Structural studies of α2-macroglobulin from the horseshoe crab *Limulus polyphemus*

Michael Nicosia

Thesis for submission towards the degree of Doctor of Philosophy

December 2016

Keele University

SUBMISSION OF THESIS FOR A RESEARCH DEGREE

Part I. DECLARATION by the candidate for a research degree. To be bound in the thesis

Degree for which thesis being submitted: PhD

Title of thesis: Structural studies of α2-macroglobulin from the horseshoe crab *Limulus polyphemus*

This thesis contains confidential information and is subject to the protocol set down for the submission and examination of such a thesis.

~~YES~~/NO [please delete as appropriate; if YES the box in Part II should be completed]

Date of submission: 20th November 2015 Original registration date: 27th September 2010

(Date of submission must comply with Regulation 2D)

Name of candidate: Michael Nicosia

Research Institute: Research Centre for Life Sciences

Name of Lead Supervisor Prof. T J Greenhough

I certify that:

(a) The thesis being submitted for examination is my own account of my own research
(b) My research has been conducted ethically. Where relevant a letter from the approving body confirming that ethical approval has been given has been bound in the thesis as an Annex
(c) The data and results presented are the genuine data and results actually obtained by me during the conduct of the research
(d) Where I have drawn on the work, ideas and results of others this has been appropriately acknowledged in the thesis
(e) Where any collaboration has taken place with one or more other researchers, I have included within an 'Acknowledgments' section in the thesis a clear statement of their contributions, in line with the relevant statement in the Code of Practice (see Note overleaf).
(f) The greater portion of the work described in the thesis has been undertaken subsequent to my registration for the higher degree for which I am submitting for examination
(g) Where part of the work described in the thesis has previously been incorporated in another thesis submitted by me for a higher degree (if any), this has been identified and acknowledged in the thesis
(h) The thesis submitted is within the required word limit as specified in the Regulations

Declaration Part 1. To be bound in the thesis

Total words in submitted thesis (including text and footnotes, but excluding references and appendices) …48639…….

Signature of candidate ……Michael Nicosia…… Date …09/11/2016…

Note

Extract from Code of Practice: ***If the research degree is set within a broader programme of work involving a group of investigators – particularly if this programme of work predates the candidate's registration – the candidate should provide an explicit statement (in an 'Acknowledgments' section) of the respective roles of the candidate and these other individuals in relevant aspects of the work reported in the thesis. For example, it should make clear, where relevant, the candidate's role in designing the study, developing data collection instruments, collecting primary data, analysing such data, and formulating conclusions from the analysis. Others involved in these aspects of the research should be named, and their contributions relative to that of the candidate should be specified (*****this does not apply to the ordinary supervision, only if the supervisor or supervisory team has had greater than usual involvement*****).***

Acknowledgements

Little did I know when agreeing to take on the challenge of a PhD, the extent of the trial that lay ahead of me. I am here now at the end of my tenure as a PhD student a different person from the one that started. This PhD has changed me in every conceivable way and in all aspects of my life. And is an experience that I take a great many positives from and draw a great deal of strength from knowing that; if I can survive a PhD, I can survive anything.

I would like to thank first and foremost Prof. Trevor Greenhough my principle supervisor. Trevor has supervised me in a way that has allowed me to flourish and develop my own ideas and ways of thinking whilst at the same time knowing when to step in right the ship and give the appropriately timed encouragement. Trevor has been extremely patient with me and put up with the numerous headaches I have no doubt caused him of the course of this body of work. I thank him once more for the opportunity to learn from him and for all that he has done for me.

The work carried out here whilst exploring a new protein for the research group builds upon the techniques, work and experience previously carried out by the group particularly Trevor Greenhough, and Ian Burns. PEG-cut serum was supplied by Peter Armstrong of the Marine Resources Centre, Marine Biology Laboratory, Woods Hole, Massachusetts. I would like to thank various group members for their help, guidance and support during the course of my studies; Ian Burns, Matthew Mold, Jenny Moran, Ruben Da Silva, Robert Williams, Jamie Littlejohn, and Darius McLeary.

Finally I would like to reserve my final and most sincere thanks to Carrie Smallcombe. Without her constant support during the tough times, and putting up with me during this difficult period I would not have been able to complete this piece of work. This piece of work represents a long and hard struggle not possible without the people mentioned above and hopefully with be

evidence to those working towards a PhD of their own that there is a light at the end of the tunnel.

Abstract

This work is focused on structural studies of the innate immune protein α2-macroglobulin from the horseshoe crab, *Limulus Polyphemus*, using crystallography and structure prediction software to reveal clues about the structure and function of this key immune mediator.

The α2-macroglobulin superfamily of proteins, characterised by the presence of an internal thiol-ester bond, is seen in humans as α2-macroglobulin, pregnancy zone protein (PZP), and Complement components C3, C4 and C5. α2-Macroglobulin (α2m) is a multifunctional serum protein, whose primary function is serving as a protease inhibitor. Rather than a traditional active-site inhibition, α2m immobilises target proteases via proteolytic cleavage of its bait region resulting in structural reorganisation of α2m and molecular entrapment of the protease. The nature of the bait region sequence allows for cleavage by a wide number of proteases which thus become entrapped. Small amines such as methylamine can also induce α2m activation resulting in the same structural reorganisation seen in proteolytic activation.

The crystal structure of *Limulus* α2m was not determined during this study, however this work represents the first reports of protein crystals of *Limulus* α2m activated with methylamine. Crystals were tested at the Diamond Light Source and diffraction was detected to 6Å with a predicted orthorhombic space group of P222 and unit cell dimensions of a = 115Å, b = 141Å, and c = 338Å.

In addition to crystallographic analysis the *Limulus Polyphemus* α2m sequence was submitted to the I-TASSER server for structure prediction. I-TASSER predicted general structural homology with the human analogue although differences arise from the human model representing the activated form and I-TASSER building a native , non-activated structure for the *Limulus* homologue. The bioinfomatic analysis and structure prediction presented here provides convincing structural

models coupled with novel insights into the activation mechanism of *Limulus* α2m and how this might relate to its functions downstream.

Contents

List of Figures

Chapter 1

Chapter 4

List of Tables

Chapter 1 – Introduction to α2-Macroglobulin: Novel Protease Inhibitor

1 Introduction to α_2-Macroglobulin: Novel Protease Inhibitor

1.1 Introduction to Immunity

1.1.1 The Immune System

The immune systems of vertebrates and other highly evolved species are branched into two distinct yet interactive systems: the innate immune system, and the adaptive immune system. The antibody mediated immune system is absent in invertebrates as it evolved early on in the evolution of the vertebrate line. The adaptive immune response depends upon the recognition and assimilation of antigens which are then processed and presented to the lymphocytes. This then leads to either cell mediated clearance of the pathogen or antibody mediated clearance. The downside to this system is that upon primary presentation of an antigen the system is slow to respond with an effective response (Meyers, 1991). This is likely the reason that high order organisms have retained their innate immune system. Whereas the adaptive immune system targets specific epitopes, the innate immune system instead is non-specific relying on the recognition of non-self, non-specific motifs on the surface of pathogens as well as recognising motifs associated with damaged self-cells. This broad non-specificity allows it to be effective in the binding, recognition and clearance of a broad range of pathogens.

1.1.2 Innate Immune Mechanisms

The innate immune system itself can be divided into three branches: barrier mechanisms, cellular, and humoral. These three branches all play a part in the innate immune systems primary weapon against pathogenic invasion – the inflammatory response. Barrier mechanisms include mucus layers, cilia and skin. They act to prevent invasion by foreign pathogens by presenting a physical barrier against invasion. For example, in the lungs the cilia lining the epithelia of the bronchioles work in conjunction with a mucus layer, to trap pathogens and move pathogens out of the

airways. Whilst not fully effective (influenza has evolved to evade this host defence mechanism); the importance of this mechanism can be seen in smokers, for whom the mucus layer and cilia are damaged. As a result they suffer from respiratory infections and irritation far more than a non-smoker with a healthy airway. Other barriers include physiological barriers such as temperature and pH, which provide hostile conditions for pathogens to proliferate. The prime example of this is seen in the stomachs of new-born babies. New-borns are born with far less acidic stomach contents than adults, and as a result are far more prone to certain stomach infections that an adult would never acquire (Wood, 2006).

The cellular component of the innate immune system is made up of phagocytes and natural killer (NK) cells. Phagocytic cells include neutrophils, eosinophils, monocytes and macrophages. The process of phagocytosis involved the ingestion and breakdown of antigens. The antigens are endocytosed following adhesion to the cells surface via receptors. Large protrusions called pseudopodia extends around the attached antigen before encapsulating and internalising the agent in the phagosome. The phagosome then fuses with a lysosome to form a phagolysosome, the formation of the phagolysosome results in the combining of the hydrolytic lysosome components with the entrapped antigen. This results in the breakdown of the antigen ready for further processing. The degraded material is then either released from the cell via exocytosis or some epitopes/motifs are then further processed to be later presented on class II MHC for T helper cell recognition (Wood, 2006).

The humoral arm of the innate immune system contains molecular instigators of the immune response. Humoral components are diverse and are an ancient component of our immune systems; this can be best shown by the presence of the pentraxins in the phylogenetically ancient horseshoe crab *Limulus polyphemus*. Their importance in the immune system is clear although their full roles and mechanisms are still poorly understood (Wood, 2006).

The pentraxins as mentioned above are a highly evolutionarily conserved family of multimeric proteins (Mantovani, *et al*., 2008), characterised by the presence of the 200 amino acid long pentraxin domain in their carboxy terminal. The pentraxins can be further categorised into two groups: long pentraxins such as PTX3, and short pentraxins such as C-reactive protein (CRP) and serum amyloid P component (SAP). Pentraxins and their homologues have been found in every species in which they have been sought indicating that the pentraxins are more than just an evolutionary artefact. In humans they are typically arranged into pentameric rings; however different conformations are seen in different species, such as the SAP-like pentraxin found in *L. polyphemus* is found in both stacked heptameric and octomeric forms (Shrive, et al., 2009). Whilst CRP and SAP share sequence homology of 51% (Gewurz, *et al*., 1995), as well as a pentameric ring formation where pentraxin helices lie on one face and the other face contains the calcium binding sites, they do differ in binding capabilities. Whilst CRP can bind to phosphocholine and phosphoethanolamine in a calcium dependant manner, SAP is only capable of binding phosphoethanolamine in this manner. These epitopes are found on a wide range of ligands such as the surface of damaged host cells and bacterial surface antigens; such as the C-polysaccharide from *Streptococcus pneumoniae*, and the lipopolysaccharide of *Haemophilus influenzae*. SAP is also implicated in the binding of some carbohydrates acting as a lectin, and is also involved in amyloid plaque stabilisation in Alzheimer's disease (Gewurz, *et al*., 1995; Mold, *et al*., 2012). CRP also has the ability to bind and activate C1q of the complement cascade as well as playing a role in the clearance of nuclear debris and binding to Fcγ receptors on immune cells.

The complement system was originally named due to its complimentary activity alongside the adaptive immune system, since then it has become established as a key feature of the immune response. The complement system is made up of three pathways: classical, lectin and alternative, all of which converge on C3, the most abundant of the complement molecules, and can resultantly lead to the formation of the membrane attack complex (MAC) and ultimately cell death (Sarma, & Ward, 2011). The alternative pathway which is constitutively active is triggered

by non-self-epitopes in the form of lipids, proteins or carbohydrates (Qu, *et al*., 2009). The classical pathway can be activated by C1qs interactions with antigen bound IgG and IgM or by antigen bound CRP, before C1r and C1s are activated leading the cleavage of C4, 2 and 3 (Sarma, & Ward, 2011). The lectin pathway is activated by carbohydrate epitopes on pathogens being recognised by Mannose Binding Lectin (MBL) – a member of the collectins. Homologous to C1q – both sharing the bouquet of tulips conformation – MBL then activates the MASPs (MBL-Associated Serine Proteases) which are equivalent to the C1r and C1s (Wallis, 2007). Figure 1.1 shows the climax of the complement cascade is the formation of the MAC, made up of C5b678 and multiple C9 molecules forming a pore that destabilises the osmotic pressure of the target cell, resulting in lysis (Sarma, & Ward, 2011).

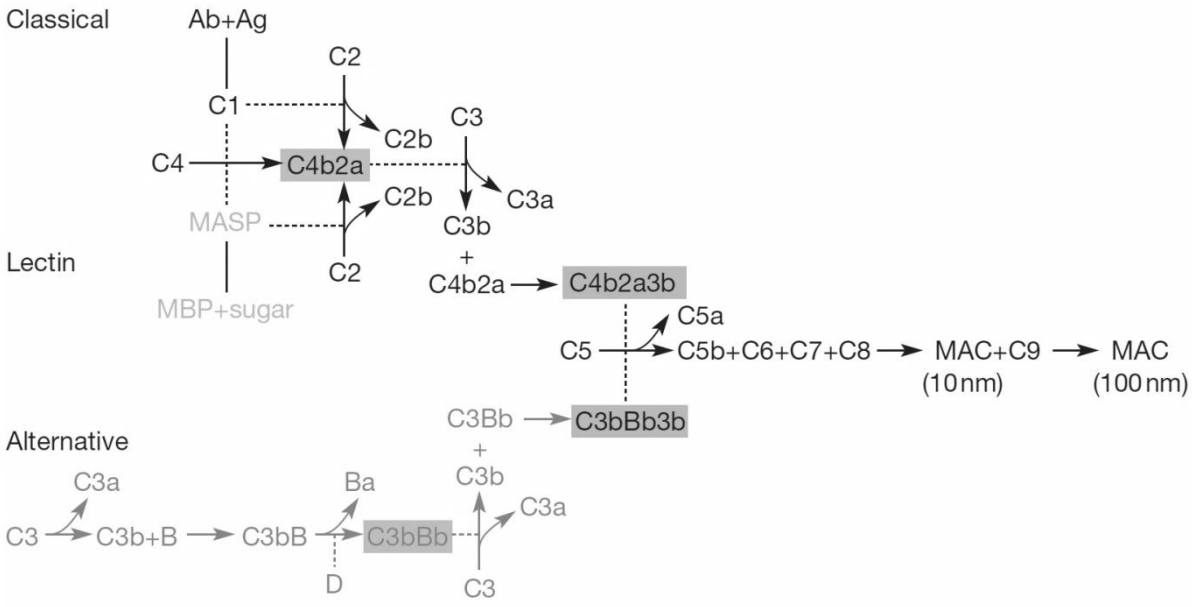


Figure 1.1. The Complement system. The classical, lectin and alternative pathways all have different origins and are all activated differently, however they all converge at C5 convertase (C4b2a3b/C3bBb3b) before following the pathway to termination at the membrane attack complex (MAC) (Wood, 2006).

Complement offers more than just cell lysis, during the cascade the various complement components are cleaved by others and some of the fragments act as major anaphylatoxins – C3a and C5a (Sarma, & Ward, 2011). These two complement components put forth a number of effects on the inflammatory response. They are capable of: acting as powerful chemoattractants

for phagocytic cells to the site of infection/injury, vasodilators inducing the contraction of smooth muscle, and inducing histamine release from mast cells (Markiewski, *et al*., 2006; Sarma, & Ward, 2011). Deficiencies in complement proteins have been associated with several pathologies including: systemic lupus erythematosus, atypical haemolytic syndrome and susceptibility to certain infections (Sarma, & Ward, 2011).

The collectins are a component of the C-type lectin superfamily containing a collagen-like domain and as with all C-type lectins a carbohydrate recognition domain (CRD). They can be distinguished into two subgroups dependent upon the number of Gly-X-Y repeats in their collagen-like domain. Those with less form bouquet like structures such as lung surfactant protein A (SP-A) and mannose-binding lectin (MBL) and those with more adopt a cruciform conformation (Weiser, *et al*., 1997). Both SP-A and SP-D exhibit specific interactions with a variety of different microorganisms and white blood cells *in vitro*. They consist of several units of heterotrimeric subunits which will normally form one of the so called stalks.

SP-D is a usually a dodecameric structure, although isolates have shown that it may form, monomers, dimers, trimers and even high-order multimers (Hartshorn, *et al*., 1996). SP-D in its dodecameric form, arranges itself in a cruciform structure with each branch of the cruciform being made up of a trimeric subunit. The individual chains are each made up 4 discrete regions; a cysteine rich N-terminus, a triple-helical collagen-like region similar to that describe in C1q, a neck region comprising of an α-helical coiled coil, and finally the C-terminus globular head region – carbohydrate recognition domain (CRD).

The primary structure of Surfactant protein-A (SPA) is similar to that of SP-D with a long cysteine containing N-terminus, a CLR, a neck region arranged again in an α-helical coiled coil and a CRD domain of 123 amino acids. SP-A shares morphology with both mannose-binding lectin (MBL) and C1q, as it has a hexameric structure. The SP-A molecule even shares the characteristic kink in the collagen-like region due to an interruption of the triplet Gly-X-Y, with SP-D and MBL. Each SP-A

subunit is made up of one SP-A1 chain and two SP-A2 chains which are closely related yet distinct (Kishore, *et al*., 2006). The few differences between SP-A1 and SP-A2 give rise to the various characteristics of the SP-A hexamer. Importantly the CRD of SP-A2 is capable of binding a broader range of carbohydrates than its compatriot SP-A1 (Oberley, & Snyder, 2003). Both surfactant proteins are key components of the host defence in the lung, capable of binding to alveolar macrophages with high affinity and promoting chemotaxis and phagocytosis of microbes (Sano, *et al*., 1999). SP-A was shown to bind to the LPS of bacteria via its lipid A region in the presence of Ca^{2+} and Na^{+} (Van Iwaarden, *et al*., 1994), whereas SP-D binds to the lipopolysaccharides of bacteria by the heptoses of the LPS core (Wang, *et al*., 2008).

1.1.3 Proteases and Their Inhibitors

Proteases are enzymes who accomplish their functions by cleaving proteins. Proteases have a range of cellular and physiological functions. Their roles include, signal transduction, defence against foreign pathogens and injury, development and proliferation, and programmed cell death. Proteases cleave proteins via a hydrolysis reaction, where a water molecule is added to a peptide bond. Peptide bonds are highly kinetically stable, taking with a half-life of 10-1000 years at neutral pH. The resonance structure of a peptide bond bestows it with a partial double bond character as demonstrated below in Figure 1.2.

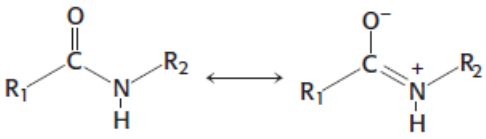


Figure 1.2. The partial double bond characteristic of peptide bonds where electrons are transferred between the nitrogen atom of the amine group and the carbon atom of the carbonyl group, leaving a negatively charged carbonyl double bonded to a positively charged amine group (Berg, *et al*. 2006)

As a direct result, the carbonyl atom is less electrophilic thus protecting it from nucleophillic attack. This in turn means that in order to cleave the peptide bond, proteases must facilitate the nucleophillic attack of a normally unreactive carbonyl group.

Proteases can be classified in two ways: by site of action (exopeptidases and endopeptidases) or the preferred method, by reaction mechanisms. Classification by reaction mechanism reveals four classes of protease:

- Serine proteases
- Cysteine proteases
- Aspartyl proteases
- Metallo-proteases

Simply put, serine and cysteine proteases act directly as nucleophiles to attack the substrate whereas, aspartyl and metallo-proteases activate water molecules as the direct substrate attacking species.

