sample. The protein concentrations were calculated according to the Beer-Lambert law equation: $A=\varepsilon lc$, where absorbance, A; is equal to the product of the extinction coefficient, ε (calculated by ExPASy, ProtPram tool); the path-length of light, l; and the concentration, c.

2.2.2 Concentration of protein samples

Buffer-exchange for NMR samples, and all the concentration steps for the variously labelled protein samples were performed in 0.5 ml, 6.0 ml or 20 ml Vivaspin™ concentrators (Sartorius Mechatronics UK Ltd, Epsom,United Kingdom), with an appropriate molecular weight cut-off membrane (3000 - 10000 Da).

2.2.3 Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SDS-PAGE analysis was employed to detect the presence of recombinant proteins and protein contaminants. Samples were mixed with reducing and non-reducing Laemmli loading buffer (Invitrogen, 4x concentrated) then boiled, followed by centrifugation if necessary, and run on a NuPAGE 4-12% Bis(2-hydroxyethyl)-imino-] tris(hydroxymethyl)-methane (bis-tris) gel at a constant voltage (200 V) for approximately 30 minutes in the NuPAGE gel system (Invitrogen). Gels were then washed in water, stained for protein using Coomassie Brilliant Blue R-250 Staining Solution (BioRad) and destained in water. For cross-linking gels a specialized gradient gel, NuPAGE 3-8% tris-acetate gel, was used for separation of high molecular weight protein bands.

2.2.4 Chromatography sytems

All chromtography steps were implemented using the ÄKTA-design™ FPLC system

(pump P-920, UV detector unit UPC-900) bar reverse-phase HPLC which was performed on an ÄKTAexplorer™ or manual sample application and elution where stated.

2.2.5 Mass Spectrometry

Spectra were collected on an Applied Biosystems DE-STR MALDI-TOF (Matrix Assisted Laser Desorption Ionisation - Time of Flight) with a nitrogen laser. Sinapinic acid was used as matrix and prepared by sonicating and vortexing 10 mg sinapinic acid in 50%v/v MeOH/H_2O mixture with 0.03%v/v trifluoroacetic acid (TFA). 0.5 µl protein sample was mixed with 0.5 µl matrix directly on the target plate. The lowest laser intensity that gave a suitable peak was used to minimize protein fragmentation. The instrument was calibrated with the external standards Cytochrome C and Horse Myoglobin (HHM). Liquid Chromatography (LC)-MS was submitted for analysis to the Scottish Instrumentation and Resource Centre for Advance Mass Spectrometry (SIRCAMS).

2.2.6 *E.coli* protein expression

2.2.6.1 Overview of Origami™ BpLysS host strain

Complement proteins and their associated domains characteristically contain multiple disulphide bonds. The natural reducing environment of *E.coli* cytoplasm however, does not support disulphide bonding and can result in protein misfolding and the formation of inactive aggregates known as inclusion bodies. Thus expression in a system that supports formation of disulphide bonds is essential. The Origami B strain of *E.coli* was utilized to obtain folded domains in the soluble phase. Origami strains benefit from mutations in both the thioredoxin reductase (*trxB*) and glutathione reductase (*gor*) genes, the combination of which greatly enhances disulphide bond formation in the cytoplasm.[117] Thioredoxin reductases are the only known enzymes to reduce thioredoxin, which in its

active thiol form reduces other proteins by cysteine thiol-disulfide exchange. Meanwhile, glutathione reductase reduces glutathione disulfide to the sulfyhdryl form, which serving as an electron donor result in reduction of disulphide bonds . Thus mutations in both the *trxB* and *gor* genes prevent recombinant protein reduction *via* the interuption of the thiorodoxin and glutaredoxin systems. The B form of the Origami strain are derived from a lacZY mutant of the conventional *E.coli* expression strain BL21(DE3) and similarly expresses T7 RNA polymerase upon IPTG induction for expression of the recombinant protein *via* the strong bacteriophage T7 promoter of the pET vector. Moreover, combining the Origami B strain with the low-level expression plasmid pLysS allows for precise control of expression levels. Whilst the Origami B strain has an additional lac permease (*lacY*) mutation allowing for uniform entry of IPTG into all cells in the population, producing a concentration-dependent, homogeneous level of induction. The pLysS plasmid expresses T7 lysozyme, which inhibits T7 RNA polymerase at the pET vectors lac promoter, until expression is induced with IPTG, when the polymerase is over-expressed.

2.2.6.2 C7-CCPs expression

C7-CCPs pET-15b vector DNA was transformed into pLysS/OrigamiB competent cells by electroporation and cells were spread on yeast tryptone (YT, Appendix B) agar plates containing ampicillin (75 μg/ml), kanamycin (15 μg/ml), tetracycline (15 μg/ml) and chloroamphenicol (15 μg/ml) for selection of the *bla*, *trxB* and *gor* mutations and the pLYS plasmid mutations respectively. Plates were incubated at 37°C overnight. Transformed cells were grown at 30°C in YT medium (natural C^{12} and N^{14} abundance) and M9 salts medium (Appendix B) supplemented with either ^{15}N ammonium sulphate and ^{15}N Isogro (^{15}N single labeling) or ^{15}N Ammonium sulphate, ^{13}C Glucose and ^{15}N, ^{13}C Isogro (^{13}C, ^{15}N double labeling). Expression was induced with 1 mM isopropyl β-D-1-thiogalactoside (IPTG) at an Optical Density ($O.D._{600}$) of 0.6 for overnight expression

at room temperature. The cells were harvested by centrifugation (15 min/6000 rpm/4°C), frozen, thawed and then lysed with BugBuster™ reagent supplemented with Benzonase Nuclease and a protease inhibitor cocktail as recommended by the supplier. The cell lysate was cleared by centrifugation (2 hrs/11000 rpm/4°C) and filtered using a 0.45 µm molecular weight cut-off (MWCO) Mini-start filter (Sartorius).

2.2.7 C7 CCPs purification

Lysate was applied to a 1 ml HiTrap chelating column (GE Healthcare) loaded with Co^{2+}/Ni^{2+} for coupling with the His-tag using phosphate wash buffer (20 mM KPO_4, 0.5 M NaCl, 10 mM imidazole) and elution buffer (20 mM KPO_4, 0.5 M NaCl, 500 mM Imidazole). Various pHs of buffer were tested to obtain the highest yield with fewer impurities. Eluted fractions were assessed by SDS-PAGE for the presence of C7-FIMs and pooled. C7-CCPs were thrombin cleaved to remove the His-tag using biotinylated thrombin (1 µl/mg protein) and 10X cleavage buffer (200 mM Tris-HCl pH 8.4, 1.5 M NaCl, 25 mM $CaCl_2$). Streptavadin-agarose was used, as recommended, to remove thrombin. To remove His-tag from solution and further purify samples, they were reapplied to the HiTrap column followed by application to either a HPLC column (Discovery® BIO Wide Pore C8-5 µm, 25 cm x 4.5 mm, Supelco) or gel filtration column (HiLoad™ 16/60 Superdex™ 75 prep grade, GE Healthcare).

2.2.8 *P. pastoris* protein expression

2.2.8.1 Overview of KM71H P. pastoris host strain**

P. pastoris is a eukaryotic yeast expression system and therefore has advantages of higher eukaryotic expression systems such as protein processing, protein folding, and post-translational modification.[118] They are easy to manipulate, have a relatively rapid growth rate and can be grown to higher cell densities than bacteria. It shares many of *Saccharomyces cerevisiae*'s molecular and genetic manipulations, whilst has the added

advantage of 10-100-fold higher heterologous protein expression levels.

As with *S.cerevisiae*, *P.pastoris* is a methylotrophic yeast, and is able to use methanol as a carbon source in the absence of glucose. In the peroxisome organelle, methanol is oxidized to formaldehyde and the harmful biproduct hydrogen peroxide using molecular oxygen and alcohol oxidase (AOX). The formaldehyde, partly leaves the peroxisome and is further oxidised to form carbon dioxide by cytoplasmic dehydrogenases, providing the organism with energy. The formaldehyde remaining within the peroxisome is used to produce cellular constituents. Whilst the hydrogen peroxide is rapidly consumed *via* the action of a catalase, producing water and the less reactive gaseous oxygen. Alcohol oxidase has a poor affinity for O2, and the cells compensate by generating large amounts of the enzyme. Consequently the AOX promoters concerned are exceptionally strong and are exploited to obtain high, inducible yields of recombinant protein.

The genome of *P. pastoris* encodes for two alcohol oxidases, AOX1 and AOX2. The two enzymes are 97% identical to each other and share approximately the same specificity.[119] However, methanol metabolism is mainly carried out through the catalytic action of the AOX1 gene product, which promotes a more rapid catabolism of methanol relative to AOX2. *P. pastoris* growing on methanol as a carbon source expresses AOX1 levels of more than 30% of total soluble protein. Regulation of the AOX1 gene involves two mechanisms. First, there is a repression/derepression mechanism, whereby in the presence of glucose the AOX1 gene is repressed. Absence of glucose on its own is not sufficient to stimulate the AOX1 expression as, secondly, the presence of methanol is required as an induction signal prior to AOX1 gene transcription.[119] There are two different methanol-metabolising *P.pastoris* phenotypes: MUT^S and MUT^+. The loss of its main methanol catabolic enzyme AOX1 results in *P. pastoris* MUT^S strains (methanol

utilization slow) that have to rely on their facility to catabolise methanol through the more slowly metabolising AOX2. Whilst MUT^+ refers to the wild-type phenotype of growth on methanol as a sole carbon source. The *P. pastoris* strain used throughout this study was KM71H (Invitrogen) of the MUT^S phenotype. Once the heterologous protein sequence has been transformed into the expression cassette and subsequently integrated in the genome, it is under the control of the strong and highly-inducible AOX1-promotor. Therefore recombinant protein expression can be regulated through the feeding of methanol to cell cultures.

Recombinant protein expressed in *P.pastoris* can be intracellular or secreted. Secretion requires a signal sequence that is provided by the pPICZαB vector in order to direct the protein to the secretion pathway. As *P. pastoris* only secretes very low levels of native protein into the medium, and is able to grow in minimal medium, secretion can be an important first step in purification of the heterologous protein.[119]
Regarding post-translational modifications benefits the *P. pastoris*-based expression system does not cause hyperglycosylation as with *S.cereviciae*. However O- and N-linked glycosylation does occur at the same sites as the native protein. Another asset of relevance to complement proteins is that native disulfide bonds patterns are generally formed in the secreted product. Usually this system expresses foreign proteins in high yields, although yields are generally not as high as for alcohol oxidase itself.[119] In conclusion the *P. pastoris* expression system is appropriate for generation of micro- to milligram quantities of pure recombinant protein with native disulfides and the option of isotopic enrichment for NMR studies.

2.2.8.2 Identification of expressing clones

5-10 colonies were tested for expression of each construct. pGAP expression trials were carried out in 50 ml sterile Falcons™, with single colonies used to inoculated 10 ml of

YT media, grown for two days and then protein expression induced with 0.5%v/v methanol for two days. Similarly for pPIC constructs, single colonies were used to inoculate 10 ml of buffered minimal glycerol (BMG, Appendix B) in a 50 ml sterile Falcon™ conical tube and incubated for two days in a shaking incubator at 30°C. However, this culture was then used to inoculate 90ml of BMG in a 250 ml baffled glass flask and again incubated for 2 days at 30°C. Protein expression was induced at the lower temperature of 15°C with the addition of 0.5%v/v methanol for two consecutive days. The flask cultures were then harvested (10 min/3000 rpm/SH-300 rotor/4°C) and resuspended in 25 ml of buffered minimal methanol (BMM, Appendix B) to induce protein expression. The cultures were incubated for a further 2 days with 0.5%v/v methanol added each day. Following expression the cells were then harvested by centrifugation (10 min/3000 rpm/SH-300 rotor/4°C) and the supernatant containing the secreted protein was filtered (Millipore, Millex sterile syringe filters, PVDF 0.22 μm, Millipore, Watford, UK). 500 μl of filtered supernatant was concentrated to 50 μl to be analysed for protein expression by SDS-PAGE (see section 2.2.3). For pGAP expressed protein, the supernatant samples were also buffer exchanged *via* the concentrator to remove the majority of the peptides present in the YT expression media.

2.2.8.3 CFF fermentor expression

The use of fermentors in recombinant protein production, allows for control and logging of pH, agitation, air and oxygen supply, thus maintaining cell constitution and generating higher protein yields than a shaker flask. In addition pronounced protein expression levels from the AOXI promoter can be achieved at growth-limiting rates acquired via fermentor expression, as opposed to excessive feeding in shakers.

Unlabelled CFF expression and ^{15}N labeled or ^{15}N and ^{13}C labeled expression for NMR structural studies was performed in 5 L and 2 L BioFlow 3000 vessels using 600 ml and

300 ml BMG shaker flask starter cultures (2days/30°C) respectively. The contents of the initial basal salts media (Appendix B), and the details of the feeding schedule, depend upon whether the protein product was to be unlabelled, ^{15}N-labeled, or ^{15}N, ^{13}C-labeled; however, for CFF double labeling conditions were used during the single labeled run in order to gauge double labeled expression levels.

To attain sterile growth conditions the fermentor vessel containing the media and all the probes were fully assembled prior to autoclaving the fermentor unit. The dissolved O2-probe was charged over night, with air bubbled continuously into the vessel through a sterile filter and agitation was set to 200 rpm in order to saturate the medium with oxygen. Prior to inoculation a relative oxygen level of 40 was maintained by varying the agitation rate (e.g. during yeast metabolism the aeration is held at a constant level but the agitation rate is increased to counteract oxygen level dropping below 40). pH and temperature probes, as well as a feeding line for a 2 M KOH solution (base feed) were also attached. The temperature and pH were set for maintenance to 30 °C and 5.0 respectively. 0.5 ml Antifoam 206 (Sigma-Aldrich) and 2.5 ml of high-purity grade fermentation trace mineral salts (PTM1 salt, Amresco®) were also added to the media (per 700 ml).

For natural abundance expression ammonium sulfate (7 g) and D-glucose (15 g) were added to the media before inoculation. This was substituted by ammonium-^{15}N-sulfate (Isotec™) for single labeling, together with D-Glucose-^{13}C (Isotec™) for double labeling. Cells were then pelleted (10 min/1500 xg/4°C) and resolubilised (20 ml 100 mM potassium phosphate, pH 6.0) for inoculation of the fermentor media. Cells were grown until nutrients were metabolised and agitation levels returned to baseline. Similarly an additional glucose (10 g) was added and metabolised, followed by a final feed of glycerol (1 g) and ammonium sulfate (1 g): all of which were isotopically

labelled for ^{15}N, ^{13}C protein expression. After ~1 day of growth the temperature was reduced to 15°C for initial induction with 0.5% (v/v of total culture volume) methanol (or Methanol-^{13}C for double-labeling). Further feeds of 0.75% to 1.5% methanol were provided over a course of 3-4days. The frequency of methanol feeds was determined on the basis of carefully monitoring the agitation-rate and dissolved oxygen curves so as to avoid overfeeding or poisoning. Every glycerol and methanol feed was accompanied by a 0.1 ml addition of PTM1-salts and all labels were added in saturated sterile filtered (0.2 µm MWCO) ddH_2O solutions.

In all cases, the supernatant was harvested by an initial centrifugation step (10min/5000 xg/4°C), followed by a more powerful spin (30min/10000 xg/4°C), before sterile filtration (0.2 µm). And, ahead of protein purification, PMSF and EDTA were added to inhibit protein degradation (final concentrations of 0.5 mM and 5 mM, respectively).

2.2.9 CFF Purification

For the initial purification step cation-exchange chromatography was employed. The pI of 6.4 calculated by ExPASY's ProtParam tool dictated that cation-exchange chromatography would be the initial purification step. Theoretical pH titration curves were calculated by the program Sedenterp, indicating that a pH of 4-5 would insure the protein is positively charged. An XK 26 column (Pharmacia, Biotech) manually packed with SP Sepharose™ (GE Healthcare) was used. The column was pre-equilibrated with 20 mM sodium acetate buffer (pH 4.0) and cell supernatant (pH 4.0, diluted 1:6 with water) was passed over the column using a peristaltic pump. Elution of the protein was then achieved by applying a 0-100% linear salt gradient using elution buffer (20 mM sodium acetate, 1 M NaCl, pH 4.0).

The N-linked glycosylation site within FIM 1 (Asn^{754}) was removed using the Endo H_f

enzyme (a recombinant fusion of endoglycosidase H and maltose binding protein, New England BioLabs®). Endo H_f cleaves within the chitobiose core (two N-acetylgulcosamine residues, Glc-NAc) of the N-linked oligosacharide chain, leaving the Endo H_f treated protein with one Glc-NAc from the core. 400-800 units of Endo H_f (10^6 U/ml) were used to deglycosylate 1 ug of protein (5 hours, 37°C) at the optimal pH for deglycosylation, 5.6.

Size exclusion chromatography was employed as the final purification step. A HiLoad™ 16/60 Superdex™ 75 prep grade column (GE Healthcare) pre-equilibrated with running buffer (20 mM sodium acetate, 100 mM NaCl, pH 4.0) was used. Following SDS-PAGE analysis selected fractions containing highly pure protein were pooled for further analysis.

2.2.10 CFF Dynamic Light Scattering

2.2.10.1 Introduction

Dynamic light scattering (DLS) is a spectroscopic technique that measures the scattering of monochramatic coherent laser light (typically 633 nm) from a molecule in solution. The light is scattered by the molecules at all angles due to Brownian motion. However, a DLS instrument only detects the scattered light at one angle (typically 90°) and therefore the intensity of the scattered light fluctuates with brownian motion and is measured as a function of time and described by an autocorrelation function. Autocorrelation is the cross-correlation of a signal with itself. Thus, it is the similarity in intensity of light scattered from the molecule as a function of the time separation between them. So for short differences in time there is a relatively high degree of correlation as the particle has not moved extensively, whilst for long differences in time there are greater changes to the molecules position and therefore a smaller correlation. In the simplest case this function is an exponential decay, and is a function of the diffusion coefficient of the

molecule in question. The diffusion coefficient is a factor of proportionality representing the amount of substance diffusing across a unit area through a unit concentration gradient in unit time and can be transformed to molecular size through the Stokes-Einstein equation (equation 1).

$$D = \frac{kT}{6\pi\eta R}$$

Equation 1: Stokes-Einstein equation. The diffusion coefficient, D, for a particle in a free volume depends on the Boltzmann constant (k), the absolute temperature (T), the viscosity of the solution (η), and the hydrodynamic radius (R) of the particle.

The size calculated is the hydrodynamic radius or Stokes radius. Specifically it is the radius of a hard sphere that diffuses at the same rate as the molecule. This radius depends not only on the size of the molecular "core", but also on surface structure i.e. the hydration shell of a protein and shape effects. Moreover scattering intensity is proportional to the square of molecular weight and therefore even small amounts of protein aggregates can be detected in the sample.

2.2.10.2 Instrumentation

To estimate the hydrodynamic size of the protein and to assess protein mono/polydispersity under various storage conditions (4°C, freeze-drying and -80°C) and various buffer conditions (20 mM sodium acetate, 0/100 mM NaCl, pH 4.0 and 20 mM potassium phosphate, 0/100 mM NaCl, pH 5.0) the Zetasizer Auto Plate Sampler Dynamic Light Scattering system (Malvern Instruments Ltd, England) was used at the Biophysical Characterisation facility at the Centre for Translational and Chemical Biology, Edinburgh. ~50 ul of sample was loaded into each well in a standard 384 well plate and samples returned to the well following averaged analysis. All samples were analysed at 25°C using a standard He-Ne laser (3 mW, 633 nm) and the scattered light detected at 90° to the incident light. All standard operating procedures and data

processing were implemented using Zetasizer APS software.

2.3 NMR Structural Studies

In this study no development of NMR method was developed. Henceforth an extensive overview of NMR theory is unnecessary,[120] however the basic principles of protein NMR are outlined in section 2.3.1. This is notwithstanding a list of the NMR experiments collected for structure determination (see section 2.3.3, Table.7) and key concepts relating to the nature of the experimental data. A guide to instrumental setup and experimental details of the suite of NMR experiments used can be found at the experimental pulse sequence repository and supporting webpage (http://nmr-linux.chem.ed.ac.uk/highfield/highfield.html) maintained by Dr Dušan Uhrín.

2.3.1 Introduction

When placed in a magnetic field, NMR active nuclei (e.g. ^{1}H, ^{15}N, ^{13}C) absorb energy at a resonant radio-frequency characteristic of the isotope, known as the excitation frequency. In a magnetic field, NMR active nuclei have a nuclear spin, I=1/2, which has an associated magnetic moment in a magnetic field, B_0, which generates two energy states separated by an amount of energy, ΔE, which is field dependent (Fig.18).

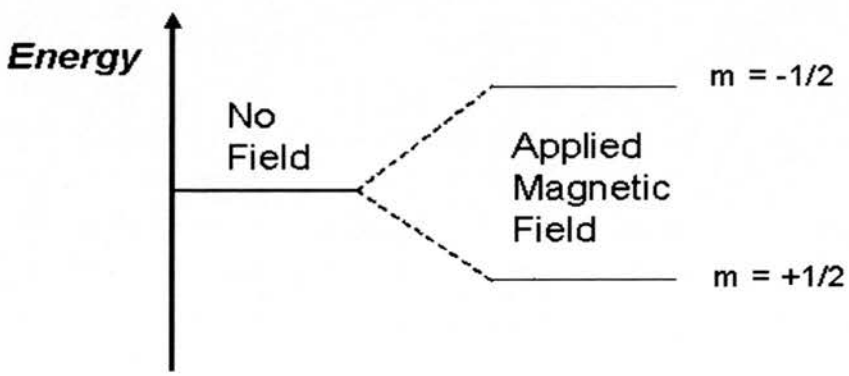


Fig. 18: Energy levels for a nucleus with spin quantum number ½.

The sum of the magnetization of the individual spins or bulk magnetization, M, will be spin aligned with B_0 (z-axis in the vector diagram Fig. 19).

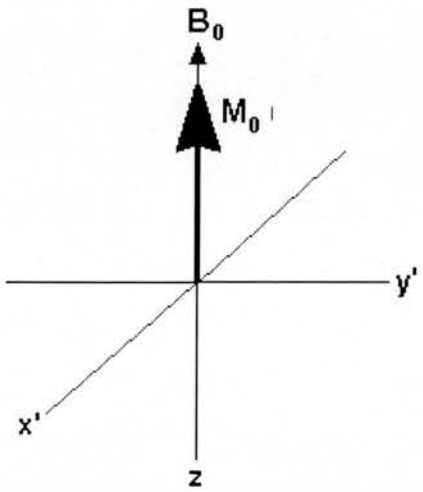


Fig. 19 Vector energy diagram. The bulk magnetization is shown spin-aligned with the external magnetic field B_0.

The force generated by B_0 on M is a torque which causes M to process about B_0, known as Larmour precession. Applying a magnetic radio-frequency pulse perpendicular to B_0 at the Larmour frequency tips M into precessing around the xy plane, which generates a measurable oscillating current in a receiver coil. When the excitation pulse is turned off, there are influences that cause the magnetization to decay and M returns to precession around the z-axis. The diminishing signal acquired by the receiver coil in the xy plane is called free-induction decay (FID). The FID which is collected and analyzed is the difference between this NMR signal or resonant frequency and the excitation frequency, or Larmour frequency as an expression of time. This complex time-domain signal, the FID, is related to the illustrative frequency domain spectrum by a mathematical process called Fourier transformation.

As the resonant frequency is influenced by the surrounding chemical environment,

chemically nonequivalent nuclei do not experience the same magnetic field due to magnetic shielding by electrons, and every NMR active nuclei in the sample will have its own characteristic resonant frequency. This frequency is proportional to the gyromagnetic ratio, γ, for a particular nucleus and the strength of the applied magnetic field and therefore in an NMR spectrum the frequency is expressed as a standardized value known as the chemical shift.

For the majority of proteins and protein domains, the vast number of protons creates a complex, overlapped 1D ^{1}H-spectrum. Therefore multidimensional NMR experiments are recorded using RF pulse sequences that exploit dipole-dipole coupling (J-coupling, Fig. 20) between covalently linked nuclei to employ multiple magnetization transfer steps which correlate the frequencies of distinct nuclei (^{1}H, ^{15}N, ^{13}C).[121] This resolves overlap by deconvolution in additional dimensions and thus increases the amount of accessible information. The size of C7-CCPs (14 kDa) is more than accessible for structure determination using a suite of standard heteronuclear NMR experiments facilitated by a ^{15}N, ^{13}C labelled recombinant protein sample. Whilst C7-CFF's size (24 kDa) inherently generates more complex spectra. Thus use of high strength magnets (see section 2.3.2) and careful processing of the FID (see section 2.3.4) are required to determine its atomic structure using the standard suite of experiments. The standard set of heteronuclear experiments for proteins of this size provide all the necessary information for assignment of the protein backbone and side-chains atoms to their associated chemical shifts as well as extraction of structural parameters for structure calculation.

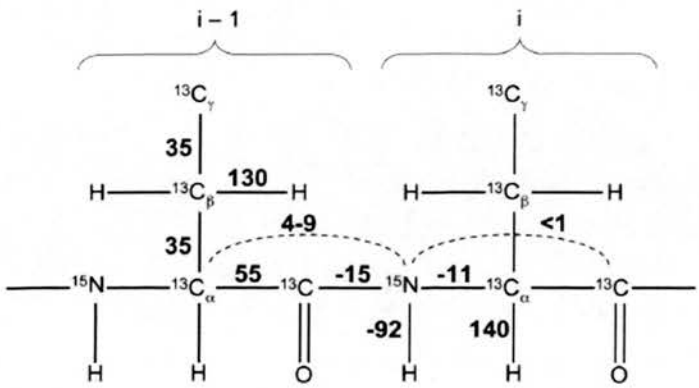


Fig.20: Through bond J-coupling constants. J-couplings (Hz) are shown between two sequential spin-systems denoted *i* and *i*-1. Adapted from paper.[121]

2.3.2 Spectrometers

A Bruker AVANCE™ 18.77 Tesla (nominal 800MHz ^{1}H frequency) spectrometer and a Bruker AVANCE™ 14.1 Tesla (nominal 600MHz ^{1}H frequency) spectrometer. Both spectrometers were fitted with a 5-mm triple-resonance cryoprobe to increase the signal/noise ratio and *z*-axis pulse field gradient coils. Topspin™ software (Bruker Bopspin 2005) was used for data acquisition and primary processing and analyses of NMR data.

2.3.3 Sample preparation, optimisation and data collection

Purified natural abundance recombinant protein samples were concentrated in spin concentrators and buffer-exchanged (20 mM potassium phosphate, 100 mM NaCl, pH 6.5) and the presence of tertiary structure identified using ^{1}H 1D NMR experiments. Concentrations for 1-D experiments ranged from 100 to 300 μM. Following this ^{15}N labelled sample conditions were tested to determine the best conditions for 3D NMR experiments by comparison of ^{15}N-HSQCs. Salt concentration (0, 50 and 100 mM), pH (4.0, 4.5, 5, 5.5, 6.5 and 7.0) and temperature (15, 20, 25, 30, 35 °C) were tested. All subsequent spectra for structure determination (Table. 7) were recorded in the optimal buffer with azide and ethylenediaminetetraacetic acid (EDTA) to maintain the sample

(20 mM potassium phosphate, 5 μM EDTA, 0.02%w/v NaN_3, pH 5.0) .

Experiment	Ref No	Experiment Type	Labelling	Dim 1	Dim2	Dim3
^{15}N-HSQC		Through-bond	^{15}N	^{1}H	^{15}N	n/a
^{13}C-HSQC		Through-bond	^{13}C	^{1}H	^{13}C	n/a
CBCA(CO)NH		Through-bond	^{15}N, ^{13}C	^{1}H	^{13}Cα/β	^{15}N
CBCANH		Through-bond	^{15}N, ^{13}C	^{1}H	^{13}Cα/β	^{15}N
HBHA(CO)NH		Through-bond	^{15}N, ^{13}C	^{1}H	^{1}Hα/β	^{15}N
HBHANH		Through-bond	^{15}N, ^{13}C	^{1}H	^{1}Hα/β	^{15}N
HNCO		Through-bond	^{15}N, ^{13}C	^{1}H	^{13}CO	^{15}N
HN(CA)CO		Through-bond	^{15}N, ^{13}C	^{1}H	^{13}CO	^{15}N
^{15}N-TOCSY		Through-bond	^{15}N, ^{13}C	^{1}H	^{1}H	^{15}N
HCCH-TOCSY		Through-bond	^{15}N, ^{13}C	^{1}H	^{1}H	^{13}C
(HB)CB(CGCD)HD		Through-bond	^{15}N, ^{13}C	^{1}H	$^{13}C_{arom}$	n/a
(HB)CB(CGCDCE)HE		Through-bond	^{15}N, ^{13}C	^{1}H	$^{13}C_{arom}$	n/a
Aromatic 13C-HSQC		Through-bond	^{15}N, ^{13}C	^{1}H	$^{13}C_{arom}$	n/a
^{15}N-HSQC-NOESY		Through-space	^{15}N	^{1}H	^{1}H	^{15}N
^{13}C-HSQC-NOESY		Through-space	^{15}N, ^{13}C	^{1}H	^{1}H	^{13}C
^{13}C-Methyl-HSQC-NOESY		Through-space	^{15}N, ^{13}C	^{1}H	^{1}H	^{13}C

Table. 7 NMR Experiments for structure determination used in this study.

2.3.4 NMR data processing

The process program from the AZARA suite of programs (Department of Biochemistry, University of Cambridge, UK) were used to process the raw free-induction decay (FID) data by Fourier transformation and by maximum entropy processing. The application of a variety of window functions to maximise resolution of the resulting frequency spectrum was included. Spectral processing with the process command requires a parameter file (ser.ref) and a script input file (scr) from the Bruker aquired data. An example of these files can be found in appendix C. Maximum entropy processing was only employed in the case of C7-CFF data to deconvolute merged spectral cross-peaks.

2.3.5 NMR Assignment

2.3.5.1 Assignment software

CcpNmr Analysis (collaborative computational project for the NMR community, CCPN) is a graphical NMR spectra analysis and assignment program. The alpha release was used for the CCPs assignment and the beta-release was used for the CFF assignment. This software was written as an extension of the CCPN data model[**] which represents all of the information used in NMR analysis and how they relate to one another (Fig. 21). The *Resonance* object in the CcpNmr Analysis data model is fundamental to the assignment process, linking all NMR information, including cross-peaks, chemical shifts and atomic assignments. It does not represent an atom or associated value, but represents the connection between an atom and the NMR parameters it gives rise to. This means that one *Resonance* can be connected to many chemical shift values associated with different experiments as for pH and temperature titrations. Also interrelated cross-peaks can be linked to a *Resonance* without knowing which atom or group of atoms it relates to as is done during the sequential backbone assignment, aliphatic and aromatic resonance assignment and NOE cross-peak assignment processes (see sections 2.3.5.3, 2.3.5.4, 2.3.5.5 and 2.3.5.2 respectively). The assignment of NMR data to a *Resonance* and the assignment of this *Resonance* to a particular atom are thus separate processes, which do not have to be performed simultaneously.

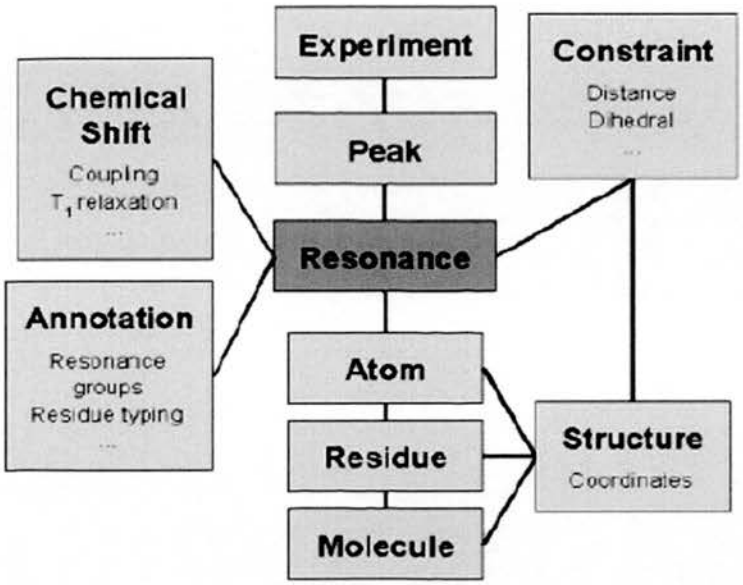


Fig. 21: Main concepts of the Data Model NMR package, and their relationships. Resonance is found in darker grey in the centre of the diagram, wher almost all relationships use the Resonance as an intermediary. Diagram from paper.[135]

2.3.5.2 General resonance assignment strategy

Assigning chemical shifts associated with cross-peaks to specific atoms in a protein generally requires a suite of three- dimensional (3D) experiments recorded on ^{13}C,^{15}N-labelled sample. The central reference spectrum for most 3D experiments is the 2D ^{15}N-HSQC, in which every cross-peak corresponds to an amide (i.e. (CO)-NH-) and thereby correlates an amide nitrogen frequency with its attached proton frequency, except proline which is devoid of an amide hydrogen. The various 3D experiments allow these correlations to be extended through-bonds to include other atoms. In this way, nuclei, as identified by their chemical shifts, may be assigned to clusters corresponding to covalently linked groups of atoms. For sequential backbone assignment three pairs of 3D experiments were used: CBCANH/CBCA(CO)NH, HBHANH/HBHA(CO)NH and HNCO/HN(CA)CO. The redundancy of information between experiment pairs allows atoms to be placed in sequential order and then matched to specific stretches of amino

acid residues within the polypeptide sequence due to the characteristic chemical shifts of some atoms.

Together with the HCCH-TOCSY experiment (see section 2.3.5.4), which yields cross-peaks for every aliphatic (attached to ^{13}C) side-chain ^{1}H, these experiments should in theory enable the assignment of all of the non-aromatic side-chain atoms. Aromatic atoms were identified by a set of 2D experiments that correlate aromatic Hδ and Hε shifts with Cβ shifts of the same side-chain (see section 2.3.5.5). All assignments are subsequently transferred to the ^{13}C- and ^{15}N-edited NOESY spectra to facilitate assignment of NOESY cross-peaks that arise from NOE transfers between non-covalently linked protons which are close in space (see section 2.3.6.1).

2.3.5.3 Sequential assignment of the backbone resonances

For the backbone experiments the first two dimensions are the amide proton (HN) and amide nitrogen (NH) frequencies and therefore the plane established by these two axes represents the ^{15}N-HSQC spectrum and act as a general point of reference in the identification of spin-systems. Whilst an extension into the 3rd dimension (H/C) results in strips of cross-peaks associated with that spin-system. For any given spin-system (one amide peak), one experiment in the pair has cross-peaks for that residue (*i*) and its previous (*i*-1) residue, while the other experiment has *i*-1 peaks only. Therefore overlaying the experiments with one another allows for the distinction between *i* and *i*-1 resonances. The *i*-1 shifts are then used to search for spin-systems with *i* peaks of the same frequency, namely the *i*-1 residue. Thus a series of connected strips, each corresponding to a unique amide cross-peak are established.

The 3-D CBCANH experiment connects the Cα and Cβ shifts of residues *i* and *i*-1 with the amide cross-peak of residue *i*. Whilst its partner, the CBCA(CO)NH experiment

connects the Cα and Cβ shifts of the *i*-1 residue with the amide cross-peak of residue *i*. Therefore spin-systems can be linked sequentially via their Cα and Cβ resonances. The CBCANH/CBCA(CO)NH pair is of particular significance as serine, threonine, glycine and alanine have characteristic Cα and Cβ chemical shift patterns. This serves as a basis for matching sequential spin systems with corresponding segments of the primary amino acid sequence. Serine and threonine Cα and Cβ shifts typically have their Cα chemical shifts up-field of their respective Cβ shifts unlike all other residues. Moreover the shifts for the Cα (53-63 ppm for serine and 56-68 ppm for threonine) and Cβ (60-68 ppm for serine and 66-74 ppm for threonine) are both relatively high making them easily distinguishable from other residues. Alanine residues Cβ shift is distincly down-field from the chemical shifts of other residues. Glycine residues on the other hand were identified from the lone Cα and its markedly high shift (42-48 ppm) in comparison to other residues' Cα shifts.

The 3D HBHANH/HBHA(CO)NH can be seen as an extension of the CBCANH/CBCA(CO)NH pair as they correlate the Hα and Hβ shifts instead of the C and Cβ shifts to the amide of residue *i*. Whilst the 3D HN(CA)CO/HNCO pair (*i*-1/*i* and *i*-1 experiments) are used to obtain the backbone carbonyl shift completing the backbone assignment. The higher sensitivity and less crowded spectra associated with the HN (CA)CO/HNCO experiments allowed for the resolution of any ambiguities from the sequential assignment with the other two experiment pairs. The magnetization pathways for the experiments are shown in Fig. 22.

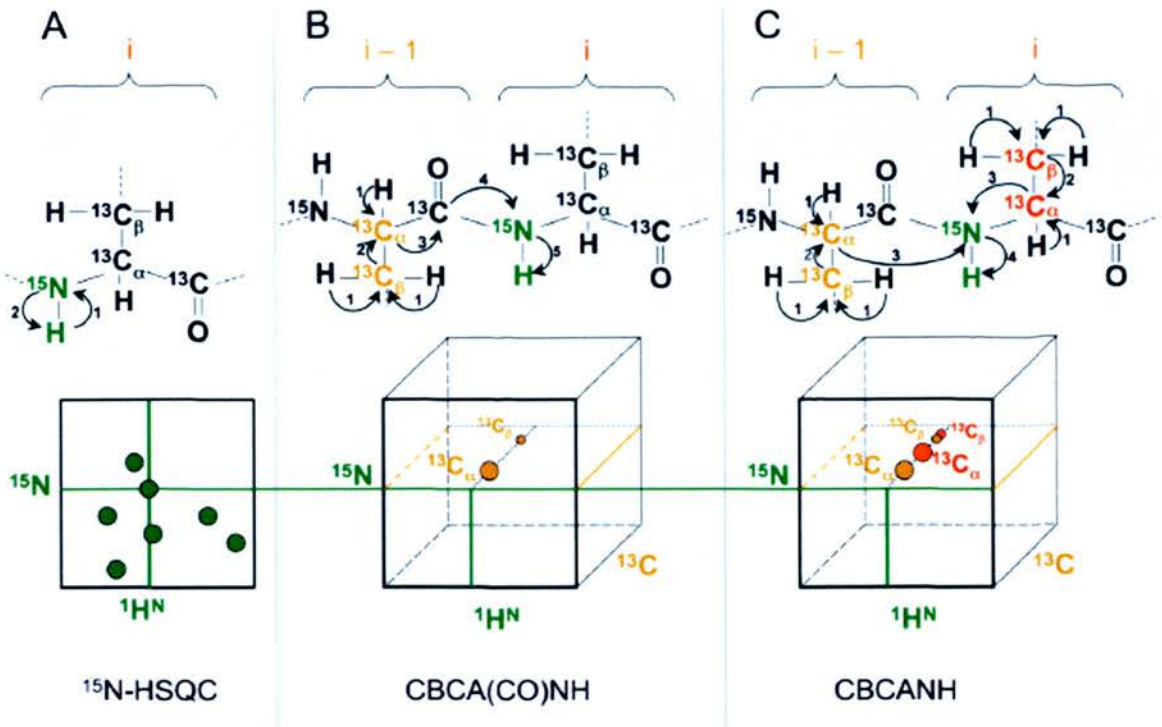


Fig. 22: Backbone assignment. (A) depicts the [1H, 15N]-HSQC, (B) the CBCA(CO)NH and (C) the CBCANH experiments. The top of each panel shows colours those atoms that are labelled during the evolution of the pulse programs and whose dimensions are detected in the spectra below. Arrows indicate the magnetisation transfer steps of the experiment. Single spectral strips are shown for (B) and (C) from the corresponding amide cross-peak.

2.3.5.4 Aliphatic side-chain assignment

Total Correlation Spectroscopy (TOCSY) experiments were employed for the assignment of the remaining ^{1}H and ^{13}C side-chain resonances. These experiments make use of the TOCSY through-bond coherence transfer to all coupled spins in a scalar-coupled network (e.g. all ^{13}C atoms in an amino acid sidechain), using isotropic mixing pulse.[130] The ^{15}N-TOCSY-HSQC correlates aliphatic side-chain proton shifts from the (i) residue with root resonances (HN and NH shifts) of the same residue (Fig. 23). This is useful for confirming and supplementing main TOCSY experiment's (^{13}C-HCCH-TOCSY) assignments and assigning side-chain NH_2 groups. Side-chain assignments

were transferred to the ^{13}C-HSQC spectrum which then served as a reference point for the assignment of the ^{13}C-HCCH-TOCSY spectrum. As the ^{13}C-HCCH-TOCSY experiment correlates each aliphatic proton of a spin-system (including prolines) to all other aliphatic protons from the same spin-system, this provides an extra level of resolution, which was ultimately critical in resolving any ambiguity in the highly overlapped methyl region.

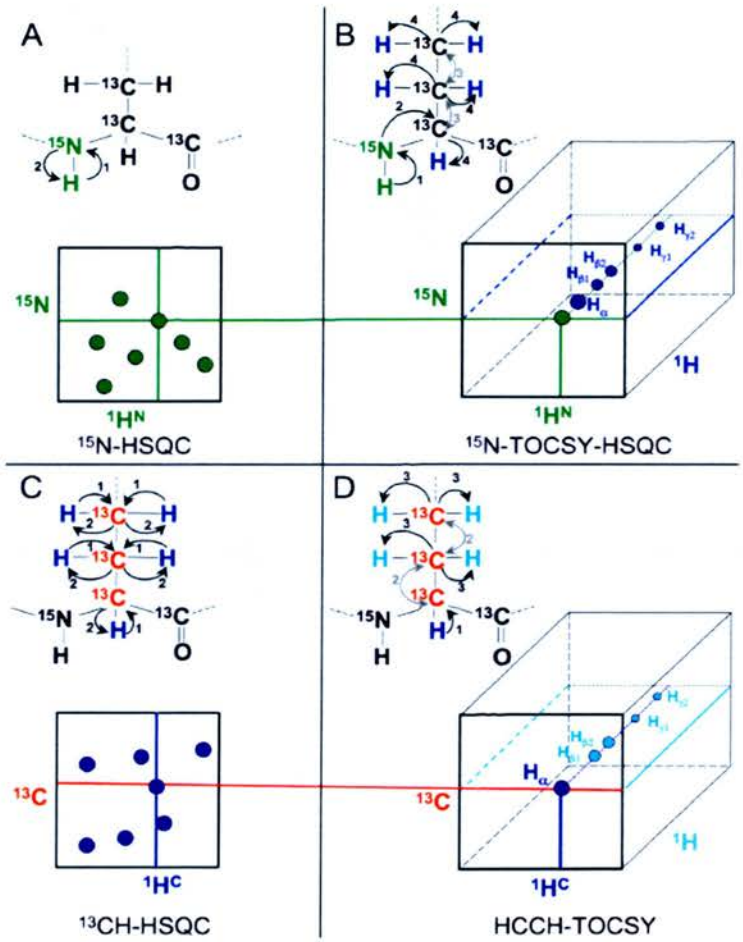


Fig. 23: Aliphatic proton assignment. (A) shows [^{1}H, ^{15}N]-HSQC, (B) the ^{15}N-TOCSY-HSQC, (C) the [^{1}H, ^{13}C]-HSQC and (D) the HCCH-TOCSY. The top of each panel shows colours those atoms that are labelled during the evolution of the pulse programs and whose dimensions are detected in the spectra below. Arrows indicate the magnetisation transfer steps of the experiment. Single spectral strips are shown for (B) and (D) from the corresponding HSQC cross-peak.

2.3.5.5 Aromatic side-chain assignment

Due to the hydrophobic nature of aromatic rings, aromatic side-chains (histidine, phenylalanine, tyrosine and tryptophan) are usually buried in the core of soluble proteins, yielding important NOE restraints that help fold the protein in the structure calculation. Relative to aliphatic protons, aromatic protons tend to be deshielded from magnetization due to aromatic ring current effects, giving them a distinguishably higher shift of 6-8 ppm. Consequently, two 2-D experiments specialised for detecting aromatic proton shift, (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE, were employed for shift assignment. These experiments correlate Cβ shifts of aromatic amino acids with Hδ ((HB)CB(CGCD)HD) and Hε ((HB)CB(CGCDCE)HE) ring protons (Fig. 24). In case of assigned Cβ shifts not significantly overlapped in the 2-D aromatic spectra, the Hδ or Hε ring protons could be correlated to the Cδ/ε shift in the ^{13}C-HSQC spectrum. Shifts that remained elusive were identified via the ^{13}C-NOESY-HSQC. This was largely based on the assumption that the most intense cross-peaks arise from intra-residue NOEs between protons within the aromatic ring.

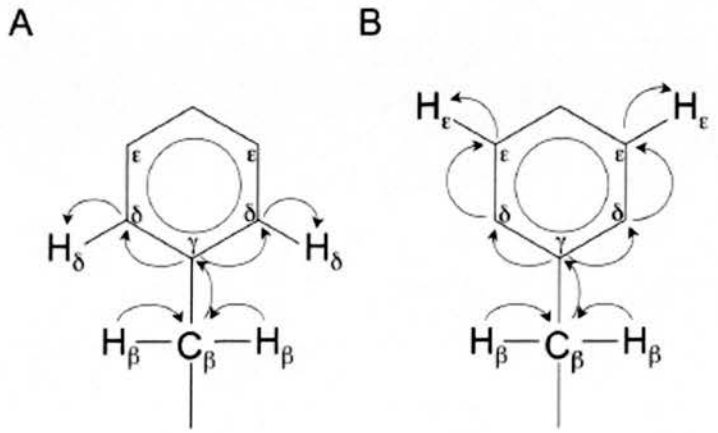


Fig. 24: Magentisation pathways of the (HB)CB(CGD)HD and (HB)CB(CGCDCE)HE experiments for phenylalanine.

2.3.5.6 Cis-trans- Proline assignment

The partially double amide bond renders the amide group planar, occurring in either *cis* or *trans* isomers (Fig. 25). The *trans* form is overwhelmingly preferred in most peptide bonds (1000:1, *trans*:*cis*), however, X-Pro peptide bonds have an appreciable number in the *cis* conformation (3:1, trans:cis) ; presumably because the symmetry between Cα and Cδ atoms of proline make the isomers nearly equal in energy. There are 10 proline residues in C7 CCPs and 13 proline residues in C7 CFF. *Trans*- or *cis*-configuration of the proline peptide bonds was determined by calculating the difference between the Cβ and Cγ shifts of each proline residue. A statistical analyses of C chemical shifts from high resolution NMR structures[136] revealed Cβ-Cγ in *trans* prolines to be 4.51 ± 1.17 ppm and in *cis* prolines to be 9.64 ± 1.27 ppm. This was confirmed by visual inspection of the ^{13}C-NOESY-HSQC as in *cis* isomers the Hα of the proline has a strong crosspeak with the Hα of the preceeding residue. While in trans conformation the Hδ of the proline has a strong crosspeak with the Hα of the preceeding residue.

Fig. 25: Proline *cis* and *trans* isomers. Each isoform has a characteristic NOE pattern whereby a strong Hα-HαNOE crosspeak (Xaa-Pro) is indicative of a *cis* isoform and a strong Hα-Hδ NOE crosspeak is indicative of a *trans* isoform.

2.3.6 Distance restraints for structure calculations

2.3.6.1 Distance restraints derived from the Nuclear Overhauser effect

Following resonance assignment, structure determination by NMR spectroscopy relies on the generation of distance restraints within the molecular system studied. The most commonly used distance restraints are derived from the nuclear Overhauser effect (NOE), which occurs between spins that have an appreciable magnetic dipole-dipole interaction (dipolar coupling) due to close proximity in space (< ~5 Å for 1 H-1 H NOE). Crosspeaks in NOE-spectroscopy experiments or NOESY experiments signify spatial proximity between the two nuclei in question and the distance between the two is proportional to the intensity of the peak.

Two NOESY experiments were used for both the CCPs' and CFF's structure calculations the ^{15}N-NOESY-HSQC and the ^{13}CH-NOESY-HSQC where magnetization is exchanged between the N/C nuclei associated with each HSQC cross-peak and all hydrogens within NOE range. Additionally, for CFF, a $^{13}CH_3$-NOESY-HSQC was recorded, producing a spectra relating only to $^{13}CH_3$ carbons, allowing deconvolution of this crowded region.

In order to calculate the correct tertiary structure, it is essential that the NOESY peaks are assigned correctly so that the distances extracted co-ordinate the correct nuclei. Thus assignments from the ^{15}N-HSQC and ^{13}C-HSQC were transferred to NOESY strips for every assigned N and C nucleus in the molecule. By overlaying the N and C-strips with the ^{15}N-HSQC TOCSY and ^{13}C-HCCH-TOCSY respectively, cross-peaks corresponding to intra-residue NOEs were completely assigned. In the case of C7 CCPs the clarity of spectra and the relatively low overlap of resonances allowed for the remaining inter-residue cross-peaks to be ambiguously assigned on the basis of chemical shifts using the structure calculation software CYANA. CFF's NOESY strips inter-residue assignment

was not wholly unambiguously assigned. A single distance restraint between a pair of protons normally results in symmetry related NOE cross-peaks, originating from each corresponding NOESY strip (Fig. 26). Using the chemical shift list information obtained from sections 2.3.5, the final dimension of an NOE cross-peak was unambiguously assigned to a lone matching Resonance candidate if there was an equivalent symmetry related cross-peak in the NOE strip from the candidate. Unambiguous assignment of proton shifts in the F2 dimension was mainly focused on amide-to-amide, amide-to-side-chain and aromatic-to-any proton cross-peaks.

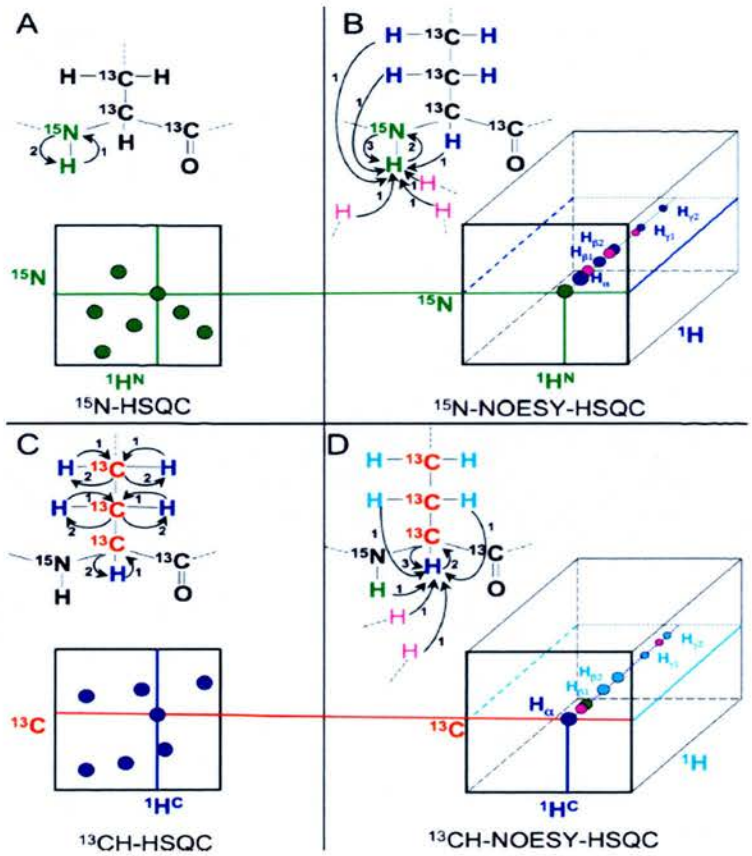


Fig. 26: NOE-cross-peak assignment. (A) shows [^{1}H, ^{15}N]-HSQC, (B) the ^{15}N-NOESY-HSQC, (C) the [^{1}H, ^{13}C]-HSQC and (D) the ^{13}C-NOESY-HSQC. The top of each panel shows colours those atoms that are labelled during the evolution of the pulse programs and whose dimensions are detected in the spectra below. Arrows indicate the magnetisation transfer steps of the experiment. Single spectral strips are shown for (B) and (D) from the corresponding HSQC cross-peak.

The amide-to-amide cross-peaks reveal characteristic secondary structural patterns, which are critical in defining structural elements in the early stages of the simulated molecular dynamics calculation. Additionally, the side-chain to amide NOE cross-peaks, in particular, from the Hα NOE strips, are readily identifiable in the ^{13}CH- NOESY-HSQC and can also serve as a guide to the secondary structure formation. However, those protons that happen to share similar chemical shifts gave rise to ambiguous distance restraints and were left unassigned (in the free ^{1}H dimension) and were treated as ambiguous by CYANA. Peak lists for distance restraints, chemical shift tables for all atoms in the molecule and the amino acid sequence were exported in XEasy format, using the FormatConverter of CcpNmr Analysis were used as input for structure calculation with CYANA.

2.3.6.2 Distance restraints derived from H-bonds

For CFF additional distance restraints inferred from intra-molecular hydrogen bonds were used for structure calculations. A modified HNCO experiment, HNCO_hbond, was used to exploit the small coupling constant associated with the slight covalent character of carbonyl-amide hydrogen bonds. Resultant cross-peaks correlate the carbonyl carbon with the amide nitrogen of hydrogen bonded pairs, with peak intensity being proportional to bond-length, with observable peaks having a proton to carbonyl oxygen distance smaller than 2.2 Å.** As with the NOESY restraint, hydrogen bonding restraints were exported using the FormatConverter of CcpNmr Analysis were used as input for structure calculation with CYANA.

2.3.7 Relaxation

2.3.7.1 Introduction

The absorption of electromagnetic radiation leaves a population of spins in the excited state; that is the distribution of spins is perturbed from the equilibrium (Boltzman) state. Following an NMR excitation pulse, it is the relaxation rate of the bulk magnetization to

thermal equilibrium that determines the lifetime of the FID and consequently dictates the line-breadth of the data; with fast relaxation rates and short relaxation times causing severe line broadening, translating into poor spectral resolution and low-signal to noise ratios. Moreover, in protein NMR, relaxation data are converted into parameters - spin-spin (R_2) and spin-lattice (R_1) relaxation rates and heteronuclear NOEs - that describe both the overall tumbling, the relative motion of domains and the extent and rate of internal motions of a macro-molecule.

Nuclear spin magnetic relaxation is a direct consequence of the magnetic coupling and the associated energy exchange of the spin system (protein) to its surrounding environment, termed the lattice i.e. the buffer solution. As the lattice modifies the local magnetic fields associated with the protein's nuclear spins, the lattice and the spin system are magnetically coupled (weakly). And due to stochastic Brownian motions of molecules in liquid solutions, this renders the local magnetic fields time-dependant. This magnetization can be resolved into two components: a nonadiabatic component perpendicular to the magnetic field (transverse) and an adiabatic component parallel to the magnetic field (longitudinal). The longitudinal component is associated with one of the two main relaxation process', namely, spin-spin (or longitudinal) relaxation. While the transverse component contributes to both of the two main relaxation processes: spin-lattice (or transverse) and spin-spin relaxation. Routinely in protein NMR the relationship between relaxation and molecular motion is exploited to study the backbone dynamics of a protein via the spin-lattice and spin-spin relaxation times (section 2.3.7.2 and 2.3.7.3 for T_1 and T_2 respectively) and also steady-state heteronuclear NOEs (see section 2.3.7.4) of backbone amide bonds.[137]

2.3.7.2 ^{15}N spin-lattice (T_1) relaxation

T_1 relaxation is the process by which the lattice acts as a thermal reservoir, accepting energy from the excited spins as they return to thermal equilibrium. There are three

principle types of magnetic interaction which contribute to spin-lattice relaxation of spin ½ nuclei. The first, dipole-dipole interactions is where the nucleus experiences a fluctuating field due to the motion of neighboring spins. The second, chemical shift anisotropy, where different orientations of the molecule exposes the spins to differing field strength as the magnetic shielding from the surrounding electron density fluctuates with rotational motion of the molecule. The third, spin-rotation interaction, results from the generation of a small electric current from molecular tumbling in turn inducing a small fluctuating magnetic field at the nucleus*. The spin-lattice relaxation time, T_1, is a time constant that characterizes the rate at which the longitudinal Mz component of the bulk magnetization ($M_{(t)}$) recovers to equilibrium ($M_{(0)}$) along the z-axis, following being flipped into the transverse M_{xy} plane, and is the time taken to recover approximately 63% of its initial value.[120]

In the recording of T_1 data the transverse components of magnetization are eliminated to prevent contributions from T_2 relaxation. The T_1 time constant is measurable by various types of experiments, however, in this project T_1 was measured using multiple inversion recovery experiments. This involves applying a series of pulse sequences of the type 180°-τ-90°, where τ is a small time which is varied from one experiment to the next. The first 180° pulse inverts Mz, which decays exponentially for time τ until application of a standard 90° pulse tips the bulk magnetization into the xy plane for detection. The T_1 value is then extrapolated from the resultant series of partially relaxed spectra as peak size is a function of T_1.

To obtain T_1 values for backbone amide nitrogens a series of interleaved [$^1H^{15}N$]-HSQC experiments, one for each different τ value were collected. Following identical processing in AZARA the individual amide cross-peaks in each spectra were picked in Analysis, excluding very weak peaks and peaks in overcrowded regions where

calculation of T_1 will be associated with a large error. Within CcpNMR, the Rates Analysis function was used to determine each amides T_1 value from an exponential fit of the intensity of the crosspeak as a function of relaxation delay times. Furthermore the Rates Analysis function was used to calculate time constant errors from the fitting error from multiple fitting routines.

2.3.7.3 ^{15}N spin-spin (T_2) relaxation

Following a radio-frequency excitation pulse, excited spins relaxing to the ground-state lose their coherent (in-phase) magnetization. This occurs not only from the conversion of transverse magnetization to longitudinal magnetization through the relaxation mechanisms described above, but also by loss of phase coherence (de-phasing) of the initially synchronous population of oscilating spins. However, the associated rate of decay of the x and y components of transverse magnetisation is characterised by the time constant T_2*. This value distinguishes it from the true transverse relaxation time, T_2, as T_2* (T_2* < T_2) also accounts for static inhomogeneities in the Mz direction of the external magnetic field that also contributing to dephasing.

To extract the T^2 value caused by *spin-spin* relaxation alone (i.e. independent of field inhomogeneity) from the T^2* a series of spin-echo experiments are performed. This involves applying a series of pulse sequences of the type 90°-τ-180°-τ. Where a 90° pulse flips the spins into M_{xy}, allowed to dephase for time τ, followed by a 180° pulse, and allowed to refocus for a period of time τ. This reverses the order of the spins with the slower ones now ahead (and falling back) and the faster ones behind (and catching up), allowing spins which were dephasing due to field inhomogeneneity to regain their phase coherence and thus eliminating their contribution to T_2. The effective rephrasing of spins results in a reappearance of the FID (spin-echo), which is smaller than the original signal the true transverse relaxation is an irreversible process and only the systematic

dephasing is reversed.

The pulse sequence used in the current study included Carr-Purcell-Meiboom-Gill (CPMG) pulse segments to reduce contributions from field inhomogeneitymolecular diffusion and chemical exchange, including conformational changes. And as with the obtaining T_1 values for backbone amide nitrogens, T_2 values were obtained from a series of interleaved [^{1}H,^{15}N]-HSQC experiments, one for each different τ value where $\tau < T_2$. Again the spectra were processed identically in AZARA and amide cross-peaks were picked in Analysis, excluding very weak peaks and overlapping peaks. Within CcpNMR, the Rates Analysis function was used to determine each amides T_2 value from an exponential fit of peak intensity as a function of relaxation delay times as well as their associated time constant errors.

2.3.7.4 Hetero-nuclear steady state [^{1}H^{15}N] NOE

The [^{1}H,^{15}N] heteronuclear NOE cross-relaxation rates are more sensitive to internal dynamics than the T_1 and T_2 relaxation rate measurements and therefore provides information regarding the motion of individual amide bond vectors in the shorter ns timescale.[137] NOE cross-relaxation occurs *via* dipole coupling of an excited spin with spins close in space. Those spins that undergo motion faster than the overall tumbling of the molecule will show a decreased NOE intensity relative to a reference spectrum. Thus the more mobile the amide bond (on a ns timescale), the lower the amide cross-peak intensity. The [^{1}H,^{15}N] heteronuclear NOE was expressed in terms of the ratio of the intensity under saturating conditions over the intensity under non-saturating conditions to reflect the impact of the negative NOE enhancement on intensity.

2.3.8 Structure calculation and refinement

2.3.8.1 Introduction

Two of the several reliable methods for obtaining 3-D protein structures from NMR data

were employed in the structure determination of C7 CCPs and CFF. The first method employed using CYANA[138] version 2.1, is the conjugate gradient minimisation of a variable target function using torsion angle dynamics. The second method employed using the structure calculation software Crystallography and NMR systems (CNS),[139] is the minimisation of a hybrid energy target function using restrained molecular dynamics in Cartesian space. CYANA was used to semi-automate the NOE assignment procedure and generate the initial non-refined structures. Following this, CNS was used to refine the structures in water solvent. Both methods involve multiple parallel searches of conformational space in order to minimise a potential energy target function as well as satisfying the experimental data set. CYANA uses torsion angle dynamics in which the empirical restraints (i.e. bond lengths, bond angles, prochiralities) are kept fixed at their optimal value to minimise a variable potential energy target function in torsion angle space subject to the experimental distance restraints. Conversely, CNS achieves this in Cartesian space by solving Newton's equations of motion for all atoms using a hybrid energy target function of empirical and the experimental restraints. In both methods, the energy minimization function reduces the target function while implementing simulated annealing. In simulated annealing, the application of repeated *in silico* heating steps followed by slow, progressing cooling steps enables adequate sampling of conformational space and the avoidance of local energy minima to reach a global energy minimum.

2.3.8.2 CYANA

The CYANA program for NMR structure calculation is based on its predecessor DYANA (DYnamics Algorithm for Nmr Application).[140] Unlike DYANA, CYANA 2.1 includes the Combined automated NOE assignment and structure determination (CANDID) software module.[141] This enables NMR structure determination of proteins by automated assignment of the NOESY spectra. The routine used for CYANA structure

calculation (Fig. 27) is an iterative approach that runs through seven cycles of NOE cross-peak assignment (by the CANDID module) followed by DYANA-driven structure calculation (torsion angle dynamics).

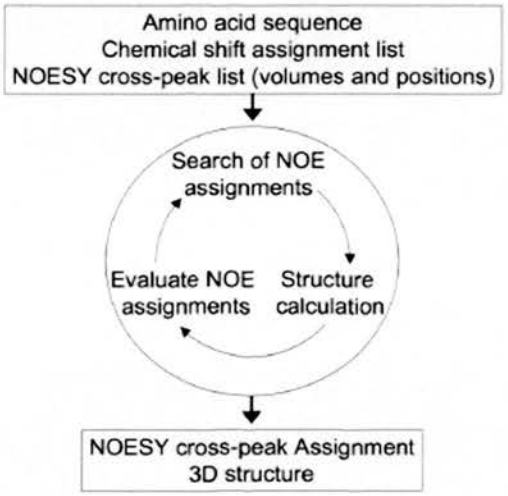


Fig. 27: General scheme for the annotated NOE assignment and structure calculation in CYANA 2.1. Figure adapted from paper.[138]

The inaugural cycle incorporates the input files into the calculation: peak lists for distance restraints, chemical shift tables for all atoms in the molecule and the amino acid sequence in XEasy format. CANDID generates a primary assignment list of one or more assignment contributions for each NOESY cross-peak (chemical shift tolerances set to 0.03, 0.04 and 0.45 ppm for the direct ^{1}H, indirect ^{1}H and $^{15}N/^{13}C$ dimensions respectively). Subsequently this list is filtered by applying a normalized score to each assignment and retaining those that are high ranking. An approach called network-anchoring is then used to remove possible noise peaks with low generalized contribution score values and to reduce ambiguity within each peak by eliminating peak contributions with low scores. The score is formulated to rate the initial, chemical shift-based assignment on the basis of how well the assigned NOE-connection is embedded in a (consistent) set of neighbouring NOE cross-peak assignments that support it. Therefore

the lowest scores are applied to single isolated constraints not mutually supported by other constraints and favours NOE cross peaks embedded in a redundant network of NOE connectivities. In the first two cycles only the Constraint Combination function of CANDID is also applied to combine long-range distance constraints, thereby reducing the influence of single incorrect distance restraints originating from NOE artefacts. Lastly the cycle is complete by a structure calculation step using the DYANA torsion angle dynamics (TAD) algorithm defined by the CYANA target function[138] with simulated annealing. A conformation that satisfies the torsion angle restraints and distance restraints in the absence of steric overlap will lead to a desirably lower target function value. During the simulated annealing, a series of 10000 TAD steps minimize the target function with the system is couples to a temperature bath and held at 9600K for the first 2000 steps and then slowly cooled down over the following 8000 steps to overcome energy barriers of local minima. All subsequent cycles incorporate the three-dimensional protein structure from the previous cycles, in addition to the complete input used for the first cycle.

As cross-peak evaluation is also driven by compatibility with the intermediate 3-D structure in proceeding cycles, it is critical that the correct polypeptide fold is found during the first cycle. This was confirmed with a backbone RMSD for each individual CCP domain (of the structures generated in cycle 1) being equal to or lower than 3Å**. Initially two sets of structure calculation were performed in parallel; one where the conserved disulfide bonds in each module were not defined, and the second with the disulfide bonds were defined. Consistency in the observed structures allowed confirmation of the defined disulfide bonds. Compliance with the remaining specified criteria[138] was essential prior to importing the NOE assignments from the final CYANA structures back in to ANALYSIS to convert the file format for compatibility with CNS.

2.3.8.3 CNS

CNS unlike CYANA allows for the refinement of protein structures in explicit solvent. This has been shown to significantly improve structural quality criteria such as backbone confirmation, the number of unsatisfied hydrogen bond donors/acceptors, packing quality and Ramachandran-plot statistics.[139] However CYANA has an explicit automated-NOE crosspeak assignment function and is therefore used to produce complete NOE assignments from the final CYANA structure. These were imported into analysis *via* the FormatConverter and the NOE cross-peak relative intensities were translated into distances in angstroms. Finally, using the FormatConverter, the distance restraints were exported in CNS format.

CNS Version 1.2 employs restrained molecular dynamics based simulated annealing in Cartesian Space to minimize a hybrid energy target function[139] for the calculated structures. The overall energy ($E_{overall}$) target function is composed of two energy terms. The first energy term refers to the experimental restraints (E_{NOE}) associated with the NOE distance restraints, where a large NOE energy term dictates a large number of NOE violations. The second term pertains to empirical restraints (E_{FF}) and is expressed as a force field described by covalent terms maintaining idealized bond lengths (E_{bond}), bond angles (E_{angles}), dihedral angles ($E_{improper}$), chiralities ($E_{improper}$) and a van der Waals repulsive term (E_{vdw}), that inhbits steric clashes between atoms based on their van der Waals radii.

Each CNS structure calculation progresses through three stages: 'random', 'regularise' and 'refine'. The first stage involves the generation of 100 reduced and non-bonded random protein structures. The random structures are calculated *via* 50 steps of energy minimization using the Powell minimization algorithm, with NOE restraints present and empirical forces set to low values. Molecular dynamics simulations are then employed to

calculate the time dependent behavior of the protein with respect to atomic motions within the molecule. At 2000 K 300 steps of molecular dynamics are simulated, followed by three time-step varying rounds of 500 molecular dynamics steps at a lower temperature of 1500 K.

The 100 'random' structures form the starting point for the 'regularise' stage of the structure calculation, whereby their local geometry are optimized. Specifically, the idealized bond length, NOE and vdW terms were reduced by Powell minimization in 400 steps, followed by another 400 steps of Powell minimization of bond angles. Chirality and planarity terms were then gradually introduced over four rounds of molecular dynamics of 250 steps each, followed by a further two rounds of molecular dynamics of 500 steps each. Next the correct handedness of the molecule was established using chiral swapping whereby the chiral moieties within the protein (methyl, methylene and side-chain amide protons) have their handedness swapped randomly. This is performed during three rounds of simulated annealing, therefore allowing the system to escape local energy minima traps by adopting higher energy conformations. The weightings of the van der Waals and NOE terms were then increased in the Powell minimization in conjunction with system cooling (2000 to 100K) over 2000 steps, prior to a final 400 minimization steps.

The random regularized simulated annealed structures obtained from the 'regularize' stage of the calculation form the input for the final 'refine' stage. Firstly, molecular dynamics simulations are preformed at high temperature (2000 K) over 10000 steps, with the NOE and H-bond derived restraint terms weightings set much higher with respect to covalent and non-bonded terms. The system was again cooled to 100 K over a total of 10400 energy minimization steps, including prochiral swapping and whereby the final 400 steps involve an increase in the weighting of the covalent and non-bonded

terms to be equivalent to the experimental terms weighting.

The 100 refined structures are then ranked and plotted in terms of NOE energy and overall energy. Those structures converging with the lowest energies (ideally >20%) were probed for violated NOE distance restraints. Violated restraints cross-peaks were then investigated and the upper distance value adjusted upon identification of peak artifacts (i.e. peak overlap and noise signal or diagonal interference) and the CNS calculation was repeated.

Following meeting the specified criteria the ensemble of structures is subjected to a final CNS refinement step in water solvent. This was achieved by mimicking the physiological environment of a protein in solution using the Lennard-Jones non-bonded potential and electrostatic force field in the RECOORD protocols (http://www.ebi.ac.uk/msd/record). Four simulated annealing and molecular dynamics rounds are used in the explicit water refinement, whilst the protein is immersed in a 7 Å shell of water molecules. The first three molecular dynamics rounds involve system heating from 100 K to 500 K over 200 steps, then 500 K maintenance over 2000 steps and cooling to 25 K over 200 steps. With the protocol concluded with a final energy minimisation over 200 steps.

2.3.8.4 Molecule visualisation programs

Individual structures and protein ensembles were viewed in the molecular visualisation programs MOLMOL[142] 2k.2 and PyMol.[143] MOLMOL was also used for the calculation of backbone root mean square deviation (RMSD) values.

2.3.8.5 Structural analysis programs

The quality of the calculated structures was assessed using Procheck[144] (Ramachandran plot statistics) and the course quality packing control module from WHATIF.[145] Surface

electrostatics were determined using the Adaptive Poisson-Boltzmann Solver (APBS)[146] within PyMOL (DeLano Scientific, San Carlos, CA) using PARSE forcefield generated PQR files calculated using the PDB2PQR server (http://kryptonite.nbcr.net/pdb2pqr/). Surface lipophilicity was determined using the MOLCAD module[147] of SYBYL v6.9 (Tripos Associates, St. Louis, MO). The buried surface area at the intermodular junctions was calculated using GETAREA 1.0,[148] being computed as: (SA Module$_i$ + SA Module$_j$) − SA Bimodule$_{ij}$. Intermodular angles were determined using the same protocol as previously described using the program XYZ. Protein interaction sites were identified using the online STP server (http://opus.bch.ed.ac.uk/stp/). Combinatorial extension[149] was employed to compare each experimentally determined CCP structure within the complement system against the closest-to -mean individual structures of CCP1 and CCP2.

2.4 Low Resolution Structural Studies

2.4.2 Small Angle X-ray Scattering

SAXS data acquisition and model production for C7-CFF was performed by Dr. Elizabeth Blackburn (Edinburgh University). Synchrotron radiation X-ray scattering data were collected on the ID14-3 BioSAXS beamline (ESRF, Grenolble) using a PILATUS 1M pixel array detector (Dectris, Switzerland) and four frames of 30 seconds exposure time. Solutions of C7-CFF (8 mg/ml) were measured at 10 °C in 20 mM sodium acetate buffer, pH 4.0 (NMR buffer conditions), at protein concentrations of 4 and 8 mg/ml. Radiation damage was observed as a significant increase in the intensities of low-angle data after the second 30-s exposure. Thus, 5%v/v glycerol was added and the sample was continually flowed through the 1.8 mm qaurtz capillary sample cell during data acquisition. Data from the detector were normalized to the incident beam intensity, averaged and the scattering of buffer solutions subtracted. The difference curves were scaled for solute concentration. All data manipulations were performed using the PRIMUS software package, with the final bead model produced with DAMMIN.[150]

2.4.3 Chemical cross-linking

2.4.3.1 The cross-linking reaction

Chemical cross-linking coupled with MS was utilised to analyse the architecture of three complement proteins: C3, C3b and C7. The chemical cross-linker BS2G-d0 (Bis[Sulfosuccinimidyl] glutarate), and its deuterated equivelant BS2G-d4 (Thermo Fischer Scientific) where used dissolved in DMSO in equal quantities for the final cross-linking reactions, to identify cross-linked peptides by their 4 Da mass difference. For trial cross-linking reactions a series of protein:cross-linker ratios 1:100, 1:300, 1:1000 and 1:3000 were performed on 100 pmol of either C3b or C7 in a final reaction volume of 100µl. The proteins were cross-linked in 10 mM HEPES pH 7.0, 200 mM potassium acetate for 1hr at room temperature. The reaction was stopped by adding 5 µl of 1M ammonium bicarbonate. The reaction mix was seperated on a NuPAGE 4-12% Bis-Tris gel using MES running buffer and Coomassie blue stain. The optimum cross-linker ratios were selected by SDS-PAGE analysis and additionally, as required in the case of C7, by MS analysis. Upon selection of the optimum cross-linking ratio final cross-linking reactions were carried out on 80 µg, 80 µg and 50 µg of C3, C3b and C7 proteins respectively, using the same conditions described above.

2.4.3.2 Sample preparation, MS analysis and database searching

Following completion of the cross-linking reaction sample preparation, MS analysis and database searching was performed by Zhuo Chen (Juri Rappsilber group, Edinburgh University). Monomer bands in the SDS-PAGE gel were excised and the proteins reduced and digested using standard trypsin digest protocols. Using the published protocol[99] cross-linked peptides were fractionated using SCX-StageTips. The peptides were then fractionated using strong cation exchange chromatography (200 X 2.1 mm Poly SULFOETHYLA column; Poly LC, Columbia, MD, USA) and were eluted directly into an LTQ-Orbitrap classic mass spectrometer. Peptides and their subsequent iontrap fragments were detected at high resolution in the Orbitrap. The raw MS files were processed into peak lists using MaxQuant[151] using default parameters. Searches were conducted using a database containing the sequences of C3, C3b, or C7 using the in-house software Xi. All identified peptides contained either a lysine residue or a protein N-terminus at the most likely linkage position as determined by observed fragments. Cross-linking candidates

were manually validated using the in-house Xaminatrix program and sorted into high and low scoring using an in-house algorithm.[152] Cross-linking reactions and analysis were repeated and only those peptides that were reproducible, had mass-errors within 6 ppm of the predicted peptide mass, had exceptional database search scores and had a well-explained spectrum with fragment information for both peptides, were considered high confidence cross-links and used for further analysis.

2.4.3.3 C7 Architecture

For the assessment of protein architecture using the high confidence cross-link list, models were required for C7's protein modules. The MACPF was homology modelled in MODELLER[153] on the X-ray structure of C8α-MACPF[49] and was provided by Dr Dinesh Soares (Edinburgh University). All other C7 module's models were produced by the online homolgy modelling server PHYRE. These protein models had reasonable alignment scores, modelling scores and WHATIF[145] coarse packing quality scores (Table. 8).

Target	Template from	% Identity	E scores	Whatif scores
TSPN	Thrombospondin TSP1	33	1.30E-011	-3.3
LDLRA	Rhinovirus	38.5	4.00E-006	-3.2
MACPF	C8α-MACPF	30	n/a	-2.6
EGF-like	Factor VII	28	1.30E-011	-3.5
TSPC	Properdin TSP2	35	3.00E-006	-2.5

Table. 8: Table of module homolgy model quality.

PROTEIN PRODUCTION, PURIFICATION & CHARACTERIZATION

3.1 Introduction and overview of constructs

The DNA encoding the C-terminal domains of C3 and C5, in addition to a series of construct coding for C-terminal modules of C6 and C7, were originally cloned into pET-15b vectors by members of Dr Ronald Ogata's laboratory. Despite expression in the Origami B pLYSs (Novagen) bacterial strain, many of the constructs, and particularly the larger ones, resulted in very low yields of soluble protein. In the current work, all constructs were re-cloned for expression in *P. pastoris*. This organism was chosen because, as a eukaryote, it promotes disulfide pattern formation in the endoplasmic reticulum. Moreover it can secrete proteins into the medium thus easing purification, and it can produce very large amounts of recombinant protein yet grow on miminal media consistent with incorporation of isotopic labels (for NMR).[154] Figure 28 illustrates the constructs used for re-cloning. Table. 9 summarises the N- and C-terminal sequences of the constructs.

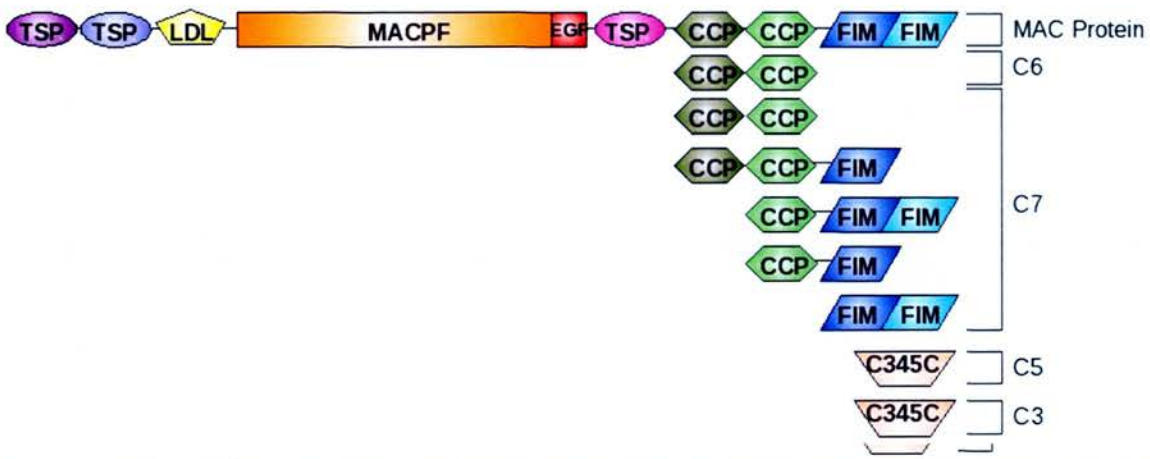


Fig. 28: Summary of constructs re-cloned into *P.pastoris* expression vectors. A schematic representation of a MAC protein is shown at the top for comparison. The modules draw below indicate the composition of the various recombinant proteins that were subjected to investigation in the current study.

3.2 Re-clone into *P. pastoris*

3.2.1 DNA Manipulation

The inserts were doubly digested from TOPO vectors (in the case of pGAP cloning) or

from PCR-amplified DNA inserts (in the case of pPIC cloning). The DNA inserts were successfully ligated into compatibly digested pPIC/pGAP vectors. pPIC/pGAP vectors were compatibly digested following amplification and extraction of the plasmid DNA from TOP10 cells. Digestion mixtures were subjected to agarose gel electrophoresis, and the appropriate plasmid or insert gel band was excised and gel extracted prior to the ligation reaction. No figure is shown as visualization of the DNA bands by UV light can damage the DNA, reducing the efficacy of the ligation process. Following ligation of the inserts into pPIC/pGAP and transformation into TOP10 cells, the resultant colonies were screened by PCR to ensure presence of an appropriately sized insert prior to amplification and extraction of the construct DNA. Many TOP10 colonies did not contain the insert, but instead contained empty vectors that allow them to grow on the agar plates containing zeocin for selection. The problem of colonies growing with empty vectors was subsequently resolved by using a higher ratio of insert to vector. Constructs containing plasmids that had been positively identified by DNA sequencing were then linearized for transformation into *P. pastoris* (strain KM71H). Transformation was successful for every pPIC construct, however not all pGAP constructs were successfully transformed (see Table.9).

Insert	5' Restriction site	First-last aa	Cloning artifact	pGAP complete cloning	pPIC complete cloning	Glycosylation	Calculated Mw	pGAP +ve expression	pPIC +ve expression
C5 C345C	XhoI	Ala^{1530}-Cyc^{1676}	EAEA	No	Yes	Yes, Asn^{1630}	17193.6	/	Yes
C3 C345C	XhoI	Ala^{1514}-Asn^{1663}	EAEA	No	Yes	Yes, Asn^{1617}	17841.9	/	Yes
C6 CCPs	PstI	Ser^{642}-Lys^{706}	EAEAA	Yes	Yes	No	14768	Yes	Yes
C7 CCPs	PstI	Glu^{569}-Gln^{690}	EAEAA	Yes	Yes	No	13689.7	No	Yes
CCF	XhoI	Glu^{569}-Ser^{768}	EAEA	No	Yes	Yes, Asn^{754}	22465.9	/	Yes
CF	XhoI	Ile^{629}-Ser^{768}	EAEA	No	Yes	Yes, Asn^{754}	15849.4	/	Yes
CFF	XhoI	Ile^{629}-Gln^{843}	EAEA	No	Yes	Yes, Asn754	23784.3	/	Yes
FIMs	XhoI	Lys^{691}-Gln^{843}	EAEA	Yes	Yes	Yes, Asn^{754}	17200.6	Yes	Yes
Fim1	XhoI	Lys^{691}-Ser^{768}	EAEA	Yes	Yes	Yes, Asn^{754}	9265.7	Yes	Yes
Fim2	XhoI	Glu^{770}-Gln^{843}	EAEA	No	Yes	No	8282.2	/	Yes

Table.9: Re-cloning summary. Summary of constructs and re-cloning success (from pET15b to pGAPzαB/pPICzαB) for all available constructs.

3.2.2 Screening for expressing clones

Successfully transformed *P. pastoris* clones were tested for protein production in "mini-scale" (see section 2.2.8.2) trials. Protein production trials for pGAP-transformed clones, in which 10 ml of YT media, were inoculated with a single colony and incubated in 50 ml Falcon tubes. As pGAP allows for constituent protein expression, which is costly for isotopic labelling of recombinant proteins, rich media was used to maximise protein production for non-NMR based applications. Subsequent analysis of the cell supernatant yielded SDS-PAGE results that were difficult to interpret due to the presence of peptone in the expression media and a consequent poor resolution of bands (Fig.29J). Nonetheless, even from these poor-quality gels it was evident that clones transformed with pGAP containing DNA encoding C7-FIM1, C7-CCPs (*i.e.* both CCP modules) and C7-FIMs (*i.e.* both FIM modules) produced proteins of the expected size and therefore glycerol stocks were stored for future analysis. In the case of pPIC-transformed clones, 50 ml of minimal growth media (BMG) in 150 ml baffled flasks were used in protein production trials to maximise aeration of the cell cultures in the hope of increasing yields of recombinant protein. After two days initial growth the cells were transferred to BMM media containing methanol to induce expression. Because these cells were grown in minimal media (to assess protein production under conditions for future isotopic labelling) the SDS-PAGE performed on concentrated media yielded clearer results. Each of the clones produced a recombinant protein of the expected size and in easily detectable (Coomassie-blue staining) quantities (Fig 29).

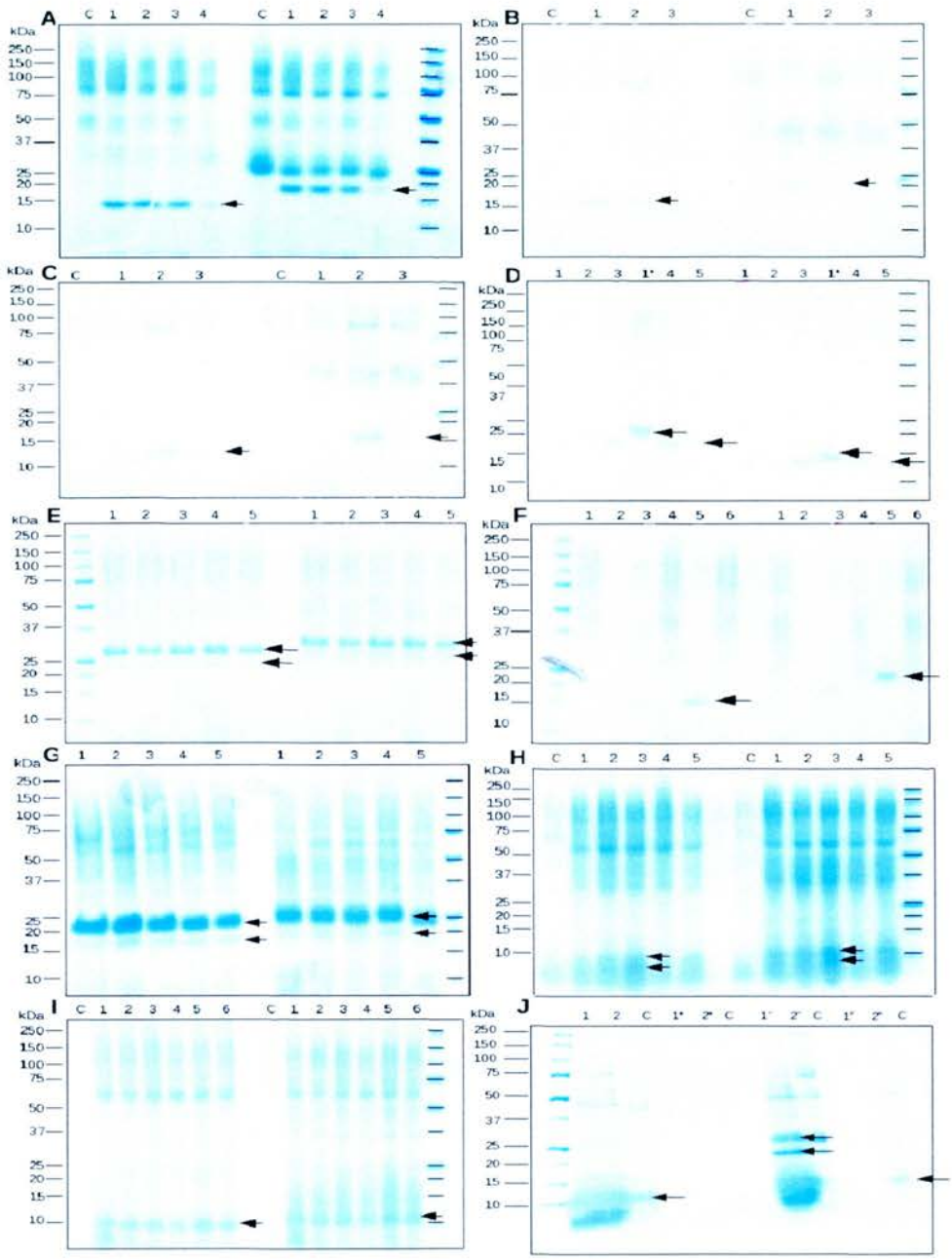


Fig.29: Mini-scale protein production trials of *P. pastoris* clones. Gels (A-I) represent the outcome of protein production trials of pPIC clones and (J) for pGAP. Non-reducing (*NR*) conditions are to the left and reducing (*R*) to the right. In Gel (I) all samples were run under reducing conditions. A: C5-C345C; B: C3-C345C; C: C6CCPs; D: C7-CCPs, D*:C7-CCF; E: C7-CFF; F: C7-CF; G: C7-FIMs; H: C7-FIM1; I: C7-FIM2; J: C6-CCPs, J*: C7-CCPs, J2": C7-FIMs. In all gels (C) is the negative control and the numbers correspond to the number of clones tested.

3.3 C7 CCPs expression, purification and characterization

3.3.1 Expression

The C7-CCPs construct was originally produced in the Origami B strain of *E. coli* freshly transformed with the suitably manipulated pET-15b vector, prior to re-cloning into *P.pastoris*. As described previously (see section 2.2.6.1) the OrigamiB pLysS expression system has been optimized for tightly regulated expression and production of soluble disulfide-containing proteins. Using expression conditions as determined by collaborators (Dr. Ron Ogata *et al*, Torrey Pines Institute, Florida), identical to those used for production of C5-C345C and C7-FIMs,[53,54] some C7-CCPs is observed in the soluble fraction but the majority of recombinant material (judging by the intensity of bands on SDS-PAGE) was found as an insoluble aggregate in inclusion bodies (Fig. 30).

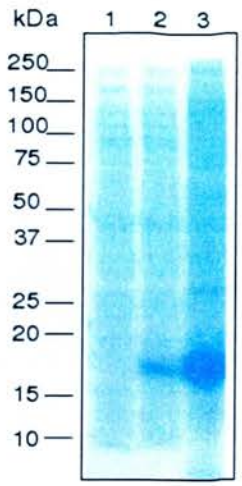


Fig. 30: Reducing SDS-PAGE of C7-CCPs recombinantly produced in the OrigamiB strain of *E. coli:* (1) media supernatant (2) cell lysate (3) cell pellet (produced by addition of 20µl of H_2O added to cell-pellet and heated at 100°. 10µl was loaded in each lane.

3.3.2 Protein purification

A protein purification procedure was optimized to obtain highly purified protein and to minimize loss. In the first purification step, C7-CCPs was purified from the cell lysate by binding via its N-terminal His_6-tag to divalent metal ions immobilized on an

Immobilised Metal Affinity Chromatography (IMAC) sepharose resin column (1 ml, GE Healthcare). Depending on the metal ion used, varying degrees of *E. coli* protein impurities will bind to an IMAC column. Moreover, changes in the pH of the buffer alters the affinity of the contaminating proteins as well as the His_6-tagged protein for the metal ion. With higher metal ion binding affinities non-specifically bound proteins compete with His_6-tagged protein for binding metal ions. This can result in reduced yields of recombinant protein and the resultant loss of protein in the flow-through. Therefore binding of C7 CCPs and protein impurities and elution from the column using an Imidazole gradient was assessed on both $Ni^{2+}$and Co^{2+}-loaded affinity resins, over a range of pH values, with the goal of obtaining strong but selective binding of C7-CCPs (Fig. 31). Non-specific binding was found to be greatly reduced (in comparison to the use of Ni^{2+}) when using Co^{2+} coupled to the resin, and a buffer pH of 6.5 provided maximum resolution in the elution positions of C7-CCPs and the impurities. These conditions were used in further purification attempts. Subsequently, 50 mM imidazole was included in the wash buffer to wash out impurities that were non-specifically bound to the column.

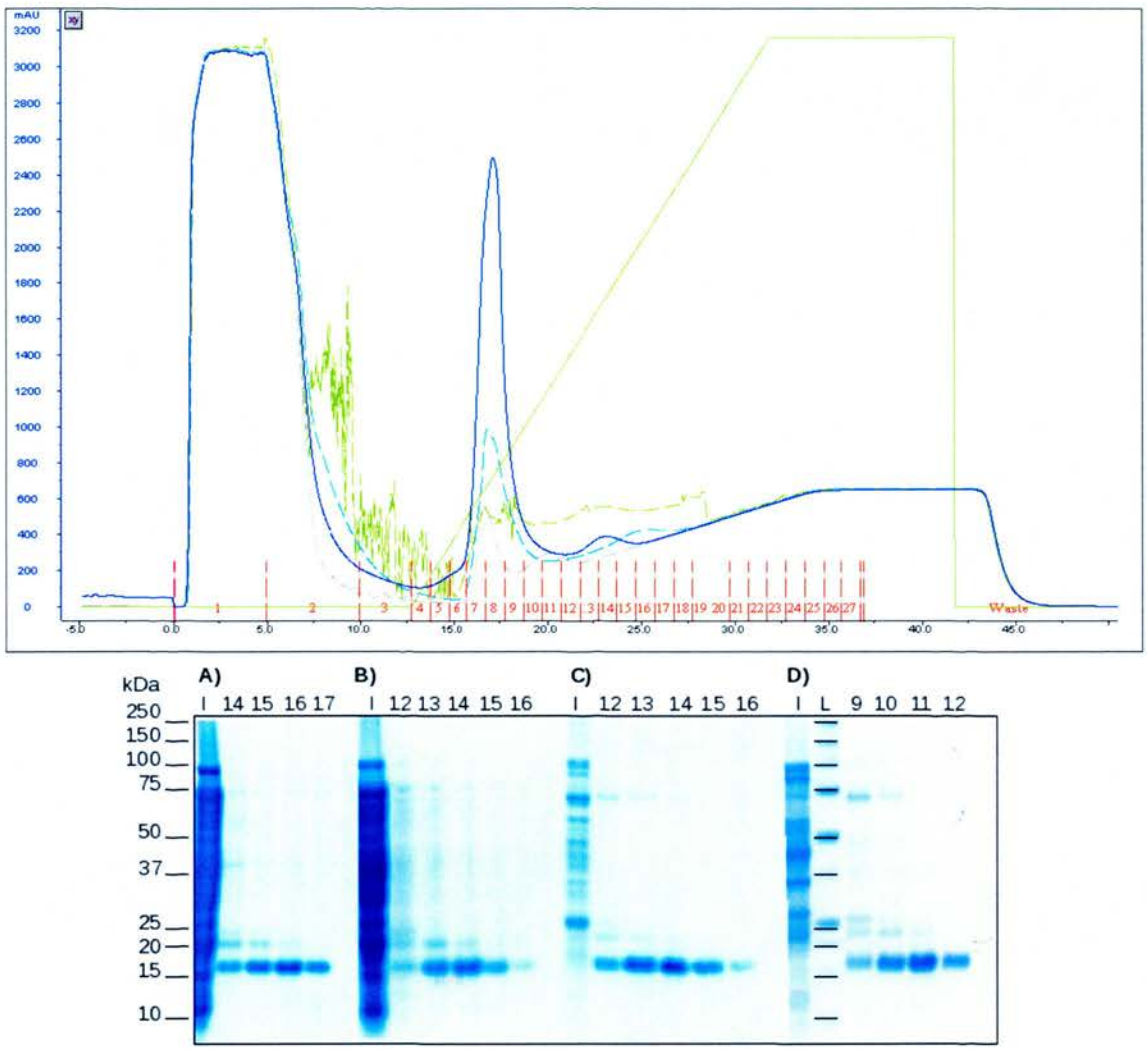


Fig. 31: Optimization of metal-chelate affinity chromatography of C7-CCPs. The top panel shows the chromatogram with UV traces of protein eluted from the Histrap column along an imidazole gradient (green-solid). Nickel was used as the metal ion at buffer pH 6.5 (light-blue-dashed line, gel A) and pH 7.2 (dark-blue line, gel B). Cobalt was used as the metal ion at buffer pH 6.5 (green-dashed line, gel C) and pH 7.2 (grey line, gel D). The bottom panel shows SDS-PAGE gels run on fractions representing the impurity peak (fractions 7/8) and the numbered C7-CCPs-containing peak fractions.

The second stage of purification was the removal of the His-tag by thrombin cleavage at the cleavage site that had been engineered in to the recombinant protein (see section 2.2.6.1). Following cleavage a second metal-chelate affinity chromatography step was

performed. This removed the cleaved His-tag and impurities, while the His tag-free C7-CCPs was recovered from the flow-through (Fig. 32). The sample still contained impurities, however, (~10% as estimated from the gel) and requires a further 'polishing' step. The efficacies of two potential polishing steps - reverse-phase HPLC and gel filtration column - were compared (Fig.33) using SDS-PAGE run on the purified fractions. Anion-exchange chromatography was also attempted, however no binding to the column was observed (data not shown). According to this comparison, gel-filtration proved more effective at removing impurities than reverse-phase chromatography as may be judged from Figure 33.

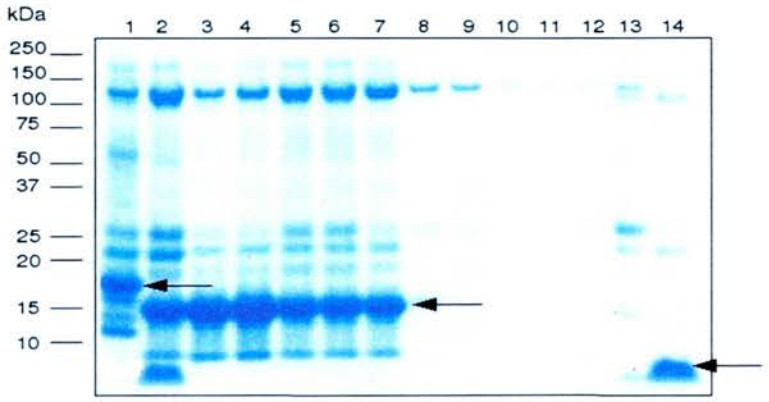


Fig. 32: Removal of the His_6-tag. Lane (1) Before proteolysis; Lane (2) Flow-through fraction 1 (FT 1); Lane (3) FT 2; Lane (4) FT 3; Lane (5) FT 4; Lane (6) FT 7; Lane (7) FT 8; Lane (8) FT 9; Lane (9) FT 10; Lane (10) elution fraction 1 (ET 1); Lane (11) ET 2; Lane (12) ET 3; Lane (13) ET 4; Lane (14) ET 5. From left to right, C7-CCPs before exposure to thrombin, C7-CCPs after cleavage, and finally the tiny N-terminal His_6 tag, are each depicted with arrows.

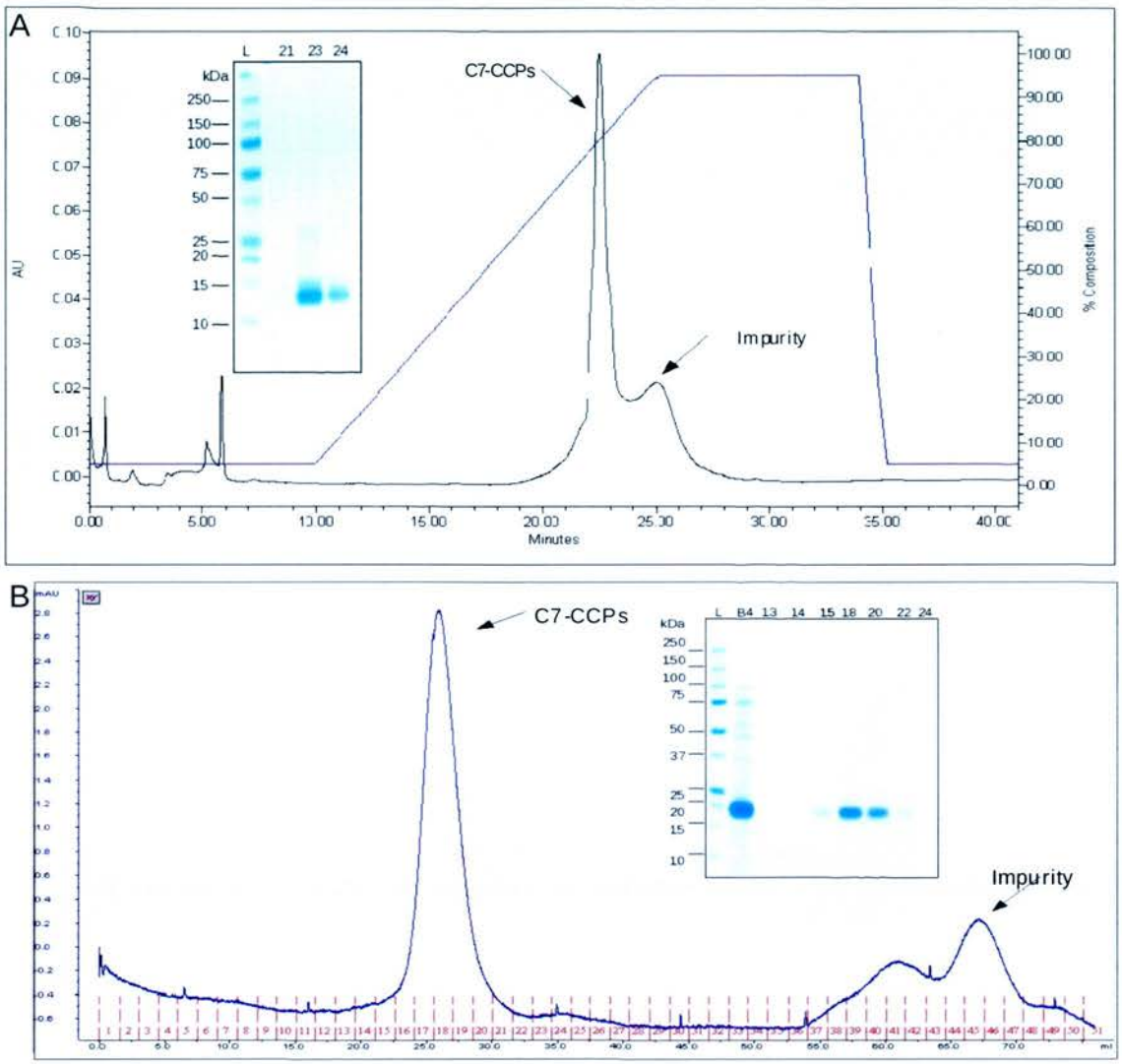


Fig. 33: HPLC and gel filtration as a C7-CCPs purification polishing step. Chromatogram showing elution of C7 CCPs samples from an HPLC column (Discovery® BIO Wide Pore C8-5 µm, 25 cm x 4.5 mm, Supelco) (A) or gel filtration column (HiLoadTM 16/60 SuperdexTM 75 prep grade, GE Healthcare) (B). Impurities and C7 CCPs are marked with arrows.

3.3.3 Characterization

Electrophoretic migration of C7-CCPs was retarded in the presence of the reducing agent, mercaptoethanol (see Fig. 34B). This is consistent with loss of compact structure by reduction of intra-molecular disulfide bonds. The folding of the recombinant protein was further assessed by inspection of their ^{1}H-NMR spectra (Fig. 34A). The spectra contain peaks in the regions around 6 ppm and 9 ppm, indicative of tertiary structure.[120] To confirm if full-length C7-CCPs had been successfully purified, samples were analyzed by mass spectrometry (MS) (see Appendix D). The recombinant protein has a predicted mass (including an N-terminal cloning artefact (GSHM)) of 13,985.1 Da (for the oxidised protein, calculated in Protparam).[33] MALDI-TOF MS yielded a mass of 13, 984.8 Da +/- 1.2

Having established a working purification strategy for C7-CCPs produced by *E. coli*, yielding ~1 mg per litre of expression media of pure, soluble, fully-folded protein, a 3L batch of ^{15}N-enriched C7-CCPs (500µg of purified protein/l of expression media) was similarly prepared for NMR optimisation. This was followed by production of a 7L batch of ^{13}C,^{15}N- C7-CCPs (300 µg of purified protein /l of expression media) for the purpose of assignment and NMR-derived structure determination.

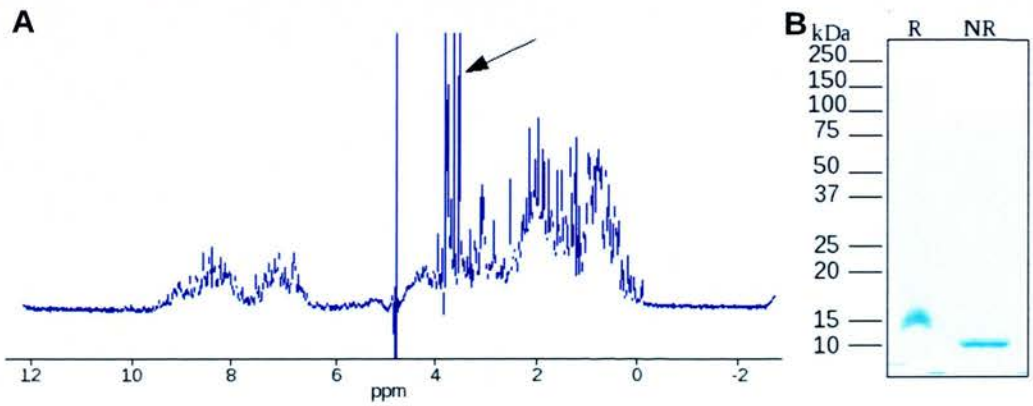


Fig. 34: C7-CCPs characterisation. (A) Shows a ^{1}H-NMR spectrum with small-molecule impurities marked with an arrow. (B) Shows the movement of protein on the gel under reducing (R) and non-reducing (NR) conditions.

3.3.4 Optimizing sample conditions for NMR

A ^{15}N-labelled C7-CCPs sample was produced and purified (as described above) and [^{1}H,^{15}N]-HSQC's were recorded (Fig. 35) for optimization of sample conditions for NMR structural characterization. Firstly the issue of spectral folding was addressed. Spectral folding is a phenomenon that occurs when the spectral width specified during experimental set-up is insufficient to detect the full range of frequencies present. Under these circumstances, NMR peaks occur in the spectra at frequencies different from their real frequencies that are outside the chemical shift range of the spectral window. These peaks may be inverted and occur within the spectra at the same distance inside the spectrum boundary as they would have been beyond it had they not been folded. To increase the digital resolution, reduce the recording time and/or enhance sensitivity of multidimensional experiments, the spectral width of the ^{15}N and or ^{13}C dimension(s) was set to a minimum whilst ensuring that folded peaks did not overlap with others, this was achieved by recording HSQCs with different sweep widths, final sweep widths of 11 ppm (^{1}H) and 24 ppm (^{15}N) were chosen.

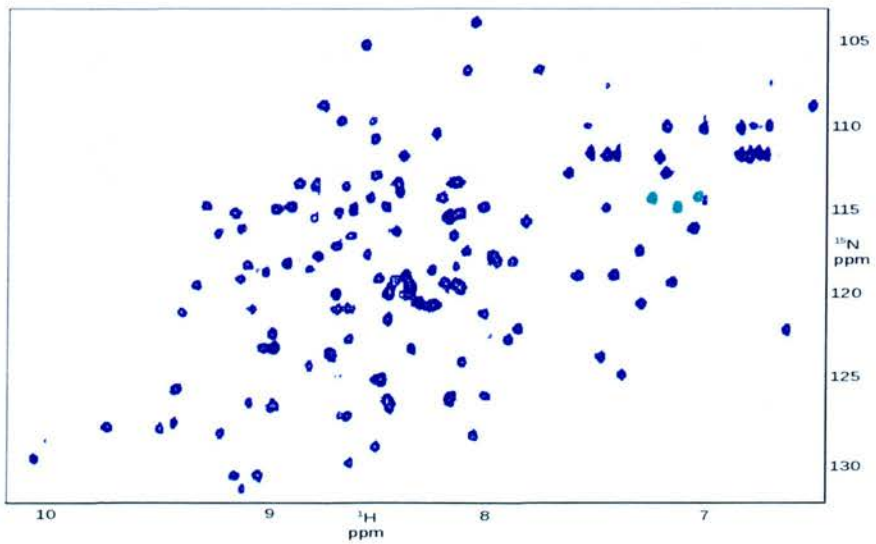


Fig. 35: [^{1}H,^{15}N]-HSQC of C7-CCPs. Spectral folding in the ^{15}N dimension was employed to increase the resolution, with three folded peaks (negative) observed in cyan.

The effects of varying salt concentration and pH of the standard potassium phosphate buffer as well as the temperature of the sample were assessed by comparison of [^{1}H,^{15}N]-HSQC spectra. A good quality [^{1}H,^{15}N]-HSQC spectrum proved to be a useful experiment to assess the suitability of conditions for further NMR experiments aimed at structure determination. At least one cross peak was expected for every residue in the protein (except proline and the N-terminal amino acid residue) as each backbone amide should give rise to a signal. In general, conditions were sought in which cross-peaks were of uniform intensity since very strong or very weak peaks suggest regions in fast or intermediate conformational exchange that are likely to pose problems later in the assignment process. Tryptophan and other residues containing NH (or -NH_3^+) in their side-chain give rise to additional peaks but these are less important, in general. The ultimate aim of sample optimization was to resolve in so far as possible a detectable peak for each backbone amide.