Chymotrypsin is often used as the model serine protease (Berg, *et al*. 2006). It plays a key role in the digestive system of mammals and other organisms, and is part of a large family of proteases within the serine protease classification. Chymotrypsin cleaves peptide bonds selectively on the carboxyl-terminal side of the large hydrophobic amino acids such as tryptophan, tyrosine, phenylalanine and methionine. It does this utilising the catalytic triad, a mechanism seen in all serine proteases. By employing serine-195 as a powerful nucleophile, thus playing a central role in the catalytic mechanism of chymotrypsin and forming part of the catalytic triad. The side chain of serine 195 is hydrogen bonded to the imidazole ring of histidine 57, which in turn is hydrogen bonded to the carboxyl group of aspartate 102. These three residues make up the catalytic triad. This arrangement leads to the positioning and polarisation of Ser-195 and its hydroxyl group. In the presence of substrate this hydroxyl group is deprotonated with the histidine accepting the

proton, this generates an alkoxide ion, on the serine, which is a far superior nucleophile compared to an alcohol. The histidine is enhanced as a proton acceptor due to its orientation which is mediated by the aspartate residue. This catalytic triad arrangement is characteristic of the serine protease family. Once the substrate protein is bound, the oxygen atom of the serine side chain instigates a nucleophillic attack on the target peptide bond. This carbonyl carbon now has four atoms bound to it in a tetrahedral arrangement. This tetrahedral intermediate is unstable and as a result passes a negative charge onto the oxygen atom of the carbonyl. The charge is stabilised by interactions via the NH group of the protein in a site termed the oxyanion hole. This tetrahedral intermediate collapses to form the acyl-enzyme in a step enabled by proton transfer from the histidine to the amino group that was formed by the cleavage of the peptide bond. This releases the amine group of the cleaved protein, which is now free from the enzyme. A water molecule now takes the place of the now absent amine group and trigger the deacylation of the acyl-enzyme intermediate, back to the tetrahedral intermediate formed earlier in the process. This is mediated by the histidine acting as a proton sink and drawing a proton away from the water molecule leaving the resultant OH^- to attack the carbonyl carbon of the acyl-enzyme, thus restoring the tetrahedral intermediate. The tetrahedral intermediate then breaks down to form the carboxylic acid product which is then released from the enzyme. This model of catalysis by chymotrypsin also applies for trypsin and elastase, homologues with 40% sequence homology. This method of catalysis has been proven using site-directed mutagenesis, to highlight the key residues, by substituting them for other residues resulting in a massive reduction in catalytic activity (Berg, *et al*. 2006).

Cysteine proteases, proteolytic enzyme with a cysteine residue as their main catalytic agent, are found in the body in the form of caspases and cathespins, as well as papain, from the papaya. Cysteine proteases share a similar mechanism to those of the chymotrypsin family. However, due to the sulphur of the cysteine residue being a much better nucleophile than the oxygen of the

serine, the aspartate residue of the catalytic triad is not required and cysteine proteases have evolved without them.

Aspartyl proteases utilise a pair of aspartic acid residues that work together, to facilitate the attack of a peptide bond by a water molecule. The water is activated and poised for deprotonation by one of the aspartic acids (in deprotonated form), whilst the other aspartic acid (in protonated form) polarises the carbonyl group of the peptide leaving it more susceptible to attack. Aspartyl proteases in the body include the blood pressure regulatory enzyme renin and the digestive enzyme pepsin. A more famous member of the aspartyl protease family is HIV-1 protease. HIV-1 protease cleaves proteins at the appropriate places to help form the infectious HIV virion thus playing a key role in the HIV life cycle.

The final class of protease inhibitor is the metalloproteinases. The active site of these proteases contains a metal ion predominantly zinc, that much like in the aspartyl proteases, activates a water to act as nucleophile molecule to attack the carbonyl group of the peptide bond. Examples of metalloproteinases are digestive enzyme carboxypeptidase A and the bacterial thermolysin.

Native proteases play key roles in the inflammatory process; be it internally in granules of phagocytic cells or externally once the granules have been exocytosed. Neutrophils alone contain at least six different serine proteases in their granules. Capable of a wide variety of functions such as, the formation of neutrophil extracellular traps (NETs) which are capable of binding Gram-positive and Gram-negative bacteria, they are key to the functional effectiveness of the neutrophilic response (Pham, 2006). The secretion of serine proteases by the neutrophils upon activation also leads to their regulation of the cytokine and chemokine responses. Given that dysregulation of the cytokine system can lead to a cytokine storm the system regulating them also requires built in control mechanisms. Matrix Metalloproteinases (MMPs) are a family of proteases that require a zinc ion present in their active site for catalysis to occur. The family consist of over 25 members that range in size from 28kDa to 92kDa and they have been shown to play a key role

in a number of both disease and healthy states such as; wound healing, angiogenesis, cancer, arthritis and atherosclerosis (Lindsey, 2004). Due to the range of roles the MMPs have they bind act upon a broad spectrum of ligands, the extracellular matrix (ECM) linked ligands such as: collagen, elastin, and fibronectin; as well as non-ECM ligands such as interleukins, α2m and Angiotensin I (Lindsey, 2004). MMP function is determined by the localisation of its target ligand and so the same MMP can promote or inhibit different processes dependent upon ligand availability and location. MMPs have been shown to be down regulated by transforming growth factor – β (TGF-β), a known chaperone target for α2m, in addition to the fact that α2m is known to bind some MMPs suggesting α2m has a role in regulating them (Overall, & Lopez-Otin, 2002).

Pathogens also employee a host of proteases in the functional processes. Some use pathogens as key virulence factors aimed at disrupting the immune response, disrupting the blood clotting system to avoid immobilisation, increasing inflammation and tissue damage to aid their proliferation (Armstong, 2006). Others such as the aspartyl protease HIV-1 protease play key roles in the life cycle of the pathogens. Proteases are key molecules in both the host and pathogens, in that they can be used to regulate vital systems, as well as being a means of attack.

Protease inhibitors occur in two classes: active-site protease inhibitors and the α_2-macroglobulin family of protease inhibitors. Active site inhibitors as their name would suggest, prevent proteolysis by binding to the active site of the protease and thus blocking its proteolytic activity. Protein protease inhibitors that inhibit the active site can be further categorised based on the class of protease they inhibit: a convention that does not apply to members of the α_2-macroglobulin family as they inhibit proteases from all the major classes of protease. Serine protease inhibitors, inhibit proteases such as chymotrypsin, elastin and trypsin, serine proteases. Cysteine protease inhibitors are capable of inhibiting enzymes such as the caspases and papain. Protease inhibitors are important for the regulation of healthy processes going on within the body. C1Inh is a serine protease inhibitor deactivates C1 of the complement cascade by

dissociating $C1r_2s_2$ to dissociate from the C1 molecule. α_1-proteinase inhibitor (α_1-PI) is the major serine proteinase inhibitor found in circulation, a member of the serpin family of proteinase inhibitors, is part of the acute phase response and inhibits neutrophil elastase reducing the tissue damaged caused during neutrophilic degranulation (Hiemstra, 2002). Secretory leucocyte protease inhibitor (SLPI) is a broadly specific serine protease inhibitor capable of inhibiting enzymes such as: neutrophil elastase, cathespin G, trypsin, chymotrypsin, chymase and tryptase. Its homologue Eladin inhibits neutrophil elastase and proteinase 3. There is one family of protease inhibitors that do not discriminate against the protease mechanism of action – the α_2-macroglobulin family.

1.2 The α_2-Macroglobulin/ Thiol Ester Domain Superfamily

The protease inhibitor family I39, is often referred to as the α_2-macroglobulin family. This family of protease inhibitors in humans includes: α_2-macroglobulin, pregnancy-zone protein (PZP), C3, C4a, C4b and C5 of the complement system, CD109 and CPAMD8 (Li, *et al*., 2004). Other higher organisms, and invertebrates share homologues with α2m as well as α_1I_3 of rats and ovomacroglobulin of avian and reptile eggs (Armstrong, 2010). Family members share a distinct structural homology with regards to their domain organisation, demonstrated in Figure 1.3. Members of the α_2m family have been shown to bind and render proteases from each major class inactive. They also differ in their mechanism. Rather than target the active site as is the case for all other protease inhibitors, the α_2m family entrap their target protease inside a molecular cage in a manner homologous to a Venus fly trap.

Pregnancy Zone Protein (PZP) is a member of the α2m superfamily that currently is shrouded in mystery. It's found in the same gene cluster as the other members of the α2m family members on chromosome 12p12-13. Unlike α2m it is arranged as a homodimer of 360 kDa (Sand, *et al*., 1985)

despite that it shares 71% sequence homology with α2m (Devriendt, *et al*., 1991). PZP is typically found at trace levels in the plasma (<0.01mg/ml), during childhood levels between genders vary little however due to the variation in female hormone levels in adulthood, with further gains seen during pregnancy as early as week 5 of gestation and peaking at term with an increase of 100-200 fold above preconception levels (Tayade, *et al*., 2005). PZP levels have also been shown to be elevated in men infected with HIV-1 (Sarcione & Biddle, 2001).The exact nature of its function like much of the innate immune system is still not fully clear but it is believed be a key mediator of foetal protection as they synergistically inhibit T cell proliferation and IL-2 production by the mother (Skornika, *et al*., 2004).

Complement components C3, C4a, C4b and C5, are α2m family members and innate immune system members, that are well characterised and play key roles in the mediation of the immune response. C3 is considered the key complement protein as it is the convergence point for all three systems within complement, classical, lectin and alternative. It is capable of interacting with a number of other complement proteins such as receptors, regulators and proteases as well as none-complement proteins such as viral and bacterial proteins (Janssen, *et al*. 2005). When the mature C3 molecule, 1641 amino acids long and 187kDa, is cleaved it results in two fragments C3a which is 9kDa in size and has anaphylatoxic effects and C3b the major fragment (177kDa), which leads to the exposure of the thiol ester domain (TED) which acts as a key binding site for the interactions of C3b. C3 is made up of a β and an α chain of 991 and 645 residues respectively that form 13 domains; 8 of which are the characteristic macroglobulin domains (MG1-8), a linker domain (LNK), an anaphylatoxin domain (ANA) a CUB domain with an inserted TED and a C345 domain, which exhibits a netrin-like fold. The MG domains are arranged in a manner similar to α2m in that they form a one and a half turn super-helical structure. When the C3 molecule is processed into C3c major domain rearrangements take place showing a molecular rearrangement and flexibility seen in α2m (Janssen, *et al*. 2005).

C4 is another key component of the innate immune system, is a 203kDa complement protein which takes part in both the lectin and classical pathways. C4 exists as two isoforms C4A and C4B, not to be confused with C4 fragments C4a and C4b, they are isoforms of the intact parent C4 molecule with over 99% sequence identity but the C4B isoform has been shown to have significantly higher cytolytic effects (Dodds, & Law, 1998). C4 is cleaved by both MASP-2 protease and C1s of the lectin and classical pathways respectively. This forms C4a which is an anaphylatoxin and C4b the major fragment which goes on to form C3-convertase (C4b2a) by binding to the C2a fragment, or is capable of acting as an opsonin for phagocytic clearance via the complement receptor 1 (CR1) expressed on their cellular surface (Van den Elsen, *et al*. 2002). The C4 molecule is made up of three chains β, α, and γ with molecular weights of ~ 75kDa, 93kDa and 38kDa respectively (Gigli, *et al*. 1977). C4 in terms of domain organisation is very similar to the C3 molecule with 8 MG domains, a LNK and ANA domain inserted into MG6, and the CUB domain with the TED inserted into it sitting between MG7 and MG8, a tetra arginine (RRRR) processing site that is in an insert of 46 amino acids of the loop between βA and βB of MG8, before the C terminus is finished off with the C345C domain (Janssen, *et al*. 2005). Again C4 shows a great deal of reorganisation, a family trait upon activation due to the inherent flexibility in its design (Kidmose, *et al*. 2012).

C5 of the complement system is a 196kDa protein, made up of a β and an α chain read from N to C terminus as seen in C3 and C4 with the addition of the γ chain in C4, is key to the complement cascade as it's cleavage leads to the formation of the membrane attack complex (MAC), the cytolytic component of the complement system. It is cleaved by C5 convertase of the classical and lectin pathways C4b2a3b or the C5 convertase of the alternative pathway $C3b_2Bb$, specifically between Arg^{751} and Leu^{752} leading to the formation of C5a the anaphylatoxic fragment, which binds to the G-protein coupled receptor C5aR triggering an intracellular signalling cascade that leads to chemotaxis and the release of proinflammatory mediators from granulocytes, and the major fragment and MAC constituent C5b, which binds to C6 and then C7 to initiate MAC

formation (Fredslund, *et al*. 2008). C5 like its fellow complement-component α2m super family members contains 8 MG domains, of which MG1-6 are arranged in the characteristic right-handed super-helix. The LNK domain and the ANA/C5a domain are inserted into MG6, and the LNK domain is packed between MG1-2 and MG4-6. MG7 then connects the MG1-6 super-helix to the CUB domain which features the inserted TED-homologue C5d, which lacks the presence of the thiol ester bond making C5 the only member of the α2m superfamily to not contain the thiol ester. Following CUB and the inserted C5d, is the MG8 domain which packs tightly with the previous two mentioned domains to form a super domain before feeding into the complement component characteristic domain of α2m family members, C345C (Fredslund, *et al*. 2008). The above mentioned CUB, C5d/TED, MG8 packed super domain is a conserved characteristic of the superfamily and serves to protect the thiol ester ,which in C5d is not present.

CD109, a member of the α2m superfamily that maintains the characteristic domain organisation and large molecular weight (170kDa) is a, glycosylphosphatidylinositol (GPI)-linked glycoprotein found on the surface of activated platelets, T-cells and a subset of haematopoietic stem cells (Lin, *et al*. 2002). Its mechanisms and structure are still poorly understood but it has been shown to be a co-receptor form transforming growth factor-β (TGF-β) and acts as an inhibitor for TGF-β mediated signalling pathways. TGF-β is also a binding target for the chaperone activity of α2m; one could propose that α2m potentially delivers bound TGF-β to its family member CD109 in an immunoregulatory manner (Man, *et al*. 2012). Although no crystal structure data exists it is believed to exhibit similar domain organisation as seen with the other family members, including the highly conserved thiol ester domain (Solomon, *et al*. 2004).

CPAMD8 – complement 3 and pregnancy zone protein-like, α2-macroglobulin domain-containing 8 – is a recently discovered member of the α2m family that is found in a number of species, such as pigs, cows, and *Fugu* sharing sequence identity of 65-85%. Its exact function is unknown but it is known to be highly expressed in the brain and highly responsive to stimulation by IL-1β and IL-6,

indicating involvement in the immune response. It is like CD109 membrane bound and like other members of the α2m family CPAMD8 has domains that may express proteinase activity. Its C-terminal domain contains a Karzal motif which is often found in serine protease inhibitors (Li, *et al*., 2004).

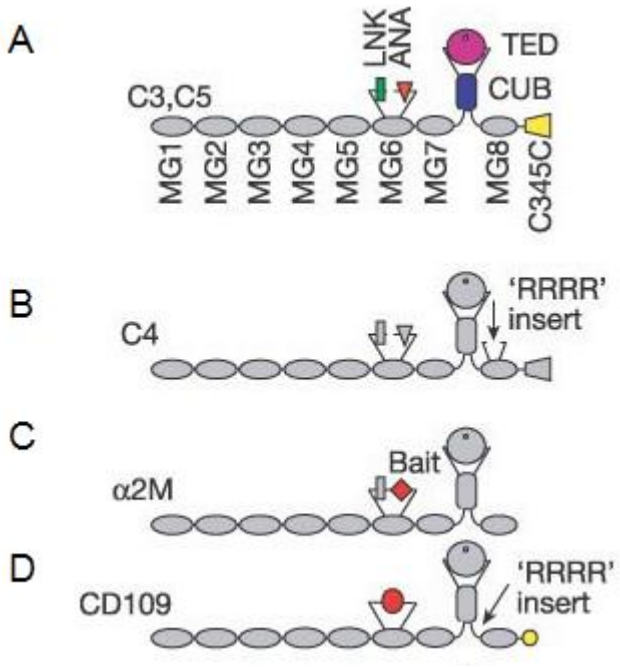


Figure 1.3. Domain schematic representations of member of the α2m superfamily. A) The domain organisation of C3 and C5 of the complement system with the labelled MG domains following through to part B), C), and D) as well as the blue CUB domain and the Pink TED/C5d, the green LNK and the red ANA domains as well as the yellow C345C domain, which are characteristic of the complement components of the α2m family, carried through to part B). B) The C4 domain organisation similar to C3 and C5 but with a tetra Arg insert in MG8. C) The domain organisation of α2m with the bait region domain (BRD) inserted into MG6 in place of the LNK and ANA domains seen in the complement proteins due to their differing functions. D) The poorly understood CD109 is believed to exhibit similar characteristics of domain organisation to its fellow family members and is known to hold a thiol ester bond in its sequence. Adapted from (Janssen, *et al*. 2005).

The structural evidence to hand has led to the development of a hypothesis that, there existed an ancestral molecule existed comprised of eight MG domains thought to have arisen from gene duplication events. This is thought to be the origin of the family due to the low levels of sequence identity between homologues in these domains thus indicating that these areas of sequence are those that have diverged furthest. The other domains are believed to have arisen from gene insertion events, which are characterised by domains existing within the loop regions of other

domains. The α2m family proteins have two such sites: MG6 and the loop region between MG7 and 8. The emergence of this as the α2m family is believed to be marked by the insertion of the CUB domain with the TED inserted within it between MG7 and MG8. The occurrence of the TED α6-α6 fold in enzymes potentially indicates that the TED fold existed separately before being incorporated into the structure. Finally to further delineate the line, the insertion of genes for the BRD in α2m, and the LNK, ANA and C345C domains of the complement proteins, marks the emergence of α2m and the complement system in the innate immune system more than 700 million years ago (Janssen, *et al*. 2005; Sahu, & Lambris, 2001).

1.2.1 Human α_2-Macroglobulin

Human α2m is a key protease inhibitor that can be thought of more as a protease binding molecule due the fact that proteases remain active once bound. It is capable of binding a broad range of proteases due to its bait region containing many cleavage sites for proteases of all classes. Implicated in a number of disease pathologies, α2m is a key protein in the human innate immune system.

1.2.1.1 Synthesis and Gene Organisation

The human α2m gene is approximately 46kb long and spans 36 exons which range in size from 21 to 229 bp with introns ranging from 145 to 7.5kb. A single copy gene in the human genome it is located on chromosome 12p12-13, as part of a gene cluster containing α2m, an α2m pseudogene, and PZP (Matthijs, *et al*. 1992; Borth, 1992). The α2m gene has been shown to contain mutations in the bait region as well as the thiol ester domain (TED); however despite this no functional changes were present (Poller, *et al*. 1992). α2m is synthesised in a number of cells, hepatocytes, astrocytes, monocytes, macrophages and lung fibroblasts all included, this broad range of cell types suggests that α2m gene expression is under the control of cell-type specific regulatory elements; with the 5’ flanking region of exon 1 of the gene containing regulatory sequences

homologous to the IL-6 response element as well as to the HP-1 element which is metal responsive, potentially important given the role of Zn^{2+} in the role of some cytokine binding (Borth, 1992). Il-6 has been shown to induce α2m synthesis in human neuronal cells, which matches with the IL-6 response region homology mentioned in exon 1. It is clear from a number of studies investigating α2m and its homologues from other species that it's synthesis in several 'normal' tissues is under humoral control (Shi, *et al*. 1990; Gaddy-Kurten, & Richards, 1991; Ramadori, *et al*. 1991). The synthesis and expression on cellular surfaces of the α2m receptor LRP-1 is also under humoral control and has been shown to be regulated by insulin in adipocytes.