High salt concentrations can cause difficulty in tuning the NMR probe and will reduce the sensitivity of both traditional and cryo-probes. Furthermore salt absorbs the radio-frequency field applied to the sample during experimentation, which can lead to a smaller bandwidth of excitation, reduce the homogeneity of the applied pulse and cause temperature changes in the sample during the experiment.[38] However, low salt concentrations may have detrimental effects on the behaviour of the protein in solution, such as precipitation or aggregation. For C7 CCPs there was a loss of sensitivity and resultant intensity of peaks in the HSQC spectrum (Fig. 36) with an increase in salt concentration. The highest quality HSQC spectrum was obtained with no salt in the sample.

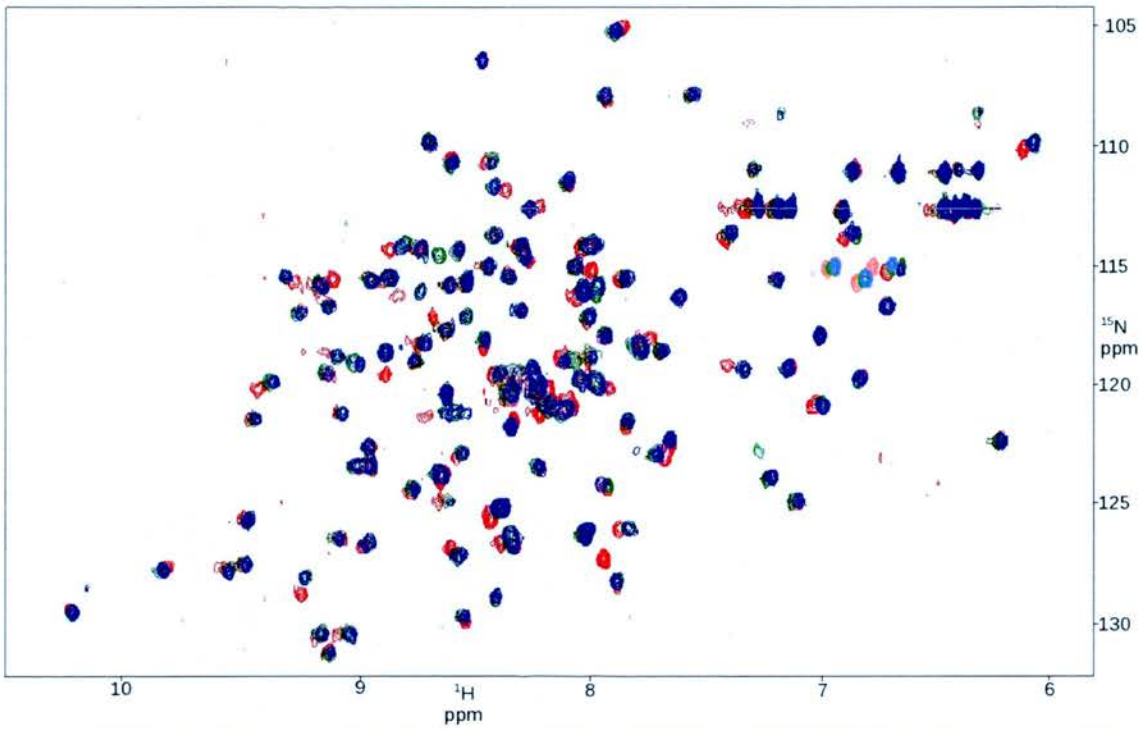


Fig. 36: Comparison of C7-CCP's' [^{1}H, ^{15}N]-HSQC at different salt concentrations: 0 mM NaCl (blue), 50 mM NaCl (green), 100 mM NaCl (red).

The optimal pH of the solvent for any NMR sample is unlikely to exceed 7.0. This is because the intrinsic exchange rate of a solvent-exposed backbone amide proton increases as a function of pH,[155] causing broadening of amide proton line-widths. Therefore, although it is desirable to mimic physiological pH conditions, (about 7.2 in blood plasma) it is often advantageous to maintain an acidic pH for NMR studies. Hence it is important to ensure the protein conformation is not disturbed or even disrupted at lower pH values. In the case of C7-CCPs, although some detectable peak movements occurred, there were no drastic changes to the NMR "fingerprint" at lower pH values. It was decided that pH 5.5 and 6.5 are too close to the theoretically calculated pI of 5.9 for C7-CCPs and should not be used so as to avoid isoelectric precipitation effects. On

balance, pH 5.0 was selected and, fortuitously, peaks were resolved at this pH that had been overlapped at higher pHs.

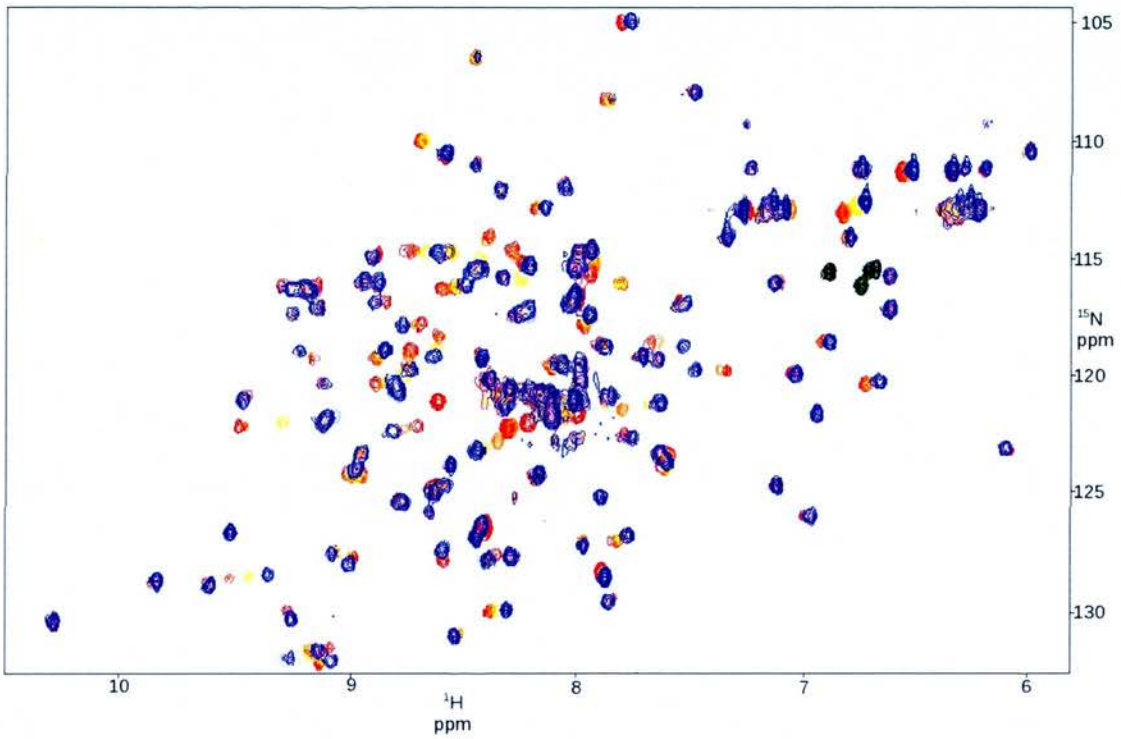


Fig. 37 Comparison of C7 CCP's' [^{1}H,^{15}N]-HSQC spectra at various pHs: pH 5.0 (red), pH 5.5 (orange), pH 6.5 (yellow), pH 7.0 (light blue), pH 7.2 (dark blue).

The temperature of the sample may also affect the strength of the NMR signal as well as sample stability. A temperature titration (range: 15-30°C) revealed that temperature changes had little effect on the C7-CCPs chemical shifts, although traceable movements of some peaks were observed (Fig. 37). The peak intensity grew weaker with a decrease in temperature, with signals at 15°C and 20°C being significantly weaker than other temperatures. The choice of 25°C seemed a reasonable compromise between sample stability, signal quality and physiological relevance. It is also consistent with the NMR studies of the C7 FIMS, which could facilitate future investigations that aim to merge CCPs and FIMs structures together.

Thus the optimal conditions for C7-CCPs were 20 mM potassium phosphate buffer, no salt, and pH 5 at 25°C. Under these optimised conditions, 3D spectra have been obtained of a quality worthy of resonance assignment.

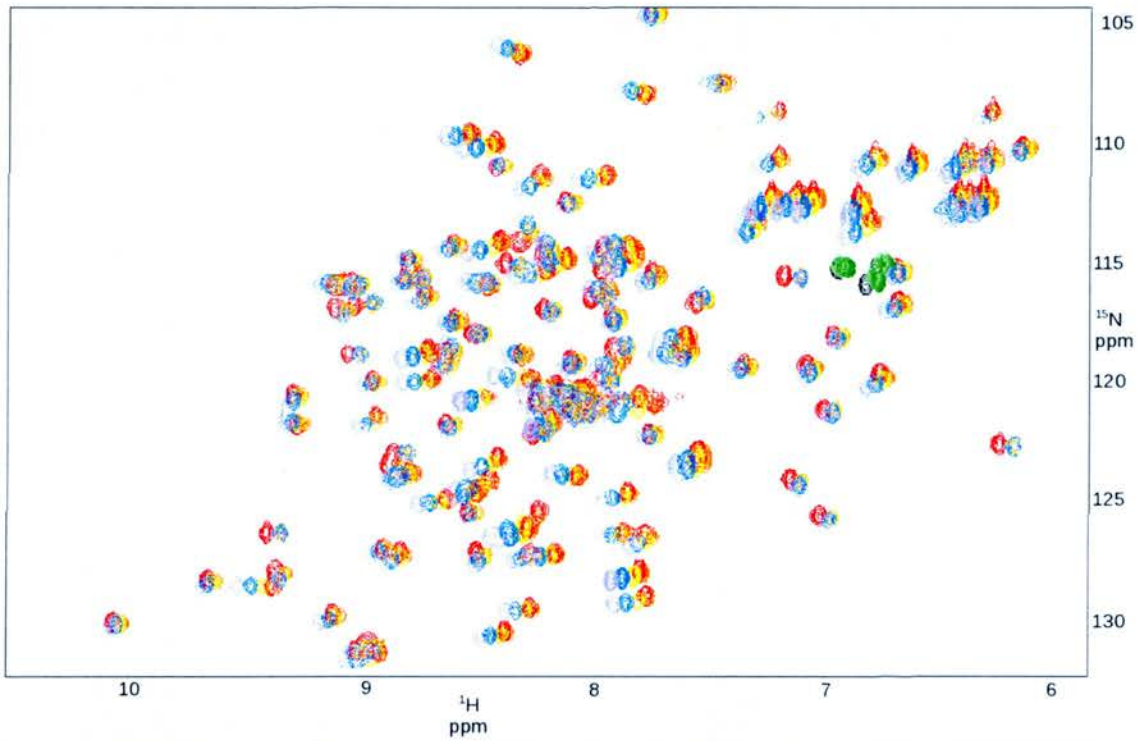


Fig. 38: Comparison of [^{1}H,^{15}N]-HSQC spectra of C7-CCPs collected over a range of temperatures. Pale blue (15°C), blue (20°C), yellow (25°C), orange (28°C), red (30°C).

3.4 C7 CFF expression, purification and characterization

3.4.1 Expression

In "miniscale" protein-production trials, all five clones picked for CFF expression produced a strong protein band (following SDS-PAGE with Coomassie blue staining) at ~30 , and a weaker band at ~24 kDa, with no evidence of protein degradation (Fig. 29). It is reported that both *P. pastoris* and human cells add approximately 5 kDa of sugars per N-linked glycosylation site and both express a mixture of glycosylated and non-

glycosylated forms of a protein.[118] Therefore the higher molecular weight band likely corresponds to CFF that is N-glycosylated at Asn^{754} in FIM1, which is the single consensus site (NXT, where X is any amino acid) in CFF. The lower band is thus likely to correspond to non-glycosylated CFF. This was later confirmed, by deglycosylation with the endoglycosidase enzyme EndoHf.

The strength of expression, the stability of the secreted protein to proteolysis and evidence of a straightforward harvesting procedure in that the protein bound to loose SP Sepharose™ beads (GE Healthcare) prompted progression to expression in a fermentor. Yields of non-isoptopically enriched material were high at 15 mg/l of cell culture. Using a protocol for feeding *P.pastoris* that was well established within the laboratory (see section 2.2.8.3) the overproduction of ^{15}N-C7-CFF and ^{13}C,^{15}N-C7-CFF, both in double labeling conditions, yielded 8 mg/l and 10 mg/l of cell culture respectively. The natural abundance batch of C7-CFF provided samples for establishing a purification procedure and for protein characterization. The ^{15}N-C7-CFF sample was used for optimizing NMR conditions and the subsequently produced double-labeled batch of C7-CFF was sufficient for the task of assignment of nuclei and collection of NOESY data.

3.4.2 Purification

Initial purification and harvesting from the supernatant was achieved *via* cation-exchange chromatography wherein strong binding of the target protein to the resin was observed, with CFF eluting from the column at 50-60% of the salt gradient (~500 mM NaCl). As with the "miniscale" production trial, a mixture of both glycosylated and unglycosylated recombinant protein was produced, resulting in the appearance of a double band in SDS-PAGE analysis of the eluted fractions. Following treatment with the endoglycosidase, EndoHf , the sample was reanalysed by SDS-PAGE to ensure

completion of the deglycosylation reaction, which was judged by the absence of double protein bands and the observation of a single band at ~24 kDa (Fig. 39). The presence of substantial proteinaceous impurities (judging from the SDS-PAGE) at this stage dictated the use of a size-exclusion based "polishing step". Indeed, gel filtration chromatography resulted in removal of the majority of impurities, with several small peaks eluting prior to the main C7-CFF peak. However, care was a taken in the selection of fractions for further analysis as the first fractions eluted from the column produced smeary bands upon SDS-PAGE. These were thought to be attributable to residual N-glycosylated C7-CFF.

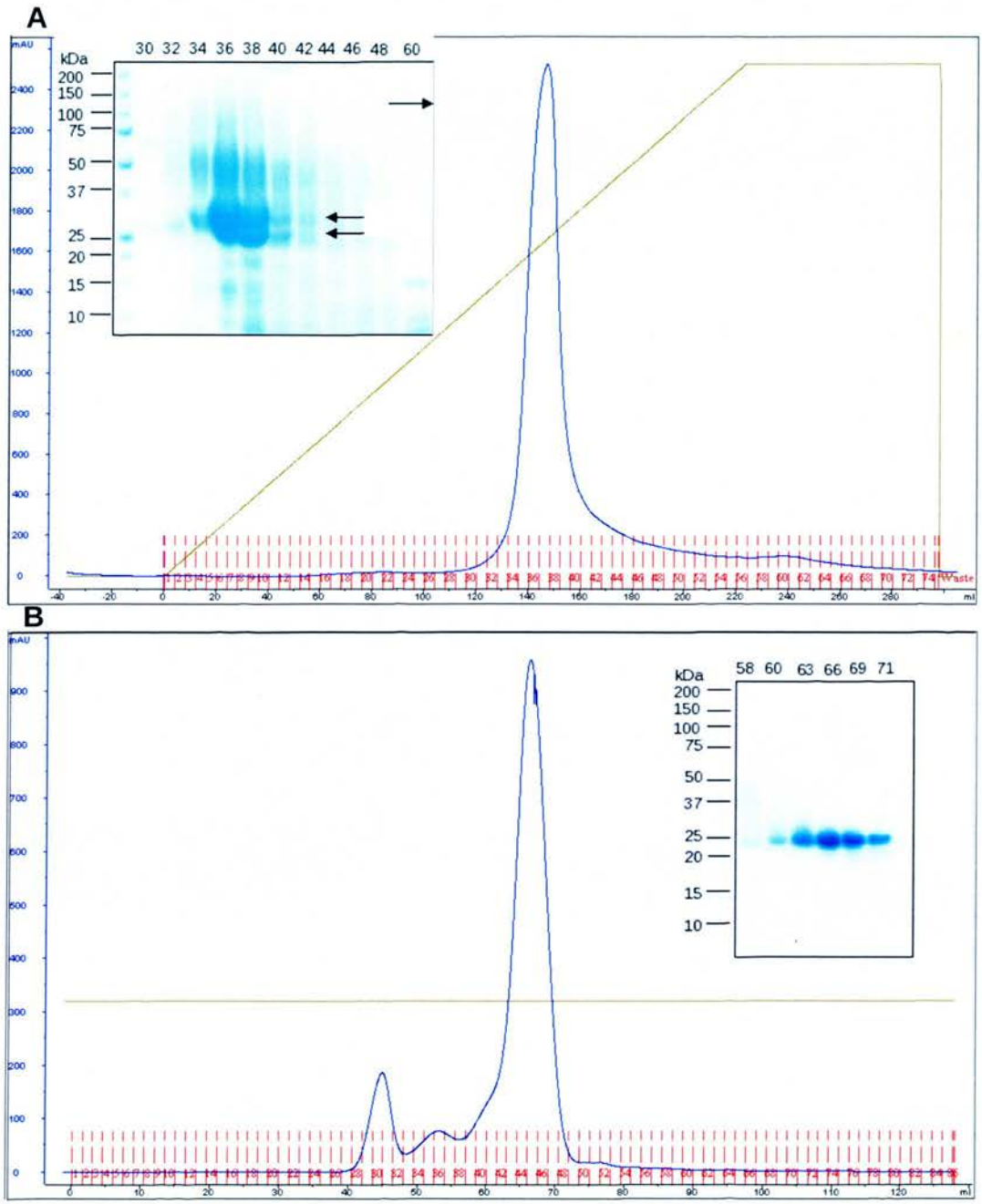


Fig. 39: C7-CFF chromatographic purification. (A) Cation-exchange chromatography of C7-CFF using a salt gradient (yellow line, 0-500 mM Nacl). SDS-PAGE analysis is shown with labels refering to elution fractions. (B) Size-exclusion chromatography of C7-CFF, with SDS-PAGE gel depicting elution fractions.

3.4.3 Characterization

Electrophoretic migration (Fig. 40B) of the C7-CFF was greater in non-reducing SDS-PAGE samples compared to reduced samples, indicative of the expected compact structure that is disrupted by reduction of the 11 intra-molecular disulphide bonds (two in the CCPs component, four in FIM1 and five in FIM2). The folding of the recombinant protein was further assessed by 1D NMR spectroscopy (Fig. 40A). Peaks within such spectra, in the regions around 6 ppm and 9 ppm, are indicative of tertiary structure.[120] To confirm whether the full-length C7-CFF had been successfully purified, samples were analyzed by mass spectrometry (see Appendix D). The recombinant protein has a predicted mass, including an N-terminal cloning artefact (EAEA) and residual *N*-Acetylglucosamine (220 Da, GlcNac from glycosylation), and with fully oxidized cysteines, of 23,982.3 Da (calculated in Protoparam).[33] LC-MS produced a mass of 23, 982.48 Da +/- 2.1, corresponding the fully oxidised mass.

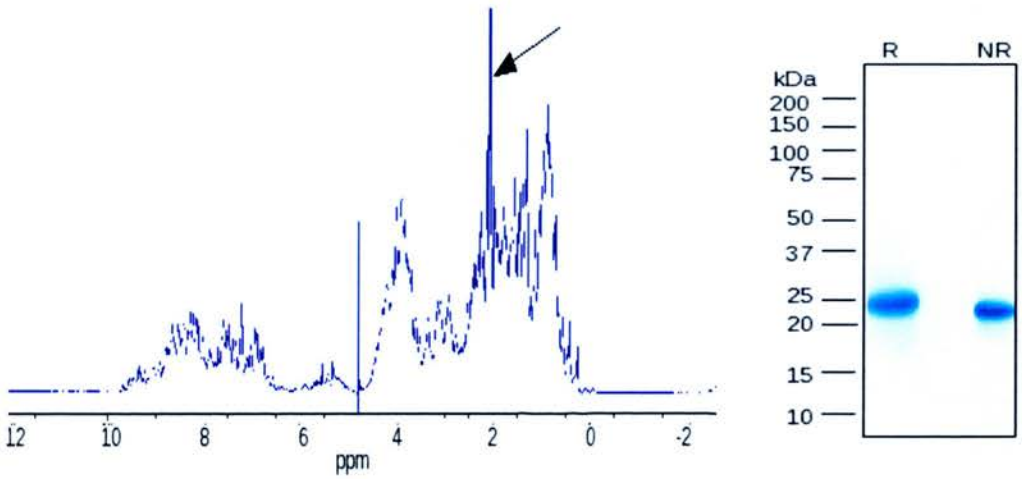


Fig. 40: C7-CFF characterisation. (A) Shows a ^{1}H-NMR spectrum with small-molecule impurities marked with an arrow. (B) Shows the movement of protein on the gel under reducing (R) and non-reducing (NR) conditions.

Aggregation issues had been previously reported for C7-FIMs, produced in OrigamiB at concentrations above 300 µM (personal communication). While the additional presence of the CCP module and/or production in a eukaryotic system may overcome these

aggregation issues, DLS was used to detect the presence of aggregates and also to determine the best storage conditions for C7-CFF. Size distribution profiles indicate that while a small amount of aggregate (or other large impurity) is present in relatively high concentration samples (> 1 mM), the percentage of total high-molecular weight "aggregate" is < 0.1%. Moreover, in further DLS experiments it was shown that the absence of NaCl in the buffer, as opposed to 0.2 M salt used routinely in NMR experiments, had no observable effect on the monodispersity of the protein, which is ideal for NMR studies.

Regarding storage conditions, storage for a week at 4 °C, -20 °C and -80 °C visibly altered the profile of the size-distribution graphs. For samples stored below 0 °C, the aggregate peak increased in intensity and its uniformity was lost. However the percentage of aggregate still remained minimal at ~0.1%. The Stokes radius or hydrodynamic radius R_H was calculated as 2.85 nm and the estimated molecular weight of 39 kDa (calculated using an empirical mass vs. size calibration curve). The high purity and monodispersity of the sample, along with the highly reproducible raw data (see Fig. 41) gave a degree of confidence to these often disregarded values. Thus the molecular weight overestimate is considered indicative of a non-globular, somewhat extended molecule, with a representative hard sphere of radius 2.85 nm (Stokes radius) diffusing at the same rate as the hydrated molecule, being much larger than the effective space the molecule likely occupies itself.

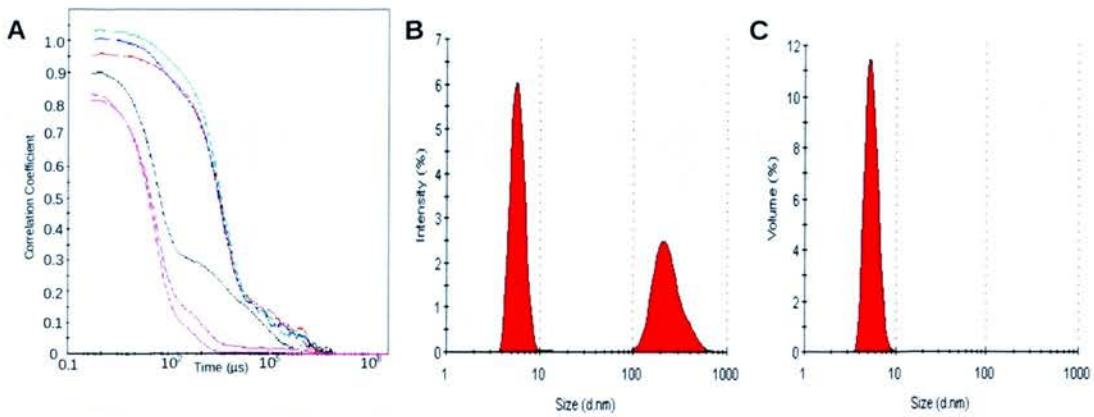


Fig. 41: DLS analysis of CFF. (A) Shows correlation coefficient curve with time. In order of best fit: 300µM and 1 mM samples (pink-solid and pink-dashed respectively), one month storage at 4¡C (black), freeze-dried (red), -20¡C storage (blue), -80¡C storage (green). (B) Shows the size distribution curve by signal intensity and (C) by intensity for the 1 mM, samples with no salt.

3.4.4 NMR optimization

As with C7-CCPs a ^{15}N-labelled C7-CFF sample was produced and purified for optimization of salt concentration, pH and temperature for NMR structural characterization. The larger size of the C7-CFF (24 kDa), and a possible extended module conformation as indicated by DLS, adversely effects the signal-to-noise ratio during NMR data acquisition, which could make full resonance assignment challenging without deuterating (triple labelling) the sample.[118] Thus optimising the NMR conditions turned out to be essential for prevention of overlap of amide proton-nitrogen correlations while maintaining the good signal-to-noise needed for successful resonance assignments.

Spectral folding was employed. Unfortunately, the inherently crowded spectra did not allow for as extensive a minimisation of the spectral width as was achieved for C7-CCPs (Fig. 42). As indicated by DLS, the global fold of CFF was not affected by changes in salt concentration, hence all buffer conditions during optimisation were devoid of salt to prevent signal damping, enhancing the signal to noise ratio. Regarding the choice of sample temperature (Fig. 43), peak intensity grew weaker with a decrease in

temperature, but 25°C was selected to maintain consistency with C7-FIMs and C7-CCPs NMR studies. A pH titration (Fig. 44) revealed that pH changes over the range tested did not significantly alter the appearance of the HSQC spectrum, indicating the overall fold of the protein backbone is unchanged. At pH 4.0 the best signal to noise was obtained and consequently more peaks could be resolved and identified in the crowded central portion of the HSQC. Moreover, at neutral pH values, initial signs of degradation were observed in natural-abundance samples that had been stored at room temperature for a month (data not shown). This phenomenon had been observed previously in the laboratory (for expression of CCP modules in *P. pastoris*) when it was also found that lowering the pH to 4.0 inhibited putative neutral *P pastoris* proteases. The stability of the backbone, resolution and dispersion of peaks, and enhancement of protein stability, as well as being far from the isoelectric point (6.4) indicated an optimum pH of 4.0. Spectra at pH 7.0 and pH6.0 were not recorded due to the proximity to the pI. Therefore the final conditions for NMR characterisation were 20 mM NaAcetate, no salt, and pH 4 at 25°C.

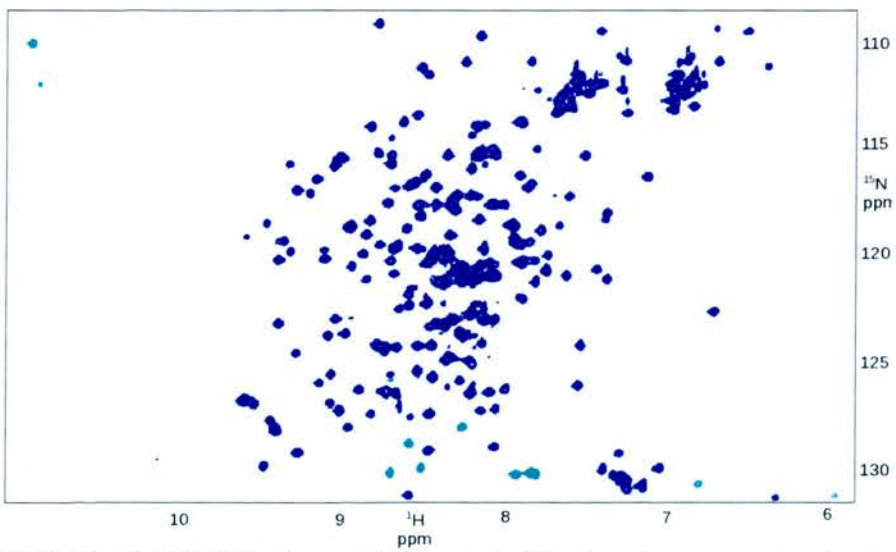


Fig. 42: [^{1}H,^{15}N]-HSQC of C7-CFF. Spectral folding in the ^{15}N dimension was employed to increase the resolution, with seven peaks (negative) observed in cyan.

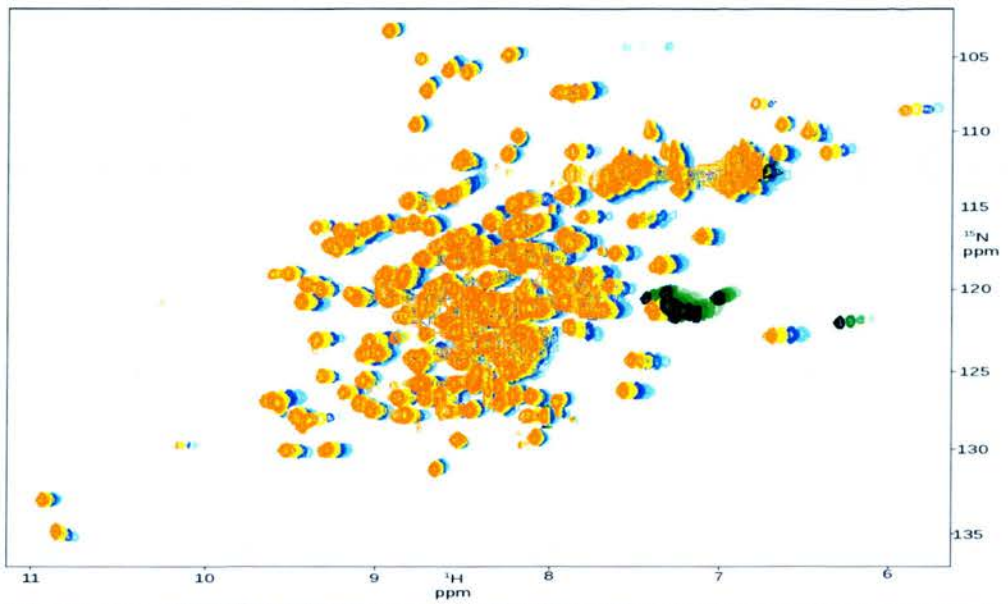


Fig. 43: Comparison of [^{1}H,^{15}N]-HSQC spectra of C7-CFF collected over a range of temperatures. Pale blue (15°C), blue (20°C), yellow (25°C), orange (28°C).

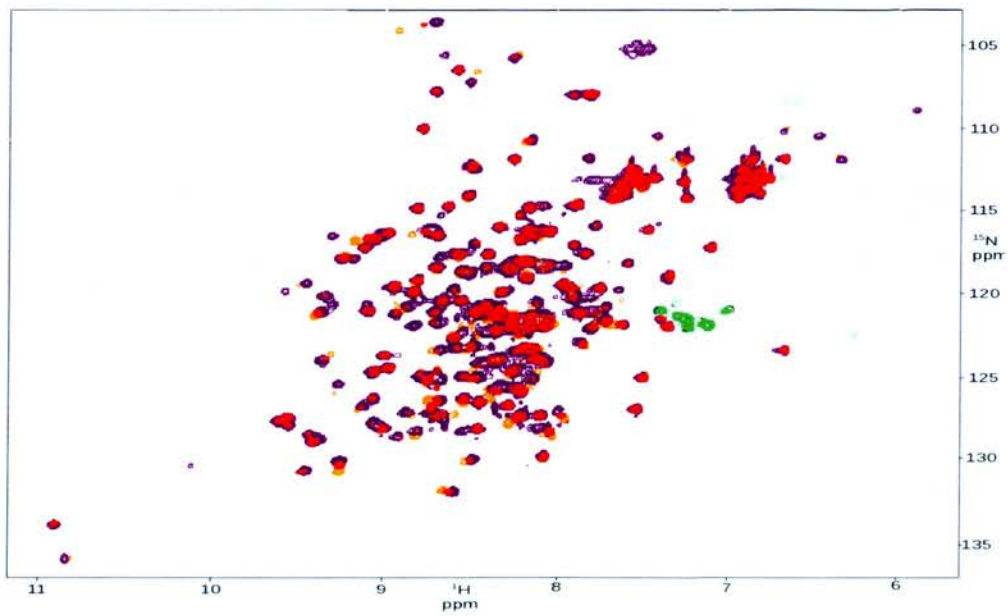


Fig. 44: Comparison of C7 CCP's' [^{1}H,^{15}N]-HSQC spectra at various pHs: pH 5.5 (orange), pH 5.0 (blue), pH 4.5 (purple), pH 4.0 (red).

NMR STRUCTURAL STUDIES

4.1 Overview

The C-terminal modules of C6 and C7 have been hypothesised to act as swinging molecular arms upon binding to C5b (or C5bC6). In order to investigate the structure and flexibility of the putative molecular arm of C7 (consisting of the following sequence of modules: EGF-TSPC-CCP1-CCP2-FIM1-FIM2) by NMR, a 'divide and conquer' strategy was adopted.[156, 157] Having successfully optimised purification of C7-CCPs (*i.e.* CCP1-CCP2) and C7-CFF (*i.e.* CCP2-FIM1-FIM2) produced in *E. coli* and *P. pastoris*, respectively, both proteins were produced as [^{13}C,^{15}N], double-labelled samples for characterization by NMR. The NMR assignment and the NMR-derived calculation of their 3-D structures was achieved on the basis of data collected using a conventional set of multidimensional double- and triple-resonance NMR experiments (see Methods 2.3.3). The larger size of C7-CFF (24 kDa), relative to C7-CCPs (14 kDa), and its extra module, justified additional analysis of the bigger protein by SAXS. Based on structures of C7-CCPs and C7-CFF, an initial model of C7-CCFF (*i.e.* CCP-CCP2-FIM1-FIM2) was produced. The backbone dynamics of both C7-CCPs and C7-CFF were probed by ^{15}N relaxation measurements. In addition, 2-D NMR was used to assess the overlapping module pairs C7-ET (*i.e.* EGF-TSPC) and C7-TC (*i.e.* TSPC-CCP1) (gifts from Dr R Ogata, Torrey Pines Medical Institute) to provide further valuable insights into the "shoulder" region of the molecular arm. The function insights affored by these structural studies are discussed in Chapter 6.

4.2 NMR-derived 3-D solution structure of C7-CCPs

4.2.1 NMR data

A standard suite of NMR experiments (outlined in section 2.3.3) was used for both backbone and side-chain resonance assignments, and to provide the NOE-derived distance restraints needed to calculate the 3-D solution structure of C7-CCPs. Excluding the N-terminal cloning artefact, the good quality of spectral data for C7-CCPs allowed for

completion of 95% of the backbone assignment and 95% of the side-chain assignment (Table. 10). The amide of Ile^{629} is the only residue missing from the assignment list. This residue was expected (on the basis of a sequence alignment) to occupy the short linking sequence between the two modules.

The extent of backbone amide assignment for C7-CCPs is illustrated in the [$^{1}H^{15}N$]-HSQC of Figure 45. This figure also shows the assignment of all Asn (four), Gln (four), Trp (two) and Arg (three) side-chain resonances. The NMR data allowed for the determination of a *cis* or *trans* conformation for each of the ten proline residues within C7-CCPs. This was achieved by visual inspection of Xaa-Pro cross-peaks and also by the difference between the Cβ and Cγ shifts (outlined in section 2.3.5.6). Pro^{644} and Pro^{681} were identified as the *cis* isoform by both methods and were therefore treated as such in structure calculations.

Category	Available	Assigned	% Assigned
Element C	596	529	88.76
Element H	748	714	95.45
Element N	163	130	79.75
Amide	247	232	93.93
Backbone	505	484	95.84
Backbone non-H	387	368	95.09
Side Chain H	630	598	94.92
Side Chain non-H	372	291	78.23
Residue Ala	5	5	100
Residue Arg	3	3	100
Residue Asn	4	4	100
Residue Asp	4	4	100
Residue Cys	8	8	100
Residue Gln	5	5	100
Residue Glu	9	9	100
Residue Gly	14	13	92.86
Residue His	3	3	100
Residue Ile	3	3	100
Residue Leu	9	9	100
Residue Lys	8	8	100
Residue Met	6	6	100
Residue Phe	5	5	100
Residue Pro	10	10	100
Residue Ser	11	11	100
Residue Thr	5	5	100
Residue Trp	2	2	100
Residue Tyr	3	3	100
Residue Val	12	12	100
All Residues	129	128	99.22

Table.10: Assignment report for C7-CCPs by CcpNmr Analysis. Table generated using the quality

reports function, and it includes the N-terminal GSHM cloning artifact.

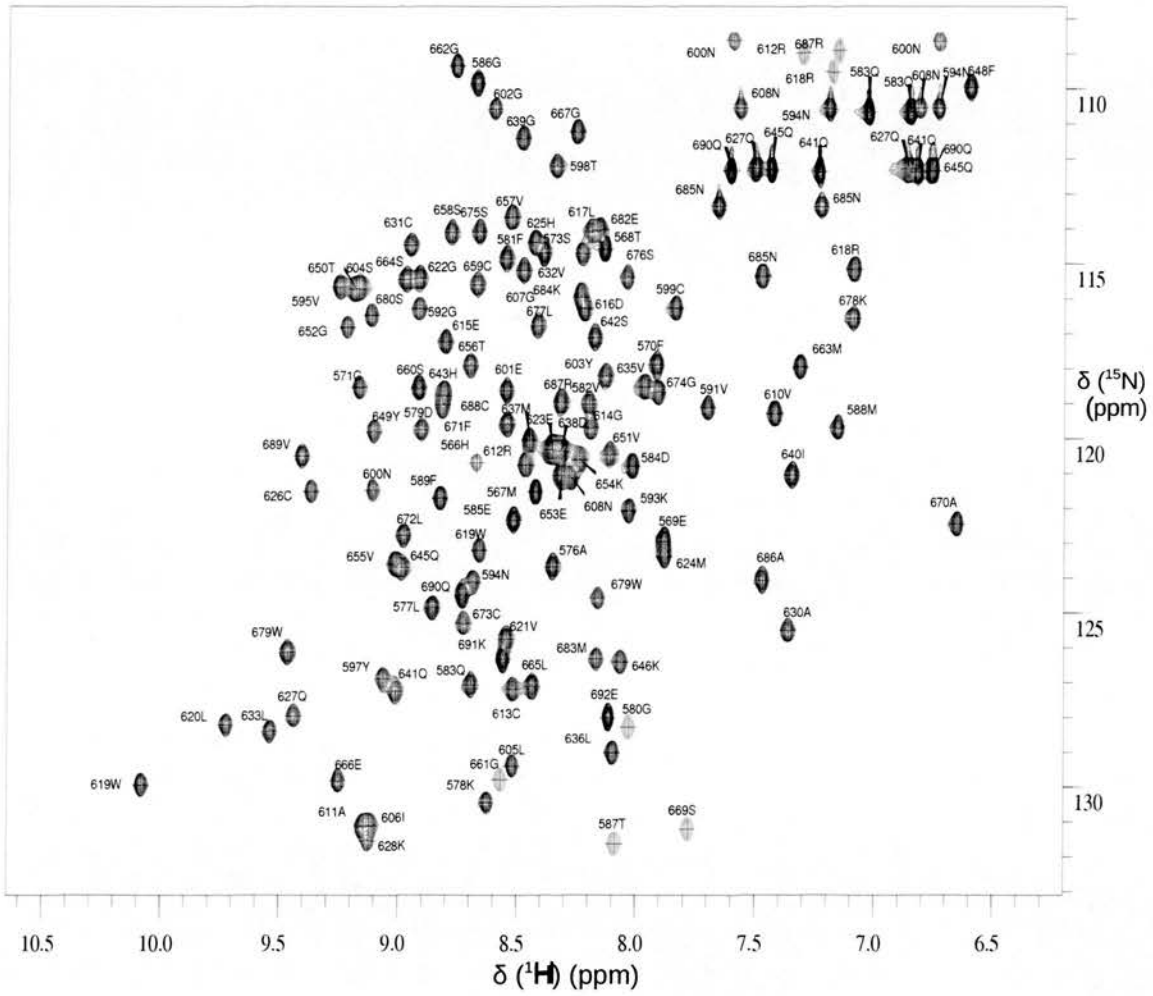


Fig. 45: [^{1}H^{15}N]-HSQC of C7-CCPs. Recorded on an 800 MHz (^{1}H frequency)) NMR spectrometer equipped with a cryoprobe. Backbone ^{1}H-^{15}N correlations are labelled sequence-specifically. Cross peaks connected by horizontal lines correspond to Gln and Asn side-chain amides, light grey cross peaks arise from folded Arg side-chain resonances, and Trp side-chain resonances are marked with an asterix.

4.2.2 Structure calculation

Table. 11 summarises the outcome of the automated assignment process *via* the CANDID algorithm in CYANA, as described in section 2.3.8.2 of the Methods. According to the published procedures[138] there are five criteria for measuring the reliability and success of a calculation performed in this way. All five of these were met in the calculation of C7-CCPs structures.

First, more than 90% of the ^{1}H chemical shifts should be assigned. In the case of C7-CCPs, this value was 95% (Table. 11) *i.e.* a total of 714 out of 748 ^{1}H atoms assigned in all.

Second, the cross peak list should be a "faithful representation" of the NOESY spectra, furnishing a large number of NOE cross peaks (4849 total and 3498 unique ones in the present case) for calculations.

The third criterion states that the average DYANA target function (as explained in section 2.3.8.2) should be <250 $Å^2$ in the first cycle and <10 $Å^2$ in the final of the seven cycles, with ≥80% of all the originally picked NOESY peaks assigned and ≤20% of initial long-range assignments being eliminated during the calculation. The target function should approach zero as the available experimental and torsion angle restraints become satisfied, with minimal steric overlap between non-bonded atom pairs.[138] In the case of calculation of the C7-CCPs structure, the target functions are well within limits, with 87% of all NOESY peaks assigned and no long-range assignments eliminated over the course of seven cycles.

The fourth criterion is that in the first cycle, the average backbone root mean square deviation from the mean (RMSD) of the calculated ensemble of structures should be

smaller than 3.0 Å after the first cycle. For C7-CCPs structures of the first cycle, this value was higher, at 5.0 Å for both modules. But it was 1.6 Å for each of the individual modules (calculated in MOLMOL[142]). The large RMSD for the CCP-CCP pair in the first round arose mainly from tilt variance between the two modules. Two bent conformations predominated: a relatively 'open' conformation and a more 'closed' conformation where the modules were brought into close proximity along the length of their long axis. No NOEs, however, were found in support of the more closed conformations, either by manual inspection of the NOESY spectra or by subsequent rounds of CYANA calculations during which the closed conformation was eliminated. As all other criteria fit, the backbone RMSDs of individual modules was less than <3 Å, and the overall backbone RMSD was progressively reduced in subsequent cycles, this was not considered to detract from the validity of the final calculated structures.

Finally the RMSD "drift" between the first and last cycle indicates the extent to which the first cycle captures the "true" polypeptide fold. This is encapsulated by the criterion stating that the RMSD between the mean structures of the first and last cycle should be below 3 Å, and this was easily met with an RMSD of 1.3 Å.

Cycle	1	2	3	4	5	6	7	Final
Peaks:								
selected	4849	4849	4849	4849	4849	4849	4849	
assigned	4423	4450	4326	4335	4280	4224	4207	
unassigned	426	399	523	514	569	625	642	
with diagonal assignment	1	1	1	1	1	1	1	
Cross peaks:								
with off-diagonal assignment	4422	4449	4325	4334	4279	4223	4206	
with unique assignment	1195	2702	3106	3162	3356	3491	3498	
with short-range assignment |i-j|<=1	2874	2804	2718	2697	2656	2607	2608	
with medium-range assignment 1<|i-j|<5	560	486	424	436	419	414	414	
with long-range assignment |i-j|>=5	988	1159	1183	1201	1204	1202	1184	
Upper distance limits:								
total	3262	3096	2894	2862	2768	2675	2756	2844
short-range, |i-j|<=1	1729	1570	1454	1395	1335	1275	1236	1280
medium-range, 1<|i-j|<5	917	816	403	409	390	375	384	395
long-range, |i-j|>=5	616	710	1037	1058	1043	1025	1136	1169
Average assignments/constraint	5.59	2.35	1.38	1.37	1.29	1.22	1	1
Target Function								
Average target function value	178.21	83.38	178.06	27.64	14.17	5.2	7.22	2.43
RMSD (residues 564..692):								
Average backbone RMSD to mean	5.64	3.15	1.34	1.23	1.03	1.13	1.1	1.01
Average heavy atom RMSD to mean	6	3.4	1.53	1.47	1.28	1.34	1.3	1.21

Table. 11: Report on structure calculation of C7-CCPs by CYANA. Statistical tracking of cross peaks picked, assignments to specific NOEs, distant-restraint deduction, target function values and RMSD values are shown (over all CANDID cycles).

Following completion of all seven cycles, a total number of 3498 unique NOE-derived restraints were generated. As shown in Table. 11, these constitute 2608, 414 and 1184 upper-distance bounds corresponding to intra-residue, medium-range (between residue i and residues (i+1 to i+4)), and long-range (between residue i and residues i+(> 4)) restraints, respectively. The Format Converter of the CcpNmr software[135] was then utilised to convert the list of upper-distance restraints generated by CYANA into a format compatible for subsequent CNS structure calculations. This was carried out because CNS has been found (according to several members of the Barlow group, personal communication) to provide superior water refinement in the final set of calculations.

The 100 structures thus re-calculated using scripts in CNS were ranked in order of NOE energy and in order of total energy (Fig. 46). The 40 structures with lowest-NOE energy and lowest overall energy converged well. However inflection points for structures 27 and 35 prompted a smaller selection, involving just the 20 lowest-energy structures for refinement in water solvent. The final 20, water-refined, NMR structures overlay very well, with low RMSD values for both the CCPs individually, and a respectable RMSD for the modules pair (Fig. 47).

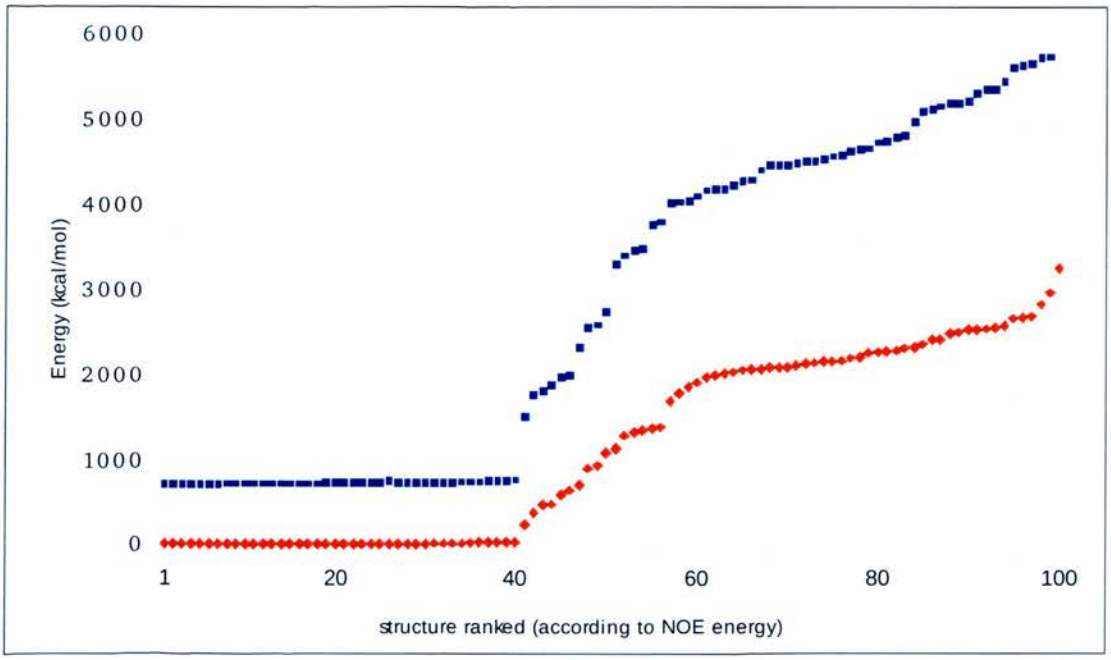


Fig. 46: Energy plot of the final, ranked 100 CNS-re-calculated C7-CCPs structures. The NOE energy is shown in red triangles and the total energy is shown in blue squares.

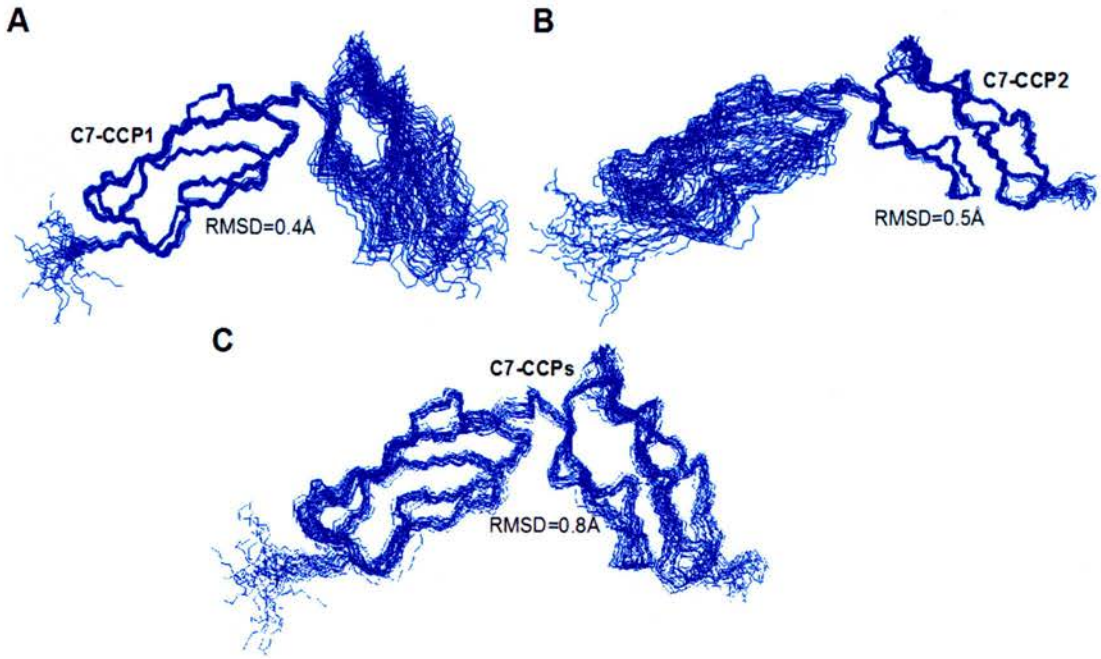


Fig. 47: Backbone overlay of the ensemble of 20 water-refined C7-CCPs structures. (A) Backbone overlay on CCP 1; (B) Backbone overlay on CCP 2; (C) Backbone overlay for both modules. MolMol[ref] was used for visualisation and calculation of the RMSDs. RMSD values are as indicated.

4.2.3 Structure description and quality analysis

Structure determination reveals that both CCP modules of C7-CCPs have archetypal CCP-module characteristics, although CCP2 also displays some atypical features (Fig. 48). Combinatorial extension[149] was employed (Dinesh Soares, University of Edinburgh) to compare each experimentally determined C7-CCP structure within the current set of 48 atomic resolution CCPs structures within the complement system (see Appendix E). Comparisons with the closest-to-mean individual structures of C7-CCP1 and C7-CCP2 are most similar to decay accelerating factor (DAF)-CCP1[158](RMSD=1.6Å) and the recently solved mammalian control protein (MCP)-CCP3[159](RMSD=2.2Å) respectively (Fig. 48). Out of all 48 known structures only 3 for CCP1 and 11 for CCP2 had backbone RMSDs greater that 3.0Å indicating that the structures are not deviant modules.

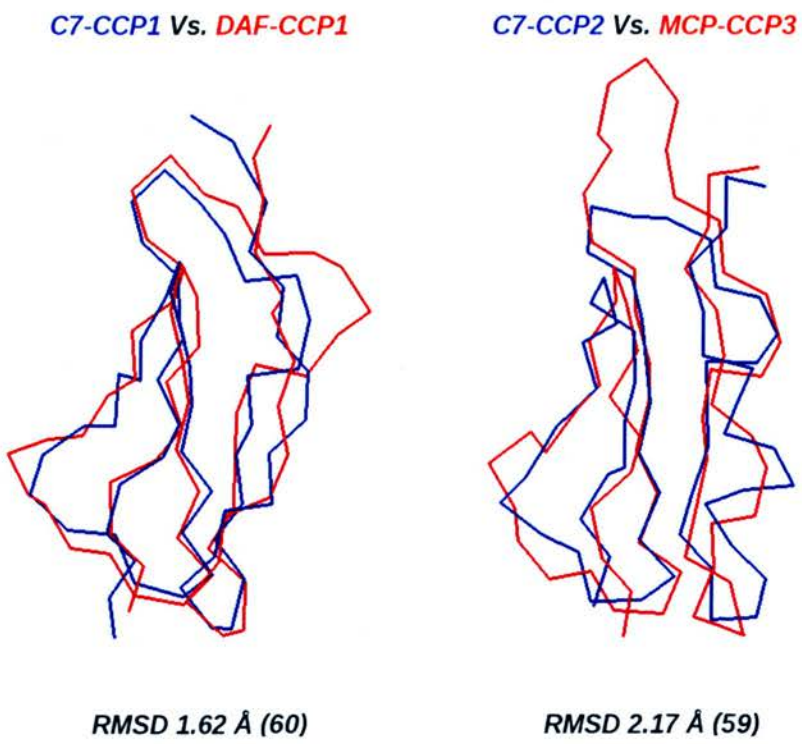


Fig. 48: Comparisons of C7-CCP1 and 2 with the most similar atomic structures in complement. Cartoon respresentations (MOLMOL[142]) showing a backbone overlay of C7-CCP1 and 2 (blue) and DAF-CCP1 and MCP-CCP3. RSMDs of the overlay are shown as well as the number of residues used in the comparison in brackets.

Overall CCP2 appears more spherical (and less ovoid) than CCP1 and is ~5 Å shorter (30 Å versus 35 Å). Both are stabilised by a cross-brace of Cys(I)-Cys(III) and Cys(II)-Cys(IV) disulfide linkages that occur close to either end of each module. Each module features five elongated stretches of amino acid residues; these pass back and forth in approximate alignment with the long axis of the prolate module and are linked together by turns and loop that occupy the poles of the module. Segments of the five extended stretches from β-strands for parts of their lengths. These β-strands in turn form small anti-parallel β-sheets in a sandwhich-like arrangement (Fig. 49A, B). The β-strand network (as defined by the criteria for secondary structure embedded within the program

MOLMOL[142]) appears more extensive in CCP1, which exhibits six (B, D, E, F, G and H) of the maximum (across the CCP module family) eight strands (A-H),[94] while CCP2 has only three regions that are classified as β-strands (B, D and F). CCP2 additionally contains a canonical three-residue 3_{10}-helix[1*] within the C-terminal of its five extended stretches of residues (occupying a region that lies between cannonical strands G and H and often forms a bulge in other CCP modules).[94] This feature was present within all 20 members of the final ensemble of calculated C7-CCPs structures.

Like most CCP modules, both modules in C7-CCPs have hypervariable loops that occur immediately after strands B (in CCP1: Q^{583}DEGPM^{588} and in CCP2: H^{643}PQKPF^{648}). The hypervariable loop of CCP2 converges better across the ensemble of 20 lowest-energy structures than that of CCP1; this is also reflected in the approximation of motion in this region (see section 4.2.4) derived from ^{15}N relaxation measurements. To obtain a summary of the angles of orientation of CCP1 with respect to CCP2, average values of tilt, twist and skew (defined as in Fig. 49C) were calculated for the final ensemble of 20 structures[ref] (Fig. 49C). The C7-CCPs have a 84° tilt between the modules, are twisted by 20° with respect to one another, and display only a small mutual skew angle.

1* In 3_{10} helices the carbonyl group in residue i and the nitrogen of the amide group in residue i+3 are hydrogen bonded. Canonical 3_{10} helices have three residues per turn, with an angle of 120° between consecutive residues, a helical rise per residue of 1.93–2.0 Å, and a helical pitch of 5.8–6 Å. In very simple terms, a 3_{10} helix is more tightly wound, longer, and thinner than an α helix with the same number of residues.

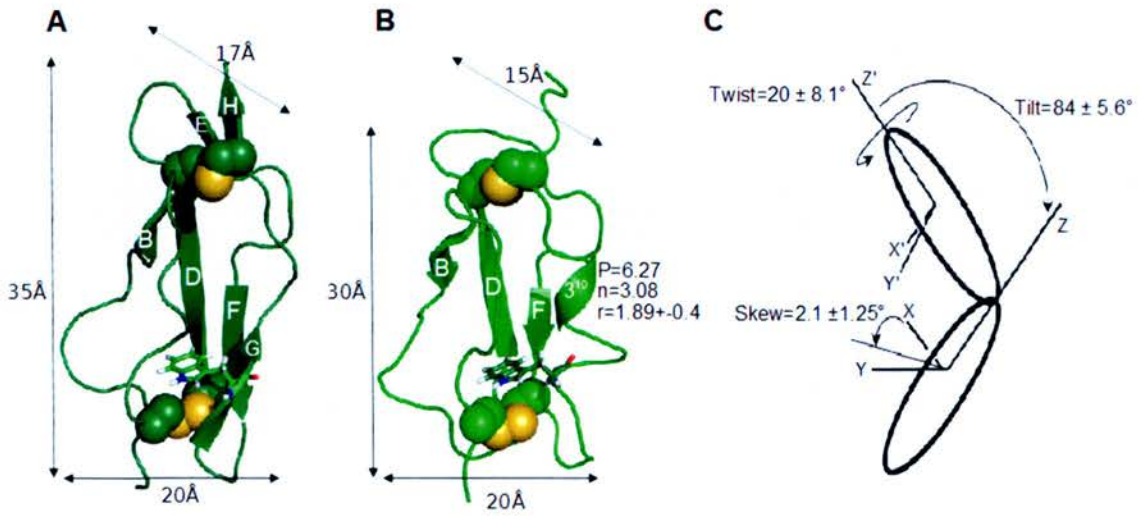
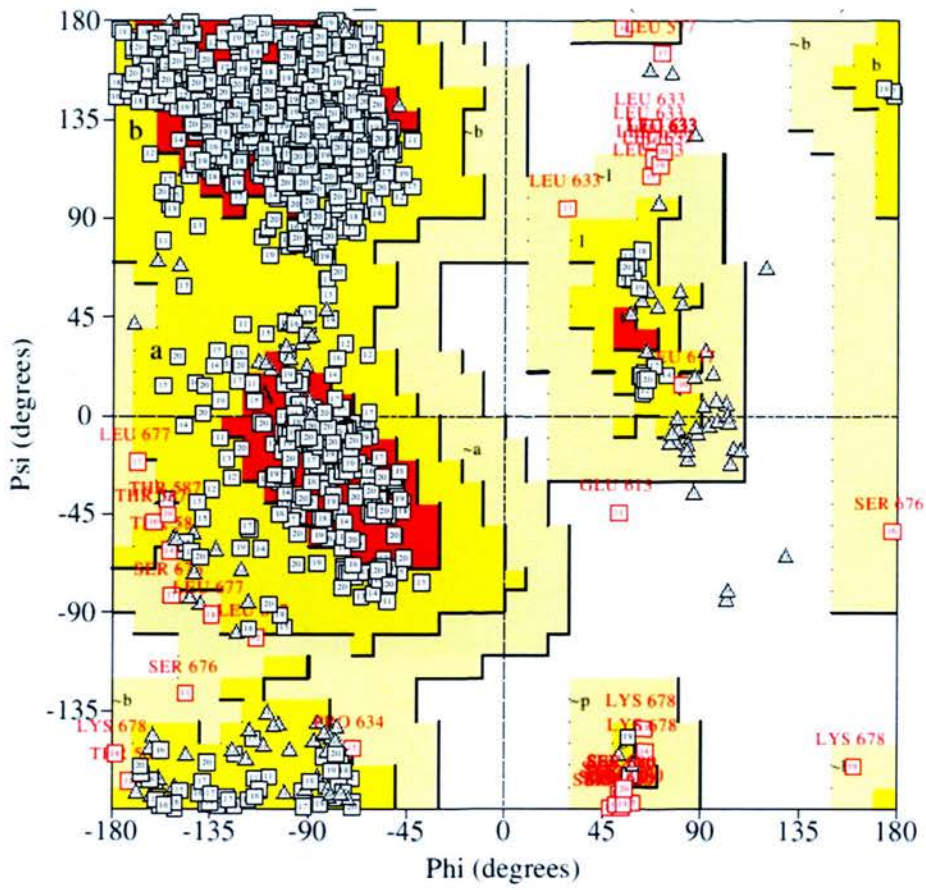


Fig. 49: Cartoon representation C7-CCPs' 3-D structure . (A) CCP1. (B) CCP2. β-strands are labelled and cysteine side chains atoms are shown as spheres. CCP dimensions are shown; average values for pitch (P), number of turns (n) and radius (r, Å) of the 3_{10} helix are also shown. (C) Average values of intermodular twist, tilt and skew calculated using XYZ as shown previously.[160]

The WHATIF "course packing quality score" was used to analyse the C7-CCPs ensemble.[145] The score gives an assessment of the normality of the local environment of atoms surrounding the individual amino acids. A WHATIF score lower than -5.0 is indicative of improper packing. The average score of the ensemble was well above this value both before (-2.2) and after (-1.5) water refinement by CNS. Subsequently, the program PROCHECK[144] was used to evaluate the ensemble. PROCHECK evaluates the stereochemical quality of protein as indicated by the extent of occupation by ϕ and ψ angles of energetically favoured regions of a Ramachandran plot (Fig. 50). 97.5% occur in the most favoured and additionally allowed regions of the Ramachandran plot, a finding which suggests reasonable stereochemical quality suggestive of reasonable stereochemical quality.



Plot statistics

Residues in most favoured regions [A,B,L]	1526	73.4%
Residues in additional allowed regions [a,b,l,p]	501	24.1%
Residues in generously allowed regions [~a,~b,~l,~p]	38	1.8%
Residues in disallowed regions	15	0.7%
Number of non-glycine and non-proline residues	2080	100.0%
Number of end-residues (excl. Gly and Pro)	20	
Number of glycine residues (shown as triangles)	280	
Number of proline residues	200	
Total number of residues	2580	

Fig. 50: Ramachandran statistics (from PROCHECK) for water-refined ensemble of the 20 lowest energy structures of C7-CCPs. Each small box represents a residue and the model number, save for glycines and prolines which are shown as triangles. Outliers are named and shown in red.

4.2.4 Relaxation Analysis

T_1 and T_2 relaxation time constants were collected for the [^{13}C,^{15}N]-C7-CCPs sample. Incremental relaxation delays used for T_1 measurements were 51.2, 301.2, 501.2, 701.2, 901.2, 1001.2, 1101.2 and 51.2. While for T_2 they were 16.96, 33.92, 67.84, 84.8, 101.76, 118.72, 135.68 and 16.96. Although extensive relaxation analysis of the protein was beyond the scope of this project, the values of T_1 and T_2 , the derived ratios T_1/T_2, and the [^{1}H-^{15}N] heteronuclear NOE measurements together provide useful insights into C7-CCPs backbone dynamics (Fig. 51). Weak peaks or those in overcrowded regions were excluded from analysis, due to their associated large error values. The highly flexible cloning artifact was excluded from analysis, and the first two and last two residues were not considered in the calculation of averages.

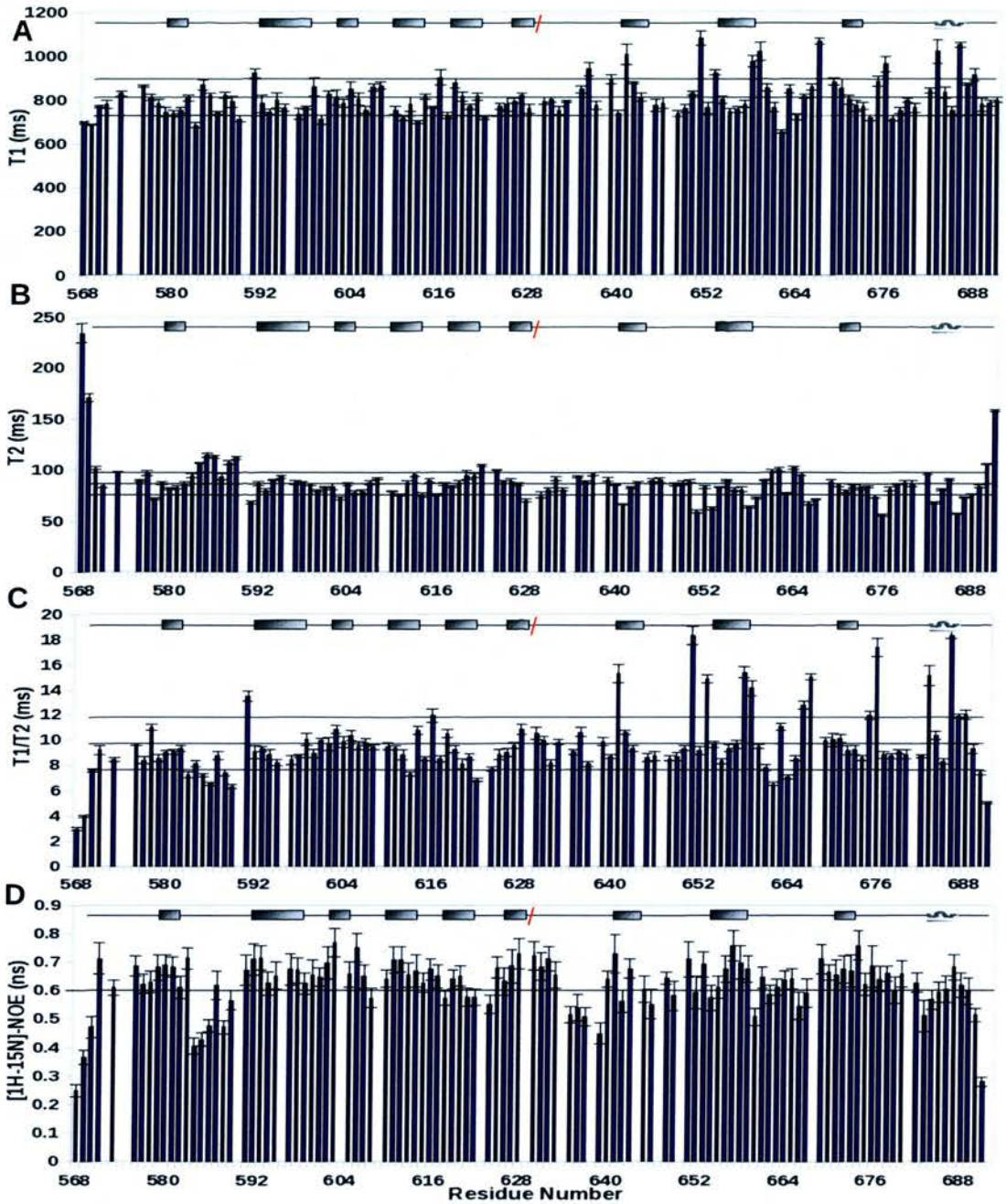


Fig. 51: Backbone amide ^{15}N relaxation measurements for C7-CCPs. From top to bottom: (A) T_1 values and associated errors at 14.1 Tesla. (B) T_2 values and associated errors at 14.1 Tesla. (C) T_1/T_2 ratios and associated errors. (D) [^{1}H-^{15}N]-heteronuclear NOE values and associated errors at 14.1 Tesla. Lines for the average and +/- one standard deviation are shown for (A), (B) and (C), the 0.6 ns NOE cut off is shown in (D). Schematic shows secondary structure with red slashes defining module boundaries.

Low T_1/T_2 values indicate backbone flexibility on the second-millisecond timescale and therefore represent relatively large and relatively slow conformational changes. Heteronuclear NOE values <0.6 indicate flexibility on the picosecond-nanosecond timescale, representing backbone librations and side-chain motions. Residues displaying flexibility on either of these timescales were mapped onto the lowest-energy C7-CCPs structure to better visualise their spatial location and distribution (Fig. 52).

As expected from the number of distance restraints between the intermodular linker residues and the modules themselves, the relaxation data show that linker residues have average backbone mobility for this molecule, *i.e.* flexibility on the time scales discussed above that is equivalent to residues located, for example, in β-sheets. The most flexible region in C7-CCPs is the hypervariable loop of CCP1, between β-strands B and D (Q^{583}DEGTMF^{589}). In this loop there is backbone mobility on both the ms and ps-ns timescales, with only one residue, Thr^{587}, having average mobility. In contrast, the hypervariable loop of CCP2 appears to lack flexibility. The last of the five stretches of residues within CCP1 also shows greater than average flexibility.

Many residues in CCP2 may be considered inflexible (on the slow timescale) on the basis of their high T_1/T_2 ratios. Half of these residues were found in regions of secondary structure (β-strands B and D and the 3_{10} helix. Some of these residues collected at the N-terminal end of the module, with side-chains engaged in hydrophobic interactions with one another (Lys^{654}, Val^{655}, Phe^{671}, Trp^{684}). Other residues that appear to be stable on the millisecond-second timescale are found in δ-turns and γ-turns[2*]. CCP2 possesses a few regions that are more flexible than average, in the first stretch of residues and in loop segments between strand D and F. Interestingly, flexibility on the nanosecond-

2* A turn in a loop segment is a structural motif in which two residues, separated by one to five residues, have their Cα atoms brought into close approach (<8Å). Hydrogen bonds are not necessary, but one or two may be present. Separation of two end residues by one, two, three, four and five residues denote a α, β, γ, δ and π-turn respectively.

picosecond timescale is observed in the lysine residue of the $M^{683}KN^{685}$ 3_{10} helix within CCP2, despite rigidity on the slower timescale and these residues forming H-bonded secondary structure.

In conclusion, although the modules have localised regions of flexibility in loops and hypervariable regions, the majority of residues, including those in the linker, are rigid on the timescales investigated.

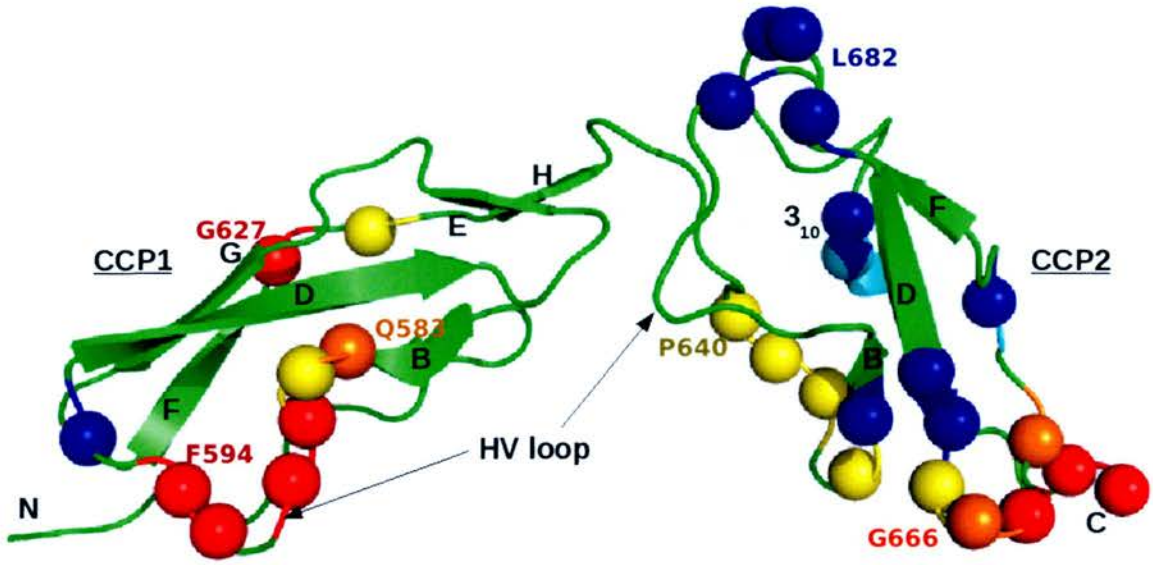


Fig. 52: A summary of backbone dynamics in C7-CCPs. Residues falling within average values for T_1/T_2 and [^{1}H-^{15}N] heteronuclear NOEs are shown in green. Residues deemed flexible on the ns-ps timescale, from their heteronuclear NOE values, are shown as yellow spheres. Residues deemed flexible, or stable, on the ms-s timescale from their T_1/T_2 ratio are shown in orange and blue respectively. Residues exhibiting flexibility on both timescales are shown in red.

4.2.5 Analysis of the intermodular interface

A degree of intermodular rigidity between CCP1 and CCP2 is reflected in the relatively low RMSD of the superposition performed over both modules of the pair (= 0.8 Å), which in turn is a reflection of the relatively high number of NOEs assigned between different modules (Table. 12) and also of NOEs between the body of each module and

the linker (84 and 21 for CCP1 and CCP2, respectively). Intermodular rigidity is also consistent with the relatively short intermodular linker (four residues between the last cysteine of one module and the first of the next) and its partly hydrophobic nature with three **large** side chains (**Q**^{627}**KI**A^{630}). A relatively rigid linker is also consistent with the lack of any unusual features of the relaxation data for this region (see section 4.2.4). Indeed a rigid bend is consistent with the intimate associations between amino acid residues within the intermodular junction formed by the C-terminal end of CCP1, the linker residues and the N-terminal loops and turns of CCP2. The side-chain of Gln^{627} within the linker is solvent-exposed and is essentially the final residue of CCP1 as it is part of the last β-strand, H, that is held in antiparallel β-sheet arrangement with E of the previous strand. Indeed the linker, a patch of residues within CCP1 (E^{601}GYS^{604}) and two patches of residues in CCP2 (Cys^{631},Val^{632} and F^{648}YT^{650}) participate in an extensive H-bonded network, comprising two H-bonds between the modules, six linker-CCP1 H-bonds and two linker-CCP2 H-bonds. This intimate H-bonded network helps to ensure a buried surface of 440 Å between the modules (calculated using GETAREA,[148] Fig. 53). The end result is that Gln^{627}, Lys^{628} and Ile^{629} (see Fig. 53B) are closely associated with CCP1, while Ile^{629} and A^{630}, are closely associated with CCP2.

CCP1			CCP2			
Residue No	Residue type	Atom type	Residue No	Residue type	Atom type	Distance (Å)
600	ASN	HB3	632	VAL	QQG	4.78
600	ASN	HB2	632	VAL	QQG	4.96
600	ASN	HB3	632	VAL	QQG	4.78
600	ASN	HB2	632	VAL	QQG	4.96
601	GLU	QB	632	VAL	QQG	3.53
601	GLU	QB	648	PHE	QD	4.83
601	GLU	QB	632	VAL	QQG	3.53
601	GLU	QB	648	PHE	QD	4.83
602	GLY	HA3	632	VAL	QG1	4.52
602	GLY	H	648	PHE	QD	5.43
602	GLY	HA3	632	VAL	QG2	4.52
602	GLY	HA2	632	VAL	QQG	4.18
602	GLY	HA3	632	VAL	QQG	3.78
602	GLY	HA3	648	PHE	HB3	5.15
602	GLY	HA3	648	PHE	HA	5.5
602	GLY	HA3	632	VAL	QG1	4.52
602	GLY	HA3	648	PHE	HA	5.5
602	GLY	HA3	632	VAL	QG2	4.52
602	GLY	HA2	632	VAL	QQG	4.18
602	GLY	HA3	648	PHE	HB3	5.15
602	GLY	HA3	632	VAL	QQG	3.78
602	GLY	H	648	PHE	QD	5.43
603	TYR	QD	632	VAL	QQG	4.52
603	TYR	QE	632	VAL	QG1	5.13
603	TYR	QE	632	VAL	QQG	4.15
603	TYR	QD	632	VAL	QG2	5.5
603	TYR	QD	632	VAL	QG1	5.5
603	TYR	H	632	VAL	QQG	5.33
603	TYR	QD	632	VAL	QG2	5.5
603	TYR	QE	632	VAL	QQG	4.15
603	TYR	QD	632	VAL	QQG	4.52
603	TYR	H	632	VAL	QQG	5.33
603	TYR	QE	632	VAL	QG2	5.13
603	TYR	QE	632	VAL	QG2	5.13
603	TYR	QD	632	VAL	QG1	5.5
603	TYR	QE	632	VAL	QG1	5.13
628	LYS	HB3	632	VAL	QQG	5.44

Table. 12: Intermodular CCP1-CCP2 NOEs. Linker residues $Q^{627}KIA^{630}$ were excluded from analysis. The distance between two atoms was determined within CYANA.

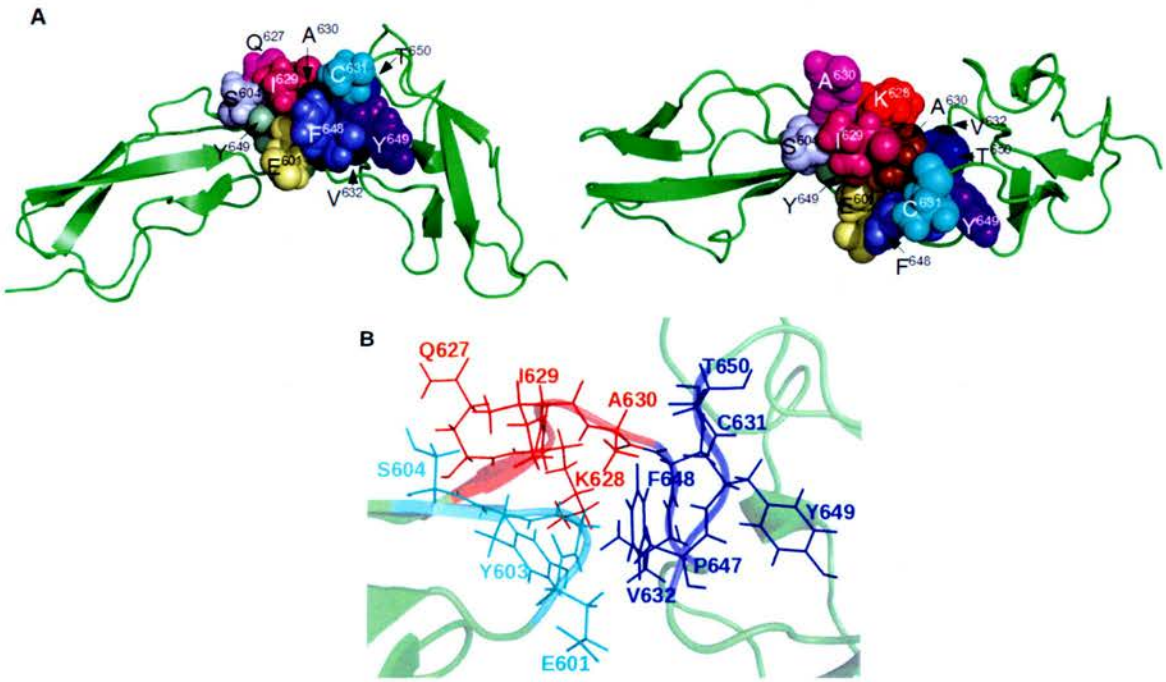


Fig. 53: C7-CCPs intermodular interface. (A) Interface shown as spheres from two views rotated by 180 ° about the y-axis. (B) A close-up of the interface residues shown as lines, CCP1 residues are in cyan, linker residues are in red and CCP2 residues are in blue.

4.2.6 Analysis of surface properties

A surface-electrostatic analysis (Fig. 54) of C7-CCPs reveals a strikingly negative concave face of the 84°-bend. In contrast the convex face presents a mix of positive, negative and neutral patches. Analysis of the lipophilicity (Fig. 55) of the surface reveals a pronounced hydrophobic patch near the N-terminal end of CCP1 (Pro^{591}, Phe^{590}, Phe^{581}, Met^{589}) Using the program STP this patch and two others (one on CCP1 and another on CCP2, Fig. 56) were identified as potential hotspots for protein-protein interaction. Interestingly, these hotspots accumulate on one face of the molecule that is between the convex and concave surfaces (*i.e.* on the "side" of the molecule). Thus, while the convex and concave faces of C7-CCPs could remain solvent exposed due to their charged and hydrophilic surface the side of the molecule carrying the three protein-protein interaction hotspots might be buried. It is unknown, however, whether these

hotspots are involved in interactions within C7 or with other proteins of the MAC complex.

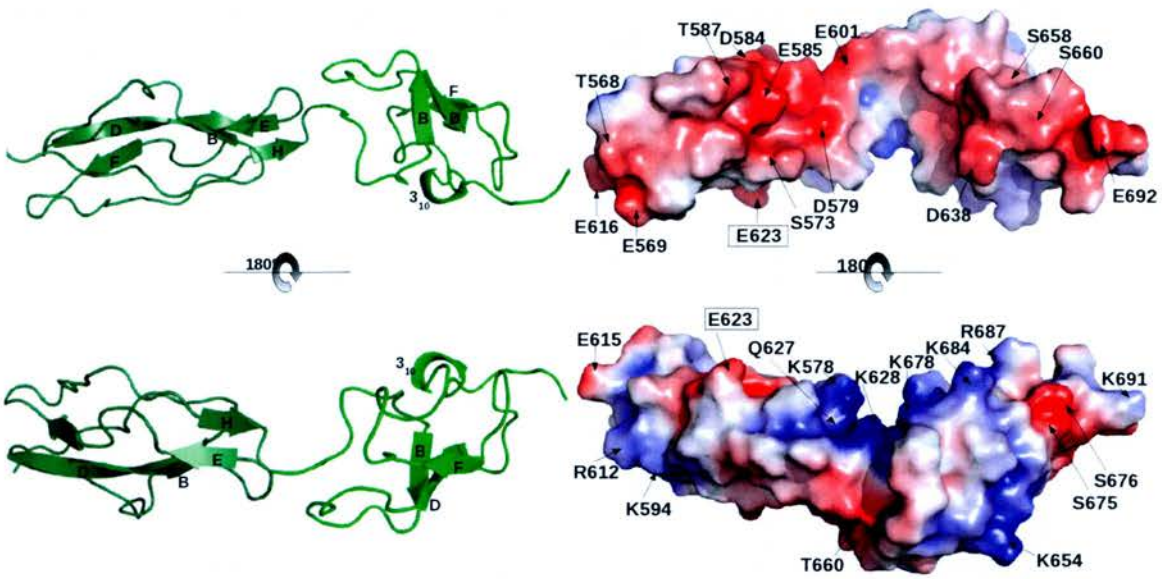


Fig. 54: Analysis of C7-CCPs surface electrostatics. Diagrams on the left are cartoon representations of CCP1 (dark-green) and CCP2 (light-green) with secondary structure as determined by Pymol.[143] Surface-electrostatics are shown with a red (negative charge), white (neutral) and blue (positive charge) colour scale, ranging from -5.0 to +5.0 k/T. The top view shows the concave face of the module pair and the bottom view shows a 180° rotation of the convex face. Glu^{623} is boxed in both views.

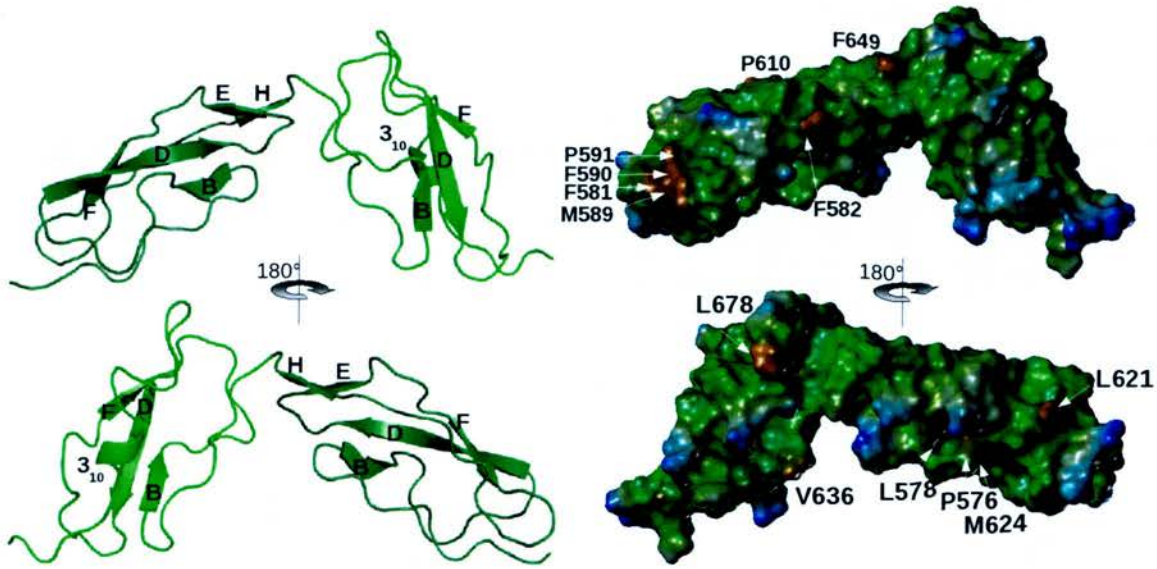


Fig. 55: Analysis of C7-CCPs surface lipophilicity. Diagrams on the left are cartoon representations of CCP1 (dark-green) and CCP2 (light-green) with secondary structure as determined by Pymol.[ref] Lipophilicity is show on a scale from tan (hydrophobic) to green (neutral) to blue (lipophobic).

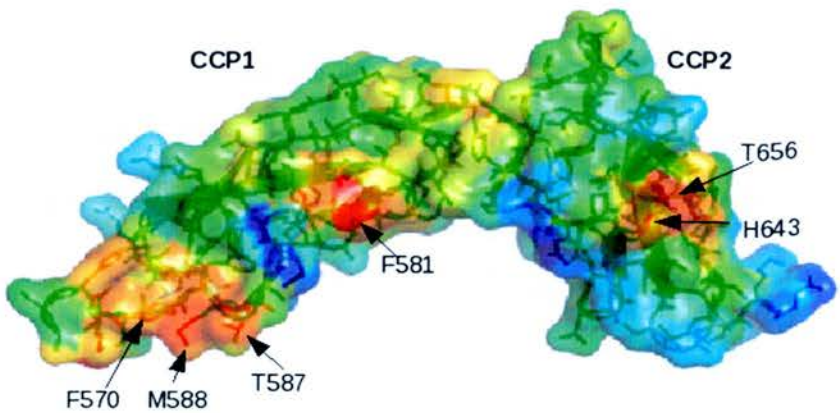


Fig. 56: Identification of ligand-binding sites using "surface triplet propensities". A surface representation of C7-CCPs with a scale going from least favourable (blue), to neutral (green), to most favourable (red) interaction site scores.

4.3 NMR structure of C7-CFF

4.3.1 Data

The same standard suite of NMR experiments used for C7-CCPs (outlined in section 2.3.3) were also used for both backbone and side-chain resonance assignments, and in

order to compile the list of NOE-derived distance restraints used as a basis for calculating the 3-D solution structure of C7-CFF. However, the larger size of C7-CFF (219 residues as opposed to 129 residues in C7-CCPs) resulted in inherently worse quality spectra than were obtained for C7-CCPs. The hydrodynamic radius (see section 2.2.10, DLS) of C7-CFF equates to a ~40-kDa protein, and when this value is compared to the MS-calculated Mwt of ~24 kDa, a somewhat extended molecule is indicated. That C7-CFF has a slower tumbling rate than C7-CCPs is borne out by its shorter average T_2 relaxation times. In fact these are not consistent across the modules of C7-CFF since they are significantly shorter in the FIMs portion than in the CCP2 portion - see section 4.3.4 for a further analysis of relaxation. The shorter relaxation times reduce the sensitivity of those pulse sequences that require long delays for coherence transfer steps.[118] This, coupled with the increased number of amino acid residues in CFF, gives rise to more complex, crowded and noisy spectra. Maximum-entropy processing was therefore employed to enhance the resolution of the spectra where necessary (see section 2.3.4). The ^{13}C-HSQC-NOESY spectra[133] were particularly overcrowded and noisy in some regions and therefore picking of NOESY peaks was shared between four spectra: a ^{13}C-HSQC-NOESY (in D_2O), the same ^{13}C-HSQC-NOESY but maximum-entropy processed, a ^{13}C-HSQC-NOESY (in H_2O), and a ^{13}C-methyl-NOESY-spectra[134] that shows strips for methyls only.

Despite the overall inferior quality of the spectral data, 86% of the triple resonance backbone and side chain assignment was completed (shown in Table 12), including 90% of the backbone. Save for two side-chain carbon resonances (Asp^{638} Cγ and Trp^{679} C') the assignment of CCP2 was substantially complete. As with the previously reported NMR assignment of C7-FIMs[53] however, three stretches of residues within FIM2 featured few or no observable resonances (Cys^{773}-Gly^{774}, P^{777}–W^{779}, and more extensively D^{783}–K^{788}). Additionally, another couple of residues, Leu^{725}-Asp^{276} had similarly exiguous

assignments. On the other hand, the assignment of Gln^{693} within the linker between CCP2 and FIMs was absent from the C7-FIMs assignment, yet was complete in C7-CFF. The extent of backbone amide assignment for C7-CFF is illustrated in the [$^{1}H^{15}N$]-HSQC of Fig. 57. This figure also shows the assignment of Asn (four of five), Gln (8 of 11), Trp (all three) and Arg (5 of 11) side-chain resonances. The NMR data allowed for the determination of a *cis* or *trans* conformation for each of ten of the 13 proline residues within C7-CFF. This was achieved by visual inspection of Xaa-Pro cross-peaks and also by the difference between the Cβ and Cγ shifts (outlined in section 2.3.5.6). Pro^{644}, Pro^{681} (CCP2) and Pro^{694} (CCP2-FIMs linker) were identified as the *cis* isoform by both methods and was therefore treated as such in structure calculations.

Category	Available	Assigned	% Assigned
Element C	989	823	83.22
Element H	1274	1070	83.99
Element N	291	189	64.95
Amide	424	378	89.15
Backbone	862	782	90.72
Backbone non-H	657	593	90.26
Side Chain H	1069	884	82.69
Side Chain non-H	623	419	67.26
Residue Ala	17	15	88.24
Residue Arg	11	11	100
Residue Asn	5	5	100
Residue Asp	5	4	80
Residue Cys	22	22	100
Residue Gln	11	11	100
Residue Glu	21	19	90.48
Residue Gly	15	15	100
Residue His	3	3	100
Residue Ile	6	6	100
Residue Leu	15	14	93.33
Residue Lys	14	14	100
Residue Met	6	6	100
Residue Phe	3	3	100
Residue Pro	13	12	92.31
Residue Ser	21	19	90.48
Residue Thr	10	10	100
Residue Trp	3	3	100
Residue Tyr	3	3	100
Residue Val	15	15	100
All Residues	219	210	95.89

Table. 13: Assignment report for C7-CFF by CcpNmr Analysis. Table generated using the quality reports function, includes GSHM cloning artifact.

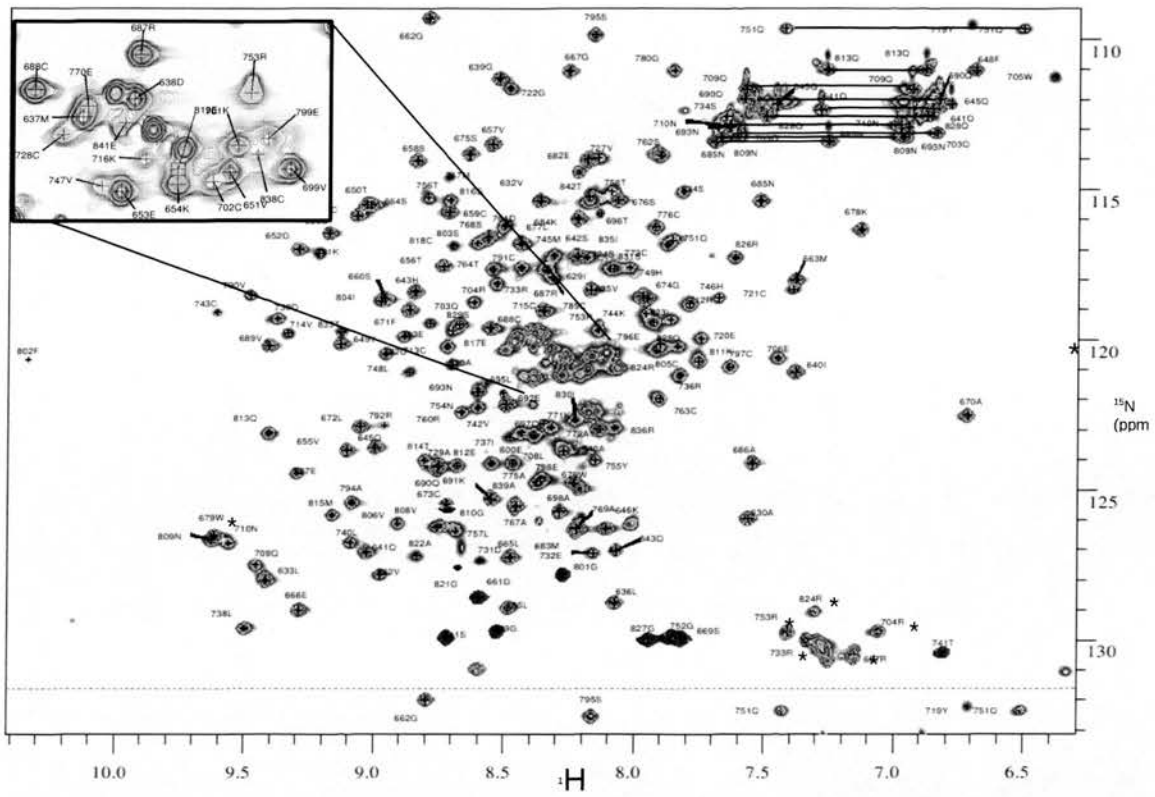


Fig. 57: [^{1}H^{15}N]-HSQC of C7-CFF. Recorded on an 800 MHz (1H) spectrometer equipped with a cryoprobe. Backbone ^{1}H-^{15}N correlations are labelled sequence-specifically. Correlations connected by horizontal lines correspond to Gln and Asn side-chain amides, light grey correlations arise from folded Arg side-chain resonances, and Trp side-chain resonances are marked with an asterix.

4.3.2 Structure calculation

Table. 14 summarises the results of the automated assignment process utilised in the current work that employed the CANDID algorithm in CYANA. As with the FIMs calculation the disulphide bonding pattern of four cysteines within the FOLN domain of FIM2 (Cys^{773}, Cys^{776}, Cys^{782}, and Cys^{789}) were the only disulphide bonds that were not clearly inferable from the NOESY data. CYANA calculations, without these four cysteines explicitly incorporated as disulphide bridges nonetheless resulted in all 20 structures adopting the Cys^{773}-Cys^{782} and Cys^{776}-Cys^{789} disulfide-bonding pattern. Thus,

these linkages were used as constraints in subsequent CNS calculations. In any case, calculations using each of three possible disulfide-bonding patterns did not affect the rest of the structure and had a negligible effect on the target function and number of restraints.

While the first two of the five CYANA criteria (listed in section 2.3.8.2) were met, full satisfaction of the final three criteria appeared to be unachievable, given the extended structure of the molecule and the complexity of the data.[138]

First, more than 90% of the ^{1}H chemical shifts should be assigned. In the case of C7-CFF, this value was just met at 90%, *i.e.* a total of 1124 of a total of 1250 ^{1}H atoms assigned in all.

Second, the cross-peak list should be a "faithful representation" of the NOESY spectra. A large number of NOE cross-peaks (10791 in the present case) were chosen initially to serve as a basis for the structure calculations. This apparently very large number of potential distance restraints includes replicates, since NOEs were picked from four different (but overlapping in data content) NOESY spectra. Moreover, attempts to "faithfully" represent the spectra, which were of overall poorer quality than for example the spectra collected for the CCPs, resulted in the picking of many spurious peaks. Consequently, 1100 peaks (about 10% of total), were eliminated from the first cycle on the grounds that they were categorised as "poor quality" according to CYANA. Poor quality peaks are generally those that border the chemical shift tolerance limits and whose assignments have low probability scores, as is indeed common of noise and artifactual peaks.

The third criterion states that the average DYANA target function should be <250 Å^2 in the first cycle and <10 Å^2 in the final of the seven cycles, with $\geq 80\%$ of all the picked

NOESY peaks assigned and ≤20% of long-range assignments eliminated from the calculation. In the case of C7-CFF, the target function started out very high and did not fall below 250 Å until the fourth cycle. Inspection of the structures produced by the first three cycles revealed that this was predominantly due to conformations in which CCP2 and FIMs, or CCP2 and the CCP2-FIMs linker, were closely associated. These cases appeared to arise from a few NOEs and these could be attributed to mis-assignments due to similarities between chemical shifts in some residues within CCP2 and other residues in FIMs. Thus, there was not a network of NOEs to support these compact conformations. Consequently the structures did not, overall, fit the experimental restraints, resulting in a high target function. Also, over 10% of the peaks were eliminated in the 1st cycle; this increased to 23% ejected peaks by the final cycle. Therefore only 77% of the picked peaks were assigned by CYANA, just falling short of the 80% criterion. On the other hand, and reassuringly, only 3%, compared to the allowed 20%, of long-range assignments were eliminated over the course of the seven cycles.

The fourth criterion is that the average backbone RMSD of the calculated structures should be smaller than 3 Å after the first cycle. CFF as a whole never meets this criterion since (as discussed below) the linkage between the CCP and the FIMs appears to be completely flexible. The CCP2 component alone satisfied the criterion with a backbone RMSD of 1.5 Å. However the FIMs component of CFF failed on this criterion, producing RMSD values of 4.1 Å and 4.3 Å for FIM1 and FIM2, respectively. Dissecting FIM1 into its KAZAL and FOLN subdomains yields RMSD values of 3.1 Å and 1.8 Å. A similar exercise for FIM2 gave values of 3.7 Å (KAZAL) and 2.4 Å (FOLN). By the second cycle all modules and subdomains had RMSDs <2.4 Å. It was decided to accept these structures despite not satisfying the fourth criterion. It was reasoned that the CYANA criteria are in fact somewhat arbitrary and empirical in nature

and arose out of experience with globular proteins. Their purpose is to ensure that a misfolded structure does not become self-perpetuating. In the case of a multiple-domain protein such as CFF with several potentially distinct folded entities it is likely to be very difficult to eliminate chance (artifactual) encounters between domains that will be supported by mis-assigned NOEs in the first round as described above. Provided these are eliminated in successive cycles, and that domains with the expected (known) folds emerge then there seems to be little danger of mis-folding. Clearly, it was important to conduct extensive manual checks to avoid missing interdomain NOEs or the mis-assignment of intradomain NOEs as interdomain ones.

The final criterion is that the RMSD "drift" between the mean structures of the first and last cycle is below 3 Å. For CCP2 this was easily met, at 2.0 Å. For the FIMs this value was 4.5 Å and 4.1 Å for FIM1 and FIM2 respectively. This is not surprising given the high starting values that may not in fact be a matter of great concern, as discussed above.

Cycle	1	2	3	4	5	6	7	Final
Peaks:								
selected	10791	10791	10791	10791	10791	10791	10791	
assigned	9690	9536	9041	8870	8616	8402	8362	
unassigned	1101	1255	1750	1921	2175	2389	2429	
with diagonal assignment	3	3	3	3	3	3	3	
Cross peaks:								
with off-diagonal assignment	9687	9533	9038	8867	8613	8399	8359	
with unique assignment	1825	5365	6236	6312	6661	6860	6929	
with short-range assignment |i-j|<=1	5664	5622	5410	5225	5066	4947	4917	
with medium-range assignment 1<|i-j|<5	1474	1309	1115	1091	1022	965	963	
with long-range assignment |i-j|>=5	2549	2602	2513	2551	2525	2487	2479	
Upper distance limits:								
total	7448	6413	5670	5360	5031	4770	4677	4737
short-range, |i-j|<=1	3641	3197	2875	2604	2430	2280	2042	2072
medium-range, 1<|i-j|<5	2306	1875	919	882	803	755	748	754
long-range, |i-j|>=5	1501	1341	1876	1874	1798	1735	1887	1911
Average assignments/constraint	6.74	2.5	1.46	1.43	1.32	1.25	1	1
Target Function:								
Average target function value	1527.82	673.37	837.49	225.89	101.38	46.16	49.16	17.39
RMSD (residues 625..843):								
Average backbone RMSD to mean	8.88	11.55	10.13	10.53	10.8	10.56	6.72	9.1
Average heavy atom RMSD to mean	9.29	11.88	10.42	10.78	11.03	10.83	6.9	9.34

Table. 14: CYANA structure calculation cycle report for C7-CFF. Statistical tracking of NOE cross-peaks assignment, distant-restraint deduction, target function values and RMSD values are shown (over all CANDID cycles).

Following completion of all seven cycles, a total of 4737 unique NOE-derived restraints was generated. These constitute 2072, 754 and 1911 upper-distance bounds corresponding to intra-residue, medium-range (between residue i and residues (i+1 to i +4)), and long-range (between residue i and residues i+ >4)) restraints, respectively. The Format Converter of the CcpNmr software[135] was utilised to convert the list of upper-distance restraints generated by CYANA into a format compatible for subsequent CNS structure calculations as was done for the CCPs. The 100 structures thus re-calculated in CNS were ranked in order of NOE energy and in order of total energy (Fig. 58). The 13 lowest-energy structures converged well. However a slight increase in both NOE and overall energy for the last three structures prompted a smaller selection involving just the

ten lowest-energy structures for refinement in water solvent. The final 10, water-refined, NMR-derived structures overlay well (although not over the whole molecule), with low RMSD values of 0.27, 0.51, 1.70 and 1.44 for CCP2, FIM1, FIM2 and both FIMs respectively (Fig. 59).

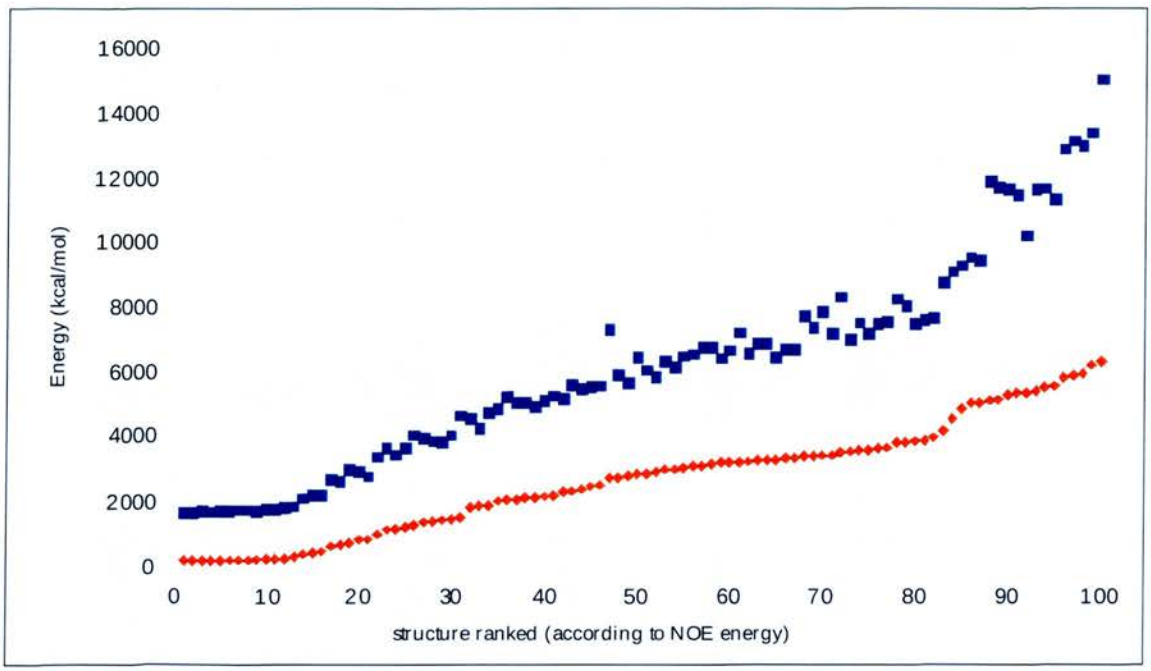


Fig. 58: Energy plot of the final, ranked 100 CNS-re-calculated C7-CFF structures. The NOE energy is shown in red triangles and the total energy is shown in blue squares.

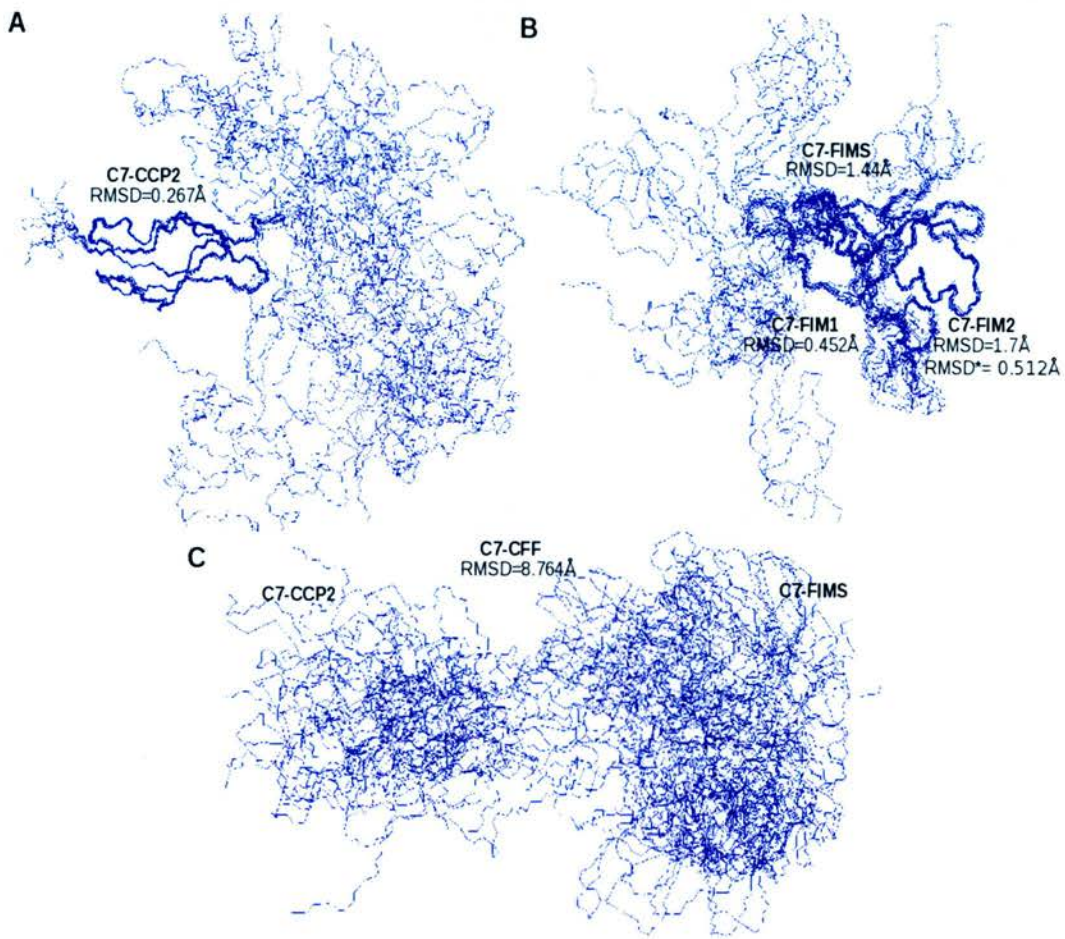


Fig. 59: Backbone overlays of the ensemble of ten water-refined C7-CFF structures. (A) Backbone overlay on CCP 2; (B) Backbone overlay on FIM1; (C) Backbone overlay for all modules. MolMol[142] was used for visualisation and calculation of the RMSD values (indicated). Domain boundaries are considered first to last cysteines of each module.