1.2.1.2 Structure of Human α2-Macroglobulin

Due to the unique mechanism of α2m, its structure has been a research interest for a number of years. Initial work utilising circular dichroism and gel electrophoresis (Frenoy, *et al*., 1977; Barrett, *et al*., 1979) gave rise to it being known that α2m is a ~725kDa, tetramer, where the smallest covalent unit is the dimer molecule, and each subunit being made up predominantly of β-sheets with a very low number of α-helices. Gel Electrophoresis work (Figure1.4) was also the first to demonstrate the two forms of α2m, fast form – reacted with a protease or a small amine, and the slower native form (Barrett, *et al*., 1979).

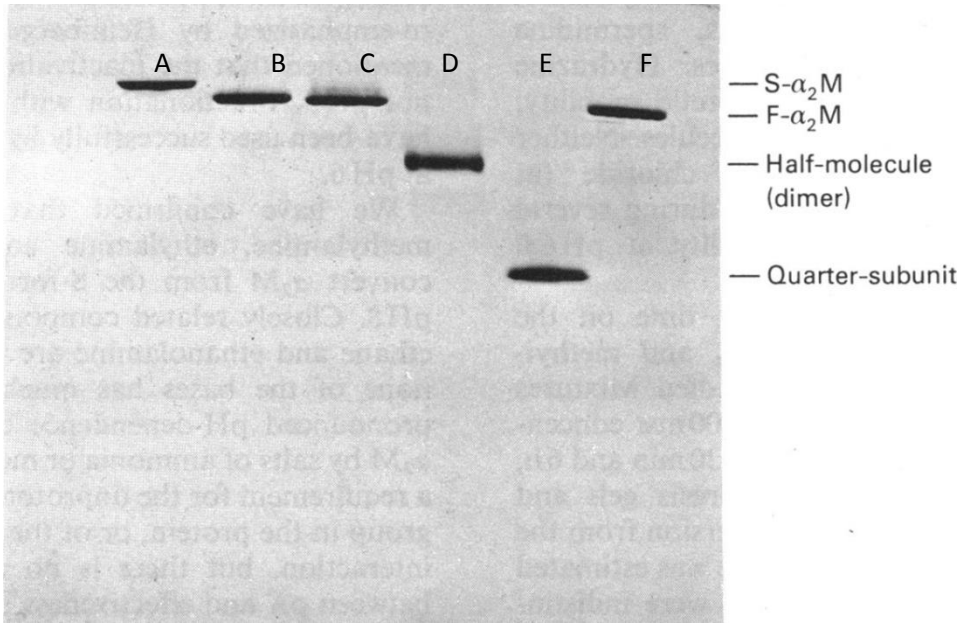


Figure 1.4. Gradient polyacrylamide gel of various α2m morphologies (Barrett, *et al*. 1979). Band A – Native/'Slow' form α2m. Bands B&C – Reacted/'Fast' form α2m, reacted with trypsin and methylamine respectively. Despite the increase in molecular weight the complex is smaller than the native form and thus can travel further through the gel, and that the overall size is the same irrespective of protease or methylamine reaction, suggesting similar mechanisms following reaction. Band D – Dimeric unit of α2m, split from tetrameric form my exposure to pH 2.0. Band E – The monomer subunit of α2m, following exposure to dithiothreitol. Band F – Reacted/'Fast' form α2m-trypsin, reduced with dithiothreitol.

Initial studies to the crystal structure of α2m yielded little success with a maximum resolution of 9Å being achieved (Andersen, *et al*., 1991; Andersen, *et al*. 1994), however this resolution was not sufficient to yield any structural data. In 2012, the crystal structure of human α2m reacted with methylamine was solved to 4.3Å (Marrero, *et al*., 2012).

As previously stated human α2m is a ~720kDa homotetramer arranged into two non-covalently linked dimers. Each of these dimers can be thought of as an active unit and thus human α2m has two active units as opposed to the one of Limulus α2m. Each individual subunit is ~180kDa in size and is 1451 amino acids long. It has been shown that each subunit of α2m has 8 glycosylation sites via the N4 of the following residues: $Asn^{32, 47, 224, 373, 387, 846, 968, 1401}$ (Sottrup-Jensen, *et al*., 1984). The bound oligosaccharides are largely heterogeneous in size, solvent exposed, and are predominantly N linked acetylglucosamine (N-GlcNAc), but also containing galactose, mannose, N – acetylneuraminic acid, and fucose. The molecule shares a great deal of homology with the

complement proteins C3, 4 &5. The α2m subunit contains seven fibronectin type 3 folded macroglobulin domains (MG), an α-helix based thiol ester containing domain (TED), the physiologically significant bait region domain (BRD), a complement protein subcomponents C1r/C1s, urchin embryonic growth factor and bone morphogenetic protein 1 (CUB) domain, and a receptor binding domain (RBD) (Marrero, *et al*., 2012; Wyatt, *et al*., 2012). The first seven domains of the human α2m monomer are of the MG class mentioned previously, and are referred to in the literature and from this point on in this work as MG1-7. The domains are approximately 110 amino acids in length, comprising seven β-strands arranged in anti-parallel β-sandwiches of 3 and 4 strand sheets. The CUB domain is comprised entirely of β-sheets and is made up of 116 amino acids arranged as two four-stranded antiparallel β-sheets. The TED domain is a 315aa helical based domain with an α-α-toroid topology, arranged of six concentric α-hairpins organised as a six fold α propeller around a central axis giving rise to a thick disc with two parallel flat sides, and features a β-hairpin between α9 and α10. The RBD also sometimes known as MG8 is a 129 amino acid C-terminal domain which is a variant of the typical Mg domain architecture with a β-α-β fold inserted, resulting in a four stranded and a five stranded twisted sheet. The BRD is a 126 amino acid long flexible domain which is roughly 85Å in length. And is more compact in structure at Gln579 – Thr705, and includes a long loop region Cys^{619} – Asp^{665} which protrudes from the monomer and is the tethering loop of the BRD (Marrero, *et al*. 2012).

The first six MG domains are arranged as a compact ellipsoidal one and a half turn right-handed, super-helix as highlighted by Figure 1.5 taken from the supplementary materials to (Marrero, *et al*., 2012).

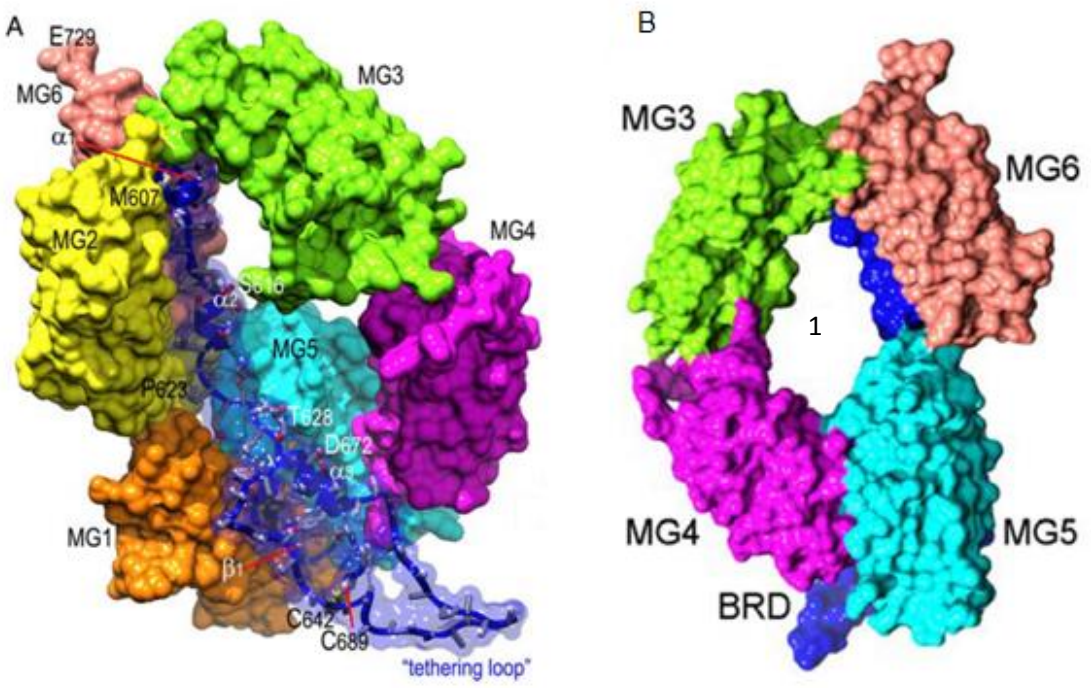


Figure 1.5. A) Internal view of the right-handed one and a half turn ellipsoidal super-helix arrangement of MG1-6 including the BRD in blue showing entrance one. B) External view of entrance one from outside of the tetramer only showing MG3-6 inclusive of the BRD. From the supplementary materials to (Marrero, *et al*. 2012).

The result is such that MG3-6 enclose a central ellipsoidal opening known as entrance 1 (Marrero, *et al*. 2012). Access to this entrance is mediated by the glycan chain that is bound to Asn^{224} of MG3, one of the encircling MG domains. MG7 closes the super-helix and forms the upper limit of the molecule. Following on from MG7 is the CUB domain which is inserted adjacent to MG2, with the TED domain inserted between β3 and β4, which sits just below the CUB domain, and consequently lies lateral to MG1 and MG2, keeping the overall structure compact. After the sequence re-joins and completes the CUB domain, the sequence then enters the RBD, which sits behind the BUB domain, tight next to MG3 and contacts TED and CUB domains (Marrero, *et al*. 2012). This overall structure leads to a convex front face and a concave back face, which leads to the formation of a large cavity at the rear of the monomer that harbours the BRD.

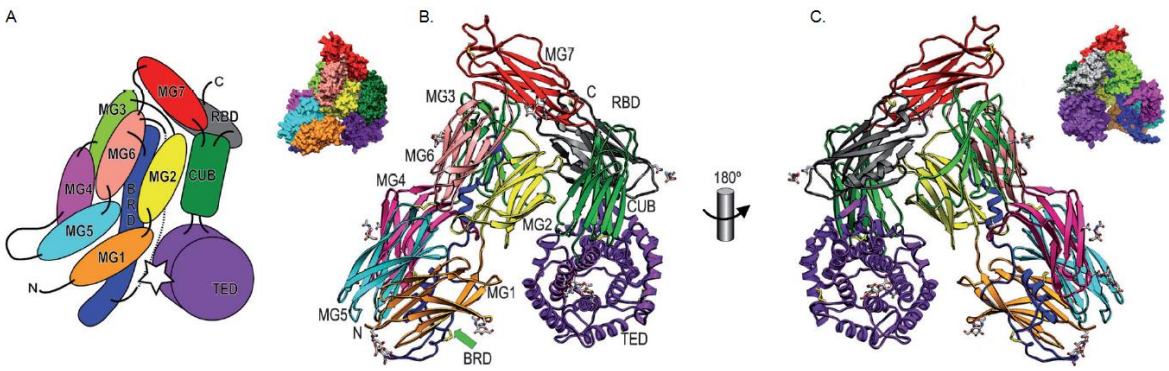


Figure 1.6. A) A schematic representation of the domain organisation/tertiary structure of a subunit of human α2m; with the bait region represented as a dashed line and the star demonstrating a potential protease cleavage site. B) Ribbon and space fill representations of the tertiary structure of an α2m subunit (convex face) with the green arrow indicating the location of the bait region. C) The back/ concave face of the α2m subunit from humans demonstrating the bait region to the rear of the molecule. Diagrams edited from (Marrero, *et al*. 2012).

The BRD interacts with MG1-3 as well as MG-5 the proximities of which can be seen in Figure 1.6A. The bait region itself, Pro^{667} - Thr^{705} sits in the lumen of the cavity formed by the other ten domains making it highly accessible to potentially proteolytic targets, and its high flexibility allowing it to twist and fit almost any active-site pocket. Native α2m in human circulation is a tetramer and key to its tertiary and quaternary structure are the disulphide bonds present. Each subunit of α2m contains 13 disulphide bonds.

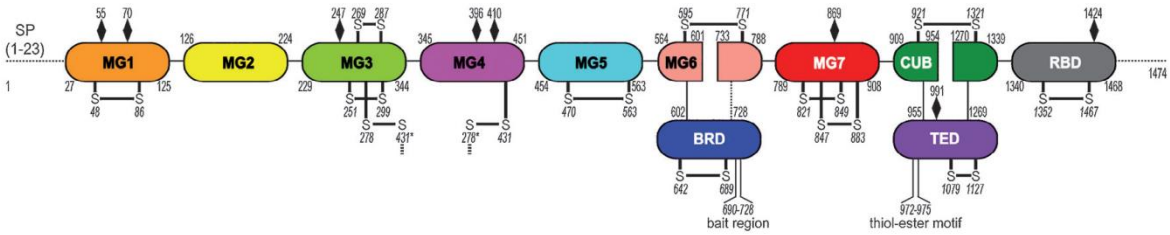


Figure 1.7. Domain schematic of human α2m depicting the residues for each domain as well as locations of N-linked glycosylation sites (shown with a black diamond), and disulphide linkage sites (Marrero, *et al*. 2012)

Eleven of which are intra-subunit between residues; Cys^{25} - Cys^{63}, Cys^{228} - Cys^{276}, Cys^{246} - Cys^{264}, Cys^{447} - Cys^{540}, Cys^{572} - Cys^{748}, Cys^{619} - Cys^{666}, Cys^{798} - Cys^{826}, Cys^{824} - Cys^{860}, Cys^{898} - Cys^{1298}, Cys^{1056} - Cys^{1104}, Cys^{1329} - Cys^{1444}, as depicted in Fig. 1.7. There are two inter-subunit disulphide bonds which provide covalent linkage between subunits via residues: 255A-408B, and 408A-255A, where A and B represent different subunits (Marrero, *et al*., 2012). The interaction surface between two

monomers of a dimeric unit is symmetrically shaped by domains MG3 and MG4, which meet at the top centre of the complex, and both subunits are linked by their back/concave faces. Similarly symmetrical contacts are made at the MG4 and TED interfaces between subunits. The dimerisation of subunits leads to the formation of entrance 2. Entrance 2 is encompassed by MG2, MG3, MG7, CUB and TED of one monomer with the disulphide linked subunit contributing MG4 to the overall structure of entrance 2.

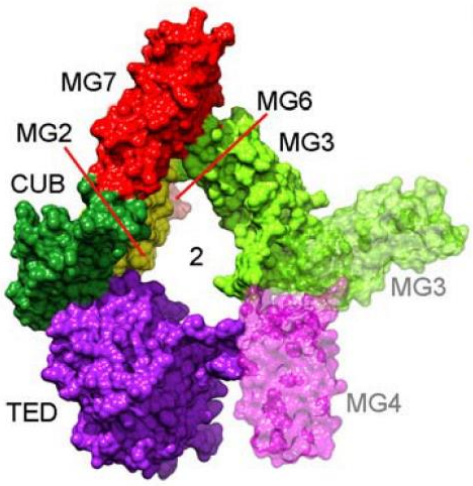


Figure 1.8. Entrance 2 between a subunit of human α2m and its disulphide linked dimeric partner including the MG3 of the disulphide linked subunit for completeness. One subunit providing MG2, MG3, MG7, CUB and TED towards the entrance (solid colours), and the disulphide linked subunit providing MG4 to the entrance super structure (semi-transparent colours). From the supplementary materials to (Marrero, *et al*. 2012).

The tetramer is a result of non-covalent interactions between the two function unit dimers, with the bottom dimer referred to as the 'vicinal' dimer, and the subunits arranged in a manner that the subunits can thought of relative to a single subunit as, disulphide linked, vicinal and opposite. As illustrated in Figure 1.8 (Marrero, *et al*. 2012).

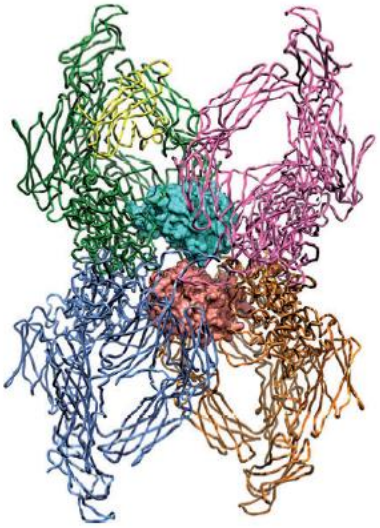


Figure 1.9. H- orientation, ribbon diagram of the human α2m tetramer, with trapped proteases demonstrated in Connelly surface in light blue and pink, as well as the RBD of the green subunit highlighted in yellow. Relative to the green subunit, the pink is the disulphide linked and completes the functional dimer. The blue subunit is vicinal to the green and the orange is opposite to the green. The green and pink dimer is vicinal to the blue and orange dimer. Diagram edited from (Marrero, *et al*. 2012)

The TED from each subunit interacts with the TED from its vicinal subunit, therefore in Figure 1.9. the TED domains from the blue and green subunits are interacting with one another as are the TED domains form the pink and orange subunits. Further quaternary interactions occur between the BRDs of opposite subunits via the BRD tethering loop.

As discussed entrance one in present in each subunit of α2m and two of entrance 2 per dimer, thus providing eight entrances to the molecule, there are though other entrances to the central 'prey chamber'. Entrance 3 is formed upon tetramerisation of α2m, this entrance is framed by the TED of one monomer, the tethering loop of the BRD of its vicinal subunit, and the MG4 and BRD of its disulphide linked subunit, illustrated in Figure 1.10. And in total there are four of these entrances bringing the total number of entrances to the prey chamber to twelve (Marrero, *et al*. 2012).

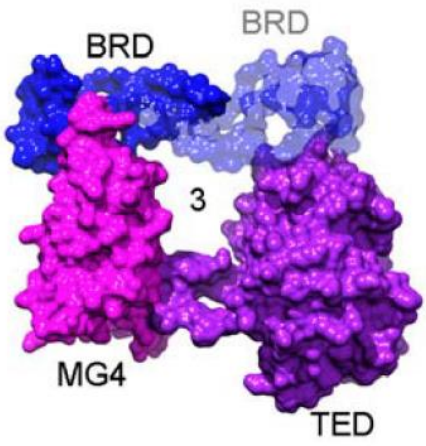


Figure 1.10. Entrance 3 of which in the whole tetrameric α2m there are four. The TED of one subunit is flanked by the MG4 and the BRD of its disulphide linked partner, and the tethered loop of the BRD from its vicinal subunit. From the supplementary materials to (Marrero, *et al*. 2012).

Access to entrance 3 much like entrance one is mediated by the glycan chains bound to the molecule at Asn^{373} and Asn^{387}. The central prey chamber is roughly 60Å in diameter which is sealed at the upper and lower limits by the MG3 and MG4 that are disulphide linked and thus mediate dimerisation. The prey chamber however isn't simply a large cavity. It is restricted at its centre by loops provided by the TEDs resulting in two halves of the prey chamber; these loops are known collectively as the cavity-narrowing belt. As a result the prey chamber has room to accommodate two 20-30kDa proteases, one for each functional dimeric unit. In addition to the prey chamber the molecular arrangement results in the addition of elongated volumes housed by each concave face of the four subunits known as 'substrate ante-chambers'. These substrate ante-chambers house small substrates (6-9kDa), small enough to pass through any of the twelve entrances described in activated α2M; subsequently providing access to the proteases and their active sites, which is evident in experimental data (Marrero, *et al*. 2012; Barrett, *et al*. 1979).

Table 1.1. Structural feature of human α2m as found in the crystal structure (Marrero, *et al*. 2012; Rehman *et al.* 2013).

Structural features of human α2m	
Quaternary Structure	Homotetramer
Molecular Weight	720kDa
Dimensions	140 x 140 x 210Å
Chain Length (per subunit)	1451
a – Alpha Helices	17
b – Beta Strands	67
Disulphide bonds	13
Prey Chamber Dimensions	60Å diameter
Prey Chamber Entrances	12

In short the human α2m tetramer is a highly ordered complex which utilised its quaternary structure to act as a molecular venus-fly trap for protease molecules of 20-30 kDa, the flexibility in its structure leading to the molecular mechanism for which it is known is vital to its function.

1.2.1.3 The α_2-Macroglobulin Receptor

The α2-macroglobulin receptor/Low Density Lipoprotein Receptor (LRP-1/CD91) is a 600kDa cell surface glycoprotein (Kristensen, *et al*., 1990). LRP-1 is a noncovalently linked heterodimer of 515kDa and 85kDa subunits, following proteolysis is the trans-Golgi (Borth, *et al*. 1994). Its action is regulated by a 39kDa receptor associated protein (RAP), which contains a heparin binding domain (Borth, 1992; Herz, *et al*.1991). LRP-1 has been found on macrophages, monocytes, astrocytes, fibroblasts, hepatocytes, adipocytes and syncytiotrophoblasts, and it's presence on the surface of early differentiated monocytes can be seen as a marker in chronic myelomonocytic leukemias (Moestrup, *et al*. 1990; Moestrup, & Hokland, 1992).

As a member of the LDL-receptor gene family, it shares the five common structural units that make up the members of this protein family. The cysteine rich ligand binding/complement repeats (CR) that occur in clusters of 2-11 individual repeats, the primary binding site for ligands of LRP-1 that have had their binding sites mapped (Herz, & Strickland, 2001;Neels, *et al*. 1999). The epidermal growth factor (EGF) precursor homology domains, these are made up of the two EGF receptor-like repeats and six YWTD domains in a propeller arrangement with another EGF following the propeller (Springer, 1998). These two motifs are repeated giving rise to 4 clusters within the receptor structure. Each cluster contains a varying amount of complement repeats, with the four clusters containing 2, 8, 10, and 11 repeats respectively (Dolmer, & Gettins, 2006). At the interface between the ligand binding clusters mentioned above and the membrane spanning region are six EGF repeats. The cytosolic side of the transmembrane region is a tail region containing two NPxY motifs that act as a molecular dock for the endocytic and cell signalling proteins (Trommsdorff, *et al*. 1998)

α2M binds to its receptor via its receptor binding domain (RBD) in a reaction that is pH dependant, where a pH exceeding an acidity of 6.8, results in no further interactions between protein and receptor (Yamashiro, *et al*. 1989). The conformational change triggered by reaction with a protease or a small amine, results in the exposure of the RBD and thus allowing the binding of α2m by the receptor and thus the clearance of the bound proteases (Enghild, *et al*., 1989; Holtet, *et al*. 1994). α2M has been shown to bind to CR3,4, & 5 of cluster II of LRP-1 (Dolmer, & Gettins, 2006) although it has also been mapped to the CRs of cluster IV as well (Neels, *et al*. 1999). The interaction between CR 3, 4 & 5 has been mapped to two lysine residues within the RBD, Lys_{1370} and Lys_{1374} (Dolmer, & Gettins, 2006). Lys_{1374} is conserved between species whereas Lys_{1370} is replaced with a glutamic acid residue in the α2M homologue from *Limulus polyphemus*, and this change may responsible for fast-form *Limulus* α2m reacting with LRP-1 but with a lower affinity than its human counterpart (Armstrong, & Quigley, 1999). Fast-form α2m is the activated form of the molecule referred to as such due to its ability to permeate through gels and

membranes faster than the native form of the molecule and thus behaving like a molecule of smaller molecular weight.

LRP-1 has been shown to recognise no fewer than 30 distinct ligands, from several families of proteins serving a broad range of functions, making CD91 a broad spectrum yet highly specific receptor protein capable of mediating several immune functions (Herz, & Strickland, 2001). LRP-1 has been shown to bind with α2M, ApoE, MMPs 9 and 13, hsp-96 but perhaps one of its most interesting interactions is that with calreticulin. It has been shown that calreticulin binds to LRP-1 on the surface of cells (Basu, *et al*., 2001). Calreticulin has also been shown to bind to the collagen region of collectins, even when surface bound via LRP-1 (Eggleton, *et al*., 1994; Vandivier, *et al*., 2002). This then allows SP-A, SP-D, MBL and potentially C1q, although this is currently disputed (Duus, *et al*., 2010), once ligand bound and in an aggregate state to present their collagen tails and producing a proinflammatory response, via upregualtion of P38 and NFκB (Gardai, *et al*. 2003).

1.2.1.4 Functional Mechanisms and Targets

A lot like the protease binding capabilities of α2m, its functions too are broad, further adding to the speculation that α2m is a key mediator of the immune response. Along with its ability to bind and trap proteins its glycosylation allows it to be bound by lectins of the host system and thus offer a protective role. It has also been shown to act as a molecular chaperone for various cytokines.

1.2.1.4.1 Protease Clearance

The importance of protease inhibitors cannot be stressed enough. Both parasites and bacteria have been shown to secrete and express proteases on their cellular surfaces. These proteases contribute greatly to their virulence. Across the plethora of invasive bacteria and parasites, every major class of protease is seen to act as a virulence factor, showing the breadth of their importance in pathogenicity (Armstrong, 2006). Invasive proteases seek to disrupt the natural defences against their invasion; such as the blood clotting system which plays an integral role in the immobilisation of the invasive agents. Proteases that attack this system seek to liberate the entrapped agents thus allowing their systemic dispersal. Some bacteria such as *Yersinia pestis*, the pathogen responsible for the black plague, go about this in an indirect manner; by releasing proteases targeted at plasminogen resulting in the activation of plasmin, they can thus trigger host fibrinolytic pathways to achieve their goals (Sodeinde, *et al*., 1992).

Table 1.2. Examples of Virulence Factors from a variety of pathogens, and their roles in pathogenesis. Table adapted from (Armstrong, 2006)

Pathogen	Protease	Role in pathogenesis	References
S. stercoralis, S. manson	Ss40	Epidermal Invasion	(Brindley, *et al*. 1995; Cohen *et al*., 1991)
Seratia, Pseudomonas	Serratia 56K, Pseudomonas alkaline protease	Cytolysis (Internal attack)	(Maeda, *et al*. 1987; Molla, *et al*. 1987)
E. histolytica	Cysteine proteases	Cytolysis (External attack)	(Reed, *et al*., 1993; Reed, *et al*., 1989a)
P. chaubaudi	P68	Intracellular invasion	(Breton, *et al*. 1992)
S. pyogenes, Y. Pestis, B. burgdoferi	Pla, OspA	Fibrinolysis	(Fuchs, *et al*., 1994; Poon-

			King, *et al*., 1993, Sodeinde, *et al*. 1992)
Pseudomonas, Serratia, Candida, C. salmositica	Elastase, Serratia 56KDa Protease	Depletion of protease inhibitors	(Khan, *et al*. 1994; Rasmussen, *et al*., 1999; Zuo, & Woo, 1997)
Streptococcus, Serratia, Candia, C. salmositica	Streptococcal C5a Protease, Candida acid Protease, Entamoeba Cysteine protease	Deactivation of complement	(Cutler, *et al*., 1993; Reed, *et al*., 1989b, Schenkein, *et al*., 1995)
Streptococci, Serratia, Candida, P. gingivalis	IgA1 protease, Serratia 56 K, Candida protease	Cleavage of immunoglobulins	(Kilian, & Reinholdt, 1986; Plaut, 1983)
E.coli, Y. pestis	Microbial plasminogen activators	Plasmin generation via plasminogen cleavage.	(Leytus, *et al*., 1981; Sodeinde, *et al*., 1992)

Table 1.2, highlights the key roles of proteases and thus the need to neutralise them. As previously mentioned there are a host of protease inhibitors at the immune systems disposal, the majority of which have extremely high specificity consequently the presence of a pan-proteinase inhibitor such as α2m and its family members is a key weapon in the arsenal of the immune system.

The protease binding activity of α2m has often been described as promiscuous due to its ability to bind a broad range of proteases, from each major category of protease. This broad specificity is due to the bait region, a 39 amino acid long loop region within the BRD (Sottrup-Jensen, *et al*.,

1989). Within this loop region are motifs of amino acids that are cleavage sites for a broad range of proteases shown below in Figure 1.11.

PQLQQYEMHGPEGLRVGFYESDVMGRGHARLVHVEEPHT

Figure 1.11. The Bait Region of the human α2m molecule. A 39 amino acid long free loop region accessible by proteases for cleavage thus activating α2m and causing the cleavage of the thiol-ester domain and ultimately the entrapment of the protease following the molecular reorganisation of α2m (Sottrup-Jensen, *et al*., 1989).

To highlight the extent at which this sequence of amino acids is capable of binding a broad range of protease inhibitors, I submitted the sequence of the bait region to the Expasy program PeptideCutter, which highlights the various cleavage sites for a range of proteases, typically used for proteomics and users of mass spectrometry but will serve the purpose of highlighting the broad nature of the bait region well in Figure 1.12.

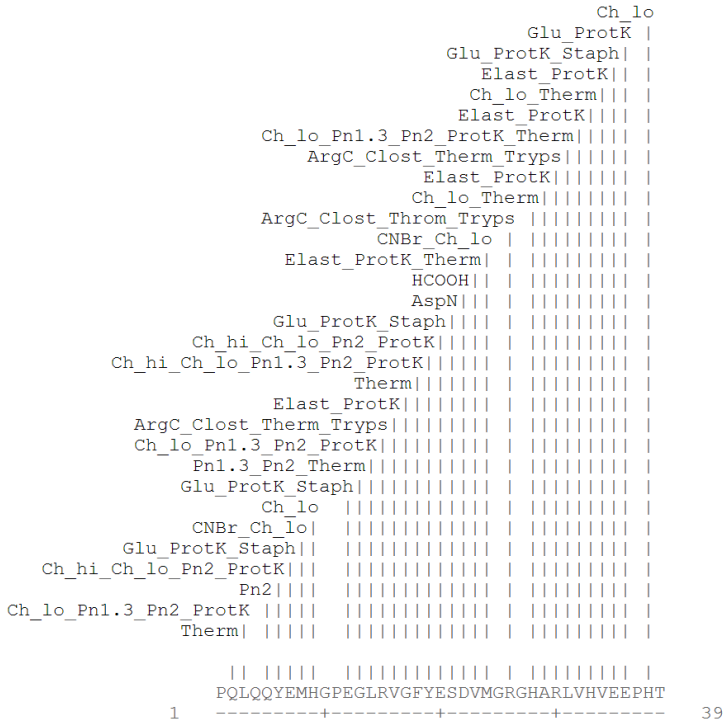


Figure 1.12. The PeptideCutter analysis of the bait region of human α2m Figure 1.10 with accompanying table explaining the data (Table 1.3). All but five residues provide cleavage sites for the proteases selected, thus highlighting just how effective the bait region is at providing cleavage sites for proteases.

Table 1.3. Protease cleavage site data from PeptideCutter analysis of the bait region of α2m.

Protease code	Protease	No. of cleavages	Position of Cleavage Sites	Class of Protease
ArgC	Arg-C Proteinase	3	15, 26, 30	Serine
AspN	Asp-N endopeptidase	1	21	Metallo
CNBr	CNBr	2	8, 24	-
Ch_hi	Chymotrypsin – high specificity	3	6, 18, 19	Serine
Ch_lo	Chymotrypsin – low specificity	12	3, 6, 8, 9, 14, 18, 19, 24 28, 31, 33, 38,	Serine
Clost	Clostipain	3	15, 26, 30	Cysteine
HCOOH	Formic acid	1	22	-
Glu	Glutamyl endopeptidase	5	7, 12, 20, 35, 36	Serine
Elast	Neutrophil elastase	5	16, 23, 29, 32, 34	Serine
Pn1.3	Pepsin (pH1.3)	5	3, 13, 14, 18, 31	Aspartyl
Pn2	Pepsin (pH >2)	9	3, 5, 6, 13, 14, 18, 18, 19, 31	Aspartyl
Prot K	Proteinase K	17	3, 6, 7, 12, 14, 16, 18, 19, 20, 23, 29, 31, 32, 34, 35, 36, 39	Serine
Staph	Staphylococcal peptidase I	4	7, 12, 20, 35	Serine
Therm	Thermolysin	9	2, 13, 15, 17, 23, 28, 30, 31	Metallo
Throm	Thrombin	1	26	Serine
Tryps	Trypsin	3	15, 26, 30	Serine

Figure 1.12 clearly shows the diverse specificity α2m has for a number of proteases. Not only do varieties of proteases cleave the bait region, note at least one from every major class, but many of them have multiple cleavage sites thus highlighting just how versatile the bait region is in terms of its specificity.

Due to the dimensions of the entrances mentioned in the previous section, proteases are limited in size to about 20-30kDa. Once they have gained access to the central prey chamber and cleaved

the bait region leading to the cleavage of the thiolester bond also a molecular rearrangement occurs where the native α2m model it is speculated narrows closing off the entrances by which the protease accessed the prey chambers thus trapping the molecule within the 'molecular cage' of α2m. Utilising this mechanism α2m is known to trap the proteinases released by both self-cells and invasive pathogens.

Among the self-proteinases α2m is known to trap are key inflammatory agents such as neutrophil elastase and chymase from mast cells, transferring the key serum iron transport protein, proteolytic members of the clotting cascade and defensins such as HNP-1 a key microbicidial protein. Foreign pathogens it is known to bind include, the collagenase from *Clostridium histolyticum* and proteases from *Trichophytaon mentarophytes, Fusiformis nodosus* and *Bacillus subtilis* (Rehman, *et al*., 2013)

The α2m receptor LRP-1 clears bound proteases primarily on hepatocytes in serum and in tissues by a host of immune cells such as: fibroblast cells, monocytes and macrophages. These complexes are formed within a matter of a few minutes after their formation and with high affinity. Once bound the α2m-LRP-1 complex dissociates readily in mild acidic conditions, which are optimal for the endosomal proteins responsible for the breakdown of the α2m, where in some cases α2m is recycled back into circulation and in other it is broken down along with the protease in question (Borth, 1992)

1.2.1.4.2 Human α_2-Macroglobulin and its interactions with Surfactant Protein-D

Human surfactant protein D (hSP-D) is a key molecule of the innate immune response to infection, primarily in the lungs (Crouch, 2000). It binds to bacterial and viral surface antigens via its carbohydrate recognition domain (CRD) in a calcium dependant manner. Each dodecamer of hSP-

D contains 12 CRDs, providing 12 potential binding sites for polysaccharides. As a result, hSP-D can form large aggregates of multiple hSP-D molecules linking multiple pathogens.

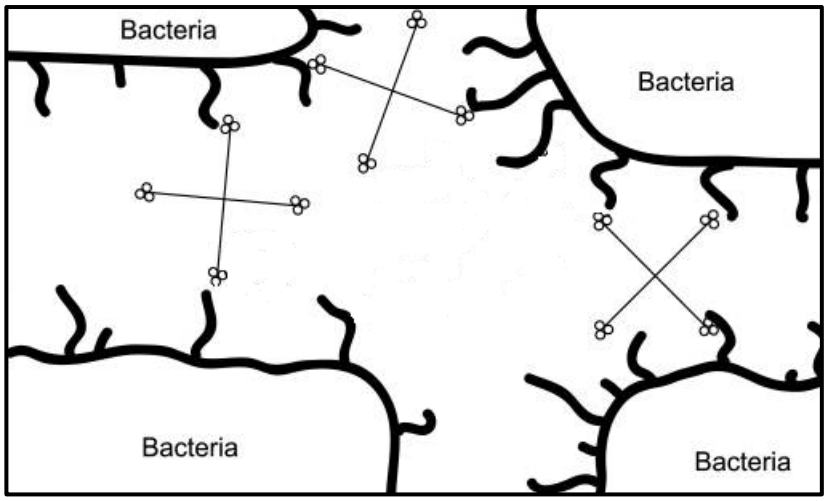


Figure 1.13. A molecule of hSP-D is capable of binding multiple pathogens thanks to its cruciform morphology. This allows large complexes to form of multiple bacteria linked by multiple hSP-D molecules, which are then targeted for clearance by the immune system (Figure adapted from Craig-Barnes, *et al*. 2010)

hSP-D binds a variety of target saccharides on the surface of its ligands, ranging from maltose, which it binds with highest affinity, to mannose and GlcNAc amongst others (Persson, *et al*. 1990). α_2-macroglobulin as mentioned earlier is highly glycosylated, with the human form having a variety of N-linked glycans such as high oligomannoses, galactose, fucose, GlcNAc and sialic acid (Arnold*, et al*. 2006). These oligomannose residues are found on a variety of molecules, perhaps most notably on the haemagglutinin of the influenza virus. It is these oligomannose residues that provide hSP-D binding capabilities to the influenza virus (Crouch, *et al*. 2011). Of the collectins, it was first shown that mannose-binding lectin (MBL), a constituent of the complement system. Binds to the oligomannose residues of human α2m found on Asn^{846}, between the bait region and the thiol ester domain (Arnold, *et al*., 2006). However more recently it was shown that hSP-D binds to α2m and that the presence of α2m actually boosts the innate immune potential of hasps (Craig-Barnes, *et al*. 2010). It was shown that hSP-D bound to the surface glycans found on α2m,

in particular those high oligomannose residues (Man_{5-7}), which make up 8% of the glycan pool. They bind these by the CRDs in a calcium dependant manner mimicking the action against pathogenic targets. Due to the high number of surface glycans found on α2m, this then facilitates a cross-linking agglutination between multiple hSP-D molecules, which can also bind to pathogens during infection, as shown in Figure 1.13. The likelihood of such an interaction occurring is far more likely during infection as α2m is found in the lungs at 0.09-2.02μg/ml but during infection that concentration increases to 220μg/ml in the acute phase response (Van Vyve, *et al*., 1995).