4.3.4 Structure description and quality analysis

Structure determination reveals that the CCP and FIMs behave as two separate globular regions connected *via* a flexible linker. Both CCP2 and the FIMs of C7-CFF have the same conformation as in the separately solved C7-CCPs and C7-FIMs respectively. The mean structure of CCP2 in C7-CCPs and C7-CFF were compared and overlay well with a backbone RMSD of ~1.8 Å. The main differences are within loop regions at the N- and C-terminal ends of the molecule that are involved, respectively, in interactions with the neighbouring CCP molecule and the first few residues of the CCP2-FIMs linker. The β-strand network of CCP2 (as predicted within the program STRIDE[161]) appears less extensive in CFF (compared to CCPs), since it exhibits two shortened strands (D and F) of the four strands identified previously (B, D, F and H) . The "missing" strands B and H are detected as single amino acid β-bridges. Strand D is particularly short (reduced from seven residues in CCPs to three residues in CFF), due to a backbone kink in the middle of these residues. The canonical three-residue 3_{10}-helix in CCP2 was, however, maintained. The less extensive secondary structure of CCP2 in CFF is likely a consequence of having determined a lower resolution structure and it may or may not reflect structural changes due to the absense of CCP1.

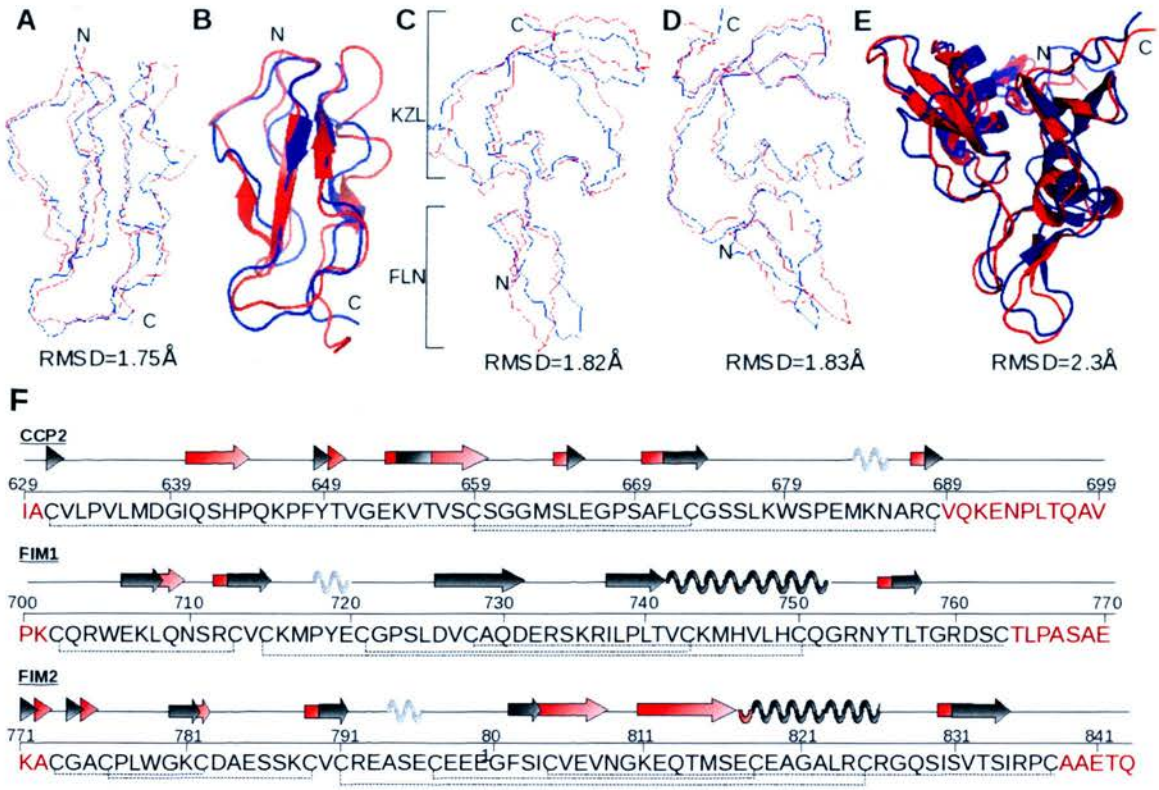
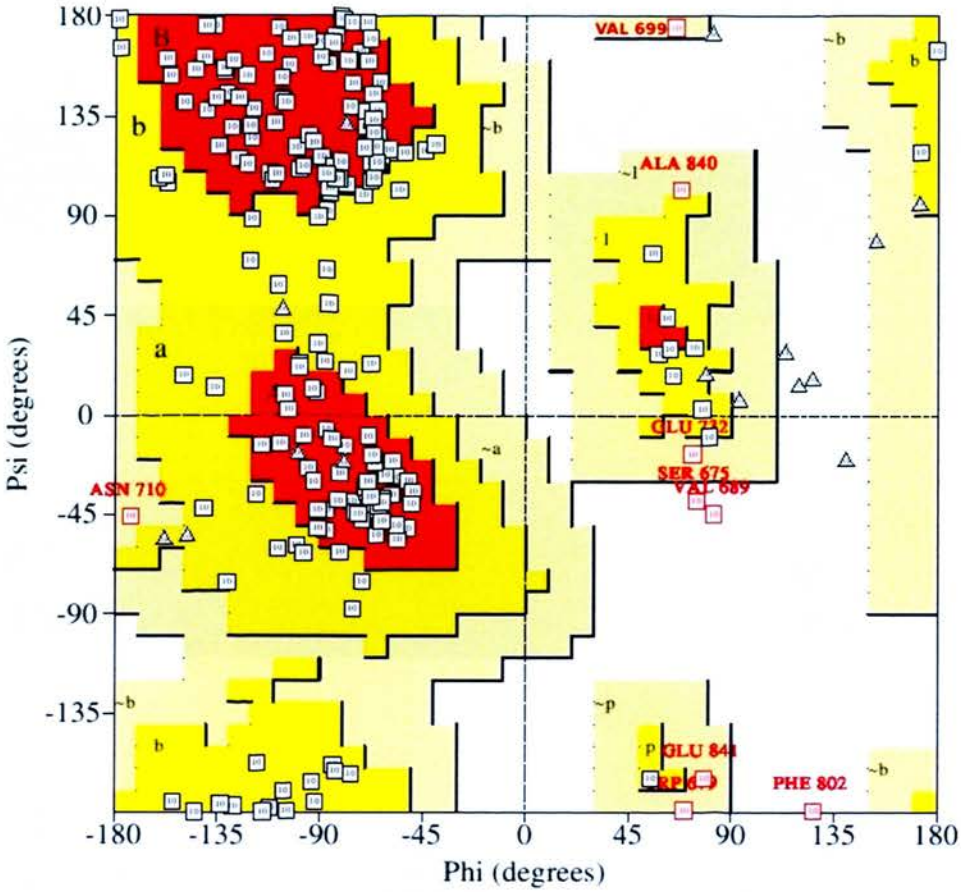


Fig. 60 Comparison of CFF modules with their previously solved counterparts. (A), (C) and (D) show respectively the backbone overlay and RMSD (MOLMOL[ref]) of CCP2, FIM1 and FIM2 extracted from the mean structure of CFF, with the mean structures of their NMR-derived counterparts solved previously (CCPs and FIMs[53]). (B) and (E) respectively depict cartoons of CCP2 and FIMs of CFF overalaid on cartoons of the previously solved structures, highlighting secondary structural elements. In all overlays FIMs(alone) is in red and CFF is in blue. (F) Shows a schematic representation of this secondary structure comparison. β-bridges and β-sheets are shown as triangles and arrows respectively. 3_{10} and α -helices are shown in light shades and dark shades respectively. Secondary structure present in CFF is shown in black and is typically extended and shown in red for FIMs (alone).

The FIMs pair component of CFF was found to exhibit the same two-fold rotational pseudosymmetry as the published structure of the FIMs in isolation. As before, the two modules are intimately associated *via* an extensive intermodular interface. FOLN and KAZAL subdomains similarly comprise each individual FIM. The backbone overlay of each individual FIM of CFF with the previously determined FIMs structures is reasonable (backbone RMSD =~1.8 Å for both modules).

The main difference between the structucture of the FIM modules in CFF and in FIMs is within the FOLN subdomains of both FIM1 and FIM2. In the case of FIM1 differences in the orientation of the FOLN subdomain relative to the FIMs structure as a whole might be genuine and could arise due to constraints induced by the attachment of the long linker to CCP2. In the case of FIM2, the FOLN domain includes the stretch of five residues (D^{783}–S^{787}), which were not assigned in either the FIMs or CFF structure calculations, suggesting the region is conformationally promiscuous. All secondary structural elements, as identified by STRIDE, that are present in the FIMs (alone) structure are also present in the CFF structure, including the 2.8 turns of an α-helix in each module. However, the secondary structure elements are generally shorter in CFF by one residues, with the exception of the first β-strand of the FIM2 KAZAL domain that is longer in CFF by one residue. The orientation of the FIMs with respect to one another is largely maintained between the FIMs (alone) and CFF structures (FIMs backbone RMSD = 2.8Å). The α-helix of FIM2, however, bends more towards FIM1 at the C-terminal end of the helix (Fig. 60) in the CFF structure.

The WHATIF "course packing quality score" was used to analyse the C7-CFF structure ensemble.[145] WHATIF scores lower than -5.0 Å are indicative of improper packing. The average score of the ensemble was well above this value both before (-2.2) and after (-1.7) water refinement by CNS. Subsequently, the program PROCHECK[144] was used to evaluate the water-refined ensemble via establishing the energetic favourability of the ϕ and ψ angles in a Ramachandran plot (Fig. 61). A total of 95% of ϕ-ψ combinations occur in the most favoured and additionally allowed regions of the Ramachandran plot, which is suggestive of reasonable stereochemical quality.



Plot statistics

Residues in most favoured regions [A,B,L]	1200	63.5%
Residues in additional allowed regions [a,b,l,p]	600	31.7%
Residues in generously allowed regions [~a,~b,~l,~p]	60	3.2%
Residues in disallowed regions	30	1.6%
Number of non-glycine and non-proline residues	1890	100.0%
Number of end-residues (excl. Gly and Pro)	20	
Number of glycine residues (shown as triangles)	150	
Number of proline residues	130	
Total number of residues	2190	

FIG 61: Ramachandran statistics (from PROCHECK[144]) for water-refined ensemble of the ten lowest energy structures of C7-CFF. Each small box represents a residue and the model number, save for glycines and prolines which are shown as triangles. Outliers are named and shown in red.

4.3.4 Relaxation analysis

T_1 and T_2 relaxation time constants were collected for the [^{13}C,^{15}N]-CFF sample. Incremental relaxation delays used for T_1 measurements were 51.2, 301.2, 501.2, 701.2, 901.2, 1101.2, 1301.2, 51.2 ms. While for T_2 they were 16.8, 33.7, 67.3, 84.2, 117.8, 134.7, 151.5 and 16.8 ms. Weak peaks, or those in overcrowded regions, were excluded from analysis due to their associated large error values. The average T_1 and T_2 values for the CCP2 component (~800 ms and ~90 ms, respectively) differed greatly to the equivalent values for the FIMs component (1400 ms and 40 ms, respectively) of CFF (Fig. 62). These markedly different vales are consistent with a highly flexible linker between CCP2 and FIMs, such that each tumble quasi-independently of the other. Thus for the purposes of the discussions below, the components were considered separately with averages taken only within module boundaries. However, this does not mean that the presence of the flexibly attached module would be expected to have no effect whatsoever.

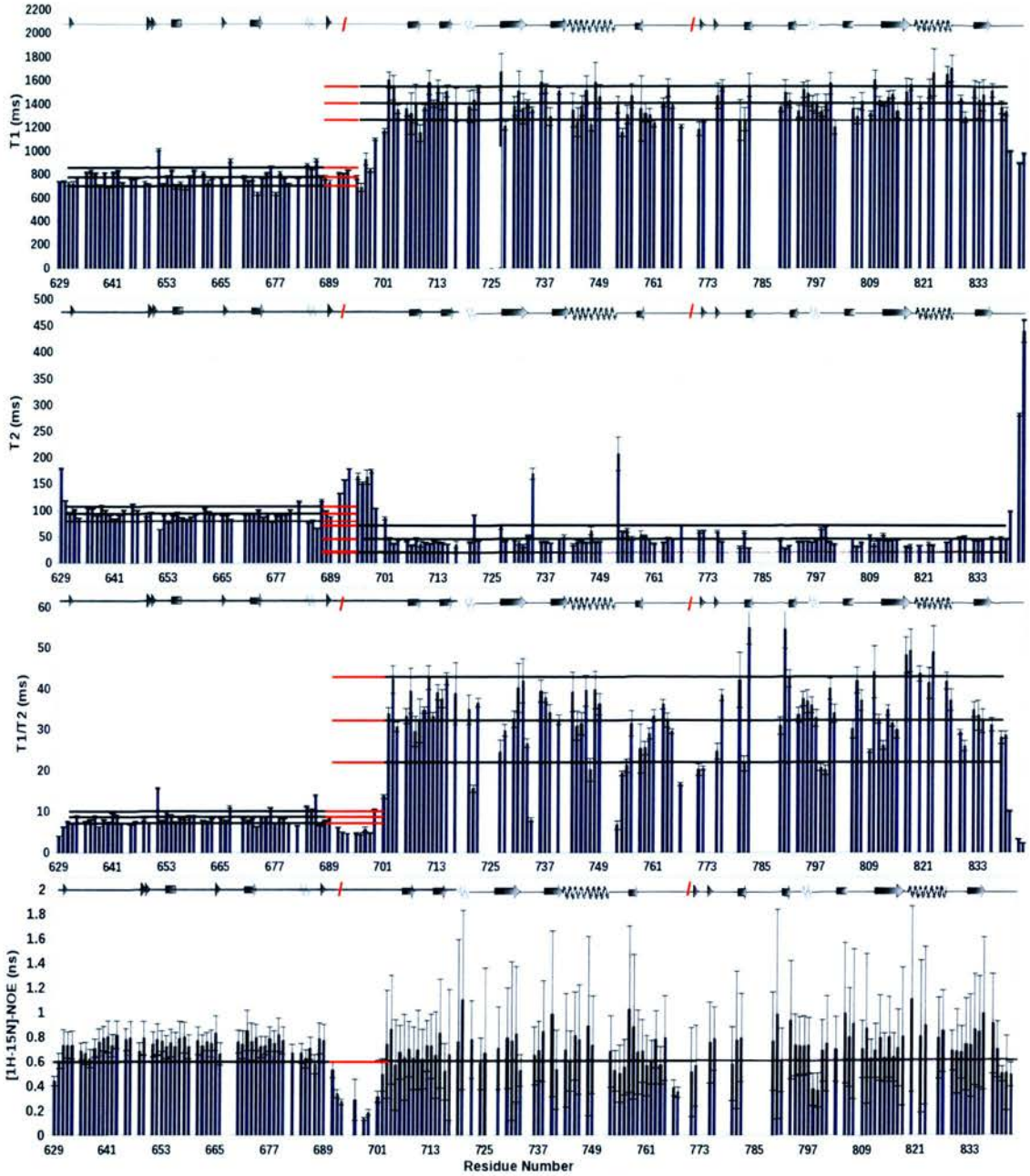


Fig. 62: Summary of ^{15}N relaxation parameters for CFF. (A) T_1 values and associated errors at 14.1 Tesla. (B) T_2 values and associated errors at 14.1 Tesla. (C) T_1/T_2 ratio and associated errors. (D) [^{1}H,^{15}N]-heteronuclear NOE values and associated errors at 14.1 Tesla . Lines for the average and +/- one standard deviation are shown for (A), (B) and (C) calculated seperately for CCP2 and FIMs. The 0.6 ns NOE cut-off is also shown by a line in (D). The red portion of each line highlights the linker between CCP2 and the FIMs. Schematic shows secondary structure with red slashes defining module boundaries.

As stated previously, low T_1/T_2 values indicate backbone flexibility on the millisecond-second timescale and represent large slow conformational changes,[118] while heteronuclear NOE values <0.6 indicate flexibility on the picosecond-nanosecond timescale, representing backbone librations and side-chain motions. The calculation of T_1 and T_2 values requires that selected peaks are not overlapped and can be integrated accurately. However, for CFF, many of the peaks (~40) were too crowded or weak, and were therefore excluded from analysis. Also those few peaks that had not been assigned (as described in section 4.3.1) were also not included. Ultimately, 50 of 60 residues of CCP2, 49 of 69 of FIM1 and 48 of 70 of FIM2 were included in the analysis. It was also recognised that the heteronuclear NOE errors in the FIMs portion of the molecule are substantially larger than for CCP2 . This is unsurprising due to the differing τ_c values, which greatly affects the efficiency of the heteronuclear NOE measurements.[121]

Residues displaying flexibility on either of slow or fast timescales were mapped onto the C7-CFF structure to better visualise their spatial location and distribution (Fig. 62). The following discussion focuses on the linker since the dynamics of the CCP2 (in CCPs) and of the FIMs (alone) have been discussed above and elsewhere [53] respectively and (especially given the high error values in the measurements for CFF) there is little point in reconsidering them in the CFF context.

The 13 residues between the last cysteine of CCP2 and the first cysteine of FIM1 display a large range of flexibility on both the millisecond-second and the picosecond-nanosecond timescales. The first two residues of the linker however (Val^{689} and Gln^{690}), are not flexible with respect to the average CCP2 NOE or T_1/T_2 values. This is likely due to side-chain interactions with the last cysteine, Cys^{688} and Met^{663} and Ser^{664} in the penultimate stretch (four of five) of CCP2 residues (Fig. 62) and is supported by NOEs. Also Lys^{691} of the linker has the potential to make ionic interactions with the carbonyl

group of Ser^{664}. Interestingly, the residues surrounding the turn between strand F and the 3_{10} helix, which contain Met^{663} and Ser^{664} was deemed flexible in the relaxation analysis of C7-CCPs (alone). These residues are therefore stabilised by the presence of the first three linker residues in the context of CFF. Flexibility on both timescales is present along the rest of the length of the linker. However inspection of the T_1/T_2 ratio reveals that the last three residues of the linker (Val^{699}, Pro^{700} and Lys^{701}) are considered rigid with respect to residues within the CCP2 component of CFF but flexible with respect to the FIMs component. The decreased mobility of these residues likely stems from hydrophobic side-chain to side-chain contacts with Leu^{708} of the β-hairpin within the FIM1 FOLN subdomain and the first residue of FIM1, Cys^{702} Again these interactions are reflected by observed NOEs.

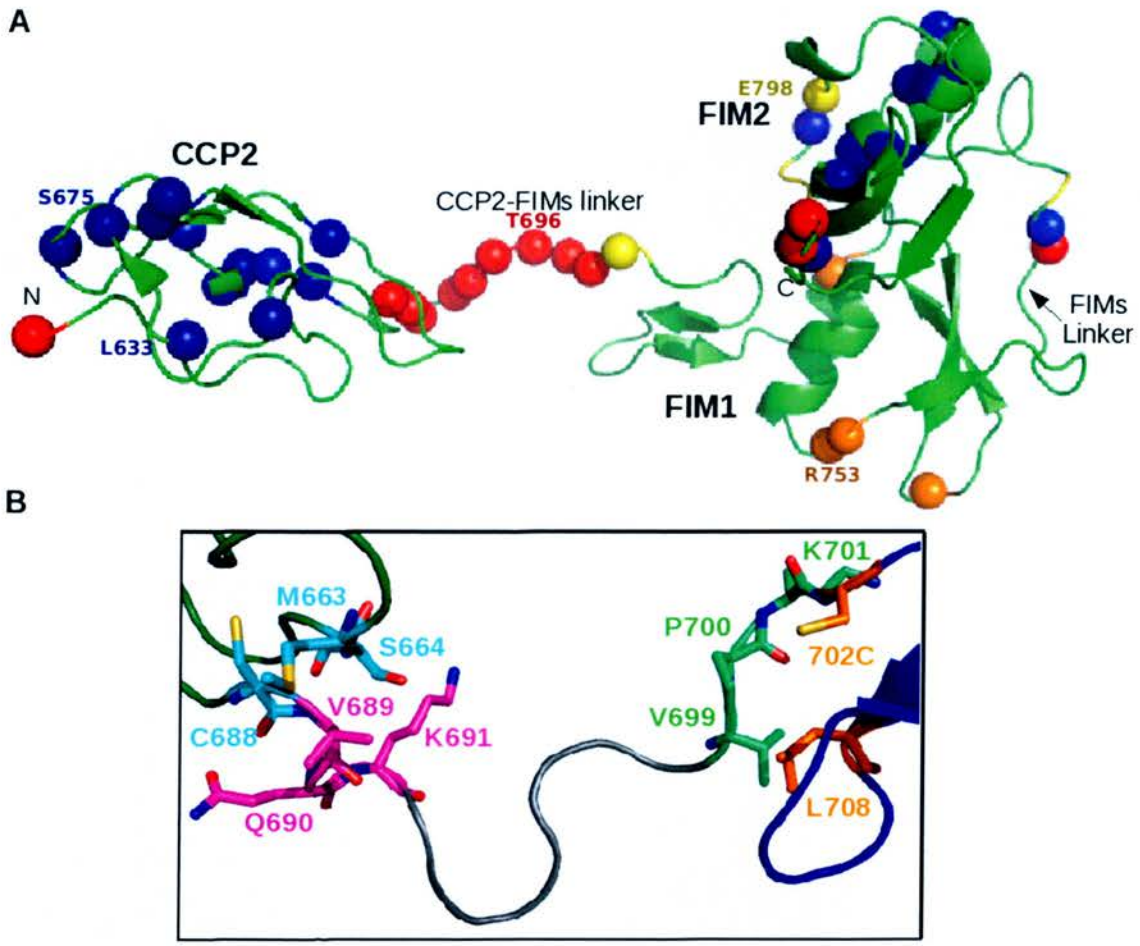


Fig. 63: Molecular dynamics of C7-CFF (A) Residues falling within plus or minus one standard deviation from the average values, on a module by module basis, for T_1/T_2 and $^1H,^{15}N$ NOE are shown in green. Residues deemed flexible, on the ns timescale, from their [$^1H,^{15}N$] NOE values are shown in yellow. Residues deemed to have particularly low or high flexiblity, on the ms-s timescale, from their T_1/T_2 ratios are shown in blue and orange, respectively. Those residues, whose backone is flexible on both timescales are shown in red. (B) Shows a close-up of the CCP2 (green) to FIMs (blue) linker (red). Linker-module interacting residues are highlighted as sticks.

4.3.5 Analysis of the intermodular interface

4.3.5.1 CCP:FIMs

As mentioned above, in the initial rounds of the CYANA-based assignment and structure calculation, cross-peaks matching to putative NOEs between FIMs and CCP2 were

identified. These did not survive subsequent rounds since they were not supported by surrounding NOEs. They may have been reassigned by CYANA, or (more worryingly) just eliminated and not used. Therefore to be as certain as possible that no genuine intermodular NOEs were being missed, all candidates (from early CYANA rounds) were visually inspected. It was thus ascertained that with the exception of NOEs between FIM1 and FIM2, no NOEs could be unambiguously assigned as intermodular.

Thus while the long flexible linker between CCP2 and the FIMs would allow for module-module interactions to occur (and it would be very surprising if transient contacts did not take place), the NMR data very strongly suggests that any such interactions are short-lived on the NMR time-scale and non-specific.

On the other hand the linker (defined as the sequence of residues between the last and first cysteines of flanking modules) is not flexible throughout its length, but only in its middle (seven-residue) section. More insight into this aspect of C7-CFF was obtained from the shape envelope derived from SAXS. The SAXS-based 3D model of C7-CFF (*Ab intio* shape determined using DAMMIN in the PRIMUS package[150]) indicates that there are indeed very likely to be some constraints on the flexion of the linker. The model depicts a large globular region and a smaller globular protrusion. Manual fitting of the water-refined C7-CFF structures to this model was intutitive with the larger region being occupied by the FIMs and the smaller region by CCP2. Five of the ten water-refined CFF structures have a relatively extended linker conformation and easily fit within the model. However the CCP2 of any single structure is insufficient to account for the volume of the smaller globular protrusion and the same is true with regard to the FIMs and the larger of the two protrusions (Fig. 64A). On the other hand, an overlay of those five NOE-derived C7-CFF structures that featured an extended linker filled the available space very nicely suggesting that the observed shape represents a structure

averaged over multiple conformations. The remaining five structures within the CFF ensemble featured less elongated linkers, with the result that CCP2 and FIMs are too close together to fit the model. Thus the SAXS-derived model indicates that there are some constraints on the flexibility of the linker and that the NMR-based structure calculations have under-restrained the structure rather than over-restraining it.

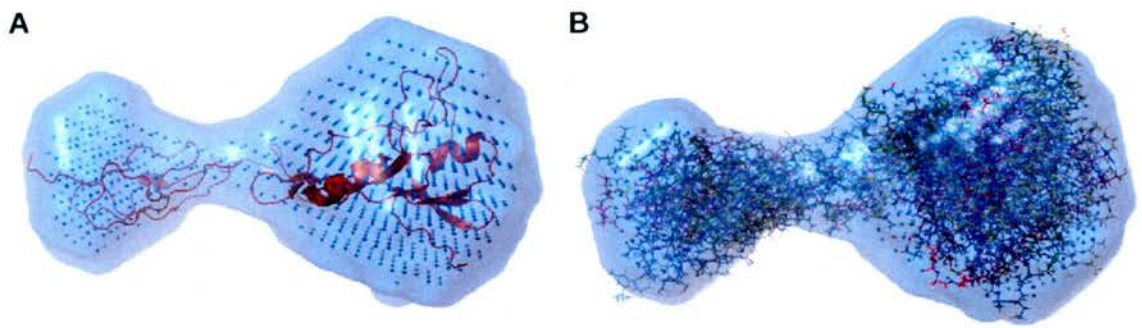


Fig. 64: *Ab initio* SAXS-based analysis of CFF. (A) shows the DAMMIN[150] derived SAXS model (courtesy of Dr. Elizabeth Blackburn) overlayed with the calculated mean structure of the CFF water ensemble. CFF is shown as a cartoon (red). (B) depicts five of the ten ensemble structures overlaid with the SAXS-derived model with CFF residues shown as sticks.

4.3.5.2 FIM1:FIM2

The inter-FIM interface within the FIMs component of C7-CFF is supported by a plethora of both electrostatic and hydrophobic interactions, including potentially module-bridging hydrogen bonds, salt bridges, aromatic-to-cysteine contacts and aromatic-to-aromatic interactions. These involve four regions of sequence in FIM1 (Arg^{704}–Glu^{706}; Pro^{723}–Leu^{725}; Val^{742}, Cys^{743}, Met^{745}, and His^{746}; and Val^{727}, Ala^{729}, and Leu^{757}) and four regions of FIM2 (Trp^{779}; Glu^{799}–Ile^{804}; Glu^{817};Gly^{821}; and Ile^{835}).

The intermodular interfaces of FIM1 and FIM2 are nearly identical between the C7-FIMs (alone) and C7-CFF structures (Fig. 65), with numerous NOEs "connecting" the two modules. All but two of the residues involved in the interface make the same contacts with the neighbouring module, Arg^{760} of FIM1 and Arg^{824} of FIM2. In C7-CFF,

Arg^{760} points away from the interface as opposed (in FIMs) to being engaged in the interface in hydrophobic interactions involving its Hβs with Val^{832} $H\gamma^{1}$. Unfortunately, the associated NOESY strips were overcrowded in Fourier transformed ^{13}C-NOESY-HSQC spectra of C7-CFF recorded in both H_2O and D_2O. Although the maximum entropy-processed spectra were clearer, a potential cross peak from Val^{832} was observed in the Arg^{760} strip but was noise-like and could not be picked. The corresponding ^{13}C-methyl-NOESY-HSQC Val^{832} strip had signal in the Arg^{760} Hβ region but this was not resolved by maximum entropy processing.

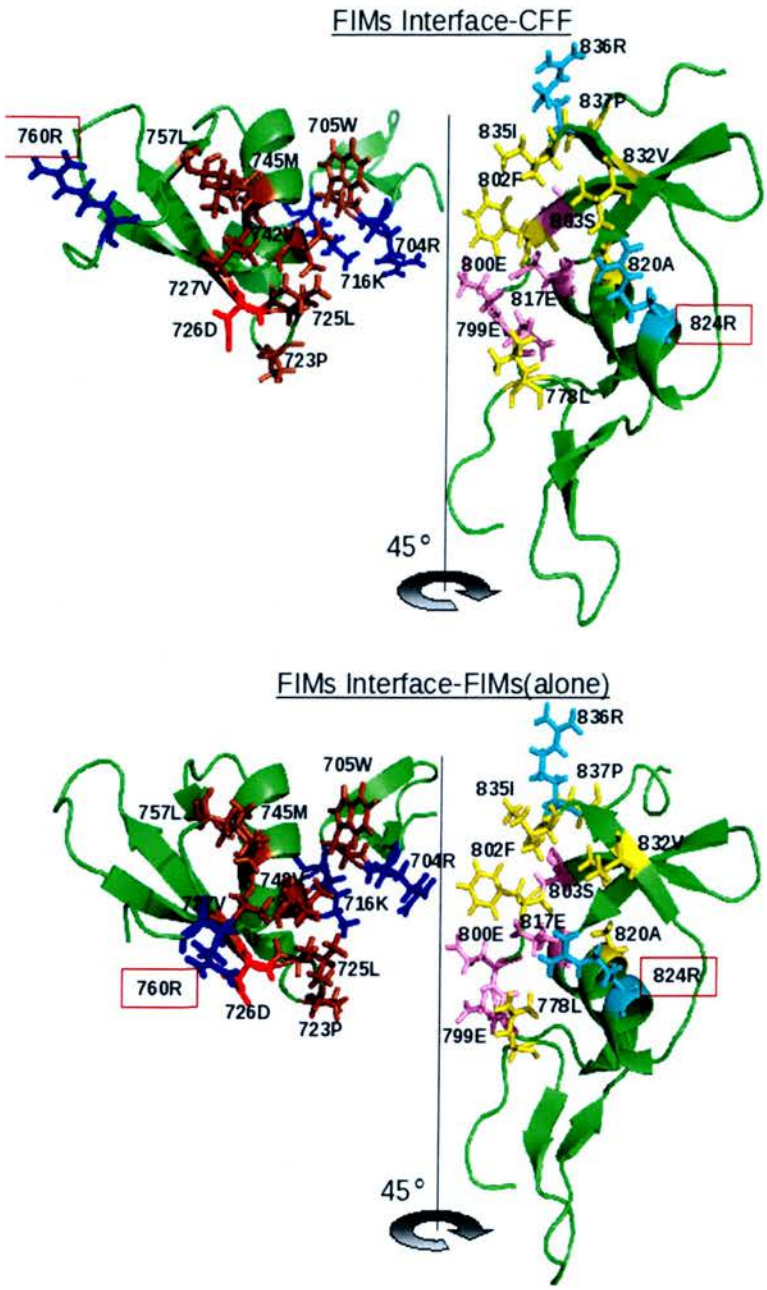


Fig. 65: The FIM1-FIM2 modular interface. (A) shows the CFF FIMs interface. (B) shows the FIMs[53] interface. In each panel FIM1 is shown on the left and FIM2 on the right with intermodular residues shown as sticks. Hydrophobic residues are brown in FIM1 and yellow in FIM2. Basic residues are Blue in FIM1 and cyan in FIM2. Acidic residues are highlighted in red in FIM1 and pink in FIM2. Boxed residues are those whose orientation varies significantly between the CFF and FIMs structures.

As for Arg^{824}, its orientation in the C7-FIMs (alone) means that it protrudes towards the centre of the interface. In C7-CFF, although the Arg^{824} side-chain faces FIM1, it does not extend so markedly towards the centre of the interface as it does in FIMs (alone). Comparison of the NOE lists indicates that the same set of NOEs for Arg^{824} are used in both calculations. What is interesting is that in both C7-FIMs and C7-CFF no intermodular NOEs are observed for Arg^{824} despite it being buried in the interface in both cases. Although the true orientation of Arg^{824} is unknown, it is conceivable that the presence of an extra residue identified as being part of an α-helix in the C7-FIMs (alone) structure, has a knock on effect with respect to the orientation of the side chain of Arg^{824}. However, the length and orientation of the FIM2 helix has no other observable effect on the intermodular interface.

4.3.6 Analysis of surface properties

For both CCP2 and the FIMs of CFF, the surface electrostatics and lipophilicity have the same characteristics as their previously solved (in isolation) counterparts. CCP2 still has one predominantly acidic face and another predominantly basic face. With regards to the FIMs, FIM1's solvent accessible face is predominantly basic, with Lys^{701}, Arg^{704}, Lys^{707}, Arg^{712}, Arg^{733}, Lys^{744}, and Arg^{753} being the main contributors. Contrastingly, FIM2's surface is predominantly acidic in nature, primarily due to several surface-exposed glutamate side-chains, namely Glu^{785}, Glu^{798}, Glu^{799}, Glu^{800}, Glu^{807}, Glu^{812}, and Glu^{841}. As in the C7-FIMs structure Glu^{798}, Glu^{799}, and Glu^{800} form a cluster of negative charge in a cleft in FIM2. The CCP2-FIMs linker contains two electronegative residues (E^{691} and T^{686}) that are sandwiched between four electropositive residues (Q^{690}, K^{691}, N^{693} and Q^{696}).

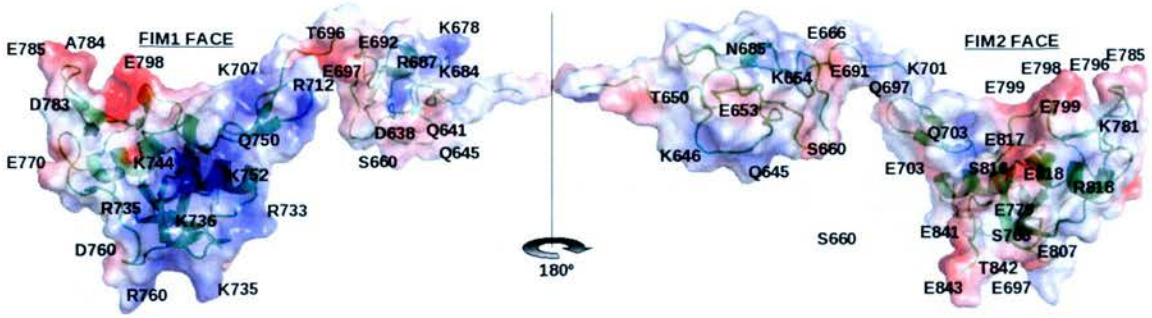


Fig. 66: Analysis of C7-CFFs surface properties. (A) Surface-electrostatics are shown with a red (negative charge), white (neutral charge) and blue (positive charge) colour scale. (B) Surface lipophilicity is represented by a brown (lipophilic) to green (neutral) to blue (hydrophilic) scale. Contributing residues are labelled in both diagrams.

The electropositive surface of FIM1 is interrupted by a patch of hydrophobicity involving residues Met^{717}, Tyr^{719}, Leu^{748}, and Tyr^{757} that is exposed in the FIM pair as discussed in the FIMs structure paper.[53] To assess the potential of this patch as a protein-protein interaction site the lowest energy CFF structure was submitted for STP analysis (Fig. 67). Indeed these residues were indicated to form a protein binding site, with the exception of M^{717}. Most striking, is the cleft that is lined with negative charge and runs along the axis of symmetry between the two FIM modules. There is an extensive protein interaction site formed between the two modules by residues W^{705}, L^{725}, M^{745}, H^{746} and L^{757} in FIM1, and I^{835} and R^{836} in FIM2. As the FIMs modules are "hinged" together at the opposite face composed of the FIMs linker and FIM2 FOLN, it is easy to envisage an closed-to-open transition upon binding C5-C345C, or alternatively open-up during MAC self-assembly.

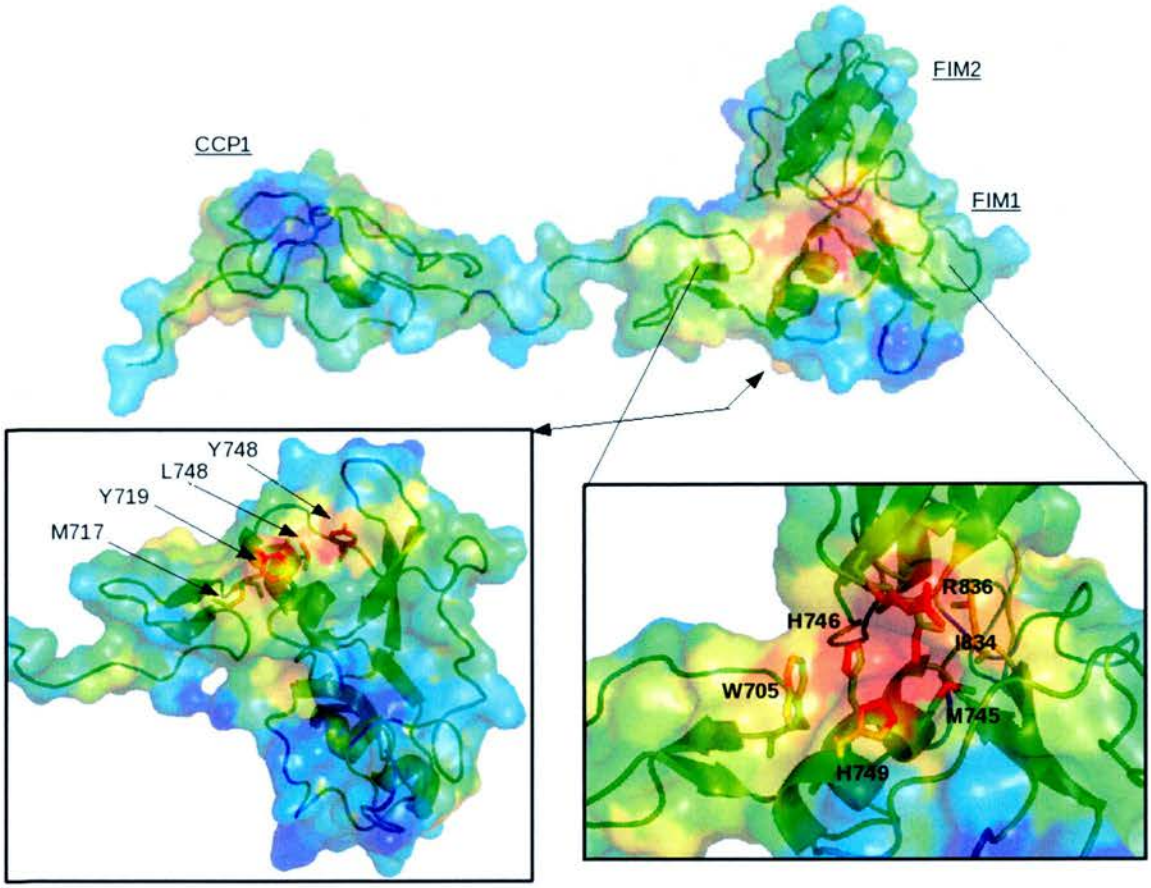


Fig. 67: Identification of C7-CFF's ligand-binding sites using "surface triplet propensities". A surface representation of C7-CFF with a scale going from least favourable (blue), to neutral (green), to most favourable (red) interaction site scores. The top view shows the whole molecule, while the panel on the left shows the backface of the FIMs (topview) and the panel on the right shows a close up of the interaction site in the FIMs cleft.

4.4 The C7 molecular arm

4.4.1 CCP1-CCP2-FIM1-FIM2

By merging the structure of C7-CCPs and that of C7-CFF one can begin to build-up a picture of the C-terminal "molecular arm" of C7. A simple overlay of C7-CCPs and the water-refined ensemble of C7-CFF provide a rudimentary reconstruction of CCP1-CCP2-FIM1-FIM2 (CCFF, Fig. 68).

The flexible nature of the CCP2-FIM1 linker likely plays a more prominent role in FIMs binding the C345C domain during formation of the membrane attack complex. Moreover, in the context of CCFF, the face at the "top" of the FIMs, that has an extensive putative protein interaction site, is distal to preceding CCP modules (Fig.68C) and may well be exposed in the intact protein. These concepts are discussed further in section 6.2 of the discussion chapter.

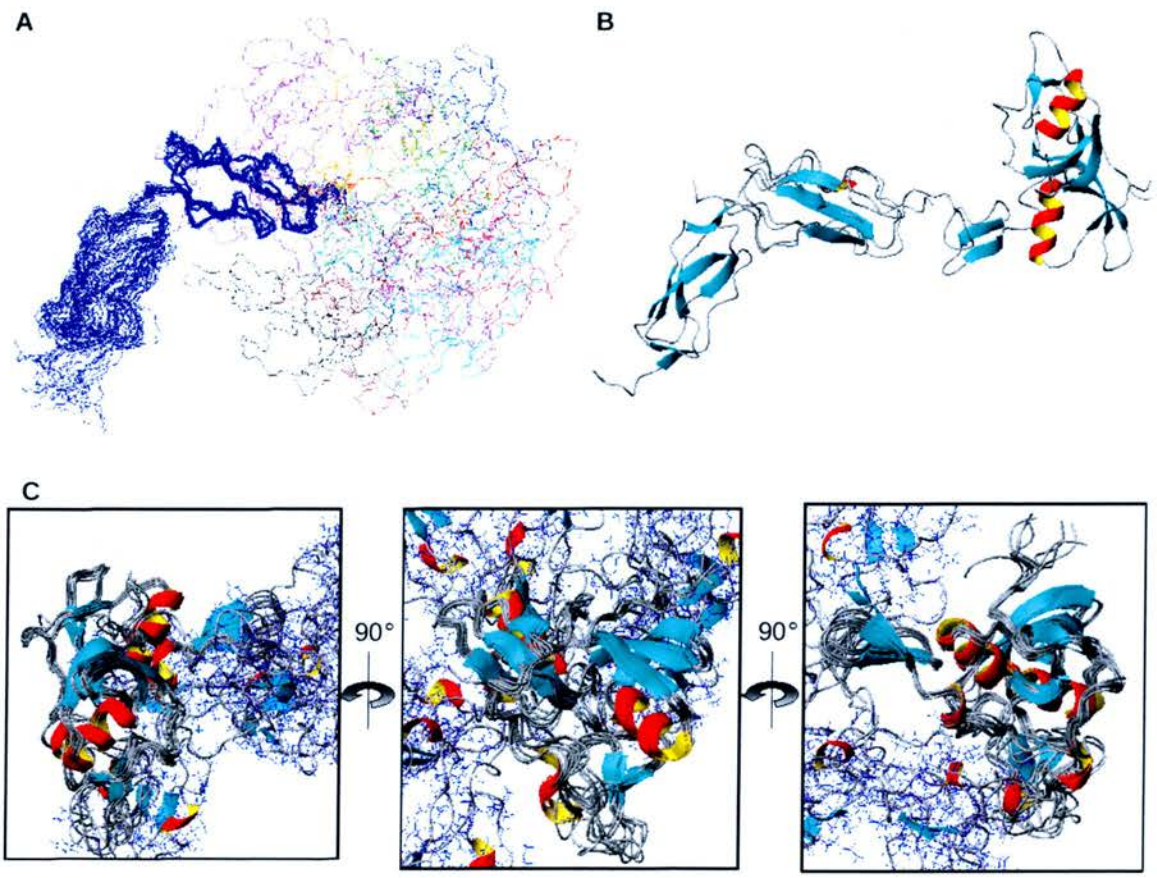


Fig. 68: Structure of CCFF. Overlays were performed in MOLMOL[142] between the water ensembles of C7-CCPs and C7-CFF. (A) Shows a ribbon trace overlaid on CCP2. (B) Shows a ribbon representation of the two lowest energy structures. (C) Shows three views of an overlay on the FIMs (ribbon), with the CCPs shown additionally as sticks to differentiate between the module types.

4.4.2 TSPC-CCP2 and EGF-TSPC

Purified ^{15}N-labelled C7-TSPC-CCP2 (C7-TC) and C7-EGF-TSPC (C7-ET) samples were produced recombinantly in *E. coli* and provided by our collaborator (Ron Ogata) for initial structural analysis by NMR. ^{1}H,^{15}N-HSQC spectra were recorded for both samples on 100-μM samples (20mM phosphate buffer, pH 5.0) (Fig. 69A, B). In both cases the samples produced good-quality HSQC spectra, with well-resolved amide cross peaks covering a wide range of chemical shifts, consistent with folded modular structures.

For C7-TC, 131 of the expected 137 peaks were resolved, while for C7-ET, 89 of 99 expected cross peaks were counted. Overlaying the two HSQC spectra with one another, and also overlaying the HSQC spectrum of C7-TC with that of C7-CCPs was an informative exercise (Fig. 69). The overlays indicate that the structure of each module is relatively unchanged by the presence of neighbouring modules. This is based on the fact that approximately half of the peaks in each comparison overlaid well.

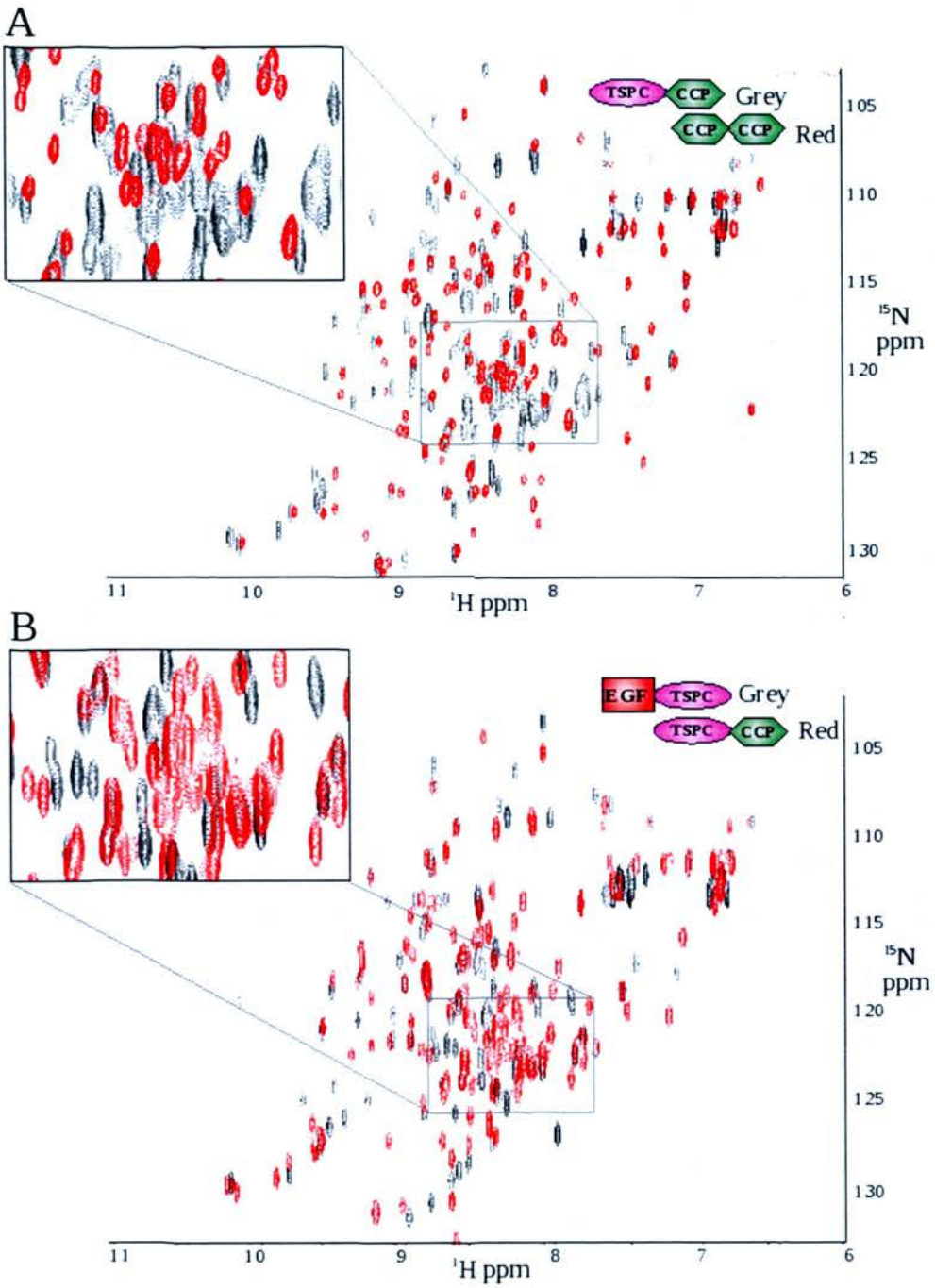


Fig. 69: [^{1}H, ^{15}N]-HSQC comparison of ET, TC and CCP module pairs. (A) shows the overlay of CCPs (red) with ET (grey). (B) Shows the comparison of TC (red) with ET (grey). The more complex central portion of each spectrum is blown-up for better visualisation.

Without resonance assignment (requiring a [^{13}C,^{15}N]-labelled samples) confirmation of the peaks identities was not possible. However in the case of C7-TC a ^{15}N-HSQC-TOCSY was recorded to better identity those peaks that overlay with peaks in the HSQC spectrum of CCPs. Using this method, 51 of a total of 57 expected CCP-assigned cross peaks were positively identified in the C7-TC HSQC spectrum. This confirms that the structure of CCP1 is relatively unchanged in the presence of the TSPC domain. Such a conclusion is consistent with the long linking sequence between them; there are 21 residues separating the (putatively) last residue of TSPC (Glu^{549}, as identified by Pfam[ref]) and the first cysteine of CCP1 (Cys^{560}).

Some CCP1-assigned peaks in the C7-TC HSQC spectrum that had moved significantly compared to the spectrum of C7-CCPs were mapped to the C-terminal end of CCP1 (Fig. 70). This is unsurprising as C terminus-proximal residues of CCP1 obviously interact with CCP2, and this second CCP module is absent in C7-TC. Other residues that were affected were the N-terminal residues of CCP1. Again, this is very much as expected since these residues are likely to interact with the first few residues of the TSPC-CCP1 linker.

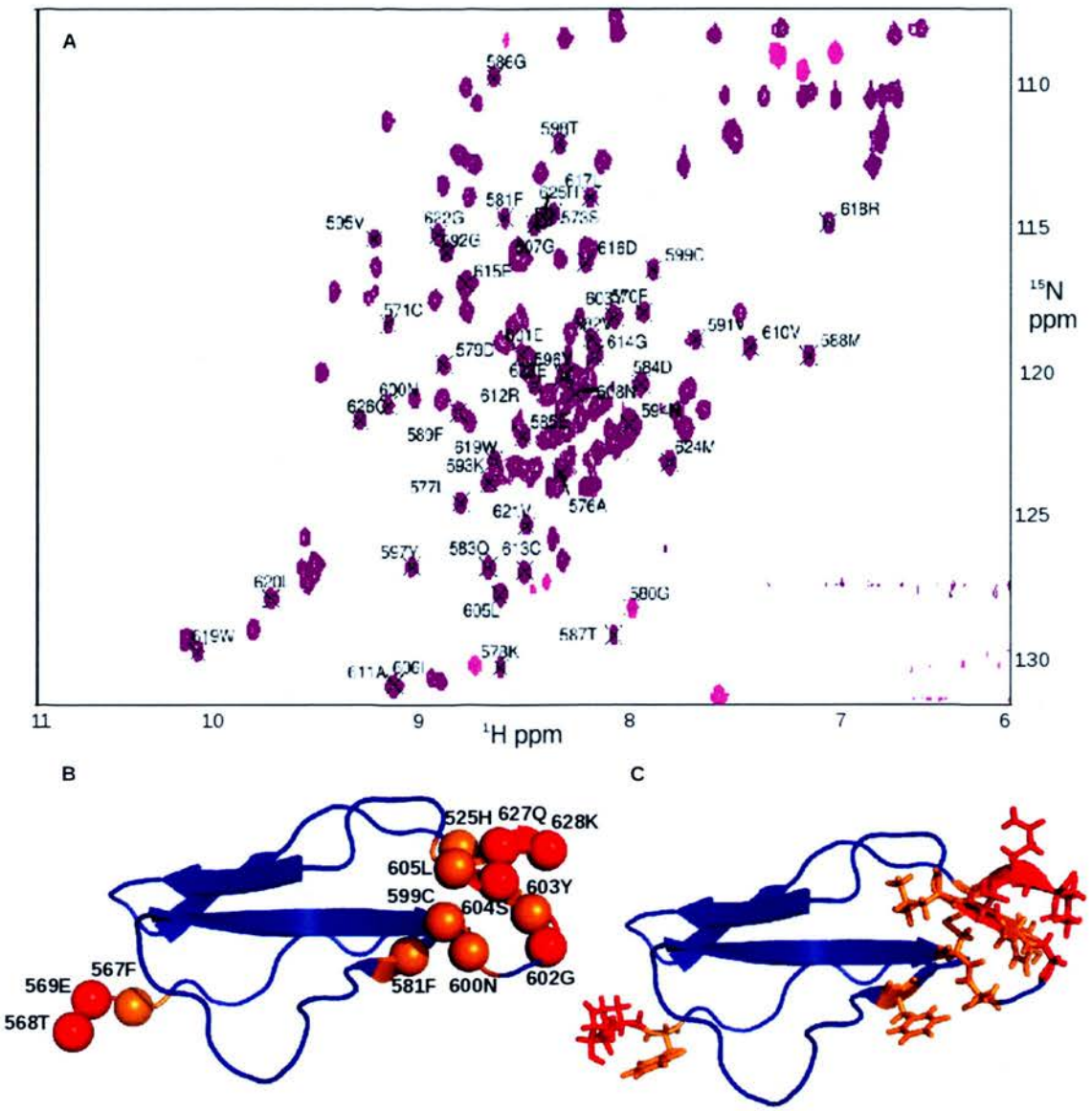


Fig.70: Analysis of the HSQC spectrum of C7-TC and chemical shift "perturbations" arising from absence or presence of the TSPC and CCP2 modules. (A) Shows the assignments of CCP1 cross peaks in the ^{15}N-HSQC spectrum of C7-TC. (B) (spheres) and (C) (sticks) highlight those residues in the CCP1 structure whose amide cross peaks had different chemical shifts in C7-TC versus C7-CCPs (orange) or could not be found in the C7-TC spectrum (red).

There is a cysteine (Cys^{571}) in the middle of the TSPC-CCP linking sequence. The sequence of TSPC has only five of the six conserved cysteine residues (Fig.71), making

Cys[571] a likely candidate as a disulfide-binding partner. Nonetheless there are two other possible binding partners. One can be found in the MACPF (Cys[433], towards its C-terminus), and the other is one residue upstream of the EGF module (Cys[455], module boundaries as outlined by EXPASY[163]). However, from comparison with other EGF sequences the disulfide bonding pattern in C7-EGF is hard to distinguish and the interdomain disulfide bond may be formed with other EGF cysteines (Fig. 71).

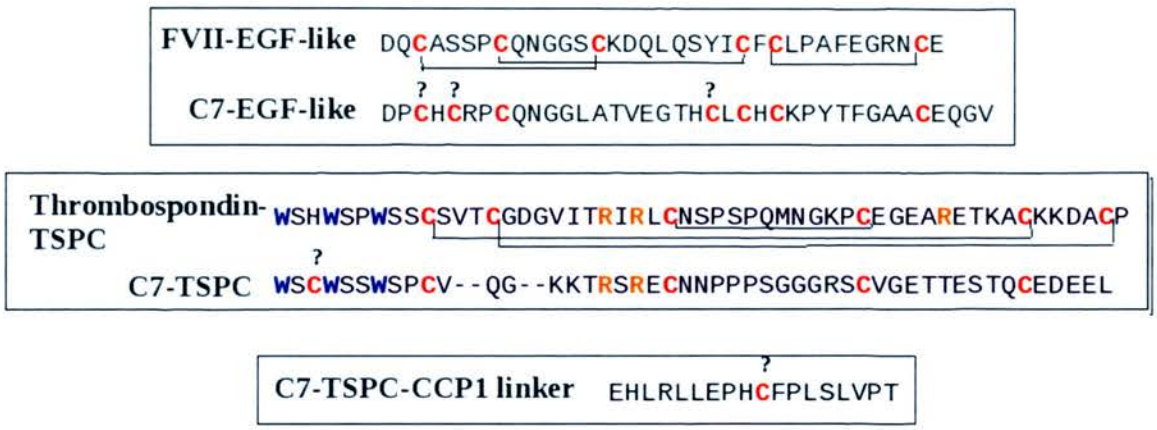


Fig. 71: Free cysteines in the EGF and TSPC domains and the C7-TSPC-CCP1 linker. The top panel shows a comparison of C7-EGF domain with Coagulation factor II, for which structures are avaialable.[ref] The middle panel shows a comparison of C7-TSPC with TSPC1 from human thrombospondin, for which a structure is available.[ref] The bottom panel shows the sequence of the C7-TSP-CCP1 linker. All cysteines are shown in red and those whose disulfide bonding partner is difficult to determine are marked with a questionmark. Well conserved tryptophans and arganines of TSPC are also coloured blue and orange respectively.

An intuitively attractive linkage would be from Cys[433] of MACPF to Cys[455] of EGF and Cys[505] of TSPC to Cys[571] in the TSPC-CCP1 linker. However, a previous study[164] isolated a proteolytic fragment in which Cys[571] (in linker) is disulfide linked to Cys[433] (in MACPF). This arrangement would leave Cys[455] (in EGF) disulfide linked to Cys[505] of TSPC. It is hard to imagine, however, that the TSPC module within the recombinant C7-ET protein (which theoretically should contain this disulfide bond) remains relatively unchanged relative to the TSPC within C7-TC (which it does as judged by comparison of HSQC spectra) despite the loss of a disulfide bonding partner (if it were really linked

to EGF). Interestingly, initial attempts at production of the C7-ET protein resulted in high-Mw multimers that were only observed under non-reducing conditions by SDS-PAGE, indicative of misguided intermolecular disulphide bonds and the presence of free cysteine(s). C7-TC production on the other hand did not share these problems, indicating that there were no free cysteines present in this preparation.
In any case, the TSPC-CCP linker is likely to be conformationally restrained due to its disulphide link with either the TSPC (as experience with the recombinantly produced proteins suggest) or to the MACPF (as per Di Scippio *et al*[164]).

By positively identifying the majority of CCP1 peaks in the C7-TC HSQC spectrum, it can be inferred that the majority of the 79 remaining peaks (out of 82 expected *i.e.* 72 for TSPC, four from the cloning artefact, and six unidentified CCP peaks) in the HSQC spectrum belong to TSPC. The overlaid HSQC spectra of C7-ET and C7-TC (Fig. 69B) show that the inferred TSPC-assigned peaks in the C7-TC spectrum overlay well with a set of cross peaks presumed to be the TSPC-derived peaks in the C7-ET HSQC spectrum. This indicates that the TSPC module structure is not greatly affected by the presence of the EGF domain. There are 15 residues in the linker between EGF and TSPC (Q^{488} to G^{502} as identified by Pfam[162]) consisting of seven small hydrophobic residues (glycines and alanines), four bulkier hydrophobic residues (valine and leucine) and four charged residues. Thus the majority of the linker is hydrophobic, and it is therefore unlikely to be highly solvent exposed in the context of full-length C7. In C7-ET however, without HSQC assignment and relaxation analysis, the flexibiltity and structure of the linker are difficult to assess.

In conclusion all of the domains behave largely as separate entities with respect to one another in solution, save for the CCPs and the FIMs where more extensive interactions (particularly in the case of the FIMs) are observed between modules of the same type .

The long linkers between EGF and TSP, TSP and CCP1, and CCP2 and FIMs, allow for a high degree of flexibility in the C-terminus of C7. While the flexible nature of the CCP2-FIMs linker was demonstrated experimentally by NMR, the extent of flexibility within the other linkers is inferred. The "free" cysteine in the TSPC-CCP2 linker might anchor it to another domain (experimental evidence suggests the MACPF domain) but the ten residues on either side might still allow flexibility between domains.

CHEMICAL CROSSLINKING

5.1 The cross-linking reaction

5.1.1 Strategy

The aim of this study was to investigate the use of intra-molecular chemical cross-links as a means of assessing domain-domain contacts and thereby complementing the high-resolution structural studies performed on selected modules of C7. As crystal structures are available for complement C3 and its activated fragment C3b,[110-113] these multiple-domain proteins were selected as candidates for determining the utility and reliability of the cross-linking analysis techniques under development by the group of Juri Rappsilber (School of Biological Sciences, University of Edinburgh). The intention was to then extend this approach to assess the tertiary structure of C7, which is currently unknown. As described in section 2.4.3.3 representative models of TSPN, LDLRA, EGF-like and TSPC domains (produced using the online homology modelling server PHYRE) were used in conjunction with experimentally derived high-resolution solution structures of CCPs and FIMs for the construction of a model of C7 architecture.

5.1.2 Optimisation and SDS-PAGE

After completion of lysine-lysine cross-linking reactions using the Bis[Sulfosuccinimidyl] glutarate (BS2G) cross-linker (Fig. 72), performed as described in section 2.4.3.1 on samples of C3, C3b or C7, the treated proteins were subjected to SDS-PAGE in attempts to separate the various cross-linked oligomers from the target monomer (Fig. 73A, B and C). A series of cross-linker:protein ratios (1:100, 1:300, 1:1000, 1:3000) were assessed in the case of C3b. A ratio of 1000:1 was considered most promising as almost no free α or β-chain remained, under reducing conditions, using this ratio (Fig. 73A). This indicates covalent cross-linking between the chains had occurred in the majority of molecules. The higher ratio of 1:3000 was not selected, as the band corresponding to cross-linked monomeric protein band was significantly smeary and irregular in comparison to bands obtained with the lower ratios. To maintain consistency between analyses of C3 and C3b, the 1:1000 protein:cross-linker ratio was used for

cross-linking within C3, and similar results were obtained (Fig. 73B). Likewise, for C7, a range of protein:cross-linker ratios was scrutinised (Fig. 73C). In this case, however, there is only one chain and therefore assessment of the extent of cross-linking could not be based on observation of inter-chain cross-linking. Analysis by SDS-PAGE, however, indicated the success of cross-linking for all ratios on the basis of a weak band corresponding to a dimer that was absent in the minus-cross-linker control. Therefore, a range of cross-linking ratios were explored by MS. A ratio of 1:1000 was again selected for further analysis as it produced the most extensive list of cross-linked peptides.

Bands (visualised by staining with Coomassie blue) corresponding to the monomer species were cut out of the gel and in-gel trypsin digested. The tryptic peptides were then analysed by LTQ-Orbitrap-MS/MS to allow identification of any cross-linked peptides. The experimentally derived mass spectra were matched to all potential cross-linked peptides by searching a protein-specific, purpose-built database and a scoring algorithm was applied to assign a degree of confidence to the identification of candidate cross-linked peptides.[152] Furthermore, to assess the reproducibility of the cross-linking experiment two separate cross-linking reactions were performed for all three proteins. Determination of a dependable cross-link list was carried out by Zhuo Chen (Rappsilber group) with only those cross-linked peptides that were reproducible, had mass-errors within 6 ppm of the predicted peptide mass, had exceptional database search scores and had a fully explicable mass spectrum with fragment information for both peptides, were considered high confidence cross-links. This final set of cross-links was used for comparison with the x-ray structures of C3 and C3b and then an equivalent set were used to determine the relative location of domains in C7.

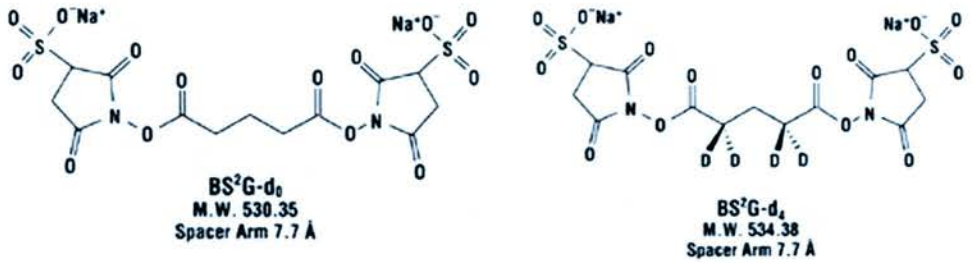


Fig. 72: BS2G (Bis[Sulfosuccinimidyl] glutarate) cross-linkers. Non-deuterated (left) and deuterated (right). Images adapted from Thermo Fischer Scientific.

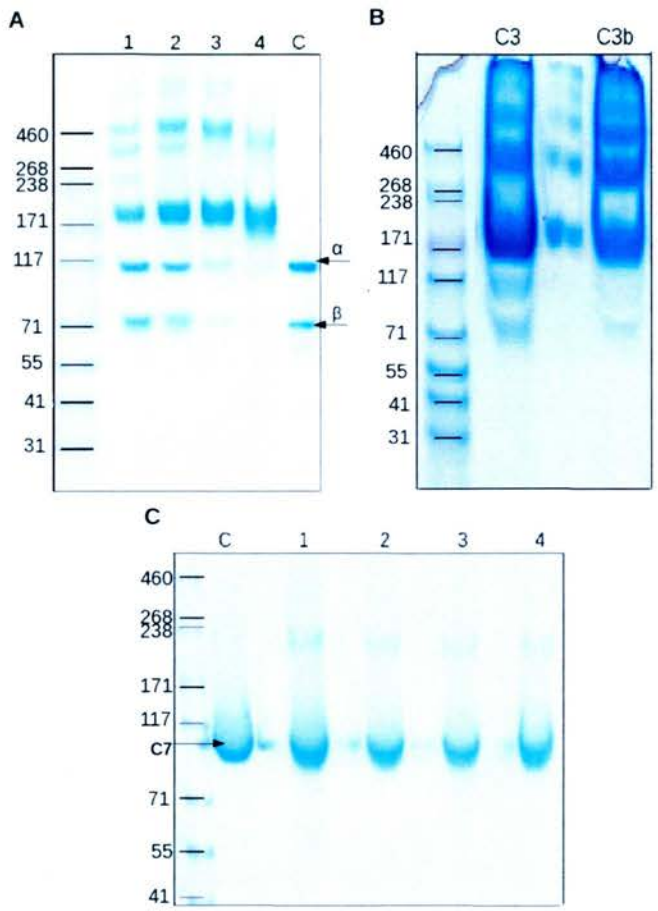


Fig. 73: SDS-PAGE analysis of C3, C3b and C7 cross-linking reactions. (A) C3b:cross-linker ratio assessment. 1=1:100, 2=1:300, 3=1:100, 4=1:3000, C=C3b control (no cross-linker). (B) C3 and C3b cross-linking reaction 1:1000 ratio. 80 µg of C3/C3b was loaded in each lane. (C) C7:cross-linker ratio assessment, 1=1:100, 2=1:300, 3=1:100, 4=1:3000, C=C7 control (no cross-linker) 50 µg of C7 was loaded in each lane.

5.2 The C3 to C3b structural transition

The cross-links identified on the basis of the analysis of the MS data (Table 15 and 16) were then evaluated by referring to the crystal structure of C3 (2A73[110], Fig. 74), the generally accepted crystal structure of C3b (2IO7,[111] Fig.75A) and the disputed C3b crystal structure (2HR0,[113] Fig.75B). Evaluation consisted of determining the distance between cross-linked lysines in the structure and comparing this to the cross-linking distance permitted by the 7.7-Å spacer arm of the cross-linker BS2G. Given the likely mobility of many lysine side chains, this distance was measured between the Cα atoms, and the maximum inter-Cα distance compatible with cross-linking was set at 24 Å (*i.e.* from lysine-A Cα to lysine-B Cα).

5.2.1 C3 analysis

In the case of C3, a total of 36 cross-links were identified with high confidence (Table. 15). Many of these lie wholly within a domain - there are three in LNK, one in CUB, two in TED, three in ANA and five in the C345C domain, all of which are consistent with the crystal structure of C3 (Fig. 74). Also compatible with the structure are a further eight cross-links within and between domains of the MG ring.

C3 monomer crosslinks				
From (K)	To (K)	From domain	To Domain	Distance (Å)
155	502	MG2	MG5	14.61
155	812	MG2	MG6 (α)	10.41
241	428	MG3	MG4	17.82
241	607	MG3	LNK	16.44
241	608	MG3	LNK	17.67
289	682	MG3	ANA	15.49
365	608	MG4	LNK	40.56
418	633	MG4	LNK	15.87
205	608	MG2	LNK	14.8
249	289	MG3	MG3	16.93
249	305	MG3	MG3	10.16
289	305	MG3	MG3	14.4
566	584	MG6 (β)	LNK	5.28
607	610	LNK	LNK	6.27
607	615	LNK	LNK	12.35
608	615	LNK	LNK	9.77
879	1526	MG7	C345C	23.78
927	1436	MG7	MG8	18.5
930	1436	MG7	MG8	13.42
959	1306	CUB	CUB	14.45
1001	1436	TED	MG8	13.56
1368	1497	MG8	ANCHOR	12.11
682	692	ANA	ANA	11.95
688	692	ANA	ANA	8.33
685	692	ANA	ANA	11.92
692	1071	ANA	TED	15.63
688	1071	ANA	TED	19.85
721	1071	ANA	TED	19.93
1113	1139	TED	TED	19.95
1203	1244	TED	TED	13.85
1504	1497	ANCHOR	ANCHOR	15.14
1551	1599	C345C	C345C	12.86
1522	1535	C345C	C345C	11.22
1526	1535	C345C	C345C	13.44
1551	1595	C345C	C345C	13.56
1600	1595	C345C	C345C	11.01

Table. 15: C3 cross-linked peptides identified with high confidence. Domain names are colour-coded as shown previously. Distances between cross-linked lysines that exceed the maximum of 24 Å in the crystal structure are highlighted in red.

More structurally informative are the multiple cross-links between domains. There are four cross-links that demonstrate the location of the LNK domain with respect to the MG ring (Fig. 74D). While three of these four cross-links, between MG4 and LNK, fall within the maximum cross-linking distance, another MG4-LNK cross-link connects residues that are 40 Å (Cα-Cα) apart in the crystal structure. This could well reflect movement of the LNK domain as it is thought to be highly flexible.[110] There are, in addition, single cross-links between the MG domains and C345C (Fig. 74B), anchor (Fig. 74B) and TED (Fig. 74C) that are consistent with the crystal structure and describe the location of these domains with respect to the MG ring. One of the four remaining cross-links connects the C345C with the anchor (Fig. 74B), while the final three connect the ANA to the TED domain (Fig. 74C). As may be appreciated from Figure 74, the experimentally derived cross-links, realistically, reflect the conformation within the two superdomains of C3 *i.e* the MG-ring and CUB-TED-MG8, as well as their location with respect to one-another and the location of the anaphylatoxin and C345C domains.

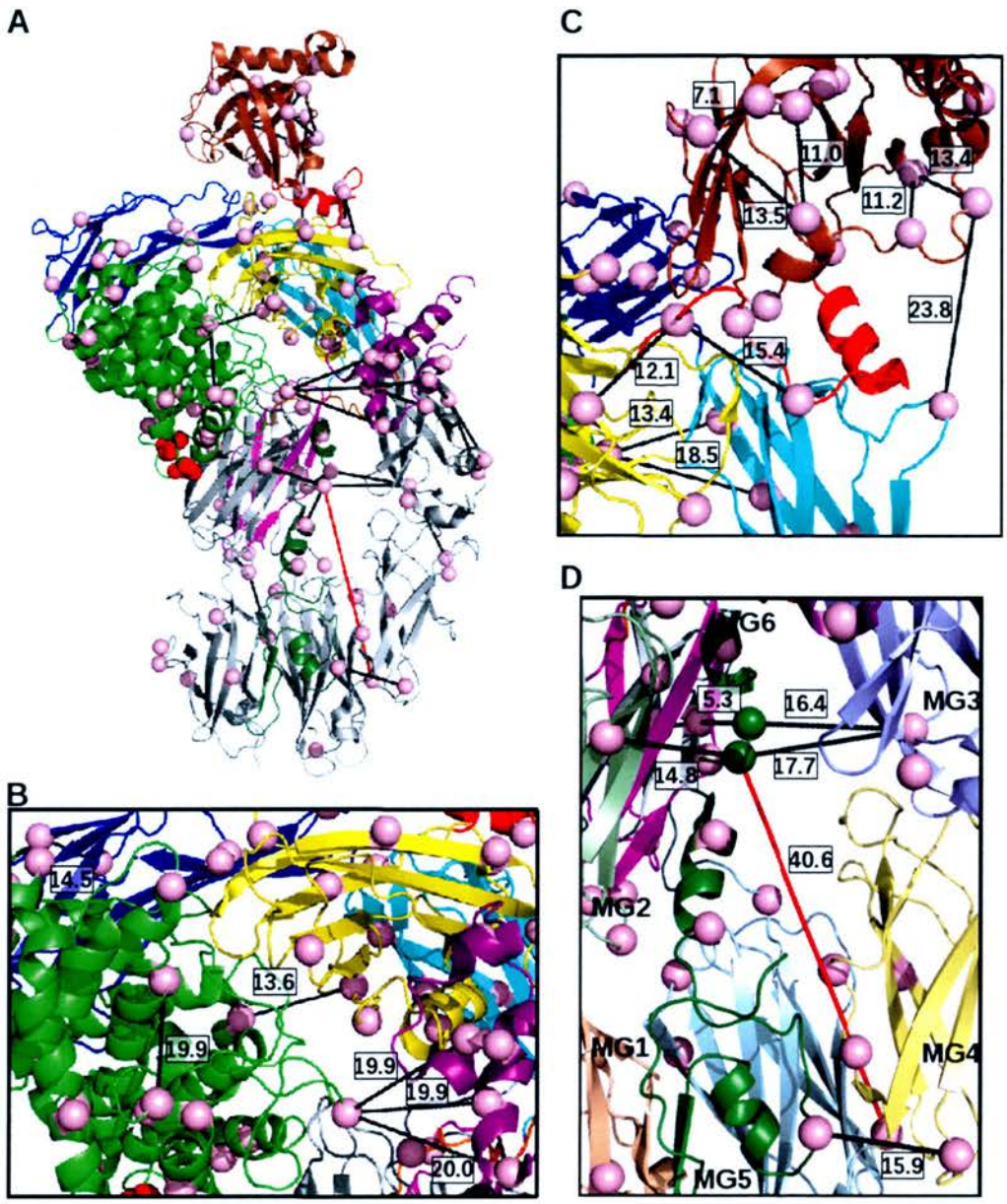


Fig. 74: Cross-links mapped to structures of C3. All diagrams (A)-(D) are of C3 (PDB ID:2A73) in a ribbon representation. (A) depicts the full molecule; (B), (C) and (D) are close-ups of the TED C345C region and LNK respectively. From N to C-terminus, MG1-MG6β are in grey, the ANA is in purple, MG6α in magenta, MG7 in cyan, the CUB in blue, the TED in green, MG8 in yellow, the anchor in red and C345C in brown. All lysines Cα are shown as light-pink spheres and residues of the thioester are shown as red spheres in the TED. In (D) LNK lysine Cα's are in dark-green, with the labelled MG domains differentially coloured for better visualisation. Cross links are shown within 24 Å are shown in black, and those exceeding 24 Å are shown in red.

5.2.2 C3b analysis

In the case of C3b a total of 32 cross-links were identified with confidence (Table. 16). Seven of these lie within and between members of the MG ring and are in agreement with both the 2I07 and 2HR0 crystal structures. Similarly, non-MG intra-domain cross-links are found within the LNK, anchor, C345C and CUB domains. Cross-links to the MG domain are also found for the LNK and C345C domains; these are consistent with both the crystal structures. Thus the relative locations in the protein of the MG ring, the LNK domain, the anchor and C345C are adequately reflected by the cross-links. While all of these cross-links are consistent with the 2I07 structure, a cross-link within CUB corresponds to a Cα-Cα distance of ~40 Å in 2HR0. Indeed, in 2HR0 the CUB domain is structurally disorganised and lacks secondary structure. Either the cross-linker is trapping an unusual conformation of the otherwise unfolded CUB domain, or (more likely) 2HR0 is not an accurate reflection of the true structure.

C3b monomer crosslinks					
From (K)	To (K)	From domain	To Domain	Distance (Å)-2I07	Distance (Å)-2HR0
23	486	MG1 (N-term)	MG5	14.22	13.85
65	1050	MG1	TED	15.03	74.99
66	1203	MG1	TED	38.34	68.03
155	812	MG2	MG6 (α)	10.19	10.81
249	289	MG3	MG3	16.67	16.82
263	891	MG3	MG7	9.65	9.76
289	305	MG3	MG3	14.69	14.41
289	1431	MG3	MG8	17.34	15.5
359	622	LNK	MG4	14.45	15.42
566	584	MG6 (β)	MG6 (β)	5.32	5.3
607	610	LNK	LNK	6.88	6.25
607	615	LNK	LNK	14.22	12.29
608	615	LNK	LNK	12.09	9.72
749	861	α-NT (α N-term)	MG7	/	15.9
749	879	α-NT (α N-term)	MG7	/	10.91
749	1381	α-NT (α N-term)	MG8	/	21.68
749	1526	α-NT (α N-term)	C345C	/	24.31
749	1535	α-NT (α N-term)	C345C	/	23.41
749	1589	α-NT (α N-term)	C345C	/	19.82
879	1526	MG7	C345C	23.45	21.47
879	1535	MG7	C345C	18.56	15.52
879	1589	MG7	C345C	19.14	16.54
904	1504	MG7	ANCHOR	13.27	13.01
959	1306	CUB	CUB	14.77	40.6
1071	1381	TED	MG8	140.02	140.12
1419	1431	MG8	MG8	9.63	8.01
1497	1504	ANCHOR	ANCHOR	11.4	15.55
1497	1589	ANCHOR	C345C	8.23	13.06
1522	1535	C345C	C345C	13.07	11.2
1526	1535	C345C	C345C	13.96	13.5
1551	1599	C345C	C345C	8.09	7.13
1595	1600	C345C	C345C	11.2	11.03

Table. 16: High confidence cross-linked peptides identified in C3b. Domain names are colour coded as shown for C3. Additionally the α-NT is shown in orange. Distances between cross-linked lysines that exceed the maximum of 24 Å in both the 2I07 and 2HR0 structure crystal structures are highlighted in red.

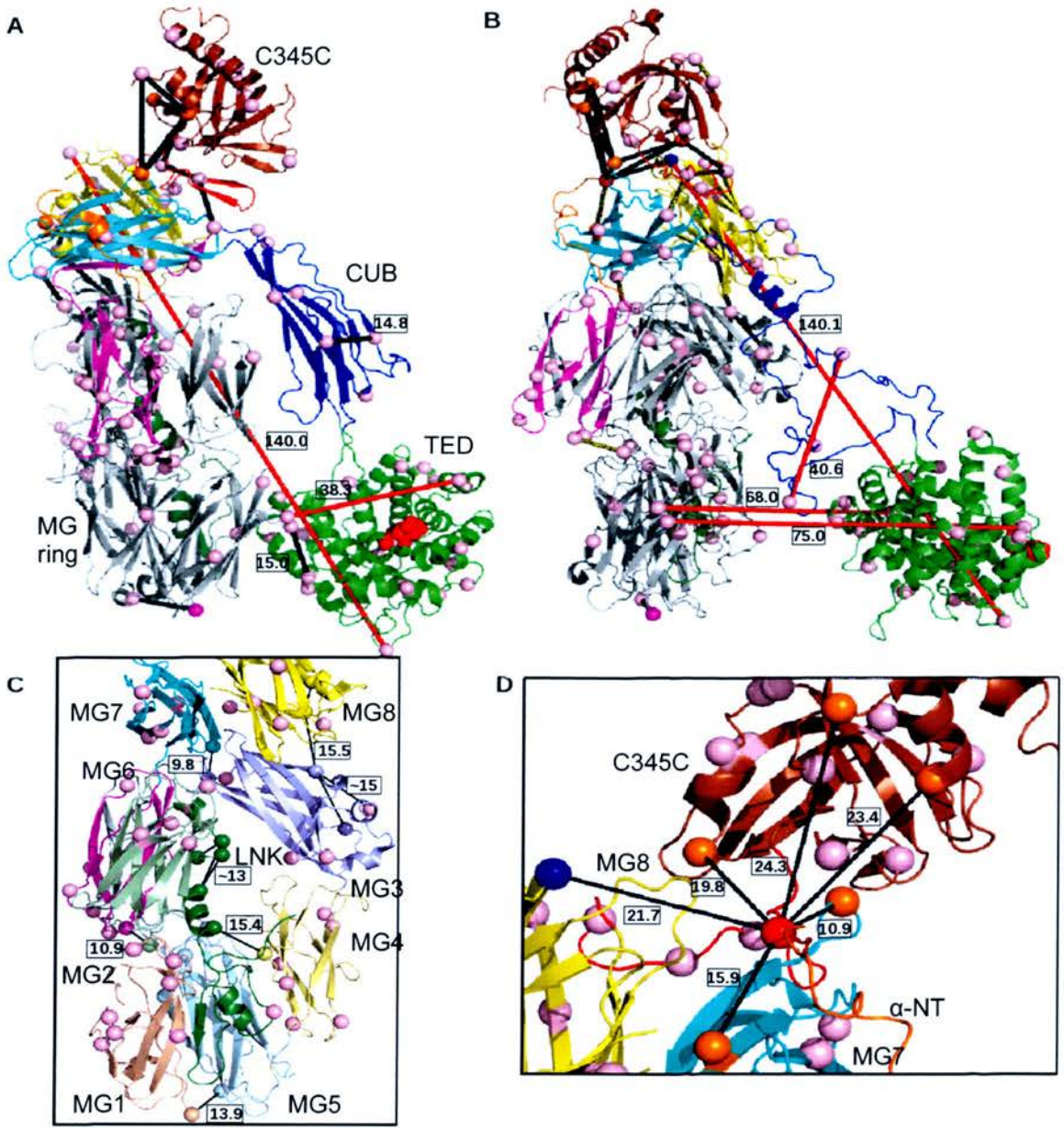


Fig. 75: Cross-links mapped to structures of C3b. ***D***iagrams (A) and (C) are of C3b (PDB ID:2I07) and (B) and (D) are of C3b (PDB ID:2HR0) in a ribbon representation. (A) and (B) depict the full molecule, (C) is a close-up of the MG ring/LNK cross-links and (D) is a close-up of cross links that are absent in 2I07. From N to C-terminus, MG1-MG6β are in grey, the α-NT in orange, MG6α in magenta, MG7 in cyan, the CUB in blue, the TED in green, MG8 in yellow, the anchor in red and C345C in brown. All lysines Cα are shown as light-pink spheres, save for the α-NT and MG8 lysines missing from 2I07 and N-terminal residue shown in orange, blue and magenta respectively. In (C) LNK lysine Cα's are in dark-green, with the labelled MG domains and their lysines differentially coloured for better visualisation. Cross links are shown within 24 Å are shown in black, and those exceeding 24 Å are in red.

Remarkably, six cross-links were identified that highlight the use of chemical cross-linking as a complementary technique to crystallography (Fig. 75D). These cross-links provide for structural assessment of regions that are absent from the crystal structure.

The N-terminal end of the α-chain, the α-NT domain (residues 727-729), and a portion of MG8 (residues 1372-1380) is missing in the 2I07 structure (although present in the 2HR0 structure). Five of these cross-links between the N-terminal residue S^{749} and MG7 and C345C are well within cross-linking distance in 2HR0. Moreover, the sixth S^{749} cross-link connects the α-NT to the portion of MG8 that is missing in the 2I07 structure.

Therefore, the cross-linking data confirms the location of the N-terminal end of the α-chain.

There is one difficult-to-explain cross-link, that connects the "top" (MG8) and "bottom" (TED) of the C3b molecule as seen in the crystal structure (standard view, Fig 75 A,B). These two domains form part of the dimer interface observed in the crystal structure[165] of a C3 dimer. Therefore it is conceivable that a putative C3b dimer shares similar contacts, which would indicate that contamination by dimeric C3b of the monomeric C3b band has probably occurred during the SDS-PAGE gel. This is not surprising given the poor resolution of the high-molecular weight cross-linked protein bands (see Fig. 73B).

Another explanation could be that the C3b used in the experiment is contaminated with other forms of C3. This is conceivable considering the C3b purchased from Complement Technology Inc was derived from pooled human plasma and claimed to be only 95% pure by SDS-PAGE. Complement Technology Inc confirmed that there were trace amounts of antibodies, C5 and complement factors in the sample, however the presence of other forms of C3 was not ascertained. It could similarly be argued that this cross-link is physiologically relevant and that the CUB-TED-MG8 superdomain permits movement

of the TED towards the MG8 domain. So far however, no C3b conformation has been identified that would support this cross-link. Attempts were made by collaborators to purify C3/C3b monomers by chromatographic methods, however they were unsuccessful. Of particular interest are two TED-MG1 cross-links. One of these (Lys^{65}-Lys^{1050}) satisfies the 2I07 structure with a cross-linking distance of 15 Å. However, in 2HR0 this distance was measured at ~75 Å, far exceeding the maximum cross-linking distance, and can not be explained without a significant movement of the CUB-TED-MG8 superdomain in solution. The second TED-MG1 cross-link (Lys^{65}-Lys^{1203}) is inconsistent with the distance in the structure (it is 38 Å) but this could be explained simply by a rotation of the TED with respect to the CUB, via the two six-residue long linkers that pass between them.

The 2IO7 structure is consistent with cross-links within the CUB domain that is located between the translocated (with respect to C3) TED and MG1. Conversely, these CUB-CUB and TED-MG1 cross-links are not consistent with the 2HR0 structure. At first glance the cross-linking data supports the 2I07 over 2HR0 with respect to the TED domains location in solution. The 2HR0 conformation brings many lysines of the CUB within cross-linking distance to lysines of the TED. Yet no cross-links were observed between the CUB and TED. In communications[114, 115] discussing 2I07 versus 2HR0, it was argued that the 2HR0 structure depicts C3b in a more open conformation that might be physiologically relevant. EM studies[83] indicate that in C3b the CUB-TED-MG8 arm swings between different conformations. When the 2HR0 and 2IO7 structures were overlaid with models representing the two main classes of C3b conformations observed by EM, 2IO7 was deemed to be in the class I conformation and 2HR0 is somewhere between the two. As no cross-links were identified to support the CUB-TED-MG8 conformation of 2HR0, it would have to be the case that this structure represents a short-

lived intermediate in solution that crystallises but that cannot be "captured" by cross-linking.

Thus this study of C3 and C3b has shown that a combination of cross linking and mass spectrometry provides a powerful means of observing domain rearrangements, and of differentiating between various molecular architectures, in big multiple domain proteins. Thus, this strategy was next applied to C7.

5. 3 C7 architecture

In a sample of full-length C7 (Complement Technologies Inc.), 19 high-confidence cross-links were identified. Three main groups of cross-links may be considered: those within the MACPF domain (eight), those between MACPF and the smaller modules (eight) and those between the smaller modules (three) (Table. 17). The eight MACPF-MACPF cross-links are all within cross-linking distance in a model of the C7-MACPF built by homology with a crystal structure of the C8α-MACPF (Fig. 76). The presence of these cross-links thus cross-validates the model of C7-MACPF. One of these cross-links connects the d1 region and the d1/d3 β-sheet core, while the remaining seven cross-links are located in the d3 helical cluster region. On the other hand, the MACPF-module cross-links are all located in the d1 region of the MACPF, which, combined with the cross-links between smaller modules (TSPC-CCP1, FIM1-LDL and FIM1-CCP2), and an absence of MACPF-MACPF cross-links in this region, indicates that the smaller N and C-terminal modules are predominantly packed around the d1 region of the MACPF (Fig. 76).

From (K)	To (K)	From (domain)	To (domain)	Distance (Å)
221	420	MACPF	MACPF	9.81
244	323	MACPF	MACPF	16.49
323	354	MACPF	MACPF	17.31
323	361	MACPF	MACPF	17.27
325	354	MACPF	MACPF	17.98
325	361	MACPF	MACPF	18.05
338	354	MACPF	MACPF	10.72
338	400	MACPF	MACPF	10.1
99	159	LDLRA	MACPF	/
167	684	MACPF	CCP2	/
159	706	MACPF	FIM1	/
167	735	MACPF	FIM1	/
167	706	MACPF	FIM1	/
167	735	MACPF	FIM1	/
167	771	MACPF	FIM1-FIM2 linker	/
167	788	MACPF	FIM2	/
99	735	LDLRA	FIM1	/
517	593	TSPC	CCP1	/
684	706	CCP2	FIM1	/
691	706	CCP2	FIM1	/

Table. 17: High confidence cross-linked peptides from C7. Domain names are colour coded as shown previously. Distances shown are for MACPF-MACPF cross-links distances in the C8 α-MACPF homology model.

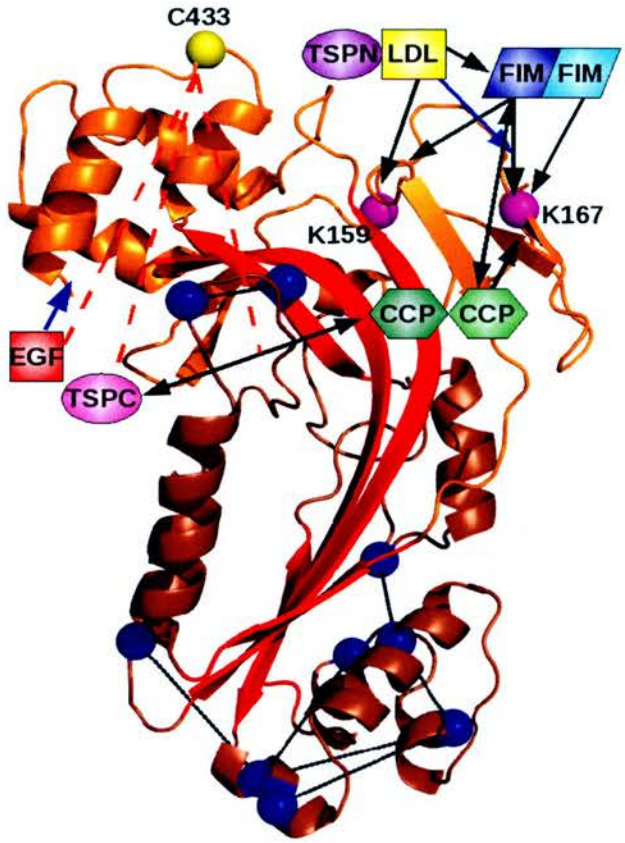


Fig. 76: C7-MACPF inter and intra-modular cross links. C7-MACPF model is shown with d1 region in orange, central β-sheet in red and helical clusters in brown. Schematic representations of the modules are coloured as shown previously: purple, yellow, red, magenta, dark-green, green, blue and cyan for TSPN, LDLRA, EGF-like, TSPC, CCP1, CCP2, FIM1 and FIM2 (from N to C-terminus) respectively. Key MACPF-module cross-linked lysines Cα's are shown as spheres (mahenta), MACPF intramodular cross-linked lysines Cα' s are blue spheres. Cross-links are shown as black arrows or lines, peptide-links are shown as blue arrows and potential cysteine-links are shown as red-dotted lines.

Of the eight identified MACPF-smaller module cross-links, six link the MACPF to FIM1, FIM2 and the linker between them. The cross-links map to one face of the FIMs with FIM1 cross-linked lysines located in both the FOLN and KAZAL sub-domains while in FIM2 they are located in the KAZAL domain only (Fig. 77). Five of the six cross-links involve the same residue (Lys^{167}) in the MACPF. Attempts to orient the FIMs domain with respect to the MACPF (by re-orientating the independent modules in

PYMOL and assessing cross-linking distances using the measurement wizard) were successful (Fig. 77). However, if all cross-link inferred distances were to be satisfied simultaneously, the FIMs would occupy a fixed position with respect to the MACPF. The cross-linked lysines are primarily situated around the cleft between domains, thus confirming this region as a site for protein-protein interactions. Although the cross-linking-inferred distance restraints are satisfied by a model in which the FIMs remain compactly folded and pack snugly against the relevant portion of MACPF, there are other explanations. For example, the FIMs may be flexibly attached, with cross-linking capturing a range of different conformations. Or the FIMs could adopt an opened up conformation in this setting, and thus wrapped around the N-terminal end of the MACPF, they could more easily satisfy all of the distance restraints. A flexible or open attachment is more probable, as the FIMs in the 'fixed orientation (Fig. 77) do not allow incorporation of LDL-FIMs and CCP2-FIMs cross-links (Fig. 78 and 80) into the model. To do so requires a rotation of the FIMs and the violation of one or more of the cross-link-derived (MACPF-FIMs) distances.

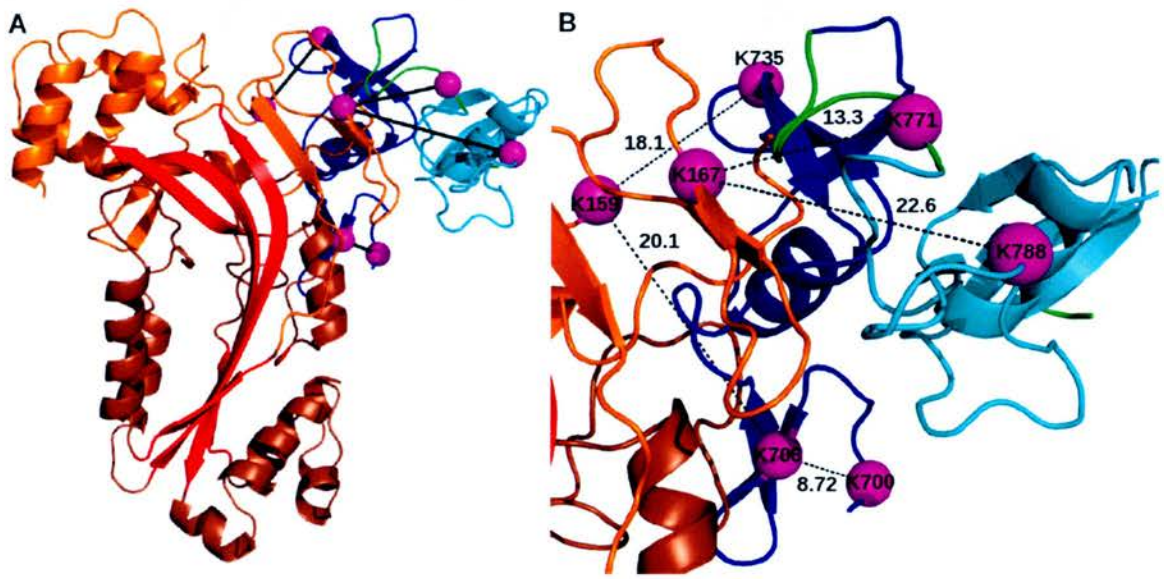


Fig. 77: Cross-linking inferred location of C7-FIMs with respect to the C7-MACPF. (A) The MACPF colour coded as in Fig. 76. Cross-linked lysine Cαs between FIM1 (blue), FIM2 (cyan), and the FIM1-

FIM2 linker (green) and the MACPF are shown as spheres (magenta). A close-up view is shown in (B).

The solitary MACPF-CCP2 cross-link also locates CCP2 (Lys^{684} - Lys^{167}) on the MACPF. The same residue in CCP2 also has a link connecting this module to FIM1 (Lys^{706}) which also has a cross-link to Lys^{167}. Thus, despite the long flexible linker described by the NMR structural analysis of C7-CFF (see section 4.3.4), in the context of full-length non-complexed C7, the CCPs and the FIMs can be found relatively close in space as can be seen in Fig. 78. No cross-links that involved Lys^{691} and Lys^{701} in the CCP-FIMs linker were observed. Therefore whether the linker is found close to the main body of the molecule or remains loose is undetermined. Although no cross-links linked CCP1 to the MACPF, the rigidity of the CCP pair as determined by NMR (see section 4.2.4) is likely to restrict the CCP modules movement with respect to one another. An attractive option is that the concave face formed by the 84° bend between the CCP modules allows the modules to wrap around the MACPF (Fig. 78C). Thus they act as a bridge between FIMs located at the N terminus of the MACPF in the d1 region, to the C-terminal end of the MACPF from which the EGF-like domain and TSPC protrude (Fig 78).

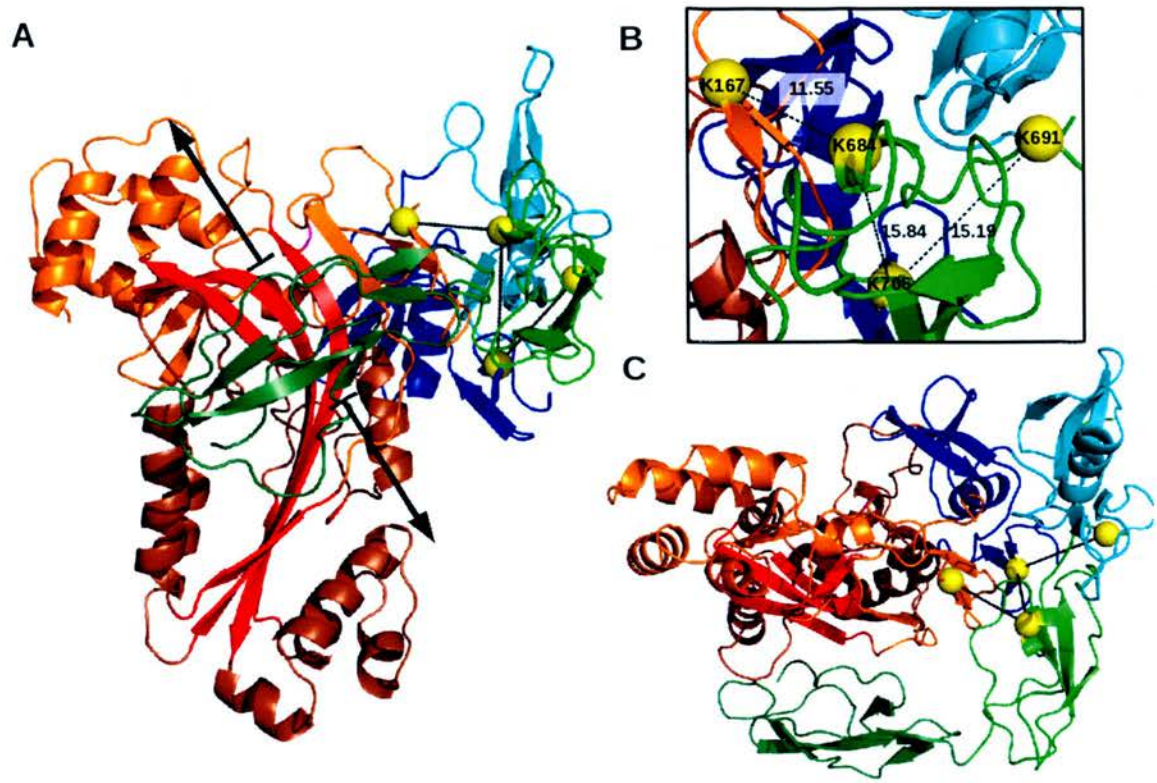


Fig. 78: Cross-linking inferred location of C7-CCPs with respect to the C7-MACPF and C7-FIMs. (A) The MACPF colour coded as previously shown, with cross-linked lysine Cαs shown in yellow. Arrows indicate that the CCP1 location is not restrained by cross-links. (B) a close-up view of the cross-links. A top or 'birds-eye-view' of (A) is shown in (C).

Looking further towards the N-terminal end of the molecular arm, neither TSPC nor EGF-like undergo cross-linking to the MACPF. Note, there are two neighbouring lysines (Lys^{516} and Lys^{517}) in TSP-C. Highlighting these residues on a model of TSP-C (Fig. 79) reveals their proximity to the C-terminus of the module. As Lys^{517} is found within cross-linking distance of CCP1 (despite the 23 residues between the last cysteine of TSPC and the first of CCP1), these lysines may be within cross-linking distance of lysines on MACPF but be unable nonetheless to form cross-links. This could be for a number of reasons. For example, there may be incompatible Lys-Lys side-chain orientations or the CCP module may simply obstruct bridging by the cross-linker as shown in Figure 79C. In this theoretical representation of TSPC, the long TSPC-CCP2 linking sequence can

easily accommodate a disulfide bond, while maintaining the TSPC and CCP1 in cross-linking range. By extension, this would locate the EGF-like module close to the free cysteine of the MACPF for disulfide bonding, and the EGF-like domain would therefore not be packed against the putative TMHs (see below). Alternatively, if the cysteine in the TSPC-CCP1 linking sequence were disulfide bonded to the MACPF (as suggested by Di Scippio[REF]) the TSPC would be have to be sandwiched between the MACPF and the CCP1 in order to keep the TSPC and the CCP1 within cross-linking range.

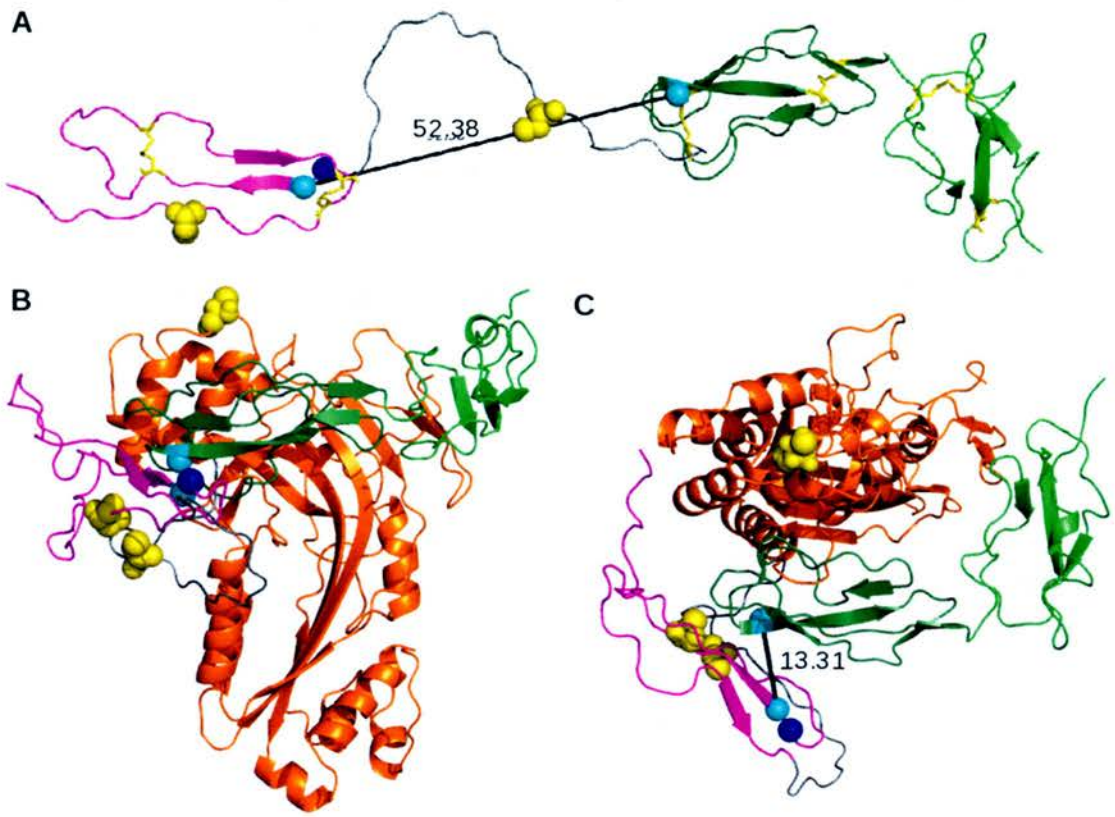


Fig. 79: Cross-linking inferred location of TSPC with respect to CCP1. (A) An extended N to C-terminal view of TSPC (magenta), linker (grey), CCP1 (dark-green) and CCP2 (green). Free cysteines are shown as yellow spheres, disulfides as sticks, cross-linked lysines as cyan spheres and the 'free' lysine of TSPC as a blue sphere. In (B) the MACPF is coloured entirely in orange for simplicity and a 'birds-eye-view' is shown in (C).

The EGF-like module on the other hand contains only one lysine, and in the homology model this is located at the C-terminal end of the module. However, EGF-like domains

are hard to model due to the existence of many similar but different sub-types and so the fold of the module and the true location of its lysine cannot be ascertained reliably. Furthermore, given that we have detected no cross-links for this lysine, there is no reason to assume its location is restricted to the d1 region of MACPF. Instead it may occupy the position relative to the MACPF domain that resembles its equivalent in perforin (Fig. 81). In perforin the EGF-like domain is packed against the TMH regions of the MACPF. The aforementioned variable nature of EGF-like domains however makes modelling the C7 EGF-like domain on perforin difficult; the six conserved cysteines are in different positions (and there is an additional cysteine in perforin's EGF-like domain, Fig. 80). If the EGF in C7 were packed against the TMHs, the free EGF cysteine would be too far from the MACPF's free cysteine to form a disulfide bond. In this instance the TSPC and the TSPC-CCP1 linker would have to span the TMHs and the free cysteine of the MACPF to form a full set of disulfides.

C7-EGF-like	CHC-458-**RPCQNGGLATVEGTHCLCHCKPYT F**-482-GAACE
	RPC+ +G + + C C C+
Perforin-EGF-like	CS-379-**RPCRSGQHKSSHDS -CQCECQDSKV**-401-TNQDCC

Fig. 80:C7-EGF-like and Perforin-EGF-like sequence alignment using BLAST2P.[166] Highlighted in bold is the section of the sequences that were selected by BLASTP for alignment.

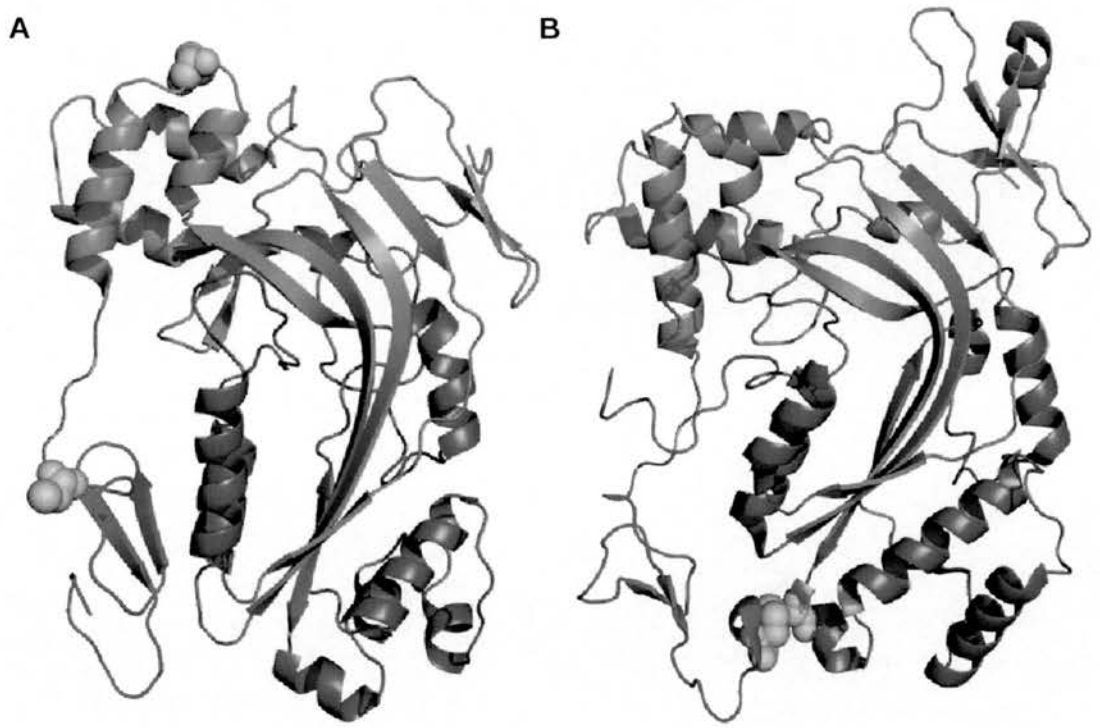


Fig. 81: The MACPF-EGF pair modelled on the perforin MACPF-EGF pair. (A) shows C7-MACPF-EGF in the arrangement found in perforin.[79] MACPFs are coloured with the d1 in orange, the d1/d3 β-sheet in pink and the TMHs in brown. The EGF-like module is in red and the EGF-MACPF linker is in grey. The free cysteines of EGF's and MACPF are shown as yellow spheres in A. The MACPF-EGF disulfide link of perforin is similarly depicted in (B).