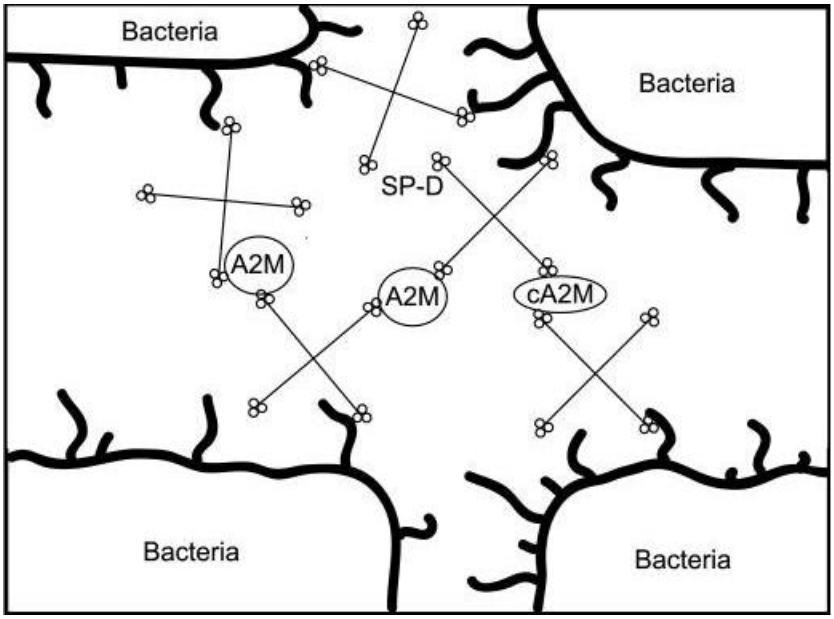


Figure 1.14. hSP-D binds the antigens on the surface of bacteria by its CRDs. The unbound CRDs are then free to bind to the oligomannose glycans on the surface of α_2 macroglobulin (A2M) and its reacted form (cA2M). The presence of unreacted a2m offers protection to hSP-D from proteolytic cleavage by elastases, whilst the reacted forms still offer structural stability to the bacterial-hSP-D-a2m complexes (Craig-Barnes, *et al.,* 2010).

Not only does hSP-D bind to α2m facilitating the formation of large aggregates but it binds to both the native and the activated forms of the molecule, as demonstrated above in Figure 1.14. This provides the hSP-D with protection from proteases, in particular elastin. Elastin is released in the inflammatory response by neutrophils as well as a defence mechanism by invading pathogens. The elastin cleaves the CRD of hSP-D rendering it inactive. The formation of α2m–hSP-D complexes results in native α2m being present and available for protease encapsulation. Once the

protease has been trapped, the α2m molecule becomes more compact following a conformational change; thus highlighting the importance of hSP-D being able to bind both native and activated forms so as not to disturb the aggregate stability. To date there have been no structural studies on the complex formed between α2m and hSP-D, in its intact form or its recombinant heads and neck region. Insights into the mechanism of binding could shed light on exactly which of α2ms surface glycans is the primary target for binding, as well as providing new evidence for this novel role of α2m in the body. The *Limulus* α2m homologue, also has high levels of mannose (31%) in the sugars that make up its surface glycans possibly pointing towards high oligomannose residues; ultimately implying that hSP-D could also bind to *Limulus* α2m (Iwaki, *et al*. 1996).

1.2.1.4.3 α_2-Macroglobulin and its Interactions Cytokines

α2m is known to bind to several cytokines, some of which have their biological activity inhibited whilst bound to α2m, whilst others maintain their biological activity (Rehman, *et al*., 2013). Different cytokines bind to the native and activated forms with varying affinities. Among the cytokines that bind to α2m are: interleukin-6 (IL-6), interleukin -1β (IL-1β), tumour necrosis factor-α (TNF-α),and various growth factors including, transforming growth factor β (TGF-β) and nerve growth factor (NGF) which binds to α2m with highest affinity (Rehman, *et al*., 2013).

Free IL-6 in serum is readily broken down by roaming proteases. When bound to α2m IL-6 is bound in such a manner as its biological activity is preserved whilst offering it protection from the proteases in serum. The role of α2m is key in delivering IL-6 to lymphocytes, hepatocytes and haematopoietic stem cells, thus triggering a host of defence reactions such as: haematopoiesis and the release of acute phase proteins like C-Reactive protein. IL-1β produced in activated macrophages as a protein precursor before being cleaved into its final form by caspase 1. IL-1β

along with TNF-α and IL-6 is a key mediator of the acute phase response another key component of the innate immune system. IL-1β has been shown to bind to the activated form of α2m only and does so in a manner that preserves it's activity, it does so in a Zn^{2+} dependant manner; this is due to the formation in the presence of Zn^{2+} of free sulfhydryl groups that act as cytokine binding sites on α2m itself (Athippozhy, *et al*. 2011; Borth, & Luger, 1989; Rehman, *et al*. 2013). TNF-α produced predominately by monocytes and macrophages is another key cytokine as previously mentioned in triggering the acute phase response as well as, as its name suggests, the ability to inhibit tumourigenesis, induce fever and apoptotic cell death. TNF-α binds with strong affinity to activated-α2m, but has been shown to also bind to the native form, and maintains biological activity regardless of the form of α2m it is bound to (Gourine, *et al*. 2002).It is thought that α2m again acts as a guardian of the cytokine protecting it from proteolytic degradation, something that is seen when TNF-α comes into contact with the proteases released by polymorphonuclear neutrophils The TNF- α/ activated-α2m complex has been shown to be endocytosed by the α2m-receptor LRP-1 and thus a role is suggested for the clearance of and processing of circulating cytokines (Gourine, *et al*. 2002).

One of the key growth factors that binds to α2m is platelet-derived growth factor (PDGF); a potent chemokine for pro-inflammatory cells and those involved in wound repair, making it a major component of the inflammatory response. Human α2m is capable of covalently binding to two molecules of PDGF, and has been hypothesised to reduce the amount of PDGF released at the site of inflammation; suggesting a regulatory effect on inflammation for α2m (Solchaga, *et al*. 2012; Rehman, *et al*.2013). Transforming growth factor β (TGFβ) is a protein that mediates a broad range of cellular processes and as a cytokine is known to have a role in immunity, cancer, and acting as a chemokine for mesenchymal stem cells thus coordinating bone formation. It is activated from its latent form by proteases (Hinze, *et al*., 2012). When bound to α2m the TGF-β molecule reverts back to a latent form. The α2m interaction with TGF-β has been shown to

protect the lens from cataracts, as TGF-β has been shown to induce cataractous changes and that α2m is the principle inhibitor of these effects (Reneker, *et al*. 2010).

Hepcidin a major peptide hormone involved in the role of iron metabolism has been shown to bind to α2m. It is thought that α2m facilitates the role of Hepcidin in regulating transmembrane iron transport by presenting Hepcidin to LRP-1 for endocytosis and passage into the cells (Rehman, *et al*. 2013). Leptin is a hormone released from adipose tissue and regulates appetite and energy expenditure and resultantly, bodyweight. It has recently been discovered that activated α2m is the leptin-binding protein, the rate at which α2m to binds to leptin resulting in its rapid clearance from circulation is thought to seriously affect it's bioavailability and thus indicates yet another key regulatory role for α2m (Birkenmeier, *et al*., 1998; Rehman, *et al*. 2013).

1.2.1.4.4 α2-Macroglobulin and its role in Alzheimer's Disease and Other Pathologies.

As such a key mediator of the immune system it is no great surprise that α2m is involved in a number of disease pathologies.

Alzheimer's disease (AD) is a crippling neurodegenerative disease that affects individuals as are as 49 through to their 80's. Its symptoms start from short-term memory loss and as the disease progresses can lead to confusion, mood swings, and ultimately long-term memory loss. It is thought to be caused by a number of both environmental and genetic risk factors, such as metal consumption and the apolipoprotein E (APOE) allele ε4 (House, *et al*. 2004; Selkoe, 2001). It should be noted, as mentioned APOE is a common ligand for the α2m receptor, and has been linked with late onset AD (Kang, *et al*. 1997). A key neuronal transmembrane protein called amyloid precursor protein (APP) undergoes proteolysis in AD leading to smaller fragments. Two such smaller fragments are beta amyloid Aβ and tau amyloid, which become misfolded and Aβ

production leads the formation of neuronal plaques or fibrils of these misfolded proteins. The proteases responsible for the splicing of the APP molecule are α-,β- and γ- secretase, and the formation of Aβ is due to repeat actions of both β- and γ-secretase producing Aβ isoforms of varying lengths the most common being Aβ40 and Aβ42, which are 40 and 42 amino acids in length respectively. During the formation of the Aβ plaques reactive oxygen species are produced which cause the neurotoxicity responsible for the neurological symptoms seen in AD (Selkoe, 2001). It has been shown that α2m binds to Aβ with high affinity, and it has been shown that activated α2m/Aβ complexes can be internalised via the α2m receptor (Narita, *et al*. 1997). α2m itself though has been shown to be contribute towards neurotoxicity and neuroprotective effects on cultured neurones. The formation of Aβ fibrils and the neurotoxicity associated with Aβ aggregation are both prevented by the binding of α2m to Aβ (Du, *et al*. 1998; Hughes, *et al*. 1998). The post-internalisation processes remain as yet unclear, but it is known that Aβ internalised by α2m does not follow the conventional lysosomal degradation pathway as it has been shown that internalised Aβ via α2m in microglial cells is subsequently released intact from the lysosomes (Chung, *et al*. 1999). In cells not expressing LRP-1, such as neuroblastoma cells, it has been shown that the presence of activated α2m increased the Aβ toxicity in those cells, due to the ability of α2m to bind to TGF-β, returning TGF-β to its latent form and thus interfering with its neuroprotective mechanisms (Fabrizi, *et al*. 1999). Studies utilising knockout mice have shown that an absence of α2m to have no effect on brain function, but increased levels have shown to have either neuroprotective or neurotoxic effects; suggesting there are other factors involved that determine which pathway α2m follows (Kovacs, 2000).

Human group G streptococcus makes up a portion of the typical flora seen on the skin, pharynx and the female genital tract. Despite this it is responsible for a number of severe infections; strain G148 expresses protein G on its surface, an IgG binding-protein which binds via both the Fc and Fab regions, thus inhibiting the binding capability of IgG. It is hypothesised that utilising protein G bacteria coat themselves in IgG thus allowing them to evade the immune system and vastly

increasing its pathogenicity (Sloan, & Hellinga, 1999). α2m has been shown to bind Protein G regardless of its activation state and also binds protein G-related α2-macroglobulin binding protein (GRAB), which is an important virulence factor for group A streptococci (Sloan, & Hellinga, 1999; Rehman, *et al*. 2013). This could be a case of intelligent immune evasion on behalf of the evolution of the streptococci; which is dependent on releasing a mass of protease as well as triggering neutrophilic protease release resulting in increased tissue damage aiding the virulence of the bacteria. The fact that it expresses an α2m binding protein, which can protect the bacteria from host proteases such as neutrophil elastase that would otherwise seek to degrade virulence determinants on the surface of the bacteria (Nyberg, *et al*. 2004).Levels of α2m can vary dependent upon various disease states such as elevate levels in: diabetes, nephritic syndrome and chronic liver diseases (Ritchie, *et al*. 2004). A rare form of α2m known as cardiac isoform α2m (C-α2m), a 182kDa serum protein, has been shown to play a key role in cardiac hypertrophy and is now resultantly used as an early biomarker in myocardial infarctions (Annapoorani, *et al*. 2006). Due to cardiac involvement in 25-40% of HIV cases C-α2m is also used as an early diagnostic marker in HIV patients with cardiac manifestations (Ramasamy, *et al*, 2006; Ramasamy, *et al*. 2010).

To summarise to merely describe α2m as a protease inhibitor does not give the molecule the credit it deserves. Not only is α2m a pan protease inhibitor with a unique binding mechanism, it is also an immunoregulatory protein capable of up or down regulating the immune response and playing key roles in many of today's big diseases.

1.2.2 α2-Macroglobulin of the Horseshoe crab *Limulus Polyphemus*

In a species such a *Limulus* that relies entirely on an innate immune system, proteins such as α2m play key roles in a multitude of systems, acting as protective agents and immunoregulatory proteins (Swarnakar, *et al*. 2000)

1.2.2.1 Synthesis of α2-Macroglobulin in *Limulus Polyphemus*

The human immune system has a plethora of blood cells at its disposal, however *Limulus* has only one – the granular amebocyte (Armstrong, *et al*. 1990); whose principle role it to act as a thrombocyte (akin to the human platelet) forming a cellular plug of adherent cells at the site of injury thus clotting the blood (Bursey, C. R. 1997), as well as the release of exocytotic vesicles. These vesicles contain the structural zymogen protein of the clot as well as an assembly of proteases that cleave the zymogen resulting in the activation of the protein and the formation of thrombogenic fibrils. Another protein exocytosed in the process, is α2m which is thought to control and limit the activity of proteases within the clot, but not those involved in the clotting cascade itself as it shows no activity towards those. Due to the relatively limited amount of study on α2m from *Limulus*, other sites of α2m are not known. It is known however, that the expression of α2m mRNA is seen in several tissues of the Japanese horseshoe crab (*T. tridentatus*) including; heart, hepato-pancreas, stomach, intestine, coxal gland, brain and skeletal muscle (Iwaki, *et al*. 1996). Due to the closely related nature of these two species of horseshoe crab, maybe best demonstrated by the 86% homology in α2m molecules, it could be suggested that α2m is also produced at the same sites in *Limulus* (Iwaki, *et al.* 1996).

1.2.2.2 Structure of *Limulus* α2-Macroglobulin

Limulus α2m is a 370kDa protein made up of a homodimer of 185kDa subunits of which 165kDa is contributed by the peptide sequence of 1482 amino acids (inclusive of a 25 amino acid long signal peptide) and the remainder by the glycosylation sites (Iwaki, *et al*., 1996). The molecular weight of the molecular and it's individual subunits have been verified using SDS-PAGE, sedimentation equilibrium and chromatographic techniques, after original misconceptions about its molecular weight being 480-550kDa leading to the belief it may be trimeric in nature (Armstrong, *et al.* 1991). Two individual subunits are held together with disulphide bonds to form the functional dimer (Husted, *et al*. 2002).

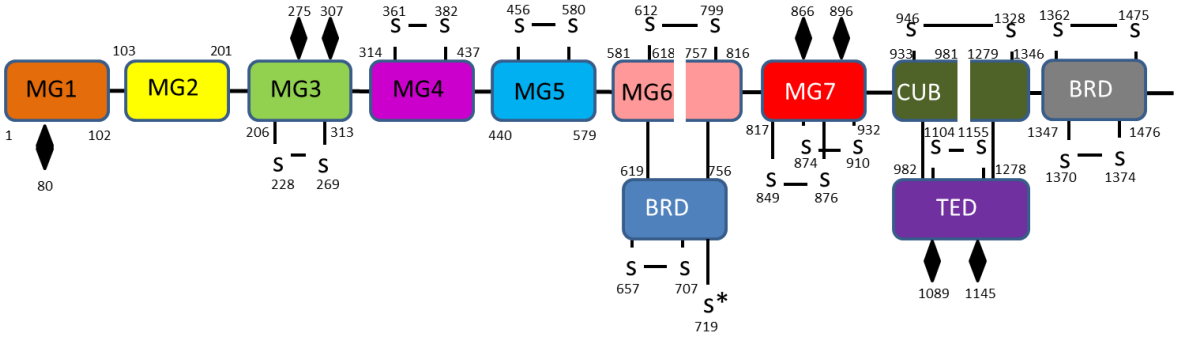


Figure 1.15. Domain schematic representation of *Limulus* α2m, depicting domain boundaries, N-linked glycosylation sites (indicated with a black diamond), and disulphide bond sites with the intersubunit disulphide at Cys^{719} shown with an *.

Intrasubunit disulphide bridges occur between, Cys^{228}-Cys^{269}, Cys^{361}-Cys^{382}, Cys^{456}-Cys^{580}, Cys^{612}-Cys^{799}, Cys^{657}-Cys^{707}, Cys^{849}-Cys^{876}, Cys^{874}-Cys^{910}, Cys^{946}-Cys^{1328}, Cys^{1104}-Cys^{1155}, Cys^{1362}-Cys^{1475}, and Cys^{1370}-Cys^{1374} as depicted in Figure 1.15; in addition to a single intersubunit disulphide bridge between, Cys^{719} from one subunit to the same residue in the other subunit (Husted, *et al*. 2002).Each subunit has seven potential N-linked glycosylation sites; Asn^{80}, Asn^{275}, Asn^{307}, Asn^{866}, Asn^{896}, Asn^{1089}, and Asn^{1145}; with an estimated 31 Man, 19 GlcNAc, 3 GalNAc, 3 Gal and 5Fuc residues per subunit (Husted, *et al*., 2002; Iwaki, *et al*., 1996). The two main characteristics of the α2m molecule, the bait region and the thiol-ester are seen at residues Pro^{697} - Thr^{735} and Cys^{975} - Glx^{978} respectively (Armstrong, & Quigley, 1999). Transmission electron microscopy showed that native α2m was made up of two clearly defined dimers, and that chymotrypsin reacted *Limulus* α2m showed a much more compact structure where each subunit was then indistinguishable (Armstrong, *et al*. 1991). Initially based on sequence homology work it was proposed that *Limulus* α2m had four key functional domains. At the N terminal end the unique region identified as a C8γ domain, due to modest homology with C8γ of the human complement system, the bait region domain, the thiol ester domain and the receptor binding domain at the C terminal end (Armstrong, *et al*. 1999). However, structure prediction work completed in this thesis suggests that *Limulus* α2m actually represents human α2m and its family members with regard to domain organisation as demonstrated in detail in Chapter 5. As of yet there have been no published crystallographic studies on α2m of *Limulus*, but as the human model was finally published in 2012

(Marrero, *et al*., 2012), this may provide suitable starting conditions for the crystallographic structure solution of *Limulus* α2m as NMR cannot be considered for the structure solution of *Limulus* α2m due to its size.

1.2.2.3 Functions of *Limulus* α2-Macroglobulin

Very much like it's human counterpart the principle role of α2m in *Limulus* is to act as a pan-protease inhibitor; in fact it may be even more vital a role in *Limulus* as it is the only known circulating protease inhibitor (Armstrong, & Quigley, 1999; Quigley, & Armstrong, 1983). Whereas the human molecule is thought of as a dimer of dimers with one dimer being one active unit, the *Limulus* homologue is a single dimer and thus a single active unit but in its protease binding activity is very similar. The bait region of *Limulus* α2m is 38 amino acids long and has 18% homology with its human counterpart.

```
human           PQLQQYEMHGPEGLRVGFYESDVMGRGHARLVHVEEPHT
limulus         PQYDVAFAAPQAANRIGG-GGEAGGFGGGIRKKTNKPVV
                ** :         . *:*   .:. * * .   :.::* .
```

Figure 1.16. ClustalW, alignment of the bait regions of both human and Limulus α2m highlighting the homologous amino acids. Identical residues are indicated with an *, residues with strong similarities are shown with a :, and weakly similar residues are shown with a . (Larkin, et al., 2007)

Cleavage of the bait region results in the same compaction of the molecule as it closes around it's protease prey, this has been shown using electron microscopy and with polyacrylamide gel electrophoresis; which also showed that methylamine has the same reaction in *Limulus* as it does in humans, producing a similar gel migration distance to that seen with protease reaction, and is also thought to react with the thiol-ester (Armstrong, & Quigley, 1987; Armstrong, *et al*., 1991; Quigley, & Armstrong, 1985; Quigley, *et al*., 1991). Differences in the bait region, illustrated by Figure 1.16, infer that the human and *Limulus* molecules have different target proteases, which is likely as they are two very different species, likely to come into contact with very different

pathogenic invaders. Other similarities between human and *Limulus* α2m is that bound proteases retain their catalytic abilities as shown in trypsin digest assays where internalised trypsin was still able to cleave Na-benzoyl-DL-arginine p-nitroanilide (BA*p*NA) (Armstrong, 2010). Following reaction with a protease *Limulus* α2m also internalises the protease for processing; this was shown using fluorescein-labelled proteins injected into the lumen of the heart and the clearance time was measured. Labelled trypsin complexed with α2m was half cleared in a time of around 10-15 minutes and maximally cleared within 20-25 minutes, and this is thought to take place on the *Limulus* blood cell (Melchior, *et al*., 1995; Armstrong, 2010).Currently a receptor for protease reacted α2m, in invertebrates has yet to be found, but the LRP-1 family is an evolutionarily ancient as the α2m superfamily and family members of LRP-1 have been found (Armstrong, 2010), but whether these act as a receptor for α2m in invertebrates remains to be seen. The *Limulus* blood cell does however contain a very high molecular mass protein that like LRP-1 binds to LRP receptor-associated protein (RAP) which is known to bind to the human α2m (Aimes, *et al*. 1995). This however does need to be the focus of future research as current evidence remains inconclusive.