Considering the N-terminal modules; the compact 39-aa LDLRA domain is peptide linked to the N-terminal end of the MACPF with a relatively short five residue-long linker. The MACPF N-terminus is located in the d1 region according to the homology model, and therefore the LDLRA domain is likely to be located directly adjacent to the putative FIMs:MACPF interface. This juxtaposition is consistent with the detection of a cross-link between the LDL and FIM1 , and this provides additional evidence for the location of the FIMs (Fig. 82). The location of the N-terminal module TSPN, that immediately precedes LDLRA in the sequence, remains to be established since there are no cross-links involving lysines in this domain. The relatively short linker between TSPN and LDLRA (five residues from consensus cysteine to consensus cysteine)

implies that these two modules will be close in space.

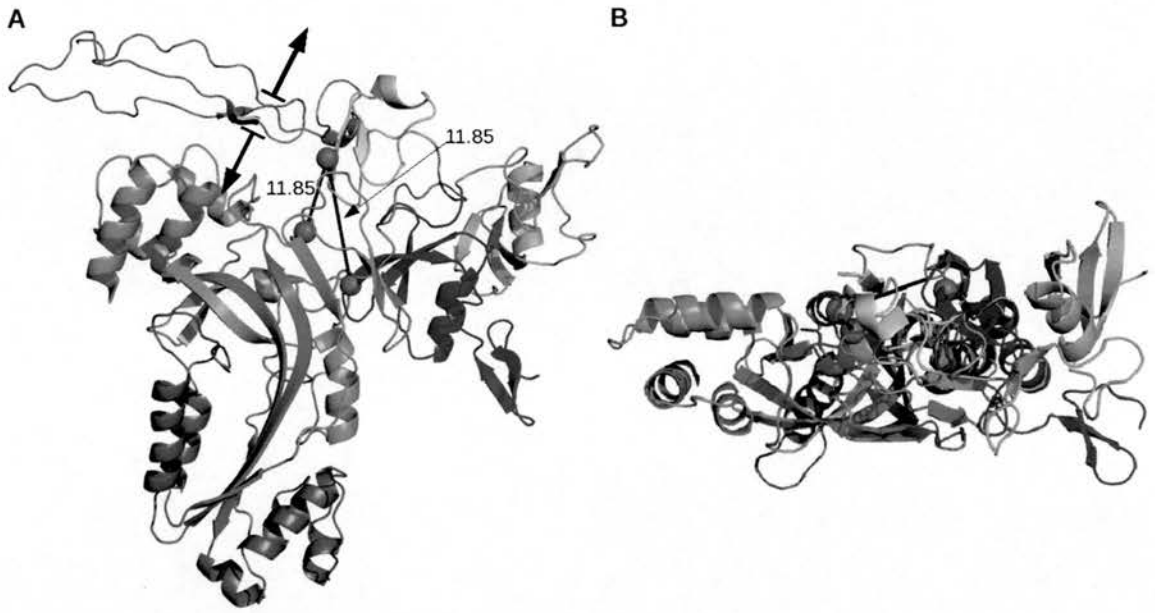


Fig. 82: Cross-linking inferred location of LDLRA with respect to MACPF and FIMs. (A) The MACPF is colour coded as previously shown (Fig. 76), TSPN is in purple, the LDLRA is in yellow and FIM1 and FIM2 in blue and cyan respectively. Cross-linked lysine Cαs are shown as green spheres. Arrows indicate that the location of the TSPN module is not restrained by cross-links. (B) shows a 'birds-eye-view' of A, without TSPN.

In summary, although no cross-links were observed for the EGF-like domain and TSPN the relative location of the other C-terminal and N-terminal domains is likely to be close to the d1 region, on the basis of either direct MACPF-module cross-links or from inspection of module-module cross-links. Models of each module (along with the NMR-derived solution structures of the CCPs and FIMs) were used to create a model representation of the tertiary structure of C7 (Fig. 83). Although the EGF-like module is shown in close proximity to the MACPF's d1 region, it similarly could adopt a different location as discussed previously. The implications of this "top-heavy" domain organisation is discussed further in section 6.2.

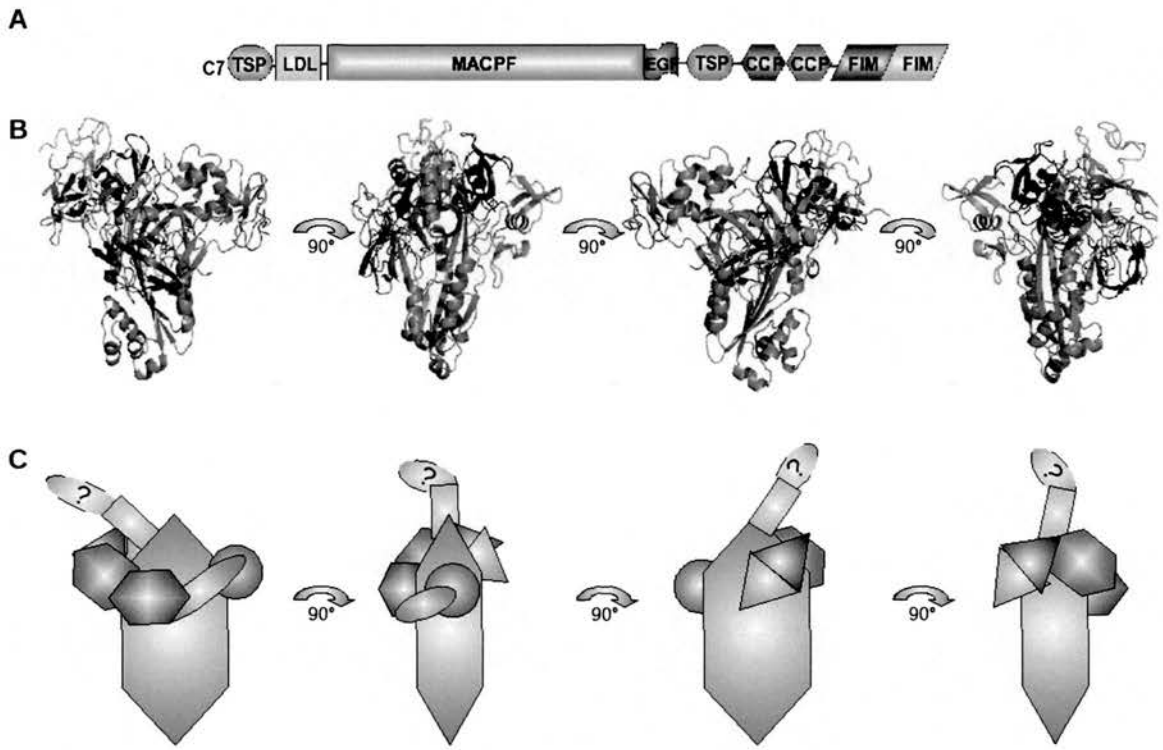


Fig. 83: Model of C7 Architecture. (A) A schematic representation of C7 to better visualise the colour scheme. (B) Four cross-link indicated orientations of C7 modules and the MACPF with linkers between modules (linkers created via Modeller[ref]). (C) Schematic representation of tertiary structure as in (B).

DISCUSSION

6.1 Critique of techniques employed in this study

6.1.1 Protein production and purification

6.1.1.1 The pET15b/OrigamiB pLysS expression system

The use of the pET15b/OrigamiB pLysS expression system for the production of the disulfide-rich protein modules from *E. coli* used in the current study made a strong contribution to the overall success of this project since it provided the C7-CCPs, C7-ET and C7-TC samples used for NMR data collection. Even employing this tailor-made system, however, half or more of the protein produced was incorrectly folded and packaged into inclusion bodies; this occurred for all of the pET15b constructs. Although there are many established protocols for refolding disulphide-rich proteins from inclusion bodies[667] - some of which have been applied to CCP modules - the precise conditions that give efficient refolding differ between protein sequences and thus there is a requirement for the sampling of a variety of conditions, with no guarantee of eventual success. In any case, the very large number of disulfides in, for example, C7-CFF (11) would be likely to pose a challenge even to the Origami system, while re-folding such a protein might well be unfeasible.

6.1.1.2 The pPICzαB/*P.pastoris* expression system

Recloning into pGAP/pPIC vectors and subsequent protein production and purification from *P. pastoris*, was a reasonable alternative that was explored in the current project and deployed successfully for the triple-module construct C7-CFF that was eventually used in the structural studies. This system has proved suitable (in the Barlow group and elsewhere) for producing numerous disulfide-rich proteins, notably intact complement factor H with 40 disulfides.[168] While *P. pastoris* growths are much slower than *E. coli* ones and the genetic manipulation required is more involved, the yeast system has advantages in that the folded protein is secreted into the medium, cell densities can be very high, and there does not appear to be issues with disulfide formation or insoluble products.

The positive results obtained for all the re-cloned constructs in "miniscale" protein production trials indicates that production in *P. pastoris* of the MAC protein truncations really was a viable alternative to production in *E. coli*. Other higher eukaryotic expression hosts, such as mammalian and insect cells, have the ability to perform complex post-translational modifications that are more similar to the human processes. However stably over-expressing genes in these cells is a more demanding task, and generally results in the production of much smaller quantities of protein.[169, 170] Furthermore, isotopic enrichment is problematic in such cells that require complex media for culture. Thus *P. pastoris* represents an excellent compromise between mammalian/insect cells and prokaryotic cell lines. In the case of C7-CFF production, fully folded protein was obtained in sufficiently high quantities for NMR-derived structure determination using double-labelled samples.

6.1.1.3 CCPs purification

With regards to CCPs (produced in *E. coli*) and its purification, the incorporation of DNA encoding the His-tag into the vector allowed standard metal-affinity batch purification. This allowed a relatively straightforward and standardised route to purified protein: metal chelate affinity chromatography coupled with thrombin cleavage of the His-tag, with the requirement of only one final 'polishing' step (size-exclusion chromatography) to remove any impurities (as visualised by SDS-PAGE). Isotopic labelling was acheived in minimal growth media, which consequently reduces cell growth and produces less protein relative to rich medium.[171] The labelling protocols used however, sufficiently overcame these issues, by using the commercially available isotopically labelled ISOGRO® growth (Sigma-Aldrich Co.) supplements in M9 minimal media. The resultant sample contained highly purified and fully folded protein, of the expected Mw (13,985.1kDa), at a yield acceptable for structure determination by NMR.

6.1.1.3 C7-CFF purification

The purification protocol used for C7-CFF was typical of methods used to isolate proteins from *P. pastoris*. The secreted nature of the recombinant proteins provides a major advantage in this respect since there are generally only a few other secreted proteins present in the medium. Thus the target protein is normally, and was in the case of C7-CFF, the strongest band on an SDS-PAGE gel and this eases identification and purification. Fortuitously the pI of C7-CFF (=6.4) permitted cation-exchange chromatography (at a buffer pH of 4.0) as the initial harvesting and purification step. This is the preferred method of choice when resolving recombinant proteins from *P. pastoris*-native ones because using anion-exchange chromatography results in co-purification of the recombinant protein with various *P. pastoris* proteins. It is also important to include such a rapid first step in the protocol since any ruptured *P. pastoris* cells in the medium are a notorious source of proteases, meaning that storage of medium can be problematic and that protease inhibitors should be used as a preventative measure. One of the post-translational modifications performed by *P. pastoris* is glycosylation. The KM71H stain of *P. pastoris* used, however, is not engineered to incorporate mammalian-type glycans (although in recent years such strains have become available[172]). Rather it will yield glycoproteins that are hyper-mannosylated. As a result, the glycosylated C7-CFF, that contains a single consensus site for N-glycosylation, must be deglycosylated leaving an N-acetyl glucosamine "stub" on Asn^{754} as detected by MS. Thus a limitation of our approach was that we were unable to investigate the potential biological role of the native glycan. Only one further, final purification step (size-exclusion chromatography) was needed to produce samples of a sufficient purity for structural determination by NMR.

6.1.2 NMR analysis of C7

6.1.2.1 CCPs

The purpose of studying the C7-CCPs was to ask the question: do these modules form a

rigid spacer arm and if so, how long is it? While it would obviously be most informative to study CCPs in the context of full-length C7 or at least the entire C-terminal arm, a divide-and-conquer strategy was adopted here based on the likely difficulties of producing recombinantly, purifying and crystallizing these larger constructs with their many domains and mutiple disulpfide linkages.

CCPs have been extensively studied by NMR (over 60 solution-NMR structures currently deposited in the PDB). They are excellent candidates for NMR due to their relatively small size (~60 aa) and compact fold based around a conserved hydrophobic core and two invariant disulfides. The C7-CCPs proved not to be an exception as even the most complex experiment recorded on this protein (the ^{13}C-NOESY-HSQC) produced high-quality spectra with well-resolved peaks. The excellent quality of the acquired NMR data was reflected in the ease of CYANA and CNS-based assignment and structure calculations, as indicated by the ease with which the CYANA -associated criteria for a reliable NMR-derived structure were satisfied, the excellent convergence of structures following simulated annealing in CNS (~40/100) and the very low RMSDs for the backbone overlay of the final NMR ensemble of 20 model structures (CCP1 = 0.4 Å, CCP2 = 0.5 Å and both = 0.8 Å).

The two CCPs had a relatively small (by the standards of specific protein-protein interactions) interface (440Å^2 buried surface area) that nonetheless seems to rigidify the junction between the domains. Residual dipolar coupling constants could be recorded in order to more thoroughly establish intermodular angles. This is not necessarily trivial, though, since it requires partial alignment of the sample (see below). The number of intermodular and module-to-linker NOEs observed in C7-CCPs (and the small RMSD value for the overlaid structures) do provide a degree of confidence in these structures derived without recourse to recording RDCs.

6.1.2.2 CFF

The structure determination of C7-CFF was required to address two principal unanswered questions. Is the closed conformation of the FIM modules (with respect to one another), which lie at the very C terminus of C7, altered by the presence of the attached CCP? More generally, what is the extent of the interactions between all three modules?

For this study we would have ideally used a C7-CCFF four-module construct, but at 30kDa this would have represented a serious challenge to standard NMR methods, particularly if it had turned to have an elongated structure (associated with anisotropic tumbling and rapid T_2 relaxation). Even the recombinant C7-CFF (~24 kDa) used in this study was on the large side for NMR-derived structure calculation without resorting to expensive and time-consuming protocols for deuteration or selective labeling.[173] In the current work [^{13}C-^{15}N]-labelling proved sufficient (although only marginally) for structure determination based on recording the standard suite of NMR experiments on an 800-MHz spectrometer for nuclear assignment and NOE-based structure calculation. Analysis of ^{15}N relaxation rates highlighted the conformational freedom in CFF (see below). This very likely improved the quality of the spectra and therefore the feasibility of our strategy. If the two module types (CCP2 and FIMs) had shared a large buried surface area then more exotic labeling would almost certainly have been necessary.

The NMR spectra of CFF were indeed hard to interpret due to the large number of cross-peaks in the spectrum and low ^{15}N T_2 values (expecially for the residues in FIMs) that caused line broadening and overlap. Within the HCCH-TOCSY spectrum, for example, strips pertaining to the FIMs component of CFF were of poor quality due to inefficient TOCSY transfer and their assignment was a tasking procedure. Maximum entropy processing allowed the deconvolution of merged resonance that was essential in picking

peaks in the ^{13}C-NOESY-HSQC spectra. A drawback, however, to using maximum entropy processing is that it can distort relative peak intensities and thus the inferred distance restraints.

The NOESY peaks submitted for automated assignment by the CYANA program were abstracted from a total of four processed NMR spectra. This was necessary because although some NOESY strips were clearer in the maximum entropy-processed spectra, some peaks were lost by this method and some strips were plagued with artifacts; hence spectra that had been processed with and without maximum entropy were used. The high spectral overlap within the ^{13}C-NOESY-HSQC experiment is likely to have prevented the picking of many lower-strength peaks that were crowded by stronger peaks. This reduces the number of restraints available and will have a consequence for convergence during simulated annealing in CNS.

Overall, the poorer quality data for CFF (compared, for example, to CCPs) is reflected in the failure to meet all of the CYANA criteria (particularly with regard to the RMSD values obtained after the first round) and in the smaller proportion of low-energy, well-converged structures within the final ensemble. The non-globular and multiple-domain nature of CFF probably also contributed to the failure to meet some of the initial CYANA criteria.

Despite this, the resultant CNS-calculated water -refined ensemble had low backbone RMSD values when overlaying on residues within individual modules (CCP2 in particular, but also, FIM1 and FIM2), good coarse-quality packing scores (a check of the normality of the local environment of individual residues) and satisfactory Ramachandran plot statistics. Moreover, comparison with structures of FIMs and CCPs solved previously indicated that the expected fold had been calculated for each of the

three domains in CFF, albeit at a lower resolution. Given that CCP2 is properly folded, it seems high unlikely that the inclusion of CCP1 in this construct (avoided due to the NMR size-limit) would have had any effect on the interaction between CCP2 and the FIMs, This vindicates the decision to work with the smaller construct (note that a CCF construct would be unlikely to be of much use on the basis that the FIMs form an intimate interface with one another and it is doubtful that FIM1 would be properly folded in the absence of FIM2).

The structure of C7-CFF (or C7-CCFF) could alternatively have been studied by X-ray crystallography. The highly flexible nature of the CCP2-FIM1 linker may, however, have proved problematic during crystalization trials. Even in the event of good quality crystal formation, X-ray techniques would not have allowed for the determination of the flexibility of the linker since it is genuinely mobile as confirmed by the ^{15}N-relaxation studies. Moreover, it is possible to selectively 'freeze' out one conformation under crystallography conditions that are normally high in salt. Indeed crystallographic conditions could also disrupt key interactions or, as seems more likely, induce closer packing of the CCP and FIMs.

Residual dipolar coupling (RDC) measurements provide information on the global folding of a protein or protein complex and long-range distance structural information that illuminates orientational restraints (as opposed to short-range distance restraints derived from NOEs). RDC's are recorded by partially restraining the orietentation of a protein in liquid crystal alignment media (*e.g.* bicelles, fillamentous phage, cellulose crystallites). The RDC values can in principal determine the relative orientation of two bond vectors even if they are at opposite ends of the molecule. However there are technical complexities encountered when using RDC for the study of a protein that exhibits reciprocal conformational freedom as in C7-CFF.[174] Thus measured RDCs may

not in any case have been sufficient to describe the orientation of the FIMs with respect to CCP2.

RDC's measurements can also provide a more extensive dynamics analysis than T_1/T_2 values and heteronuclear NOE measurements, because they additionally include time-scales close to the overall correlation time of the protein in question. This degree of dynamic analysis could in theory allow a much more useful characterisation of the amplitude of motion between the module types. However, dynamics analysis *via* RDC's requires that measurements are taken in at least five distinct alignment media. This is inevitably a time-consuming process that involves testing the success of alignment in a variety of media, and therefore requires larger amounts of labelled protein than were produced in the current study.

Recording SAXS data on the other hand, is far less time-consuming and can be performed on unlabelled protein. It proved an ideal method to assess the orientation of modules within CFF. With the FIMs together being much larger and bulkier than an individual CCP module, fitting of the NMR-derived structures to the SAXS-derived shape envelope was easily achieved and did not require the use of specialised software to obtain a good fit. The SAXS-derived shape envelope was only compatible with half of the ensemble of selected C7-CFF structures from NMR. This implies that the SAXS data could have been used as a valuable restraint in the simulated annealing, NOE-based structure calculation. There are straightforward algorithms available for doing this and it would certainly have been worthwhile if time had allowed.

In conclusion, despite the fact that the structure of C7-CFF is of poorer quality than the structures of C7-CCPs and of C7-FIMs, it is nonetheless good enough to address the questions we set out to answer.

6.1.3 Protein architecture analysis by chemical cross-linking

The study of proteins and protein-protein interactions by chemical cross-linking coupled with mass spectrometry is a low resolution, structural analysis technique that has become more widely used in current years.[98-105] This technique (as opposed to higher resolution methods such as NMR and X-ray crystallography) has, theoretically, no limit with regards to the size of the protein or protein complex to be studied. This is because the proteins are subjected to enzymatic digestion prior to MS/MS, creating peptides, which are more easily analysed. This makes cross-linking an ideal technique to study the architecture of full-length C7, which has so far evaded detailed tertiary structural analysis. Moreover, with the increasing repertoire of cross-linkers and continued advances in methodology for mass spectrometric analysis of cross-linked peptides, cross-linking has the potential to generate extensive lists of various intermolecular or intramolecular upper distance bounds, thus enhancing other kinds of structural analysis. In its current state of development however, such applications have not been routinely implemented. There may be several reasons for this, such as the requirement for specialized reagents, the necessity of access to, and expertise in the use of, high-end tandem mass-spectrometers and, in particular, the need for sophisticated tailored software that can find and identify cross-linked peptides in a very highly populated mass spectrum.[103]

6.1.3.1 C3 to C3b structural transition

The cross-linking analysis of test proteins C3 and C3b (for both of which high-resolution 3-D structures are available), proved reasonably accurate and informative with regards to the organisation of domains. The findings are largely consistent with the crystal structures, proving the reliability of the hybrid cross-linking:mass spectrometry method. Furthermore the cross-linking data sheds light on the dramatic conformational changes that accompany the C3 to C3b transition. Moreover, two polpeptide segments that were

absent from the 2I07 structure - (the N terminus of the α-chain and a portion of MG8 (that was present in 2HR0) - were involved in inter-domain cross-links. This highlights cross-linking use as a tool to augment (as well as confirm) structural analysis by X-ray crystallography.

6.1.3.1 C7 Architecture

Unlike in the case of C3 and C3b, no crystal structure of C7 has been determined to date. Thus C7 presents an interesting challenge for reconstruction of the tertiary structure of a mutiple-module protein using cross-linking derived distance restraints, in combination with computational models or atomic structures of the individual protein modules. Initial attempts were made using scripts written for the program CNS to generate a low-resolution model of C7 that was consistent with all of the cross-linking data. It soon became clear, however, that the number and coverage of cross-link derived distance restraints was too small to allow convergence on a consensus structure. For example, there were no detectable cross-links involving lysine residues in TSPN or EGF despite the likely proximity of EGF to MACPF. This limitation could possible have been overcome using cross-linkers with different functional groups and spacer-arm lengths possibly coupled with the use of proteases with different cleavage sites. Indeed, "zero-length" *N*-hydroxysuccinimide ester-based cross-linking was carried out on C7, as was protease digestion using GluC (as opposed to trypsin). However, the in-house software for identifying cross-linked peptides is still under development.

Despite the lack of a complete 3-D structural model of C7, the cross-linking studies proved very useful in two respects. First, numerous intra-domain cross-links confirmed the homology-based model of the C7-MACPF. Second, the pattern of intermodular cross-links, while inadequate to define an absolute configuration, clearly showed that most of the C-terminal arm of C7 wraps around the "top" of the MACPF domain, away

from what is likely (by analogy with the CDCs and perforin) to be the region that initially docks onto a membrane surface, The implications of these findings discussed in the following section.

6.2 Structure based model of MAC formation

6.2.1 Modules of the molecular arm

This structural investigation of the C-terminal modules (EGF, TSPC, CCP1, CCP2, FIM1 and FIM2) of C7 provides insights that help to elaborate current hypotheses based around the existence of a molecular arm that articulates during the remarkable, enzyme-free, process of MAC self assembly.

We propose that the CCPs form the rigid part of an arm that hinges on one or both of the EGF and TSPC modules to deliver the C-terminal FIMs pair to their binding site on C5-C345C. Displacement of the arm (or "safety catch") from its original position on C7 (in response to binding of C5b6) would, according to the hypothesis, allow the triggering of a major structural rearrangement of the MACPF domain that alters its characteristics from those of a soluble protein to those of a membrane-associated one. The activated C7 protein thus develops an affinity for the membrane and the nascent C5bC6C7 complex then acts as landing pad for C8, thus promoting further steps on the irreversible path to MAC assembly. Activation of C7, by analogy with the CDCs, may involve immediate release of its cryptic and potentially membrane penetrating β-hairpins; alternatively this may occur after C7 has docked onto the membrane surface. Something similar presumably occurs in C6 that has the same domain structure as C7.

Based on observations of other proteins that carry domains similar to the FIMs, it was speculated that the two FIM modules open up to provide a large specific surface area for the C5-C345C interaction.[53] Thus the putative swinging arms of C6 and C7 are fundamental to both driving self-assembly forward, but they are equally important as

"safety catches" on the MACPF domains of C6 and C7, ensuring they do not become activated inappropriately. The current work thus set out to test various predictions of this hypothesis.

6.2.1.1 FIMs

The solution structure of the FIMs module pair (alone) was solved previously. But this work did not reveal whether or not the closed conformation of FIMs (involving intimate mutual association of the two modules) is affected in any way by the presence of the neighbouring CCPs. This is important because the FIMs form the C-terminus of C7 and hence their only neighbour (within the sequence) is CCP2 so there is a good chance that the structure of FIMs in the context of C7-CFF (or C7-CCFF – too big for NMR, see above) will recapitulate their structure in the freely swinging arm of intact C7.

The solution structure of CFF, in conjuction with relaxation studies and SAXS analysis, confirm that the FIMs have indeed adopted their closed conformation in the context of C7-CFF. Moreover, it was ascertained that the seven-residue central section of the linker between CCP2 and FIM1 is highly flexible or indeed disordered in the context of C7, on several timescales. Thus the FIMs and CCP2 of C7-CFF behave largely as separate modules in solution. In the context of C7 the linker may be packed alongside the MACPF but the current findings are consistent with the idea that once released from MACPF (if indeed this happens) the arm would have a flexible joint between the "forearm" (CCPs) and the "fist", *i.e.* the FIMs, which may become an open "hand" when binding C5-C345C. Note that the C345C domain is likewise flexibly attached to the body of C3 (or C5). This makes sense in terms of an induced fit mechanism whereby the initial encounter between the C345C domain and the FIMs is followed by an adjustment in which complementary faces are engaged.

6.2.1.2 CCPs

The structure of CCPs, on the other hand, reveals a rigid, bent rod-like structure that is 6.6nm in length (directly measured from N-terminus to C-terminus). A significant amount of energy would appear to be required to articulate around such a joint. Therefore, the structural data bears out the notion of rigidity in the arm. This is a requirement of our model because a completely flexible arm would presumably be less good (it would incur more entropic costs) at direct delivery of FIMs from their site of residence on C7 to their binding site at the top of C5. The CCPs would thus act as rigid spacers, providing the arm with the reach needed to carry out its goal.

6.2.1.3 EGF-TSPC

Initial ^{15}N-HSQC based NMR investigations into TSPC-CCP1 (C7-TC) reveal that very few cross peaks of CCP1 are perturbed by attachment of TSPC its N-terminus. Thus TSPC and CCP1 are unlikely to be in physical contact with each other. Furthermore the structure of CCP1 is not modulated in any way by the presence of TSPC. Thus, as with the CCPs and FIMs, we may infer from looking at the sequence alignment, that TSPC and CCP1 are joined by a long, potentially flexible, linking sequence. This is in agreement with most simplistic notions of the swinging molecular arm that would require a pivot point at either end of the rigid CCP-CCP section, analogous with the "elbow" and the "wrist" of a real-life arm. Although the extent and location of chemical shift changes induced in TSPC by the presence of EGF (in the ET HSQC spectrum) could not be ascertained directly due to the absence of any resonance assignments, it is clear that the structure of TSPC is largely unperturbed by attachment to the EGF domain. Again, a flexible linker is implied. However this cannot be unambiguously determined without further structural analysis.

6.2.1.4 Summary

Thus in summary (Fig. 84), one can describe the FIMs as the 'hand' of the molecular arm

that is closed like a fist in the C7-CFF context and hence is very likely to be closed also when these modules are part of C7. The FIMs may open-up, however, on binding the C345C domain of C5 but we have no evidence for that suggestion. Continuing with the arm analogy, the CCP2-FIM1 linker is the flexible 'wrist' that joins the FIMs to the rigid 'forearm' formed by the CCP pair. The CCP pair is connected to the 'upper arm' or TSPC via another, presumably flexible, linker between them, corresponding to the 'elbow'. And the TSPC links to the EGF domain via a "shoulder" joint.

Finally the EGF domain may be fused to the MACPF domain in an analogous fashion to that seen in the experimentally determined structure of perforin[79] (where a MACPF domain is also followed immediately next to an EGF domain). These observations relate to the free arm since they are based purely on a dissect-and-rebuild approach to the C-terminal domains performed in the absence of the rest of the C7 molecule. To shed light on the manner in which these domains are organized in the intact protein, we rely on the cross-linking studies.

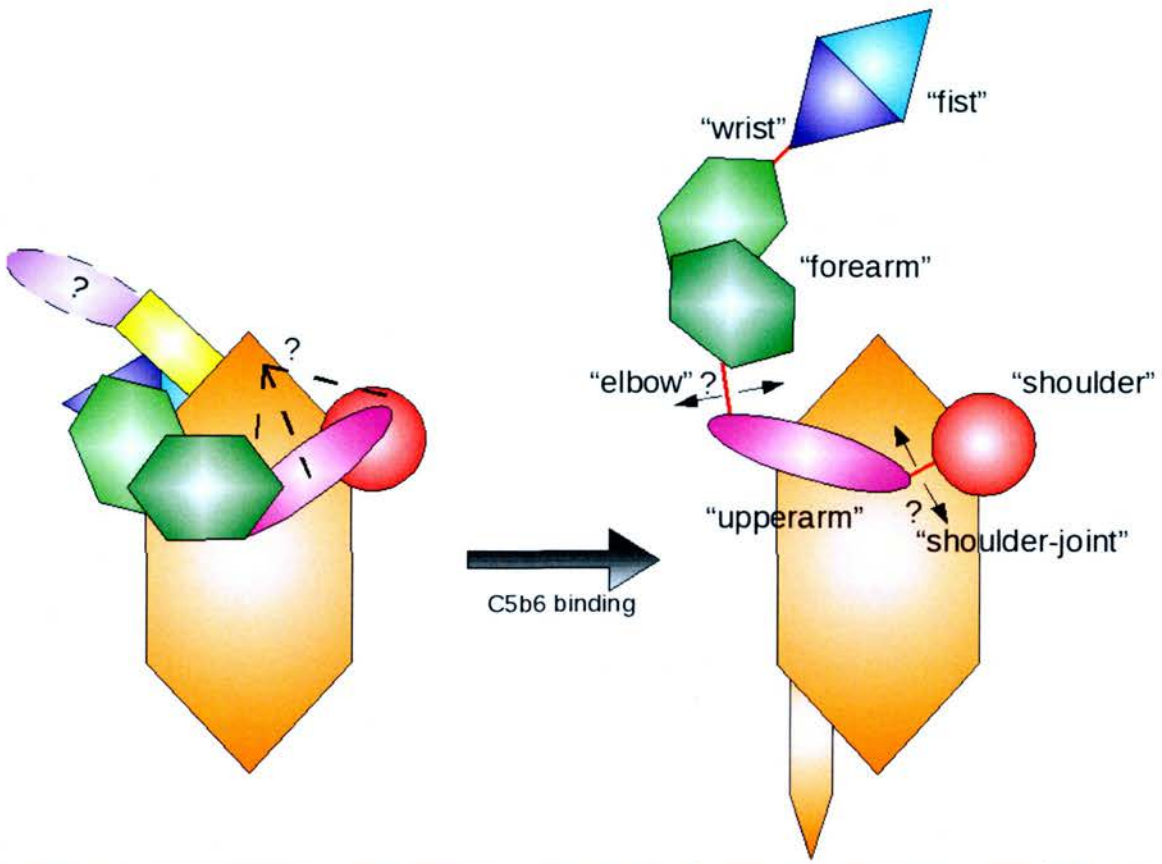


Fig. 84: C7 structural transition. The cartoon image on the left is a schematic representation of C7 in the "arm-folded" conformation; the image on the right depicts C7 in the "open-arm" conformation. Modules are labelled with the parts of an arm in an analogical manner and coloured as previously. Dotted black lines show potential disulfide links to the MACPF. Solid red lines denote linkers between modules. Extents of conformational freedom that are currently unknown are shown with double-headed arrows.

6.2.2 C7's molecular arm

The cross-linking data give an indication as to the location of the C-terminal "molecular arm" with respect to the MACPF domain. The arm appears to fold around or embrace the "upper" portion of the MACPF domain as observed in the standard view with the membrane proximal regions at the bottom. This part of the C7-MACPF domain is known as the d1 region by analogy with CDCs.[74] The N-terminal modules LDLRA and TSPN are also situated around this region as indicated by the detection of an LDLRA-MACPF cross-link.

Early electron microcopy studies have provided some very low-resolution images of the C7 molecule. Note that the staining methods used to record the EM image dehydrate the molecule and can result in exposure of hydrophobic regions and molecular distortions. Nonetheless, negatively stained images of single molecules can be very useful as emerged in studies of C3, C3b and C3(H_2O).[83] In the case of C7 it is worth remembering in this respect that C7 is proposed to undergo conformational changes (upon opening out of the C-terminal molecular arm in our models) that exposes hydrophobic regions for binding to the target membrane. Thus it is concievable that the EM staining methods have forced the molecular arm into an open configuration (or more precisely, trap a rare conformation of C7 in which the arm is dissociated). To better visualise this a comparison of the EM image and the C7 reconstruction is shown in a similar conformation (Fig. 85). One can see from the compatibility of dimensions that the globular region is clearly the MACPF, and the molecular arm in the EM images is of a length that can accommodate the C-terminal modules (and is too long to be formed by the two N-terminal modules). Thus the EM images do not conflict with our C7 reconstruction and likely represent the C7 molecule in an arm-open configuration.

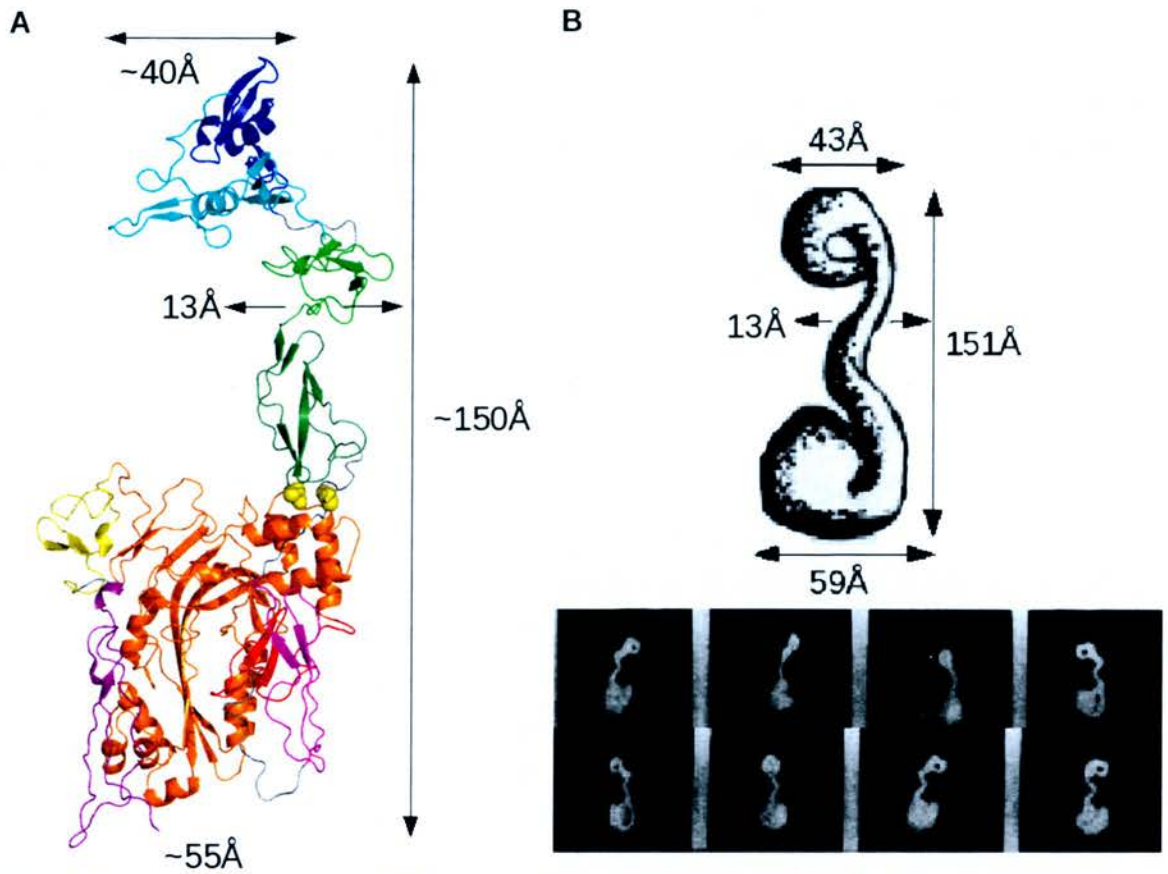


Fig. 85: Structural comparison of C7 model with its EM image. (A) shows the C7 reconstruction in a conformation representative of the EM image[96] in (B). Distances were measured for equivalent dimensions as in (B) from the two furthest separated points and are as shown. Colors for C7 modules are as shown previously. The cysteine of TSPC-linker and of the MACPF are brought into close-proximity in this representation and are therefore shown as yellow spheres.

Thus the cross-linking data indicate that in solution and prior to incorporation into the nascent MAC complex the molecular arm is a "folded" one. The degree of flexibility between modules in this context is difficult to ascertain, however the occurrence of TSPC-CCP2 and CCP2-FIM1 cross-links indicate that the modules are in a closer proximity than we observe in our dissect-and-rebuild strategy. It would not be at all surprising if interactions between individual modules and MACPF restrain the potential for flexibility in the molecular arm, especially considering the MACPF is disulfide linked to either the EGF-like domain, TSPC or the TSPC-CCP1 linker. None of these

observations are inconsistent with our model (Fig. 86) in which "release" of the molecular arm from the C7-MACPF upon binding the C5b6 complex reveals faces of the modules that were previously in close contact with the MACPF. The wide spread of cross-links in the FIMs modules, locating to one region of the MACPF indicate that the FIMs still display some flexibility with respect to the MACPF. Moreover the MACPF cross-links in the FIMs modules span the cleft between the two FIMs that form a potential protein binding site. It may result that an interaction of the cleft residues with the MACPF is replaced by their interaction with C5-C345C in the formation of the MAC. One can envisage a dynamic equilibrium between open and closed forms (*i.e.* forms in which the arm is free versus forms in which the arm is folded up around the top of the MACPF). The binding of C5b and provision of the C345C as an alternative tight binder for FIMs presumably then stabilizes the open form sufficiently for the C7 helices in d1 (that are equivalent to the helical clusters of the CDCs, C8 and perforin) to rearrange themselves, perhaps (by analogy with these homologues) into β-hairpins that have affinity for membranes.

It is also important to consider that C5 binds C7 in the pre-activation complex discussed in the Introduction (see section 1.2.3.2). In our model, the location of the FIMs and the flexible wrist could allow for an interaction between the FIMs and C345C without a gross movement of the molecular arm. Not until a complementary surface in C5b6 (as opposed to C5) is displayed for binding the MACPF/molecular arm modules would the energetic cost of the molecular arm unfurling be compensated for.

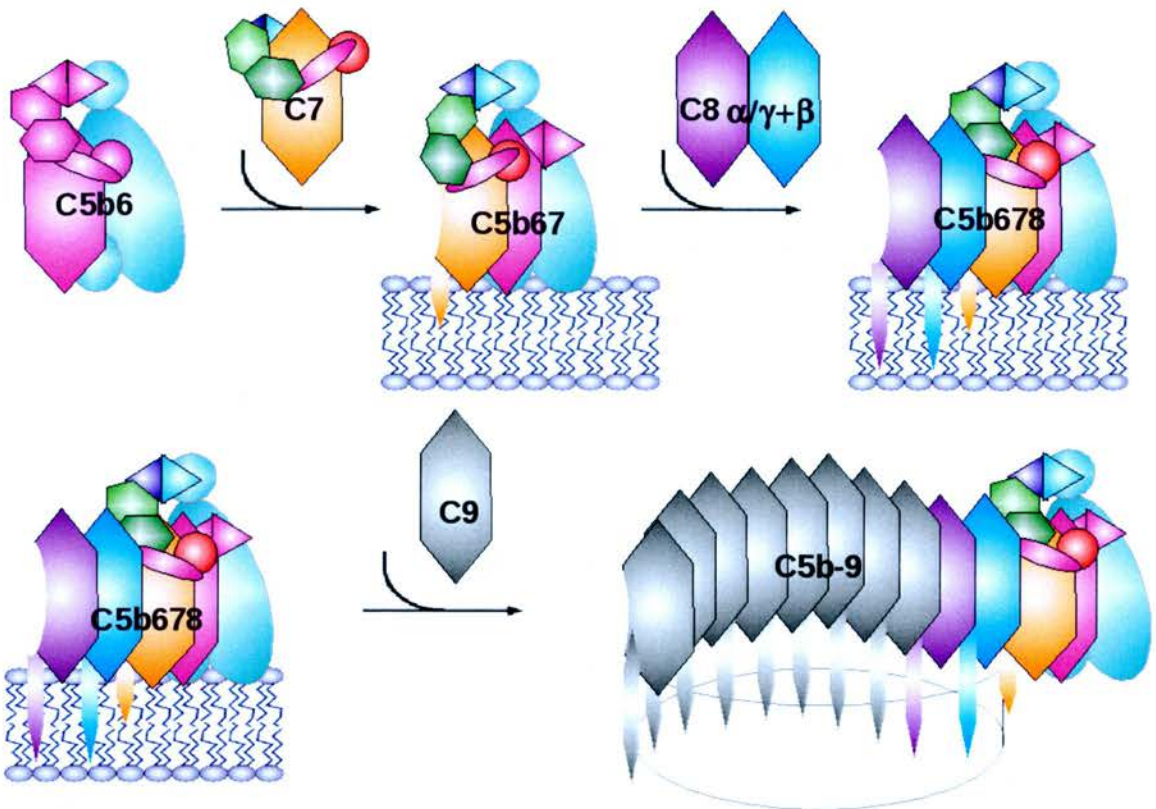


Fig. 86: Model of MAC assembly. MAC proteins are shown as schematic representations. C5 is in teal with the top circle representing C5-C345C and the lower circle representing C5d; C6 is in magenta; C7 is colour-coded, module by module, as shown previously; C8α-γ is in purple; C8β is in blue and C9 is in grey.

6.3 Future Work

6.3.1 High resolution structural analyses of C7

6.3.1.1 EGF-TSPC

The next logical step in our dissect-and-rebuild strategy would be to devote more study to the TSPC and the EGF-like domain. Initial ^{15}N-HSQC based structural studies reveal that C7-TSPC-CCP and C7-EGF-TSPC constructs produced by our collaborators (Ogata *et al*) have structures that should be solved with relative ease by NMR. Their sizes (ET=~10 kDa and TC=~15 kDa) are well within limits to solve their structures using the standard suite of experiments employed to solve the structures of C7-CCPs and C7-CFF. The sensitivity and spectral complexity issues encountered during the structural

determination of C7-CFF should not arise.

However, prior to solving these new module-pair structures, it would be worthwhile to determine the disulfide bonding pattern of C7. This is intrinsically difficult for proteins such as C7 that have clusters of cysteine residues. In traditional enzymatic cleavage/mass spectrometry techniques even the most non-specific proteases cannot easily cleave between cysteine residues that are very close neighbours.[175] The plethora of disulphide bonds (28 for C7 in total) and their clustering in relatively small disulphide rich protein modules will make disulfide mapping difficult as already found for C7-FIMs.[53] The contemporary literature contains several mass spectrometry based methods that may overcome such problems.[175, 176] Following the establishment of as much disulphide bonding information as possible for C7, it will be possible to check whether these same linkages exist in the recombinantly produced ET and TC protein module pairs.

6.3.1.2 MACPF

The high quality C7-MACPF model (WhatIf score -1.65) indicates that the C7-MACPF contains the same three subdomains (d1,d2 and d3) as the C8α-MACPF. How the EGF domain interacts with the MACPF remains to be seen. It may be that the EGF is elongated, flexible and packs against the TMHs as in perforin or it may be that C7-EGF has a different location with respect to the MACPF and/or has a structure that is more highly ordered with more secondary structural elements like the EGF-like domain of Heregulin-α.[177] Atomic resolution structural analysis of MACPF-EGF would identify any structural differences with C8α-MACPF and determine the true fold of the EGF domain. Due to the size of such a protein (~41kDa) solving the structure would be very challenging and therefore crystallization would likely be the method of choice.

6.3.2 Chemical cross-linking and complement

6.3.2.1 C7 Architecture

A more exhaustive cross-linking analysis of full-length C7 may provide insights into the MACPF-EGF arrangement and, of course, a more detailed description of C7 as a whole. Achievement of a more extensive list of cross-link lists can be achieved *via* the use of different proteases to produce a larger variety of peptide fragments, and to identify cross-linked peptides whose trypsin-digested counterparts did not "fly" in the mass spectrometer. As mentioned earlier, different cross-linkers could also be used. The expanding repertoire of available cross-linkers incorporates various combinations of functionalities *e.g.* for binding to amine, sulfhydryl, carboxyl, and carbohydrate groups, and non-specific photoreactive cross-linkers. Careful selection of cross-linkers could allow resolution of regions were there was previously a dearth of cross-links.

An exhaustive set of cross-link derived distances combined with solving the atomic structures of the individual modules and module pairs should allow a simulated annealing (e.g using CNS) or rigid-body docking (e.g. in HADDOCK[178]) type approach to structure calculation for C7. There are likely to be localized structural differences between modules solved separately and the same modules in the context of C7, but any gross changes in structure should be highlighted by intra-modular cross-linking distances.

However, the tertiary structure model produced would require validation by another, orthogonal, low-resolution solution-based structural technique such as SAXS.

6.3.2.2 MAC

The structural transition suggested to occur in the formation of C5b from C5 could also be investigated (by analogy with the successful work done on C3b). Taking cross-linking in the membrane attack complex to the next level and looking to the future and more advanced software and hardware, it would be exciting to establish the relative location of

domains with respect to one another for C6, C8 and C9 both as individual proteins but also in the various complexes that can readily be isolated: C5b6, C5b67 etc. Moreover, such work need not be limited to isolated complexes. Quantitative techniques allow meaningful cross-linking:mass spectrometry studies to be done in mixtures and even in plasma.

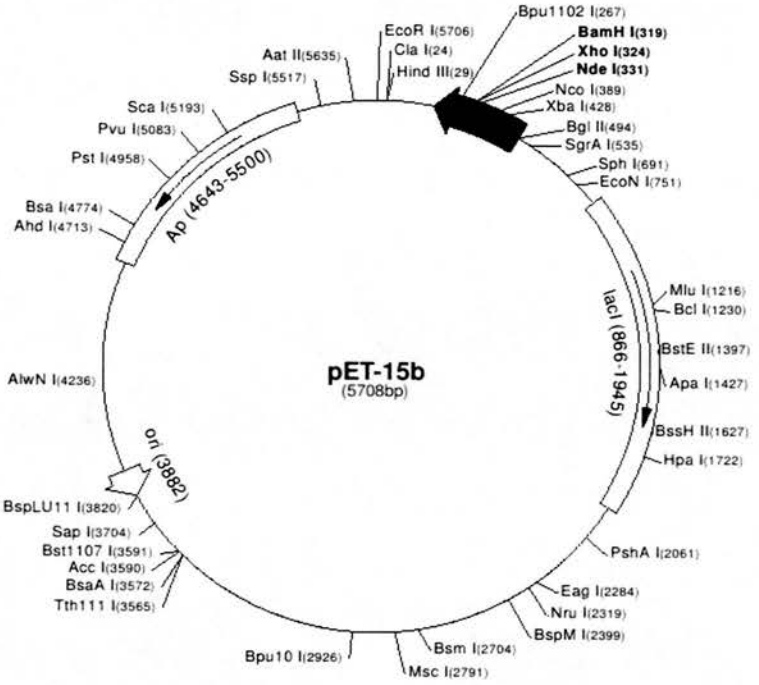
Carbohydrate/lipid-protein (target membrane to MAC) cross-linking of MAC complexes bound to either living cell membranes or artificial micelles should also be considered as these could highlight those regions of the MAC that associate with the target cell membrane.