Another family of proteins *Limulus* shares, with humans is the pentraxins; homologues are found for both C-reactive protein (CRP) and serum amyloid P component (SAP), but limulus has an additional member of the pentraxins, the sialic acid binding protein limulin (Armstrong, & Quigley, 1999). Limulin is the sole cytolytic protein of the *Limulus* immune repertoire (Harrington*, et al*., 2008, Armstrong, *et al*., 1996) which, taking place in a Ca2+ characteristic of its family, leads to membrane permeabilisation. This takes place when limulin, which is present in *Limulus* serum at concentrations of 30-50nM, inserts itself into the plasma membrane, via sialic acid binding, forming a hydrophilic pore (1.7nm). This pore too small for cytosolic proteins to pass through allows the passage of water into the cell, causing osmotic swelling leading to cell lysis (Swarnakar, *et al*., 2000). α2m regulates the limulin-based cytosolic system seen in *Limulus*; but only the activated form of α2m has any effect as native has been shown to have no effect (Swarnakar, *et*

al., 1995). The binding of activated α2m to limulin results in the negation of its cytolytic activity. This could form part of the pathogenic invasion strategy, to release proteases, activating α2m and thus disabling the cytosolic system.

1.2.3 Conclusions, Aims and Objectives

The emergence of α2m and its superfamily over 700 million years ago is a key moment in the evolutionary timeline (Sahu, & Lambris, 2001). This family of proteins that are present in a diverse range of species from humans and other mammals to fish, reptiles, nematodes, insects, arthropods and other invertebrates, are vital immune mediators (Armstrong, 2010). α2m superfamily members in humans make up part of the repertoire of the humoral component of the innate immune system as well providing links to the acquired/adaptive immune system (Janssen, *et al*. 2005; Kidmose, *et al.* 2012; Fredslund, *et al*. 2008). α2m itself is a multifaceted, multifunction protein capable of mediating the immune systems response to a variety of attacks as well as modulating it's functions;, in its ability to protect immune molecules and deliver cytokines to the location of pathogenic invasion (Rehman, *et al*. 2013). The solution of the crystal structure of human α2m has generated a great many clues about its molecular mechanisms as well as comparisons to its fellow family members in humans (Marrero, *et al*. 2012). Already used as a molecular marker (cardiac isoform), the uses of α2m as a therapeutic agent are set to develop further as we understand it more, potentially reverse engineering it so that it's bait region contains a cleavage site for a protease that we may wish to negate the effects of (Rehman, *et al*. 2013). Such information has yet to be revealed about the homologue from *Limulus polyphemus*, the American horseshoe crab, where the α2m molecule also forms an additional role by helping mediate the cytosolic system within *Limulus* with the pentraxin protein limulin which serves as the MAC-esque homologue forming a pore in cell membranes leading to cell lysis (Swarnakar, *et al*. 2000). As an invertebrate *Limulus* lacks any acquired immune system and as such relies heavily on its innate immune system, of which there are many human homologues, to

protect it from infection. α2m is one such innate immune system and understanding its structural and functional mechanisms may lead to a better understanding of the human molecule as well as how the innate immune system evolved.

Crystal structure data now exists for several α2m family members in humans, α2m (Marrero, *et al*. 2012),and C3, C4 and C5 of the complement system (Janssen, *et al*. 2005; Kidmose, *et al*. 2012; Fredslund, *et al*. 2008). Yet there not been any published crystallographic studies of the α2m homologue from *Limulus*. The aim of this thesis is to fill the existing gaps in our knowledge about this superfamily, building on the works of Prof. Peter Armstrong, Prof. James Quigley, Assoc. Prof. Lars Sottrup-Jensen and others, using x-ray crystallography and structure prediction software to reveal clues about the structure, mechanisms and function of α2m from *Limulus*. Focussing on predicted domain homologies, particularly in the highly conserved TED region of the protein, as well as domain arrangement and quaternary structure.

Chapter 2 – Isolation and Purification of α2-Macroglobulin from the serum of the Horseshoe crab, *Limulus polyphemus*

2.1 – Introduction to the Isolation of alpha-2 macroglobulin of the Horseshoe crab, *Limulus polyphemus*.

α_2-Macroglobulin (α_2m) was isolated from serum of *Limulus polyphemus* using a combination of the polyethylene glycol (PEG) cut procedure, affinity chromatography for the phosphoethanolamine (PE) binding proteins, and size exclusion chromatography for the removal of hameocyanin. The PEG cut procedure is used to facilitate the removal of large quantities of the haemocyanin present in serum as well as increasing the concentration of the remaining serum proteins. Both of the pentraxin-like proteins bind to PE on the affinity column whereas only C-Reactive Protein (CRP) will bind to phosphocholine (PC). Using PC in conjunction with PE, and manipulating the binding affinity of the two pentraxin-like proteins allowed not only for the separation of the pentraxins from the serum but also separate them individually from one another. The pentraxin depleted serum was then run down a size-exclusion chromatography column on our in house fast protein liquid chromatography (FPLC) system. This separates the remaining proteins (haemocyanin and α_2m) by their molecular weight into fractions. These fractions were then concentrated and used in gel electrophoresis, to monitor the efficiency of the protocol and to monitor the purification process. For ease of visualisation a schematic of the purification process has been produced in Figure 2.1. Purified protein was then used in crystallisation trials. Various crystal conditions were trialled utilising Molecular Dimensions Structure Screens as well as existing conditions from the literature. α_2m was used in crystal trials with methylamine acting as a ligand by cleaving the thiol-ester bond.

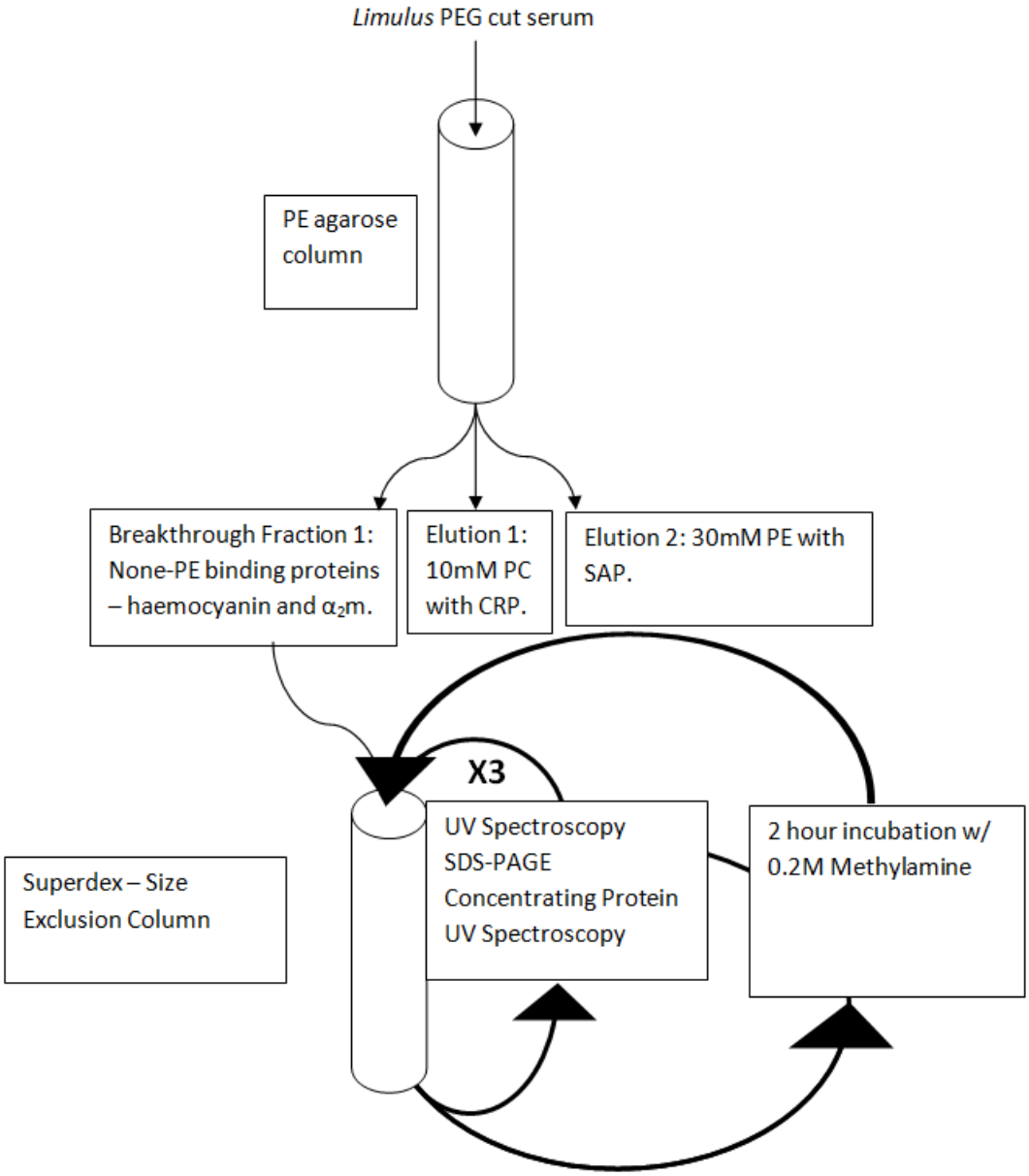


Figure 2.1. Schematic representation of the purification protocol for purifying *Limulus* α2m from PEG cut serum. Following removal of pentraxins via affinity chromatography on a PE agarose column, the breakthrough fraction which is now pentraxins depleted undergoes three rounds of size exclusion chromatography using a Superdex column to remove the remaining amounts of haemocyanin. SDS-PAGE and UV spectroscopy were used between size exclusion chromatography experiments to monitor and assess progress. After three size exclusion runs, the haemocyanin was removed and the isolated *Limulus* α2m was incubated with 0.2M methylamine/10mgml^{-1} α2m for a two hour period before the fast and slow forms are then separated on the size exclusion column to ensure homogeneity of the sample for crystallisation trials.

2.1.1 Obtaining *Limulus polyphemus* plasma and the PEG cut procedure.

The methods discussed here (Armstrong, & Conrad, 2008) were not carried out by myself, however they're included for completeness in order to understand the history of the serum and the proteins used.

Polyethylene glycol (PEG)-cut serum from *Limulus polyphemus* was generously provided by Professor Peter Armstrong, of the Marine Resources Centre, Marine Biological Laboratory, Woods Hole, Massachusetts. Prior to bleeding the animals are chilled in a cold room (4°C) for one hour to minimise the risk of blood clotting. Typically a single horseshoe crab can yield 50-200ml of haemolymph. This is extracted by penetrating the heart, via the hinge, with a sterile 14 gauge needle to a 1-2cm depth (Armstrong, & Conrad, 2008).Once the blood is collected it is immediately centrifuged at 1000rpm for 5 minutes, to pellet the granular amebocytes, that contain the clotting agents.

In *Limulus*, the oxygen transporting molecule is haemocyanin, a molecule similar to our own haemoglobin except with copper as its oxygen binding atom; and is the source of the serums blue colour. Another key difference from haemoglobin is that it is a free serum protein and not bound by the blood cells. As a result in order to study other *Limulus* serum proteins the majority of the haemocyanin needs to be removed. This was done via the PEG-cut procedure. The PEG-cut procedure utilises principles of molecular crowding to selectively precipitate out proteins – in this case haemocyanin. PEG in this case and in a similar way to it behaviour in crystallographic experimental setups interacts with the proteins via an excluded volume effect, whereby the presence of PEG minimises the available volume for the protein to reside (Herzfeld, 1996; Atha, & Ingham, 1981). The result is that the protein concentration following the addition of the PEG is increased, increased well beyond the point of supersaturation. As this achievement of supersaturation isn't gradual as seen in the vapour diffusion experiments of crystallographers the result is the precipitation of the protein. Larger proteins and proteins with already high

concentrations are those to precipitate first (Herzfeld, 1996; Atha, & Ingham, 1981). Given that haemocyanin is the most abundant serum protein and also the largest, it is the protein that is precipitated out in the greatest quantity. To 1.5 litres of serum pheylmethylsulfonylfluoride (PMSF) – a serine protease inhibitor – was added to a concentration of 1mM, as well as PEG 8000 to a concentration of 3%. This preparation was then incubated for 3-6 hours at 4°C before being centrifuged at 40,000g for 30 minutes, to form a pellet containing PEG and haemocyanin. PMSF and PEG were added to the supernatant again to a final concentration of 1mM and 10% respectively. The supernatant now containing 1mM PMSF and 10% PEG 8000 was incubated at 4°C for approximately 1 hour and then centrifuged at 40,000g for 10 minutes to form another pellet. This pellet contains PEG, the pentraxins and the protease inhibitor α_2-macroglobulin, and was redissolved in calcium free buffer (50mM Tris pH 7.5, 150mM NaCl and 0.2mg/ml NaN_3) with 1mM PMSF and left overnight at 4°C. Calcium Chloride was then added to a final concentration of 5mM. The now PEG-cut serum has had the majority of the haemocyanin removed but a lot of it still contaminates the serum and so PEG-cut serum retains its blue colour.

2.1.2 Introduction to Affinity Chromatography

Affinity chromatography, a technique pioneered in 1968 (Cuatrecases, *et al*., 1968), is a method of macromolecular separation based upon a molecules binding affinity to an immobilised media bound ligand. Once the target molecule is bound it allows for non-ligand binding contaminants to be washed through – and if necessary collected. The target molecule is then eluted from the column with either a competitive ligand or an agent capable of disturbing protein-ligand interactions, in elution buffers (Catsimpoolas, in Heftmann, 1983). Using known ligands of the protein is a specific way to elute the protein from the media. Whereas using analogues of the ligand or new ligands to compete with the media can be especially helpful if multiple proteins are bound to the media, but have different binding affinities for the elution ligands (Villems, &

Toomik, in Kline, 1993). Using none specific agents such as detergents, chelating agents, changes in pH or salt concentration etc. are a good method of purging the media of all the bound proteins – helpful if only one protein has bound to the media (Cuatrecases, *et al*., 1968).

Affinity chromatography in this instance was not used to isolate the protein of choice but two isolate two contaminants – the pentraxins, CRP and SAP. The PEG-cut serum provided by Peter Armstrong contains the serum proteins in high concentrations, the most abundant proteins being; haemocyanin, c-reactive protein, serum amyloid P- component and α_2-macroglobulin (Armstrong, & Quigley, 1999). The use of affinity chromatography utilises the binding affinity of the pentraxins to phosphocholine (PC) and phosphoethanolamine (PE), and the none binding of the haemocyanin and the α_2-macroglobulin to these molecules. Like their human counterparts, *Limulus* CRP and SAP differ in their binding affinities allowing for their separation via specific ligand elution. CRP will bind to both PC and PE (Volanakis, & Kaplan, 1971; Schwalbe, *et al.* 1992; Robey, & Liu, 1981). Conversely SAP will bind only PE (Shrive, *et al*. 1999; Tharia, *et al.* 2002). Importantly both molecules bind in a calcium dependant manner, meaning that they will only bind PE and PC in the presence of calcium ions acting as a cofactor. As a result calcium ions must be present in both the wash buffer and the elution buffers to ensure protein binding to the column and to ensure successful elution from the column when the ligands are added respectively. As both pentraxins bind to PE, they can be separated from PEG-cut serum using PE linked beaded agarose which is packed into a 25ml column. Buffer is pumped over this column at a low flow rate of 0.5ml/min, which prevents the agarose from compacting within the column which could lead to reduced flow rate and a reduction in surface area for pentraxins to bind to the PE. Due to the high stability and inert surface of agarose (Roe, in Harris, & Angal, 1989), the pentraxins do not interact with it preferentially over PE.

The pentraxins were eluted using an isocratic elution of single ligand buffer. An isocratic elution being one that does not change in composition during the elution, in this case 10mM PC and

30mM PE. As the serum was ran onto the PE agarose column the none PE-binding proteins – haemocyanin and α_2-macroglobulin were collected in the breakthrough fraction. These were then concentrated and subjected to three runs of gel-filtration chromatography prior to α_2-macroglobulin being used in crystal trials.

2.1.2.1 Materials and Procedures

Materials, Equipment and Chemicals:

- Acetate filter - 0.2µm pore size – (Sartorius)
- Biologic Low Pressure System and HP controller (BIORAD)
- O-Phosphoethanolamine agarose beads (Sigma)
- Calcium Chloride dehydrate (Sigma)
- Ethylenediaminetetraacetice acid (EDTA) (Sigma)
- Phosphocholine chloride calcium salt tetrahydrate (Sigma)
- O-Phosphoethanolamine (Sigma)
- Sodium Azide (Sigma)
- Sodium Chloride (VWR)
- Tris(hydroxymethyl)methylamine (VWR)

Buffers:

- Calcium wash buffer: 50mM Tris pH 7.4, 10mM $CaCl_2$, 150mM NaCl.
- Phosphocholine elution buffer: 50mM Tris pH 7.4, 10mM $CaCl_2$, 150mM NaCl, 10mM Phosphocholine chloride.
- Phosphoethanolamine elution buffer: 50mM Tris pH 7.4, 10mM CaCl2, 150mM NaCl, 30mM O-Phosphoethanolamine.
- EDTA buffer: 50mM Tris pH 7.4, 150mM NaCl, 10mM EDTA.
- Regeneration buffer A: 200mM Tris pH 7.4, 500mM NaCl, 10mM EDTA.

- Regeneration buffer B: 50mM Tris pH 7.4, 500mM NaCl, 10mM $CaCl_2$.
- Regeneration buffer C: 50mM Tris pH 7.4, 150mM NaCl, 10mM $CaCl_2$, 0.02% NaN_3.

2.1.3 Introduction to Size Exclusion Chromatography

Once the pentraxins were removed, size-exclusion chromatography (SEC) also known as gel filtration, was used to separate the remaining haemocyanin and α_2-macroglobulin; following the removal of the pentraxins via affinity chromatography as mentioned above. This was performed on an AKTA Explorer FPLC system using a Superdex HiLoad 16/60 column.

SEC separates molecules according to differences in their size, as they pass through a special medium filled column. Unlike other forms of chromatography, in SEC molecules do not bind to the medium meaning that buffer composition has a negligible effect on peak resolution. The medium is made up of specially selected spherical particles, chosen for their inertness and stability, which form a porous matrix which allows molecules below a certain size to be adsorbed. The interparticular pores are filled with a buffer which is often referred to as the stationary phase, and the buffer that passes around the outside of the particles being the mobile phase. As a sample is applied to the column, it moves with the mobile phase in and out of the pores of the matrix. The larger molecules are restricted by their size as to how deep into the matrix they can penetrate; as a result they do not interact with the column for as long as smaller molecules that is capable of moving deeper through the media. The detection of protein molecules is done using a 280nm wavelength lamp and analysis of the absorbance spectra of the solution as it exits the column. As a result when the fractions are collected at the end of a sample run, molecules come off the column in order of size, with the larger molecules coming off first having not interacted with the column for as long a period as the smaller molecules that come off later. This entire separation process takes place in just one column volume allowing a quick turnaround time if required.

SEC affords a great deal of flexibility, with regard to the environmental conditions the target molecules are subjected to. It can be performed at either room temperature or in a cold room, molecules can be purified in any chosen buffer, the separation can be performed in the presence of a variety of molecules such as essential ions, cofactors etc.

2.1.3.1 Materials and Procedures

Materials, Equipment and Chemicals:

- Acetate filter - 0.2µm pore size – (Sartorius)
- AKTA Explorer FPLC System (GE Healthcare)
- Superdex 200 HiLoad 16/60 Column (GE Healthcare)
- Calcium Chloride dehydrate (Sigma)
- Ethanol (VWR)
- Sodium Azide (Sigma)
- Sodium Chloride (VWR)
- Tris(hydroxymethyl)methylamine (VWR)

Buffers:

- Calcium wash buffer: 50mM Tris pH 7.4, 10mM $CaCl_2$, 150mM NaCl.

Procedure:

Effluent from the above mentioned affinity chromatography procedure was concentrated and injected onto the AKTA Explorer FPLC system, via the injection valve, typically at volumes of approximately 500µl and within a concentration range of 3-30 mg/ml. The system was brought out of storage by purging the tubes with deionised water before the column is washed with filtered and degassed deionised water prior to the equilibration step. During equilibration the

filtered and degassed Calcium wash buffer was ran through the column and the entire system for several column volumes to ensure uniformity within the system as well as within the Superdex media. The sample was then added to the system as described above and the system ran at a flow variety of flow rates until a flow rate of 0.2 ml/min was settled upon for optimum resolution of the chromatogram. As the void volume of the column is approximately 40 ml, sample collection was not started until just under that volume had passed through the column. At that point using the attached auto-sampler the effluent was collected in 24 3.5ml fractions A and B 1-12. Samples were then labelled and stored at 4°C for further analysis.

2.1.4 Concentrating Proteins

The effluents from both the affinity and size-exclusion chromatography experiment can be very large volumes (as much as 50 ml) as a result these solutions often need to be reduced in volume whilst minimising any losses in protein to provide a more appropriate working concentration and volume. This was done to allow for following purification steps, analysis by gel electrophoresis and ultimately crystal trials. This is done using molecular weight cut-off centrifugal filter devices. The use the centrifugal force to push the solution through a filter membrane of a stated porousness designed to withhold molecule over a stated molecular weight – 10kDa Amicron Ultra 15ml and 4ml filters. Samples were spun at a speed of typically 3000g and 4°C until the desired volume was achieved. The filtrate should be devoid of protein whilst the retentate should be of a much higher concentration than first measured.

In order to assess the effectiveness of this as well as to quantify the amounts of protein being produced in the lab I needed to measure the concentration of the effluents. This is done using the UV absorbance spectrophotometry, measuring the absorbance unit for full scale deflection (AUFS) of a solution at both 280nm a 320nm wavelengths of light. Absorbance at 280nm is provided

primarily by the aromatic rings of both Tryptophan and Tyrosine, as well as some absorbance from disulphide bonds. Absorbance of 320nm is used as a measure of background absorbance as peptide chains do not absorb at this wavelength. Using the online program Protparam from Expasy (http://web.expasy.org/protparam/) I can give the sequence of my target protein and determine the Molecular Extinction Coefficient (MEC) which is based on the absorbance of 1M solutions of both Tryptophan and Tyrosine as well as their abundance in the sequence of the target protein, this coefficient is equal to the absorbance of a 1M solution of the protein. When the MEC is divided by the molecular weight of the protein we are provided with the absorbance equal to 1mg/ml. Following that the AUF_{280nm} of a sample is then subtracted by the AUF_{320nm} to obtain the absorbance contributed solely by the protein before dividing by the MEC to give the mg/ml concentration for the sample.

2.1.5 Introduction to Gel Electrophoresis

Gel electrophoresis is a lab technique that allows for the separation of mixtures of DNA, RNA or proteins by their electrophoretic mobility - a function of their size to charge ratio. Upon application of the electric current the biological molecules migrate through pores of variable size toward the positive electrode, and they migrate through at a rate determined by their electrophoretic mobility. The variable pore size allows for adjustments in migration rates, smaller pores will only allow smaller molecules to travel through with ease, larger pore sizes allow smaller molecules to travel much more quickly as well as enabling the mobility of larger molecules that would be limited by smaller pore sizes.

Two types of gel medium are available for such separations; agarose and acrylamide. Agarose is a linear polysaccharide of galactose - 3, 6-anhydrogalactose repeats cross-linked with hydrogen bonds and is primarily used in gel electrophoresis of nucleic acids due to its relatively larger pore

size at low percentages (Gaal, *et al*., 1980; Andrews, 1986). Acrylamide gels on the other hand are capable of producing much smaller pores making them far more suitable gels for protein separation. Agarose gels simply set and become solid as the solution cools; acrylamide on the other hand needs to be polymerised in order to set. Acrylamide is polymerised initially to polyacrylamide in the presence of N, N' -methylenebisacrylamide (bis-acrylamide) which acts as a cross-linking agent (Shi, & Jackowski, in Hames, 1998). To further polymerise the acrylamide ammonium sulphate is decomposed by N, N, N', N'-tetramethylene diamine (TEMED), the result is the production of free radicals that promote the polymerisation of acrylamide monomers with one another as well as bis-acrylamide. Acrylamide composition is defined using two terms C% and T%. C% is the percentage of bisacrylamide (cross-linker) relative to the total, whereas T% is defined as the total percentage concentration of both acrylamide and bisacrylamide. These two values can be adjusted to vary the resolving power of the gel in addition to the use of correct buffers and optimum pH values (Shi, & Jackowski, in Hames, 1998).

As mentioned above, electrophoretic mobility is a function of both size and charge, in order to separate proteins by size alone a method must be employed to ensure charge uniformity between proteins. This is done by utilising sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE), where the proteins are pre-treated with SDS; an anionic detergent that coats the length of the protein ensuring charge uniformity, of negative charge, along the chain as well as denaturing the protein to a linear form. In addition to SDS a second denaturant is added – β-mercaptoethanol, which reduces disulphide bonds that holds both tertiary and some quaternary structures together. The result of these treatments is linear polypeptide chains with a net negative charge; this allows them to be separated by size only as their charge is uniform and prevents the three-dimensional packing of the molecule, as some molecules are packed tighter than others, from misleading researchers about their size. When composing an SDS-gel the gel can be formed in two 'layers', the resolving and stacking gels in a system known as a discontinuous system. Firstly the resolving gel is poured, it is typically a higher concentration than

the stacking gel and thus separates the proteins by their molecular weight, then the stacking gel is poured, the stacking gel is of a low concentration and this allows the proteins to concentrate into a single band prior to entering the resolving gel resulting in greater resolution (Shi, & Jackowski, in Hames, 1998). Continuous buffer systems are rarely used in protein separations as they often yield fuzzy and poorly resolved bands; they are commonly used for the separation of nucleic acids. Sample buffer is then added to the proteins, the buffer contains the optimum resolution conditions for the gel in question as well as a dye to help visualise the electrophoresis front, in order to prevent the over-running of the gel and resultantly having proteins of interest migrate out of the gel. This sample is then heated to further denature the proteins to ensure maximum migration through the gel. If it is of interest to separate proteins by size and charge then native PAGE gels can be used. In native gels proteins are prepared in a non-denaturing and non-reducing sample buffer, as well as the gel being prepared without SDS. The lack of denaturants and reducing agents results in the maintaining of the secondary structure of the protein as well as it native charge density (Shi, & Jackowski, in Hames, 1998). Once both the gel and the samples have been prepared the samples are then loaded into the wells of the gel and an electric current is applied. The chloride ions already present in the gel migrate from the anode towards the cathode and do so faster than the proteins do resulting in an ion front. The glycinate ions, provided by the Tris-glycine running buffer, form an ion front behind the proteins. The leading ion (Cl^-) moves through the gel ahead of the proteins, the trailing ions (glycinate) follow until they overtake the proteins and establish a linear voltage gradient that the proteins then sort themselves along in accordance with their size and charge. In SDS-PAGE gels the principle is the same, where the proteins are stacked between the leading and trailing ion fronts. When these fronts meet the interface between the stacking and resolving gels the % acrylamide increases greatly, restricting the migration of the proteins. The proteins are then only separated by size as their charge-to-mass ratio is the same following SDS treatment and the result is proteins separated into band

patterns according to their molecular weights (Shi, & Jackowski, in Hames, 1998). In order to visualise the proteins they must be stained post-separation.

2.1.5.1 Materials and Methods

Materials, equipment and chemicals for gels:

- Sodium dodecyl sulphate (Sigma)
- N, N, N', N'-tetramethylene diamine (TEMED) (Sigma)
- Ammonium persulphate (VWR)
- Tris(hydroxymethyl)methylamine (VWR)
- Acrylamide and bisacryalmide (Sigma).

Loading buffer:

- Glycine (Fischer)
- Glycerol (Sigma)
- β-mercaptoethanol (Sigma)
- Bromophenol blue (VWR)
- Molecular weight markers (Biorad)

Running buffer:

- Tris(hydroxymethyl)methylamine (Fischer)
- Sodium dodecyl sulphate (Sigma)
- Glycine (Fischer)

Procedure:

SDS-PAGE was used to analyse the success of various rounds of separation of proteins from the plasma of *Limulus polyphemus*. SDS-PAGE took place following every chromatographic step of purification to guide the process. Typical gels were made up a 12.5% acrylamide resolving gel with a 4% acrylamide stacking gel at a thickness of 1.5mm as per Table 2.1 and in reducing samples only β-mercaptoethanol was added to 6%. The samples prior to application to the gel were treated with the loading buffer also shown in Table 2.1 and then heated at 95°C for two minutes to further denature them. Gels were then loaded into their cassettes and electrophoresis tanks, running buffer was added, samples were loaded and the gels were ran at a voltage of 200V for 40-50 minutes.

Table 2.1. Table showing the components used to make and run two 1.5mm SDS-PAGE gels

SDS-PAGE Components	
12.5% Resolving gel	6.66ml acrylamide/bisacrylamide (30%T 2.6%C)
	5.1ml deionised water
	4ml 1.5M Tris HCl pH8.8
	160µl 10% SDS
	10µl TEMED
	100µl 10% Ammonium persulphate
4% Stacking gel	1.33ml acrylamide/bisacrylamide (30%T 2.6%C)
	6ml deionised water
	2.5ml 0.5M Tris HCl pH 6.8
	100µl 10% SDS
	10µl TEMED
	100µ 10% Ammonium persulphate
Sample Buffer	125mM Tris-HCl pH6.8
	4% SDS
	0.02% bromophenol blue
	(reducing samples contained 5% β-mercaptoethanol)
Running Buffer	3.13g Tris
	14.42g Glycine
	1g SDS

Materials for staining:

- Instant Blue (Coomassie based from Expedeon)

Once the gel had been running for 40-50 minutes at a voltage of 200V, it was removed from the tank and left in Instant Blue a Coomassie based stain on a rocker table for approximately an hour.

2.2 Isolation and purification of α2-Macroglobulin from the serum of *Limulus Polyphemus*

As mentioned previously, PEG Cut serum was provided by Prof. Peter Armstrong of the Marine Resources Centre, Marine Biological Laboratory, Woods Hole, Massachusetts. This serum was then applied to the phosphoethanolamine linked agarose affinity chromatography column in order to remove the pentraxins from the serum. Prior to the initial chromatography run PEG-cut serum was ran on an SDS-PAGE gel to assess its contents shown in Figure 2.2.

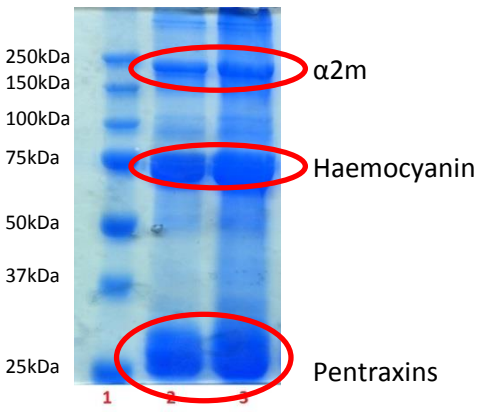


Figure 2.2. SDS PAGE analysis of PEG-cut serum supplied by Professor Peter Armstrong. The gel was a 7.5% SDS PAGE gel with all samples ran under reducing conditions. Lane 1 contains molecular weight size markers – Biorad All Blue Precision Markers from top to bottom – 250kDa, 150kDa, 100kDa, 75kDa, 50kDa, 37kDa and 25kDa, smaller markers (20kDa, 15kDa and 10kDa) migrated out of the gel. Lanes 2 and 3 contains PEG cut serum at a protein content of 28µg and 56µg respectively.

Protein content for the gel was extremely high. This was done to ensure that even proteins present in the smallest quantity would be visualised. It is clear from the gel that three major bands are present in both lanes. The band at the top lies between the markers for 250kDa and 150kDa, the middle band runs equivalent to the 75kDa band and broad band at the bottom of the gel lies just above the 25kDa marker band in lane 1. It is known that the three most abundant proteins in the serum of *Limulus Polyphemus* are in order: haemocyanin, the pentraxins and α_2-macroglobulin (Armstrong, & Quigley, 1999). It is known that haemocyanin is an enormous molecule capable of forming hexamers or oligiohexamers of subunits of around 75kDa (Martin, *et al*., 2007). Under SDS-PAGE reducing conditions the dissociation of these oligomers will yield various strands of approximately 75kDa as seen in the middle band produced in lanes 2 and 3 of the gel shown in Figure 2.2. *Limulus* pentraxins of which there are CRP and SAP homologues, can be found as stacked octomeric and heptameric rings of repeating 25kDa subunits (Shrive, *et al*. 2009). Based on this information it was concluded that the band that travelled furthest through the gel and was in line with the 25kDa marker was representative of the pentraxins which could not be distinguished from one another on the gel. α2-Macroglobulin has been shown to have a molecular weight of 185kDa per subunit in a homodimer arrangement (Iwaki, *et al*. 1996). Being the largest subunit of these three most abundant proteins, the band that travelled the least and fell between markers for 150kDa and 250kDa was deemed to be the α2m band. Following the assessment of the PEG cut serum to ascertain the presence of α2m PEG cut serum was then passed through the affinity chromatography column.

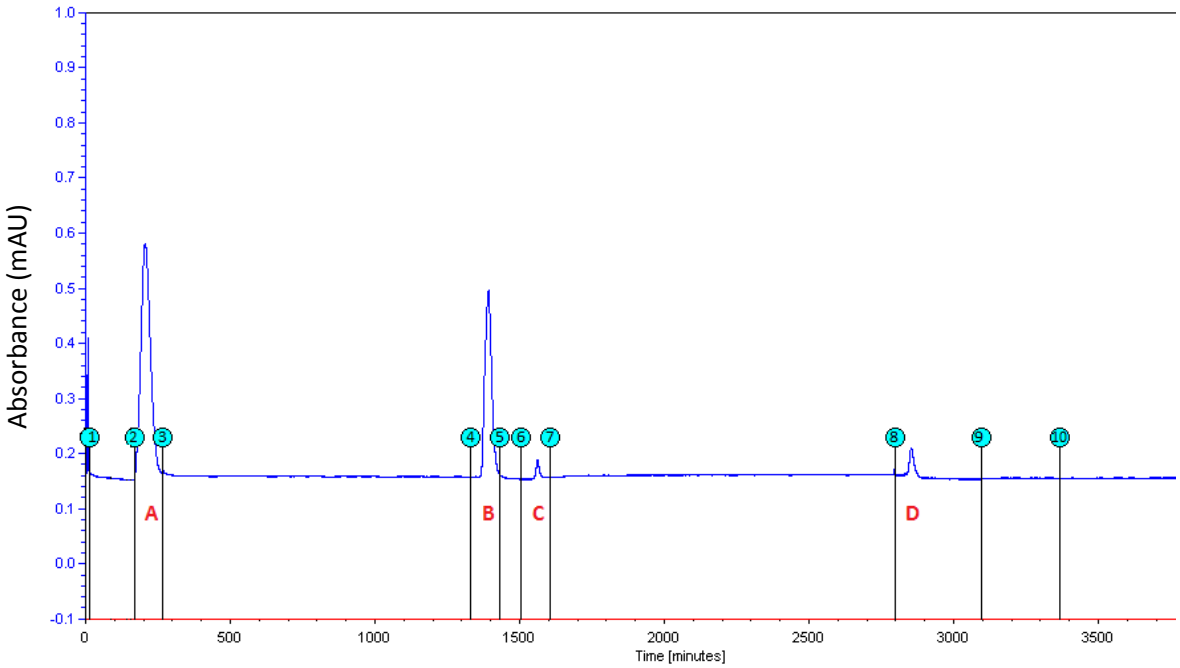


Figure 2.3. Chromatogram from affinity chromatography run to remove the pentraxins from PEG cut serum. The markers shown on the chromatogram are as follows: 1 – protein loaded onto the column, 2 – breakthrough fraction collection started, 3 – breakthrough fraction collection finished, 4 – PC applied to the column and collection started, 5 – PC collection finished and calcium buffer reapplied, 6 –PE applied to the column and collection started, 7 – PE elution collection finished and calcium buffer reapplied, 8 – EDTA buffer applied and effluent collection started, 9 – EDTA collection finished, 10 – column regeneration started.

Figure 2.3 shows the chromatogram from an affinity chromatography run of PEG-cut serum as described previously. Peak A represents the breakthrough peak of proteins that do not interact with the phosphoethanolamine-linked agarose column. In this case the vast majority of the absorbance of this peak is contributed by haemocyanin and α2m. Peak B is the phosphocholine elution peak, containing previously column bound proteins that bind to phosphocholine preferentially. In this case Peak B represents the CRP that was previously bound to the column. Peak C is the phosphoethanolamine peak where proteins that were column that bind to PE are eluted. Peak C represents the SAP peak where elution with PE 'released' it from the column. Peak D is the EDTA peak which removes any remaining column bound proteins. The breakthrough fraction was then concentrated using a centrifugal filter device prior to its application to the size exclusion FPLC column.