APPENDIX A

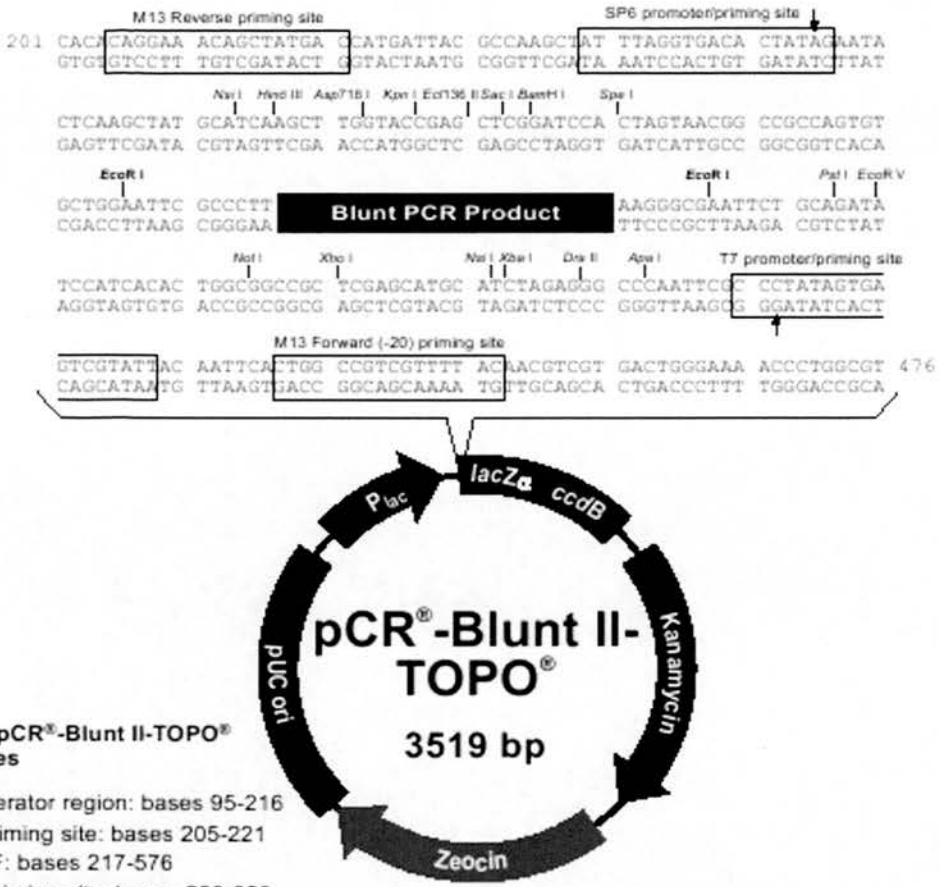
VECTOR MAPS

A) pET15b

pET-15b sequence landmarks	
T7 promoter	453-469
T7 transcription start	452
His•Tag coding sequence	362-380
Multiple cloning sites (*Nde*I - *Bam*H I)	319-335
T7 terminator	213-259
lacI coding sequence	(866-1945)
pBR322 origin	3882
bla coding sequence	4643-5500



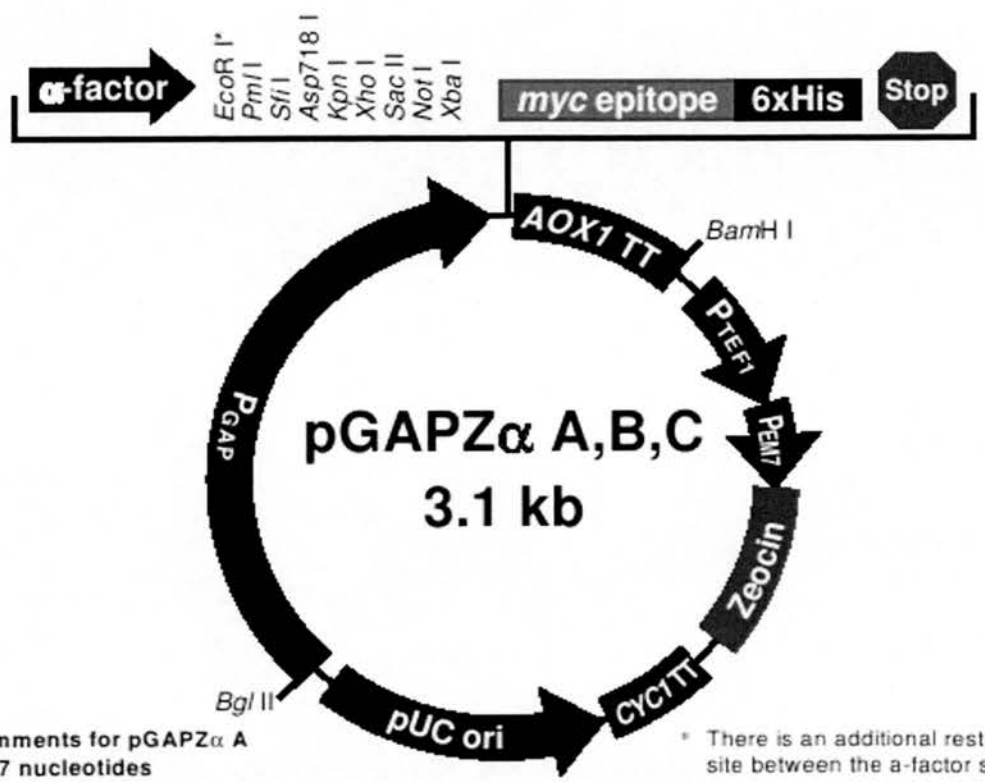
B) TOPO VECTOR



Comments for pCR®-Blunt II-TOPO®
3519 nucleotides

lac promoter/operator region: bases 95-216
M13 Reverse priming site: bases 205-221
LacZ-alpha ORF: bases 217-576
SP6 promoter priming site: bases 239-256
Multiple Cloning Site: bases 269-399
TOPO®-Cloning site: bases 336-337
T7 promoter priming site: bases 406-425
M13 (-20) Forward priming site: bases 433-448
Fusion joint: bases 577-585
ccdB lethal gene ORF: bases 586-888
kan gene: bases 1099-2031
kan promoter: bases 1099-1236
Kanamycin resistance gene ORF: bases 1237-2031
Zeocin resistance ORF: bases 2238-2612
pUC origin: bases 2724-3397

C) pGAPzαB



Comments for pGAPZα A
3147 nucleotides

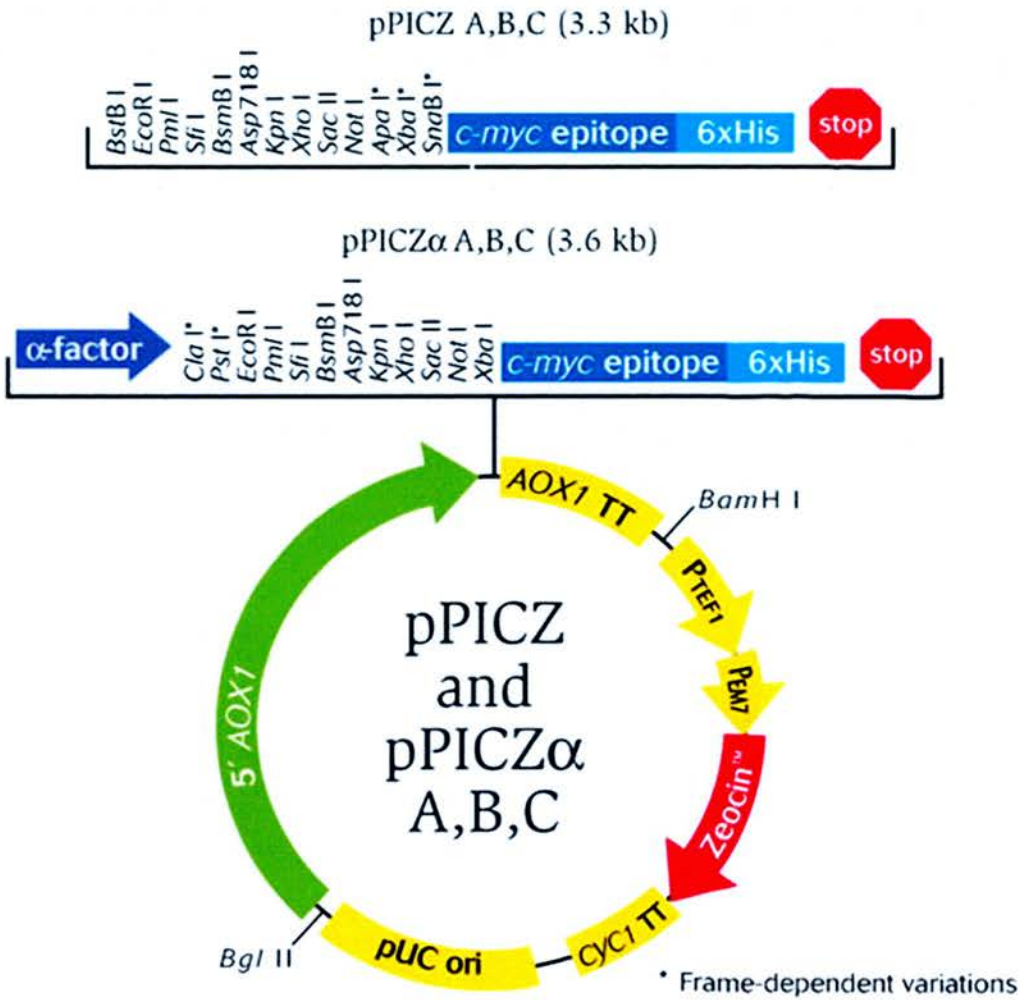
GAP promoter region: bases 1-483
α-factor signal sequence: bases 493-759
Multiple cloning site: bases 760-828
myc epitope tag: bases 827-856
Polyhistidine tag: bases 872-889
AOX1 transcription termination region: bases 893-1233
TEF1 promoter region: bases 1234-1644
EM7 promoter: bases 1645-1712
Sh ble ORF: bases 1713-2087
CYC1 transcription termination region: bases 2088-2405
pUC origin: bases 2416-3089

* There is an additional restriction site between the a-factor signal sequence and the *Eco*R I site in versions B and C of pGAPZa:

Pst I in pGAPZα B
Cla I in pGAPZα C

D) pPICzαB

Figure 1 - EasySelect™ vector maps (pPICZ, pPICZα)



APPENDIX B

BUFFER and MEDIA

Media	**Composition** (w/v unless otherwise stated)
Basal Salts for unlabelled fermentor growth	4% (v/v) glycerol 0.413% potassium hydroxide 1.5% magnesium sulphate heptahydrate 1.82% potassium sulphate 0.093% calcium sulphate 2.67% (v/v) ortho-phosphoric acid
Basal Salts for ^{15}N-labelled fermentor growth	3.125% (v/v) glycerol 1.5% magnesium sulphate heptahydrate 1% potassium sulphate 0.093% calcium sulphate 10% (v/v) 200 mM potassium phosphate pH 5
Basal Salts for ^{15}N,^{13}C-labelled fermentor growth	1.5% magnesium sulphate heptahydrate 1% potassium sulphate 0.093% calcium sulphate 10% (v/v) 200 mM potassium phospate pH 5
BMG (Buffered minimal glycerol)	100 mM Potassium phosphate pH 6 1.34% YNB 4 x 10-5 % biotin 1% glycerol
BMM (Buffered minimal methanol)	100 mM potassium phosphate pH 6 1.34% YNB 4 x 10-5 % biotin 0.5% methanol
LB (Lysogeny Broth) Lennox (Low salt (less than 5 g/L) is required for efficient selection with Zeocin™)	0.5% yeast extract 1% tryptone 0.5% sodium chloride +/- 1.5% agar

SOC (Super Optimal broth with Catabolite repression)	0.5% yeast extract 2% tryptone 10 mM sodium chloride 2.5 mM potassium chloride 10 mM magnesium chloride 10 mM magnesium sulphate 20 mM D-glucose
YNB (Yeast Nitrogen Base)	10x YNB stock (Sigma) with ammonium sulphate without amino acids
YPD (Yeast Extract, Peptone, Dextrose)	1% yeast extract 2% peptone 2% dextrose (D-glucose) +/- 1.5% agar
YPDS (Yeast Extract, Peptone, Dextrose Sorbitol)	1% yeast extract 2% peptone 2% dextrose (D-glucose) 1 M sorbitol +/- 1.5% agar +/- 100-300 µg/ml Zeocin
EDTA stock solution (ethylenediaminetetraacetic acid)	0.5 M stock adjusted to pH 8 with sodium hydroxide pellets
SDS-PAGE sample loading buffer	50 mM Tri-HCl 100 mM %-mercaptoethanol 2% Sodium-dodecyl-sulfate 0.1% bromophenol blue 10% glycerol

APPENDIX C

AZARA SCRIPTS

Ser.ref

```
!
! It was assumed that the offset in 1H dimension is set to water
! No consideration was made for frequency/pulse offsets, in particular if your pulse
program
! has statement similar to this "fq=cnst23(bf ppm):f2" then you might consider setting
O2 by
! hand into the shifted value (here it would be cnst23*BF2).

! You have used: temp=298.00K, salt conc=0.00mM, pH=5.00
! Frequencies of zero ppm were: 1:8.000100e+08 2:8.106431e+07 3:8.000100e+08
! pulse program used was <hsqcNH_wg.du>
!
! It is assumed that you will zerofill data once and reduce
!
ndim 2

file ser

int

dim 1
npts 2048
!sw 11160.714286      !modified using half_sw1 by user cclark on Fri Oct 17 13:36:35
BST 2008
```

sw 5580.357143

sf 800.013811

refppm 4.771785

refpt 1025

nuc 1H

! refppm will be o2p + 2.68 ppm this is a correct value

dim 2

npts 128

sw 1945.903872

sf 81.074052

refppm 120.212443

refpt 65

nuc 15N

Scr

! This might be not the final scr that you need,

! you will most certainly need to adjust phases

input ser.ref

output spc

!script_com 1

! complex

! conv_box 8

! avance 12 16

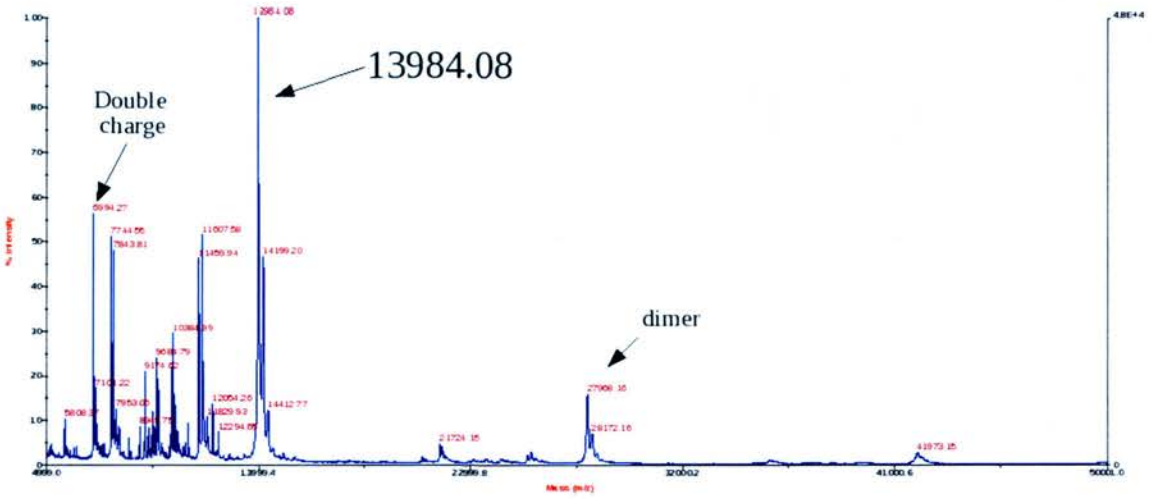
! sinebell2 90

```
! zerofill 1
! fft
! avance_phase
! phase 0 0 ! change phasing here
 ! reduce
! upper 1024 ! suggested if H(NH) in this dim, do not forget to shrink sw in spc.par
accordingly
!end_script

script_com 1
 complex
 avance 12 16
 sinebell2 90
 zerofill 1
 fft
 avance_phase
 phase -70 0 ! change phasing here
end_script

script_com 1
 complex
 ifftn
 conv_box 30
 fftn
 reduce
 upper 1024 !
end_script
```

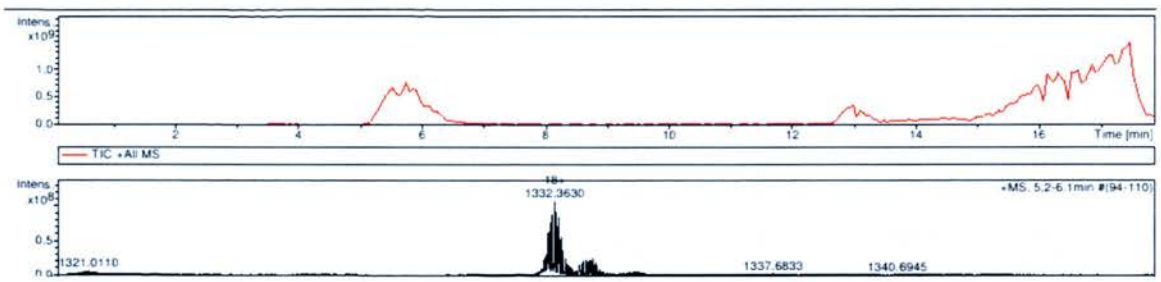
```
script_com 2
  mask_ppmm ! dim 2 States-TPPI
  complex
 !conjugate  !if this dim looks upside-down remove conjugate
  sinebell2 90
  zerofill 1
  fft
  phase 90 -180 !! States-TPPI, might need change
  reduce
end_script

script_com 1
  base_poly 8 0  ! baseline correction
end_script
```

APPENDIX D

C7-CCPs Mass spectra (MALDI-TOF)



C7-CFF Mass spectra (LC-MS)



18+ charge state

APPENDIX E

Pair-wise CCP module structural comparisons of C7-CCP1 and C7-CCP2.

Protein~module number (PDB code	C7~01 RMSD in Å (alignment length, gaps included)	C7~02 RMSD in Å (alignment length, Gaps included
C7~01		2.76 (59)
C7~02	2.76 (59)	
C1r~01 (1GPZ)	2.46 (59)	2.24 (58)
C1r~02 (1GPZ)	2.21 (34)	2.43 (35)
C1s~02 (1ELV)	2.30 (49)	2.54 (61)
C2~01 (3ERB)	2.89 (51)	2.79 (47)
C2~02 (3ERB)	3.06 (60)	2.46 (60)
C2~03 (3ERB)	2.57 (58)	2.34 (58)
C4BPα~01 (2A55)	2.99 (58)	3.58 (58)
C4BPα~02 (2A55)	2.24 (59)	2.44 (58)
CR1~15 (1GKN)	2.35 (58)	3.25 (58)
CR1~16 (1GKN)	2.41 (59)	2.35 (59)
CR1~17 (1GKG)	1.88 (57)	2.49 (57)
CR2~01 (1LY2)	2.37 (60)	2.57 (60)
CR2~02 (1LY2)	2.23 (59)	2.47 (59)
DAF~01 (1OK3)	1.62 (60)	3.11 (59)
DAF~02 (1OK3)	1.88 (59)	3.01 (58)
DAF~03 (1H03)	2.49 (56)	2.52 (59)
DAF~04 (1H03)	1.81 (59)	2.44 (59)
FB~01 (2OK5)	3.14 (57)	3.43 (58)
FB~02 (2OK5)	2.85 (60)	2.32 (60)
FB~03 (2OK5)	2.55 (58)	2.43 (58)
FH~01 (2RLP)	2.16 (57)	3.25 (58)
FH~02 (2RLQ)	2.40 (57)	2.84 (58)
FH~03 (2RLQ)	2.46 (58)	2.24 (58)
FH~04 (2WII)	2.52 (57)	2.35 (50)
FH~05 (not deposited)	2.69 (56)	2.58 (56)
FH~06 (2UWN)	2.85 (54)	3.66 (56)
FH~07 (2UWN)	2.96 (49)	2.84 (56)
FH~08 (2UWN)	2.24 (55)	2.64 (56)
FH~12 (2KMS)	2.18 (58)	2.64 (57)
FH~13 (2KMS)	3.25 (57)	3.20 (49)
FH~15 (1HFH)	2.01 (58)	3.28 (58)
FH~16 (1HFH)	2.56 (56)	3.18 (46)
FH~19 (2G7I)	2.19 (58)	2.49 (57)
FH~20 (2G7I)	2.72 (56)	2.98 (57)
MASP1~01 (3GOV)	2.48 (60)	2.37 (60)
MASP1~02 (3GOV)	2.21 (60)	2.72 (61)

***MASP2~01* (1ZJK)**	**2.62 (60)**	**2.32 (60)**
***MASP2~02* (1ZJK)**	**2.42 (60)**	**2.47 (62)**
***MCP~01* (1CKL)**	**1.84 (59)**	**3.06 (59)**
***MCP~02* (1CKL)**	**2.91 (60)**	**2.62 (60)**
***MCP~03* (3O8E)**	**2.30 (60)**	**2.17 (59)**
***MCP~04* (3O8E)**	**2.81 (59)**	**2.37 (54)**
***VCP~01* (1G40)**	**2.99 (59)**	**2.58 (58)**
***VCP~02* (1G40)**	**2.78 (57)**	**2.62 (58)**
***VCP~03* (1G40)**	**2.40 (58)**	**2.75 (58)**
***VCP~04* (1G40)**	**2.58 (57)**	**2.66 (55)**

Table 1: Pair-wise CCP module structural comparisons of C7~01 and C7~02. Comparison of individual lowest energy structures of C7~01 and C7~02 versus all other individual CCPs of known structure within the complement system based upon C_α RMSD values using structural alignment program CE. For each CCP, inclusive module boundaries were one residue before Cys^{I} and the third residue after Cys^{III}. In cases where structures have been solved by both NMR and X-ray diffraction, the higher resolution X-ray structure was used for comparison. Where both liganded and unliganded structures were available, the highest resolution unliganded X-ray or NMR structure was used. A few residues were missing in the crystal structure of C1r~02, and hence in this case, the structure with the most determined residues was employed for both modules. Colour key used in table: Blue: 0 - 1.99 Å; Green: 2.00 – 2.99 Å; Red: 3.00 – 3.99 Å; Brown: Alignment lengths < 40 amino acids. Abbreviations used in Table: C4BPα = C4b-binding protein α-chain; CR = complement receptor; DAF = decay-accelerating factor; FB = factor B; FH = factor H; MASP1 / 2 = mannan-binding lectin-associated serine proteases 1 / 2; MCP = membrane cofactor protein; VCP = Vaccinia virus complement control protein. Some residues were not present (solved) in the electron density map for the C1r~02 module crystal structure, and this explains the short structural alignment length (shown in brown).

BIBLIOGRAPHY

1.Walport, M.J. (2001) Complement. *N Engl J Med.* **344,** 1058–1066.

2.Dodds, A.W. And Matsushita, M. (2007) The phylogeny of the complement system and the origins of the classical pathway. *Immunobiology.* **212**, 233-243.

3.Nonaka, M. and Yoshizaki, F. (2004) Evolution of the complement system. *Mol. Immunol.* **40**, 897-902.

4.Hedge, G.V., Meyers-Clark, E., Joshi, S.S. and Sanderson, S.D. (2008) A conformationally-biased, response-selective agonist of C5a acts as a molecular adjuvant by modulating antigen processing and presentation activities of human dendritic cells. *Int. Immunopharmacol.* **8**, 819-827.

5.Mastellos, D. and Lambris, J.D. (2002) Complement: more than a 'guard' against invading pathogens? *Trends. Immunol.* **23**, 485-491.

6.Daha, M.R., Kooten, C. and Roos, A. (2006) Compliments from complement: a fourth pathway of complement activation? *Nephrol. Dial. Transplant.* **21**, 3374-3376.

7.Huber-Lang, M., Sarma, J.V., Zetoune, F.S., Rittirsch, D., Neff, T.A., McGuire, S.R., Lambris, J.D., Warner, R.L., Flierl, M.A., Hoesel, L.M., Gebhard, F., Younger, J.G., Drouin, S.M., Wetsel, R.A. and Ward, P.A. (2006) Generation of C5a in the absence of C3: a new complement activation pathway. *Nat. Med.* **6**, 682-687.

8.Mueller-Eberhard, H.J. (1970) The reaction mechanism of human C5 in immune hemolysis. *J Exp. Med.* **132,** 775.

9.Hammer, C.H., A.S. Abramovitz, and M.M. Mayer. 1976. A new activity of complement component C3: cell-bound C3b potentiates lysis of erythrocytes by C5b,6 and terminal components. J. *Immunol.* **117**, 830-834.

10.Silversmith, R.E., and G.L. Nelsestuen. 1986. Interaction of complement proteins C5b-6 and C5b-7 with phospholipid vesicles: effects of phospholipid structural features. *Biochemistry.* **25**, 7717-7725.

11.Preissner, K.T. (1985) The membrane attack complex of complement – relation of C7 to the metastable-binding site of the intermediate complex C5b-7. *J.* Immunol. **135**, 445-451.

12.Hammer, C.H., Nicholson, A. and Mayer, M. (1975) On the mechanism of cytolysis by complement: Evidence on insertion of C5b and C7 subunits of the C5b,6,7 complex into phospholipid bilayers of erythrocyte membranes. *Proc. Nat. Acad. Sci. USA.* **72**, 5076-5080.

13.Mueller-Eberhard, H.J. (1986) The membrane attack complex of complement. *Ann. Rev. Immunol.* **4,** 503-528.

14.Koski, K.L., Ramm, L.E., Hammer, C.H., Mayer, M.M. and Shin, M.L. (1983) Cytolysis of nucleated cells by complement: cell death displays multi-hit characteristics. *Proc. Natl. Acad. Sci. USA.* **80**, 3816-3820.

15.O'Hara, A.M., Moran, A.P., Wurrzner, R. and Orren, A. (2001) Complement-mediated lipopolusaccharide release and outer membrane damage in *Escherichia coli* J5: requirement for C9. Immunol. **102**, 365-372.

16.Rus, H.G., Niculescu, F.I. and Shin, M.L. (2001) Role of the C5b-9 complement complex in cell and apoptosis. *Immunol. Rev.* **180,** 49-55.

17.Helperin, J.A., Taratuska, A. and Nicholson-Weller A. (1993) Terminal complement component C5b-9 stimulates mitogenesis in 3T3 cells. *J. Clin. Invest.* ***91***, 1974-1978.

18.Milis, L., Morris, C.A., Sheehan, M.C., Charlesworth, J.A. and Pussell, B.A. (1993) Vitronectin-mediated inhibition of complement: evidence for different binding sites for C5b-7 and C9. *Clin. Exp. Immunol.* **92,** 114-119.

19.Yan, J., Allendorf, D.J., Li, B., Yan, R., Hansen, R., Donev, R. (2008) The role of membrane complement regulatory proteins in cancer immunotherapy. *Adv. Exp. Med. Biol.* **632**, 159-174.

20.Rother, R.P, Mojcik, C.F. and McCroskery, E.W. (2004) Inhibition of terminal complement: a novel therapeutic approach for the treatment of systemic lupus erythematosus. *Lupus* **13**, 328-334.

21.Oliveira, G.H., Brann, C.N., Becker, K., Thohan, V., Koerner, M.M., Loebe, M., Noon, G.P. and Tore-Amione, G. (2006) Dynamic expression of the membrane attack complex of the complement system in failing human myocardium. *Am. J. Cardiol.* **97**, 1626-1629.

22.22.Younger, D.S., Rosoklija, G. and Hays, A.P. (1998) Diabetic peripheral neuropathy. *Semin.Neurol.* **18,** 95-104.

23.Suhr, B.D., Black, S.M., Guzman-Paz, M., Matas, A.J. and Dalmasso, A.P. (2007) Inhibition of the membrane attack complex of complement for induction of accomodation in the hamster-to-rat heart transplant model. *Xenotransplantation.* **14,** 572-579.

24.Aronica, E., Boer, K., van Vliet, E.A., Reseker, S., Baayen, J.C., Spliet, W.G., van Rijen, P.C., Troost, D., da Silva, F.H., Wadman, W.J., and Gorter, J.A. (2007) Complement activation in experimental and human temporal lobe epilepsy. *Neurobiol. Dis.* **26**, 497-511.

25.Zerzi, Y., Kallel-sellami, M., Abelmalek, R., Laadhar, L., Chaabane, T., Makni, S. (2010) Hereditary complement C5 deficiency: study of 3 Tunisian adult cases and literature review. *Tunis. Med.* **88**, 269-276.

26.Orren, A. (2000) Molecular mechanisms of complement component C6 deficiency; a hypervariable exon 6 region responsible for three of six reported defects. *Clin. Exp. Immunol.* **119**, 255-258.

27.Fernie, B.A. and Hobart, M.J. (1998) Complement C7 deficiency: seven further molecular defects and their associated marker haplotypes. *Hum. Genet.* **103**, 513-519.

28.Kojima, T., Horiuchi, T., Nishizaka, H., Fukumori, Y., Amano, T., Nagasawa, K., Niho, Y. and Hayashi, K. (1998) Genetic basis of human complement C8 alpha-gamma deficiency. *J. Immunol.* **161**, 3762-3766.

29.Kaufmann, T., Hansch, G., Rittner, C., Spath, P., Tedesco, F. and Schneider, P.M. (1993) Genetic basis of human complement C8 beta deficiency. *J. Immunol.* **150**, 4943-4947.

30.Nagata, M., Hara, T., Aoki, T., Mizuno, Y., Akeda, H., Inaba, S., Tsumoto, K. and Ueda, K. (1989) Inherited deficiency of ninth compnent of complement: an increased risk of meningococcal meningitis. *J. Pediatr.* **114**, 260-264.

31.Botto, M., Kirschfink, M., Macor, P., Pickering, M.C., Wurrzner, R. and Tedesco, F. Complement in human diseases: Lessons from complement deficiencies. *Mol. Immunol.* **46**, 2774-2783.

32.Würzner, R.. (2003) Deficiencies of the complement MAC II gene cluster (C6, C7, C9): is subtotal C6 deficiency of particular evolutionary benefit? Clin. Exp. Immunol.133:156.

33.Ricklin, D and Lambris, J.D. (2007) Complement-targeted therapeutics. *Nat. Biotechnol.* **25**, 1265-1275.

34.Wagner, E. and Frank, M.M. (2010) Therapeutic potential of complement modulation. *Nat. Rev.* **9**, 43-56.

35.Rother, R.P., Rollins, S.A., Mojcik, C.F., Brodsky, R.A. & Bell, L (2007) Discovery and development of the complement inhibitor eculizumab for the treatment of paroxysmal nocturnal hemoglobinuria (vol 25, pg 1256, 2007). *Nat. Biotech.* **25**, 1488-1488.

36.Soltys, J., Kusner, L.L., Young, A., Richmonds, C., Hatala, D., Gong, B.D., Shanmugavel, V. and Kaminiski, H.J (2009) Novel Complement Inhibitor Limits

Severity of Experimentally Myasthenia Gravis. *Annals of Neurology* **65**, 67-75.

37.Rollins, S.A., Matis, L.A., Springhorn, J.P., Setter, E. & Wolff, D.W. (1995) Monoclonal antibodies directed against human C5 and C8 block complement-mediated damage of xenogeneic cells and organs. *Transplantation.* **60**, 1284-1292.

38.Hobart, M.J., Fernie, B.A. and Discipio, R.G. (1995) Structure of the human C7 gene and comparison with the C6, C8α, C8β and C9 genes. *J. Immunol.* **154**, 5188-5194.

39.Sandoval, A., Ai, R., Ostresh, J.M. & Ogata, R.T. (2000) Distal recognition site for classical pathway convertase located in the C345C/netrin module of complement component C5. *J. Immunol.* **165**, 1066-1073.

40.Thai, C.T. and Ogata, R.T. (2004) Complement components C5 and C7: recombinant factor I modules of C7 bind to the C345C domain of C5. *J. Immunol.* **173**, 4547-4552.

41.Discipio, R.G. (1992) Formation and structure of the C5b-7 complex of the lytic pathway of complement. *J. Biol. Chem.* **267**, 17087-17094.

42.Disipio, R.G., Linton, S.M. And Rushmere, N.K. (1999) Function of the factor I modules (FIMs) of human complement component C6. *J. Biol. Chem.* **274**, 31811-31818.

43.Arroyave, C. M., H. J. Müller-Eberhard. (1973) Interactions between human C5, C6, and C7 and their functional significance in complement-dependent cytolysis. *J. Immunol.* **111**: 536-545.

44.Würzner, R., M. J. Hobart, B. A. Fernie, D. Mewar, P. C. Potter, A. Orren, P. J. Lachmann. (1995) Molecular basis of subtotal complement C6 deficiency. *J. Clin. Invest.* **95**: 1877-1883.

45.Yu, J.X., Bradt, B.M. & Cooper, N.R. (2000) Molecular cloning of the C6A form cDNA of the mouse sixth complement component: functional integrity despite the absence of factor I modules. *Immunogenetics* **51**, 779-787.

46.Thai, C.T. & Ogata, R.T. (2004)Complement components C5 and C7: Recombinant factor I modules of C7 bind to the C345C domain of C5. *Journal of Immunology* **173**, 4547-4552.

47.Thai, C.T and Ogata, R.T. (2005) Recombinant C345C and factor I modules of complement components C5 and C7 inhibit C7 incorporation into the complement membrane attack complex. *J. Immunol.* **174**, 6227-6232.

48.Arlaud, G.J., Barlow, P.N., Gaboriaud, C., Gros, P. & Narayana, S.V.L. (2007) Deciphering complement mechanisms: The contributions of structural biology. *Molecular Immunology* **44**, 3809-3822.

49.Hadders, M.A., Beringer, D.X. and Gros, P. (2007) Structure of C8α-MACPF reveals mechanism of membrane attack in complement immune defense. *Science*. **317**, 1552-1554.

50.Slade, D.J., Lovelace, L.L., Chruszcz, M., Minor, W., Lebioda, L. and Sodetz, J.M. (2008) Crystal structure of the MACPF domain of human complement protein C8α in complex withthe C8γ subunit. *J.Mol.Biol.* **379**, 331-342.

51.Fredslund, F., Laursen, N.S., Roversi, P., Jenner, L., Oliveria, C.L., Pederson, J.S., Nunn, M.A., Lea, S.M., Discipio, R., Sottrup-Jensen, L. and Anderson, G.R. (2008) Structure of and influence of a tick complement inhibitor on human complement component 5. *Nat. Immunol.* **9**, 753-760.

52.Laursen, N.S. *et al.* (2010) Structural basis for inhibition of complement C5 by the SSL7 protein from Staphylococcus aureus. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3681-3686.

53.Phelan, M.M. *et al.* (2009) Solution Structure of Factor I-like Modules from Complement C7 Reveals a Pair of Follistatin Domains in Compact Pseudosymmetric Arrangement. *J. Biol. Chem.* **284**, 19637-19649.

54.Bramham, J., Thai, C.T., Soares, D.C., Uhrin, D., Ogata, R.T. And Barlow, P.N. (2005) Functional insights from the structure of the multifunctional C345C domain of C5 of complement. *J. Biol. Chem.* **280**, 10636-10645.

55.Dankert, J.R. & Esser, A.F. (1987) Bacterial Killing by complement-C9-mediated killing in the absence of C5b-8. *Biochemical Journal* **244**, 393-399.

56.Taylor, P.W. (1992) Complement-mediated killing of susceptible gram-negative bacteria- an elusive mechanism. *Exp. Clin. Immunog.* **9**, 48-56.

57.Stewart, J. L., and Sodetz, J. M. (1985) Analysis of the specific association of the eighth and ninth components of human complement: identification of a direct role for the α subunit of C8, *Biochemistry* **24**, 4598-4602.

58.Monahan, J.B. & Sodetz, J.M. (1981) Role of the beta-subunit in interaction of the 8^{th} component of human-complement with the membrane bound cytolytic complex. *J. Biol.Chem.***256**, 3258-3262.

59.Plumb, M.E. *et al.* (1999) Chimeric and truncated forms of human complement protein C8 alpha reveal binding sites for C8 beta and C8 gamma within the membrane attack complex perforin region. *Biochemistry* **38**, 8478-8484.

60.Scibek, J.J., Plumb, M.E. & Sodetz, J.M. (2002) Binding of human complement C8 to C9: Role of the N-terminal modules in the C8 alpha subunit. *Biochemistry* **41**, 14546-14551.

61.Slade, D.J., Chiswell, B. & Sodetz, J.M. (2006) Functional studies of the MACPF domain of human complement protein C8 alpha reveal sites for simultaneous binding of C8 beta, C8 gamma, and C9. *Biochemistry* **45**, 5290-5296.

62.Musingarimi, P., Plumb, M.E. & Sodetz, J.M. (2002) Interaction between the C8 alpha-gamma and C8 beta subunits of human complement C8: Role of the C8 beta N-terminal thrombospondin type 1 module and membrane attack complex/perforin domain.

Biochemistry **41**, 11255-11260.

63.Brannen, C.L. & Sodetz, J.M. (2007) Incorporation of human complement C8 into the membrane attack complex is mediated by a binding site located within the C8 beta MACPF domain. *Mol. Immunol.* **44**, 960-965.

64.Lovelace, L.L., Chiswell, B., Slade, D.J., Sodetz, J.M. & Lebioda, L. (2008) Crystal structure of complement protein C8 gamma in complex with a peptide containing the C8 gamma binding site on C8 alpha: Implications for C8 gamma ligand binding. *Mol. Immunol.* **45**, 750-756.

65.Slade, D.J. *et al.* (2008) Crystal structure of the MACPF domain of human complement protein C8 alpha in complex with the C8 gamma subunit. *Journal of Molecular Biology* **379**, 331-342.

66.Parker, C.L. and Sodetz, J.M. (2002) Role of the human C8 subunits in complement-mediated bacterial killing: evidence that C8 gamma is not essential, *Mol. Immunol.* **39,** 453–458

67.Brickner, A. and Sodetz, J.M. (1985) Functional domains of the alpha subunit of the eighth component of human complement: identification and characterization of a distinct binding site for the gamma chain, *Biochemistry* **24,** 4603–4607.

68.Bubeck, D., Roversi, P., Donev, R., and Morgan, B.P. (2010) Structure of human complement C8, a precursor to membrane attack. *J. Mol. Biol.* Article in Press.

69.Shatursky, O., Heuck, A.P., Shepard, L.A., Rossjohn, J., Parker, M.W., Johnson, A.E., and Tweten, R.K. (1999) The mechanism of membrane insertion for a cholesteroldependent cytolysin: a novel paradigm for pore-forming toxins. *Cell* **99,** 293–299.

70.Ramachandran, R., Tweten, R.K. and Johnson A.E. (2004) Membrane-dependent conformational changes initiate cholesterol-dependent cytolysin oligomerization and

intersubunit beta-strand alignment, *Nat. Struct. Mol. Biol.* **11**, 697–705.

71.Soltani, C.E., Hotze, E.M., Johnson, A.E. & Tweten, R.K. (2007) Specific protein-membrane contacts are required for prepore and pore assembly by a cholesterol-dependent cytolysin. *J. Biol.Chem.* **282**, 15709-15716.

72.Czajkowsky, D.M., Hotze, E.M., Shao, Z.F. & Tweten, R.K. (2004) Vertical collapse of a cytolysin prepore moves its transmembrane beta-hairpins to the membrane. *Embo Journal* **23**, 3206-3215.

73.Rossjohn, J. *et al.* (2005) Structures of perfringolysin O suggest a pathway for activation of cholesterol-dependent cytolysins. *Journal of Molecular Biology* **367**, 1227.

74.Lukoyanova, N. & Saibil, H.R. (2008) Friend or foe: the same fold for attack and defense. *Trends in Immunology* **29**, 51-53.

75.DiScipio, R.G. & Berlin, C. (1999) The architectural transition of human complement component C9 to poly(C9). *Mol. Immunol.* **36**, 575-585.

76.Dankert, J.R. & Esser, A.F. (1985) Proteolytic and chemical modification of human-complement protein C9 – Loss of poly(C9) and circular lesion formation without impairment of function. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2128-2132.

77.Rossi, V., Wang, Y.X. & Esser, A.F. (2010) Topology of the membrane-bound form of complement protein C9 probed by glycosylation mapping, anti-peptide antibody binding, and disulfide modification. *Mol. Immunol.* **47**, 1553-1560.

78.Amiguet, P., Brunner, J. & Tschopp, J. (1985) The Membrane attack complex of complement – lipid insertion of tubular and non-tubular polymerized C9. *Biochemistry* **24**, 7328-7334.

79.Law, R.H.P. *et al.* (2010) The structural basis for membrane binding and pore formation by lymphocyte perforin. *Nature* **468**, 447-51.

80.Laine, R.O. & Esser, A.F. (1989) Identification of the discontinuous epitope in human-complement protein C9 recognized by anti-melittin antibodies. *J. Immunol.* **143**, 553-557.

81.Sottrupjensen, L., Stepanik, T.M., Kristensen, T., Lonblad, P.B., Jones, C.M.,Wierzbicki, D.M., Magnusson, S., Domdey, H., Wetsel, R.A., Lundwall, A., Tack, B.F. and Fey, G.H. (1985) Common evolutionary origin of alpha-2-macroglobulin and complement component -C3 AND component -C4. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 9-13.

82.Doan, N. & Gettins, P.G.W. (2007) Human alpha(2)-macroglobulin is composed of multiple domains, as predicted by homology with complement component C3. *Biochem. J.* **407**, 23-30.

83.Nishida, N., Walz, T. & Springer, T.A. (2006) Structural transitions of complement component C3 and its activation products. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 19737-19742.

84.Li, K.Y., Gor, J. & Perkins, S.J. (2010) Self-association and domain rearrangements between complement C3 and C3u provide insight into the activation mechanism of C3. *Biochem. J.* **431**, 63-72.

85.Bhakdi, S., Tranum-Jensen, J. & Klump, O. (1980) The terminal membrane C5b-9 complex of human complement. Evidence for the existence of multiple protease-resistant polypeptides that form the trans-membrane complement channel. J. Immunol. 124,2451–2457.

86.DiScipio, R.G., Chakravarti, D.N., Müller-Eberhard, H.J. & Fey, G.H. (1988)The structure of human complement component C7 and the C5b-7 complex. *J. Biol. Chem.* 263,549–560.

87.Disipio, R.G., (1992) Formation and structure of the C5b-7 complex of the lytic

pathway of complement. *J. Biol. Chem.* **267**, 17087-17094.

88.Sandoval, A., Ai, R., Ostresh, J.M. & Ogata, R.T. Distal recognition site for classical pathway convertase located in the C345C/netrin module of complement component C5. *J. Immunol.* **165**, 1066-1073 (2000).

89.Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005) Protein Identification and Analysis Tools on the ExPASy Server; (In) The Proteomics Protocols Handbook, Humana Press, 571-607.

90.Harrington, A.E., Morris-Triggs, S.A., Ruotolo, B.T., Robinson, C.V., Ohnuma, S. and Hyvonen, M. (2006) Structural basis for the inhibition of activin signalling by follistatin. *Embo J.* **25**, 1035-1045.

91.Stamler, R., Keutmann, H.T., Sidis, Y., Kattamuri, C., Schneyer, A. and Thompson, T.B. (2008) Differential binding of N-terminal domains influences follistatin-type antagonist specificity. *J. Biol. Chem.* **283**, 32831-32838.

92.Kirkitadze, M.D. and Barlow, P.N. (2001) Structure and flexibility of the multiple domain proteins that regulate complement activation. *Immunol. Rev.* 180, 146-161.

93.Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.* **28**, 235-242.

94.Soares, D.C., Gerloff, D.L., Syme, N.R., Coulson, A.F.W., Parkinson, J. and Barlow, P.N. (2005) Large-scale modelling as a route to multiple surface comparisons of the CCP module family. *Protein eng. Des. Sel.* **18**, 379-388.

95.Discipio, R.G., Chakravarti, D.N., Müller-Eberhard, H.J. and Fey, G.H. (1998) The structure of human-complement component C7 and the C5b-7 complex. *J. Biol. Chem.* **263**: 549-560C

96.Discipio, R.G. and Hugli, T.E. (1989) The molecular architecture of human-

complement component-C6. *J. Biol. Chem.* **264,** 16197-16206.

97.Würzner, R., Mewar, D., Fernie, B.A., Hobart, M.J. & Lachmann, P.J. Importance of the third thrombospondin repeat of C6 for terminal complement complex assembly. *Immunology* **85**, 214-219 (1995).

98.Young, M. M., Tang, N., Hempel, J. C., Oshiro, C. M., Taylor, E. W., Kuntz,I. D., Gibson, B. W., and Dollinger, G. (2000) High throughput protein fold identification by using experimental constraints derived from intramolecularcross-links and mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **97,**5802–5806

99.Rappsilber, J., Siniossoglou, S., Hurt, E. C., and Mann, M. (2000) A generic strategy to analyze the spatial organization of multi-protein complexesby cross-linking and mass spectrometry. *Anal. Chem.* **72,** 267–275

100.Zhang, H., Tand, X., Munske, G.R., Tolic, N., Anderson, G.A. and Bruce, J.E. (2009) Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. *Mol. Cell. Prot.* **8**, 409-420.

101.Sinz, A. (2006) Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrometry Reviews* **25**, 663-682.

102.Petrotchenko, E.V. & Borchers, C.H. (2010) Cross-linking combined with mass spectrometry for structural proteomics. *Mass Spectrometry Reviews* **29**, 862-876.

103.Leitner, A. *et al.* (2010) Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. *Molecular & Cellular Proteomics* **9**, 1634-1649.

104.Singh, P., Panchaud, A. & Goodlett, D.R. (2010) Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. *Analytical Chemistry* **82**, 2636-2642.

105.Maiolica, A. *et al.* (2007) Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Molecular & Cellular Proteomics* **6**, 2200-2211.

106.Chen, Z.A. *et al.* (2010) Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *Embo Journal* **29**, 717-726.

107.Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567

108.Tang, C. & Clore, G.M. (2006) A simple and reliable approach to docking protein-protein complexes from very sparse NOE-derived intermolecular distance restraints. *Journal of Biomolecular Nmr* **36**, 37-44.

109.Chen, Y.W., Ding, F. & Dokholyan, N.V. (2007) Fidelity of the protein structure reconstruction from inter-residue proximity constraints. *Journal of Physical Chemistry B* **111**, 7432-7438.

110.Jansssen, B.J.C., Huizinga, E.G., Raaijmakers, H.C.A., Roos, A., Daha, M.R., Nilsson-Ekdahl, K., Nilsson, B. and Gros, P. (2005) Structures of complement component C3 provide insights into the function and evoution of immunity. Nature. 437, 505-511.

111.Janssen, B.J.C., Christodoulidou, A., McCarthy, A., Lambris, J.D. And Gros, P. (2006) Structure of C3b reveals conformational changes that underlie complement activity. Nature. 444, 213-126

112.Wiesmann, C., Ktschke Jr, K.J, Yin, J., Helmy, K.Y., Steffek, M., Fairbrother, W.J., McCallum, S.A., Embuscado, L., Deforge, L., Hass, P.E. and van Lookeren Campagne, M. (2006) Structure of C3b in complex with CRIg gives insights into regulation of complement activation. *Nature.* **444**, 217-220.

113.Ajees, A.A., Gunasekaran, K., Volanakis, J.E., Narayana, S.V.L., Kotwal G.J. and Murthy, H.M.K. (2006) The structure of complement C3b provides insight into complement activation and regulation. *Nature*. **444**, 221-225.

114.Janssen, B.J.C., Read, R.J., Brunger, A.T. & Gros, P. Crystallography - Crystallographic evidence for deviating C3b structure? *Nature* **448**, E1-E2 (2007).

115.Ajees, A.A., Gunasekaran, K., Narayana, S.V.L. & Murthy, H.M.K. (2007) Crystallography - Crystallographic evidence for deviating C3b structure? Reply. *Nature* **448**, E2-E3.

116.Sherman, F., Fink, G.R. and Hicks, J.B. (1986) The laboratory course manual for methods in yeast genetics, 1st edition, Cold Springer Harbor Press.

117.Bessetter, P.H., Aslund, F., Beckwith, J. and Georgiou, G. (1999) Efficient folding of protiens with multiple disulfide bonds in the Escherichia coli cytoplasm. *Prc. Natl. Acad. Sci. U.S.A.* **23**, 13703-12708.

118.Sambrook, J., Fritsch, E.F., & Maniatis, T. (1989) Molecular Cloning, 2nd edition, Springer press.

119.Zhange, A.L., Zhang, T.Y., Fu, C.Y., Su, D.X., Tu, F.Z. And Pan, Y.W. (2009) Inducible expression of human angiostatin by AOXI promoter in *P.pastoris* using high-density cell culture. *Mol. Biol. Rep.* **36**, 2265-2270.

120.Cavanagh, J. (2006) Protein NMR: principles and practice. *Academic Press.*

121.Sattler, M., Schleucher, J., and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients *Prog. Nucl. Mag. Res. Spec.* **34**, 93-158

122.Bodenhausen, G., and D.J. Ruben. (1980) Natural abundance nitrogen-15 NMR byenhanced heteronuclear spectroscopy. Chem. Phys. Lett. **69**, 185.

123.Vuister, G.W., and A. Bax. (1992) Resolution enhancement and spectral editing

ofuniformly 13C-enriched proteins by homonuclear broadband 13C decoupling. J.Magn. Reson. **98**, 428-435.

124.Grzesiek, S., and A. Bax. (1992) Correlating backbone amide and side chainresonances in larger proteins by multiple relayed triple resonance. *NMR. J. Am.Chem. Soc.* **114**, 6291-6305.

125.Grzesiek, S., and A. Bax. (1992) An efficient experiment for sequential backboneassignment of medium-sized isotopically enriched proteins. *J. Magn. Reson.* **99,** 20-206.

126.Grzesiek, S., and A. Bax. (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J. Biomol. NMR.* **3,** 185-204.

127.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 SequentialBackbone Assignment of Medium-Sized Proteins. *J. Magn. Reson.* **105**, 196-200.

128.Grzesiek, S., and A. Bax. (1992) Improved 3D triple-resonance NMR techniquesapplied to a 31 kDa protein. *J. Magn. Reson.* **96**, 432-435.

129.Clubb, R.T., V. Thanabal, and G. Wagner. 1992. A constant-time three-dimensionaltriple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C'chemical shifts in 15N-13C-labelled proteins. *J. Magn. Reson. Ser. B.* 97, 213-217.

130.Kay, L.E., G.Y. Xu, A.U. Singer, D.R. Muhandiram, and J.D. Formankay. (1993) AGradient-Enhanced HCCH-TOCSY Experiment for Recording Side-Chain 1H and 13C Correlations in H2O Samples of Proteins. *J. Magn. Reson. Ser. B* **101 ,** 333-338.

131.Yamazaki, T., J.D. Forman-Kay, and L.E. Kay. (1993) Two-dimensional NMRexperiments for correlating carbon-13.beta and proton.=-delta./epsilon. chemicalshifts of aromatic residues in 13C-labeled proteins via scalar couplings. *J. Am.Chem. Soc.* **115**, 11054-11055.

132.Sklenar, V., M. Piotto, R. Leppik, and V. Saudek. (1993) Gradient-Tailored WaterSuppression for 1H-15N HSQC Experiments Optimized to Retain Full Sensitivity. *J.Magn. Reson. Ser.* A **102**, 241.

133.Pascal, S.M., D.R. Muhandiram, T. Yamazaki, J.D. Formankay, and L.E. Kay. (1994) Simultaneous Acquisition of 15N- and 13C-Edited NOE Spectra of ProteinsDissolved in H2O. *J. Magn. Reson. Ser. B* **103**, 197.

134.Jordan , J.B., Kovacs, H., Yuefeng, W., Mobli, M., Lu, R., Anklin, C and Kriwacki, R. (2006) Three-Dimensional 13C-Detected CH3-TOCSY Using SelectivelyProtonated Proteins: Facile Methyl Resonance Assignmentand Protein Structure Determination *J. Am. Chem. Soc.* **128**, 9119-9128 9 9119

135.Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E.D. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins.* **59**, 697-696.

136.Chen, Y. and Bax, A. (2010) Prediction of Xaa-Pro peptide bond conformation from sequence and chemical shifts. *J. Biomol. NMR* **46**,199-204.

137.Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **28**(23), 8972-8979

138.Guntert, P. (2004) Automated NMR structure calculation with CYANA. *Methods Mol. Biol.* **278**, 353-378

139.Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination.. *Acta Crystallogr D Biol Crystallogr* **54**, 905-921

140.Guntert, P., Mumenthaler, C., and Wüthrich, K. (1997) Automated combined

assignment of NOESY spectra and three-dimensional protein structure determination.. *J Mol Biol* **273**(1), 283-298 dyana

141.Herrmann, T., Guntert, P., and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE-assignment using the new software CANDID. J. Mol. Biol. **319**, 209- 227.

142.Vriend, G., and Sander, C. (1993) Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystall.* **26**, 47-60.

143.DeLano, W. L. (2002) The PyMOL Molecular Graphics System. In., Palo Alto, CA

144.Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8**, 477-486

145.Vriend, G. (1990) A database of protein structure families with common folding motifs. *J Mol Graph* **8**, 52-56, 29

146.Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**(18), 10037-10041

147.Heiden W., Moeckel G., Brickmann J.(1993)A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comput. Mol. Des.* **7**, 503– 514

148. Fraczkiewicz R., Braun W. (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules *J. Comput. Chem.* **19**, 319– 333.

149.Shindyalov, I.N. And Bourne, P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucl. Acid. Res.* **29**, 228-229.

150.Konarev, P.V., Volkov, V.V. , Sokolova, A.V., Koch , J. and Svergun, D.I. (2003)

BIBLIOGRAPHY

PRIMUS: a Windows PC-based system for small-angle scattering data analysis, *J. Appl. Crystallogr.* **36**, 1283–1284.

151. Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26,** 1367–1137 .

152.Shindyalov, I.N. and Bourne, P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the combinatorial extension (CE) algorithm. *Nucleic Acids Res.* **29**, 228–229.

153. Eswar, N., Marti-Renom, M.A., Webb, B., Madhusudhan, M.S., Eramian, E.D., Shen, E., Pieper, E. and Sali, A. (2006) Comparative Protein Structure Modeling With MODELLER. *Curr. Prot. Bioinf.* **15**, 561-563.

154.White, C. E., Kempi, N. M., and Komives, E. A. *(1994)* Expression of highly disulfide-bonded proteins in Pichia pastoris. *Structure.* **2,** 1003-1005.

155.Connelly, G. P., Bai, Y., Jeng, M. F., and Englander, S. W. (1993) Isotope effects in peptide group hydrogen exchange. *Proteins* **17,** 87-92.

156.Barlow, P. N., and Campbell, I. D. (1994) Strategy for studying modular proteins: application to complement modules. *Methods Enzymol* **239**, 464-485.

157.Campbell, I. D., and Downing, A. K. (1998) NMR of modular proteins. *Nat Struct Biol* **5,** 496-499.

158.Kuttner-Kendo, L., Hourcade, D., Anderson, V., Mugim, N., Mitchell, L., Soares, D.C., Barlow, P. and Medof, M.E. (2007) Structure-based Mapping of DAF Active Site Residues That Accelerate the Decay of C3 Convertases. *J. Biol. Chem.* **282**, 18552-18562.

159.Cupelli, K., Müller, S., Persson, B.D., Jost, M., Arnberg, N., Stehle, T. Human Adenovirus type 21 knob in complex with domains SCR1 and SCR2 of CD46 (membrane cofactor protein, MCP). *J. Virol.* **84**, 3189-3200.

160.Soares, D. C., and Barlow, P. N. (2005) (In) Structural Biology of the Complement System. *Press, Taylor & Francis Group, Boca Raton. FL.*

161.Frishman D., Argos P. (1995) Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Genet.* **23**, 566–571.

162.Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. and Bateman, A. (2010) The Pfam protein families database. *Nucl. Acid. Res.* **38**, 211-22.

163.Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucl. Acid. Res.* **31,** 3784-3788.

164.Discipio, R. (1992) Formation and Structure of the C5b-7 Complex of the Lytic Pathway of Complement. *J. Biol. Chem.* **267**, 17087-17094.

165.Fredslund, F., Jenner, L., Husted, L.B., Nyborg, J., Andersen, G.R., Sottrup-Jensen, L. (2006) Structure of mammalian C3 with an intact thioester at 3A resolution *J.Mol.Biol.* **361,** 115-127

166.Tatuusova, T.A. and Madded, T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.*, Vol. **174**, 247-50.

167.Mayer, M. and Buchner, J. (2004) Refolding of inclusion body proteins. *Methods. Mol. Med.* **94**, 239-54.

168.Schmidt, C.Q., Slingsby, F.N., Richards, A., Barlow, P.N. (2010) Production of biologically active complement factor H in therapeutically useful quantities. *Prot. Exp.Purif.* **76**, 254-263.

169.Wurm, F.N. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotech.* **22**, 1393-1398.

170.Ikonomou, L., Schneider, Y.J., Agathos, S.N. (2003) Insect cell culture for industrial

production of recombinant proteins. *Appl. Microbiol. Biotechnol.* 62, 1-20.

171.Goto, N.K., Kay, L.E. (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy. *Curr. Opin. Struc. Biol.* **10**, 585-592

172.Roger K.B. (2003) Genetic engineering of Pichia pastoris to humanize N-glycosylation of proteins. *Trends. Biotechnol.* **21**, 459-462.

173.Morgan, W.D., Kragt, A. and Feeney, J. (2000) Expression of deuterium-isotope-labelled protein in the yeast pichia pastoris for NMR studies. *J. Biomol. NMR*. **17**, 337-47.

174.Pan, Y. and Konermann, L. (2010) Joining RDC data from fleexible protein domains. *Inverse Problems.* **26** , 502-531.

175.Foley, S.F., Sun, Y., Zheng, T. and Wen, D. (2008) Picomole-level mapping of protein disulfides by mass spectrometry following partial reduction and alkylation . *Analyt. Biochem.* **377**, 95–104.

176.Srebalus Barnes, C.A. and Lim, A. (2007) Applications of mass spectrometry for the structural characterization of recombinant protein pharmaceuticals. *Mass. Spectrom. Rev.* **26**, 370-388.

177.Jacobsen, N.E., Abadi, N., Sliwkowski, M.X., Reilly, D., Skelton, N. and Fairbrother, W.J. (1996) High-Resolution Solution Structure of the EGF-like Domain of Heregulin-R. *Biochemistry* **35**, 3402-3417.

178.Dominiguez, C., Boelens, R. and Bonvin, A.M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Amer. Chem. Soc.* **125**, 1731-1737.