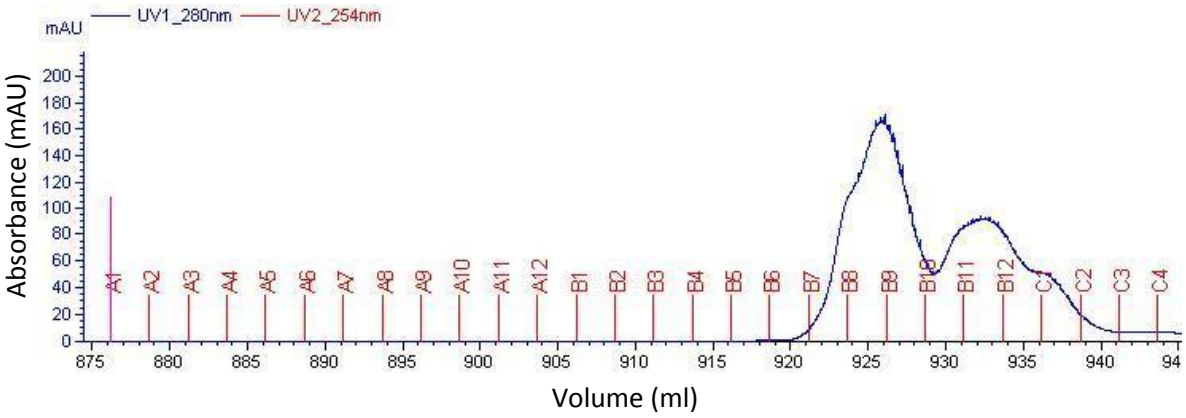


Figure 2.4. Chromatogram from size exclusion chromatography run (SEC1) using AKTA Explorer setup as described previously but with 2.5ml fractions taken from the moment of sample injection thus collecting the void volume. The blue line depicts absorbance of light by the sample detected at a wavelength of 280nm indicating protein. The pink line shown indicates the point of sample injection.

In Figure 2.4, there are two clear peaks visible. The first peak occurs almost 46ml after sample injection which indicates the void volume. A peak at this range shows a protein that was too large for the column to interact with. Given that it is known that the limitations of the column are proteins of 600kDa and that the smallest conformation of the haemocyanin molecule places it at 900kDa with the larger oligohexamer weighing 3600kDa, it was expected that this peak is the haemocyanin peak which starts in fraction B6 and finishes in B10 before another peak appears. The second peak becomes visible in B10 approximately 53ml post-injection and carries on until B12. A third peak is then seen as a shoulder peak to the second peak starting in B12 60 ml after injection and finishing in C2.

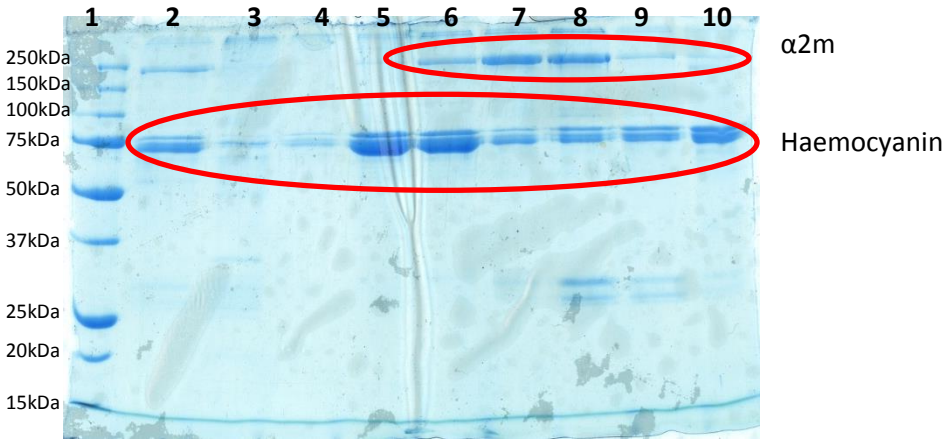


Figure 2.5. A 12.5% SDS-PAGE gel showing the fractions from the size exclusion chromatography. Lane 1 – Biorad All Blue precision markers: 250kDa, 150kDa, 100kDa, 75kDa, 50kDa, 37kDa, 25kDa, 20kDa & 15kDa

with the 10kDa band not visible. Lane 2 – contains the affinity chromatography breakthrough fraction at 0.5mg/ml. Lanes 3, 4, 5, 6, 7, 8 9 and 10 – Contains fraction B5, B6, B7, B8, B9, B10, B11, B12, and C1 respectively from the size exclusion chromatography run at concentrations between 0.2mg/ml and 0.5mg/ml.

Figure 2.5 shows the migration of protein bands from the initial size exclusion chromatography run of the affinity chromatography breakthrough fractions that showed absorbance at 280nm. For comparison some of the original affinity chromatography breakthrough fraction was held back to run in Lane 2 to assess its contents as well as seeing which fractions those contents appear in respectively. Lane 3/B6 shows a faint very high molecular weight band that has not migrated as far as the 250kDa marker has as well as a very faint band in line with the 75kDa marker. As this lane equates to fraction B6 and the fraction starts 42.5ml after the injection of the sample it is clear that this represents the end of the void volume, the point at which proteins too large to interact with the column are eluted. Lane 4/B7 contains the same bands as Lane 3 but with the 75kDa band being marginally more prominent than the larger mW band. Lane 5/B8 sees a very clear and thick band present at 75kDa. As the molecular weights of the subunits are known as well as the expected elution order it is likely that this band is for the haemocyanin subunits. Lane 6/B9 shows the same prominent band at 75kDa believed to be haemocyanin in addition to a band that lies between the 250kDa and 150kDa markers – this band is believed to be the start of the α2m elution. Lane 7/B10 shows a less intense peak at 75kDa (haemocyanin) and a more intense peak present at ~180kDa (α2m). Lane 8/B11 shows the same bands at Lane 7 with the same relative intensities but with the addition of a band in the region of ~25kDa believed to be the pentraxins. Lane 9/B12 shows a ~180kDa band (α2m) with reduced intensity relative to Lane 8, a 75kDa (haemocyanin) band with the same intensity as in the previous two lanes and a less intense 25kDa (pentraxin) band. Lane 10/C1 no longer shows the 25kDa band and shows a much fainter ~180kDa band however the 75kDa band persists.

Given that size exclusion chromatography separates proteins by size with the largest being eluted first and the smaller proteins last; the first protein eluted should be haemocyanin. This is seen as

it is the first protein to appear on the gel, but is visible from fraction B5 through to C1. The presence of the haemocyanin in lanes that clearly equate to the elution volume of smaller proteins means that the haemocyanin is either present in such large quantities that it 'leeches into the other fractions or that it is interacting with either the column or the lower molecular weight proteins in some way. The majority of the protein content of the affinity chromatography breakthrough should be haemocyanin followed by α2m and there should be no pentraxins present as they should have bound to the phosphoethanolamine affinity chromatography column. Therefore the next largest protein should be the ~370kDa α2m, which is appears to be eluted in fractions B9, B10, B11 and B12 as evidenced by the gel in Figure 2.5. The presence of the bands at 75kDa and ~180kDa were anticipated based on the principles of size exclusion chromatography and what was known about the affinity chromatography breakthrough fraction. The appearance of the double band at ~25kDa was as such a surprise. Given that it occurs in the same fractions as the α2m peak it should be noted that it is of a similar molecular weight. The heptameric stacked pentraxins have 14 25kDa subunits reaching a net total molecular weight ~350kDa which puts it in the same range as α2m. If the affinity chromatography was successful, which it appeared to be based on the affinity chromatography trace, why are there pentraxins present in the gel following the size exclusion chromatography of the affinity chromatography breakthrough fraction? Whilst present in significantly lower levels proportional to those seen in native PEG cut serum (Figure 2.2), this still presents a problem as the protein shouldn't be present as it should've bound to the PE column. However it has been shown that with PEG cut serum the addition of the serum to a PE-agarose column leads to the precipitation of some of the pentraxins (Shrive, *et al*. 2009) these precipitated proteins would still be filtered by the centrifugal filter devices and behave on both the size exclusion column and the gel as they normally would. It was shown that PEG cut serum should be incubated with the resin prior to column packing and elution from the column to minimise pentraxins precipitation.

Following the SDS-PAGE analysis of the size exclusion chromatography fractions, those shown to contain α2m (B10, B11 and B12) were concentrated using centrifugal filter devices prior to UV spectrophotometry to assess protein content (0.4ml @ 2.83mg/ml) before being injected back onto the size exclusion chromatography system for a second pass. This was done to further remove the haemocyanin which should be eluted off approximately 47ml after the sample is injected whereas α2m isn't eluted until ~54ml after sample injection.

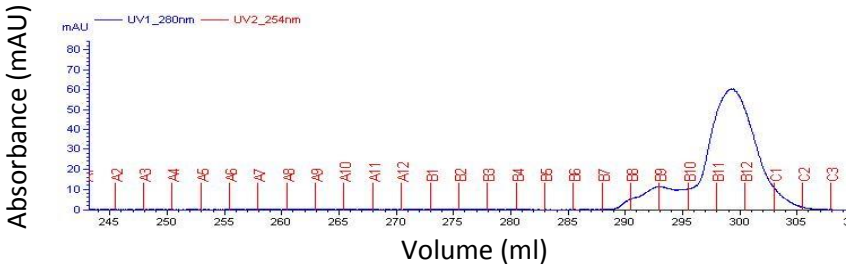


Figure 2.6. Size exclusion chromatography chromatogram (SEC2) using AKTA Explorer of a sample from fractions B9, B10, B11 and B12 of previous chromatogram collected into 2.5ml fractions at a flow rate of 0.2ml/min for optimum resolution.

In Figure 2.6 it is clear the effect of selectively reapplying fractions from the previous size exclusion run has had. The injection of the sample occurred at a volume of 242.94ml and the fractions were collected immediately thus collecting the void volume. The first peak appears in fraction B7 at approximately 289ml, 47ml after injection which is in line with the previous SEC run that saw its first peak occur at 46ml after injection. Therefore this peak is likely to be contributed by a protein that is large and did not interact with the column. As it is known that fractions B10, B11 and B12 were chosen from the first SEC run and they relate to lanes 7, 8 and 9 on the gel in Figure 2.5 it is clear that in those three lanes three bands of protein are visible; ~25kDa, ~75kDa and ~180kDa. The ~75kDa band which is present in all three lanes/fractions, is believed to be the tail end of the haemocyanin peak. This would then coincide with the appearance of a peak on this second SEC run immediately after the void volume as seen in Figure 2.6, as even the smallest oligohexamer of haemocyanin would be beyond the filtration capability of the column (~600kDa). A second peak forms in fractions B10, B11, B12 and C1 roughly 54ml after sample injection. This is

consistent with the previous SEC run, except that unlike in Figure 2.4, there is no shoulder peak present in C1 and as a result this peak doesn't extend into fraction C2.

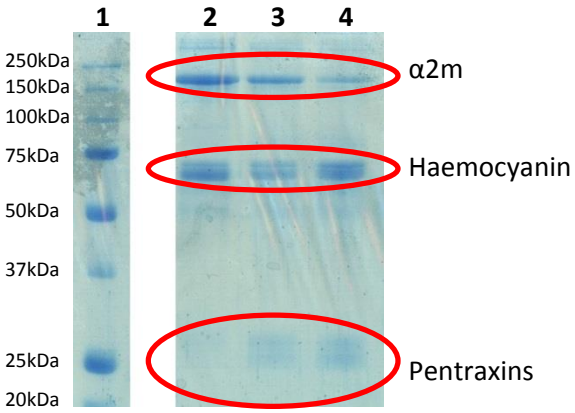


Figure 2.7. SDS-PAGE analysis of fractions of the second SEC run. In Lane 1 Biorad All Blue Precision Plus Markers (250kDa, 150kDa, 100kDa, 75kDa, 37kDa, 25kDa and 20kDa with the 15kDa and 10kDa bands not on the gel), Lanes 2, 3 and 4 are Fractions B10,B11 and B12 respectively, of the second SEC run shown in Figure 2.6, at concentrations of 0.4-0.8mg/ml. The gel was edited to show only the lanes containing α2m.

SDS-PAGE analysis of the SEC fractions shows similar results to the previous SEC run except that the relative amounts of haemocyanin to the proposed α2m band is diminished. This was further evidenced by UV spectrophotometric analysis of the fractions following collection.

Table 2.2. The calculated concentrations using UV spectrophotometry of fractions from both size exclusion chromatography runs.

mg/ml	B8	B9	B10	B11	B12	C1
SEC1	0.64	0.33	0.17	0.28	0.17	0.09
SEC2	0.01	0.03	0.04	0.016	0.08	0.00

The decline in protein content was visualised in the SEC chromatograms (Figures 2.4 and 2.6), as well as the SDS-PAGE analysis (Figures 2.5 and 2.7) and the UV spectrophotometric analysis of the fractions. The decline in protein content, primarily caused by the selective exclusion of the haemocyanin containing fractions from SEC1 has resulted in better visualisation of protein bands in Figure 2.7. It is known that the haemocyanin subunits range from 72kDa to 75kDa (Martin, *et al.*, 2007), this is clearly shown in Figure 2.7 where better resolution between bands is visible meaning haemocyanin is now clearly represented by two very close bands one that has travelled a little further than the 75kDa marker and another which has travelled the same distance.

Pentraxins are at this stage still present in small quantities with very faint bands present loosely aligned with the 25kDa size marker. Following analysis of the SEC2 fractions by SDS-PAGE fractions B10, B11 and B12 were combined and concentrated using a centrifugal filter device to a concentration of 1.24mg/ml and a volume of 0.6ml, with the concentration determined by UV spectrophotometry. The purpose of this was to remove the remainder of the haemocyanin just leaving behind α2m.

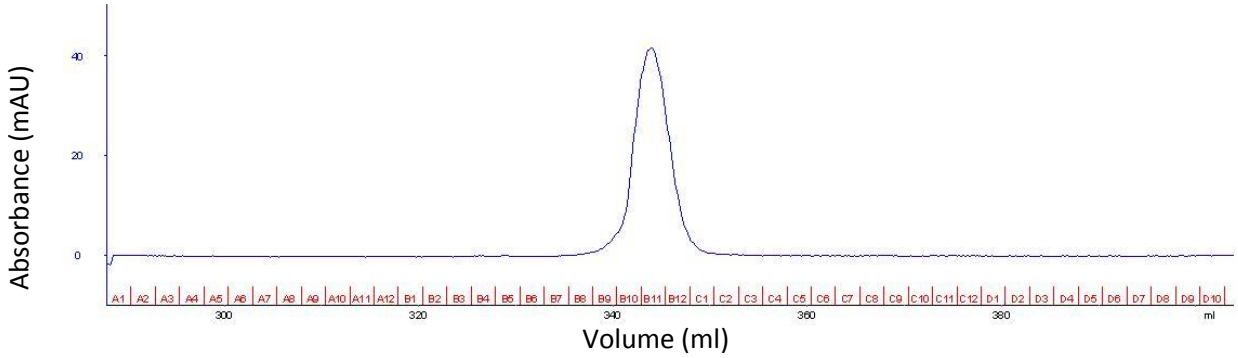


Figure 2.8. Size exclusion chromatogram (SEC3) of selected sample containing fractions B10, B11 and B12 from SEC2, which were shown to contain α2m (Figure 2.6). The blue line depicts absorbance of the solution at 280nm and thus shows the quantitative presence of protein in each fraction. Conditions were identical to the two previous SEC runs to ensure the comparability of the chromatograms; the protein was eluted in 1x Ca buffer (10mM $CaCl_2$, 150mM NaCl, 50mM Tris Base pH 7.4), at a flow rate of 0.2ml/min for optimum resolution and collected into 2.5ml fractions from the moment of sample injection.

Size exclusion chromatographic separation of the α2m containing fractions from SEC 2, yields a single peak that begins ~52ml and spans B10, B11, and B12 after the sample was injected onto the Superdex column. There is no distinguishable peak present at ~46ml which is the typical volume at which haemocyanin is eluted as shown in Figure 2.8. The absorbance does creep up a little in fraction B7 which is eluted 47ml after elution the typical range for the void volume and therefore haemocyanin.

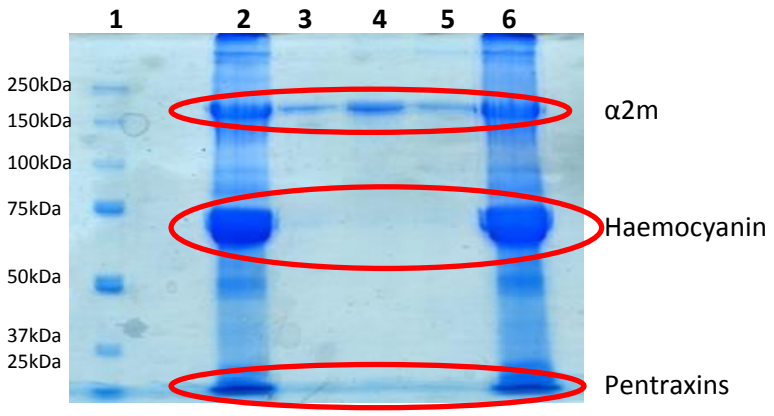


Figure 2.9. SDS-PAGE analysis of fractions B10, B11 and B12 from SEC3. Analysis was performed on a 10% SDS-PAGE gel and stained with Expedeon Instant Blue a Coomassie based stain. Lane 1 contains the Biorad All Blue Precision Plus markers (250kDa, 150kDa, 100kDa, 75kDa, 50kDa, 37kDa and 25kDa, with the 20kDa, 15kDa and 10kDa markers having migrated out of the gel), Lanes 2 and 5 contain the native PEG cut serum at a protein content of 28µg. Lanes 3, 4 and 5 contain fractions B10, B11 and B12 from SEC3 at 0.2-0.5mg/ml.

SDS-PAGE analysis of the fractions from the third size exclusion chromatography run, shows that fractions B10, B11 and B12, only contain one band that lies between the 150kDa and 250kDa markers at approximately 180kDa. Given that the known subunit size of *Limulus* α2m is ~185kDa once glycosylated it is assumed that this is the protein responsible for the visualisation of this peak. Furthermore the lack of any bands at both 25kDa – the known subunit size for *Limulus* pentraxins (Shrive, *et al*., 2009) and 72-75kDa – the known subunit size for *Limulus* haemocyanin (Martin, *et al*. 2007) it can be said with a reasonable certainty that the isolation of *Limulus* α2m from PEG cut serum is complete as no trace contaminants are visible on the gel in Figure 2.9.

2.3 Reaction of *Limulus* α2-Macroglobulin with methylamine

In order for α2m to allow proteases to access its bait region, it has a very high level of flexibility and is often referred to as being floppy. Transmission electron microscopy studies (TEM) have shown that α2m in its activated form is far more uniform in shape (Armstrong, *et al*. 1991). This molecular flexibility would no doubt pose a problem with regards to crystallisation of the molecule. As crystallisation utilises symmetry and the use of repeating building blocks, the unit cell, having an irregular protein capable of many orientations in its inactive form would be prohibitive to the formation of crystal contacts and thus crystal growth. As a result of this

activated α2m was the target molecule for crystallisation. Whilst a wide variety of potential ligands are available for α2m, thanks to its promiscuous bait region, and a broad range have in fact been reported and tried with the human α2m (Andersen, *et al*. 1991; Andersen, *et al*. 1994a; Andersen, *et al*. 1994b; Andersen, *et al*. 1995), there has been only one success of note, that is the production of diffraction spots to a resolution good enough to obtain a model, the crystallisation of human α2m reacted with methylamine (Marrero, *et al*. 2012). A protocol was devised based upon that used by the group lead by Lars Sottrup-Jensen (Andersen, *et al*. 1991; Marrero, *et al*. 2012). In their protocol (Andersen, *et al*. 1991), human α2m at a concentration of 10mg/ml is incubated with 0.2M methylamine and 10mM iodoacetamide, which is a protease inhibitor for cysteine proteases, for two hours. These concentrations were adjusted based on available concentrations of *Limulus* α2m to maintain the relative ratios. Following incubation for 2 hours, the sample was then concentrated in a centrifugal filter device to remove any remaining methylamine and iodoacetamide before being applied to the size exclusion chromatography FPLC column to separate the two forms of α2m – native and methylamine reacted.

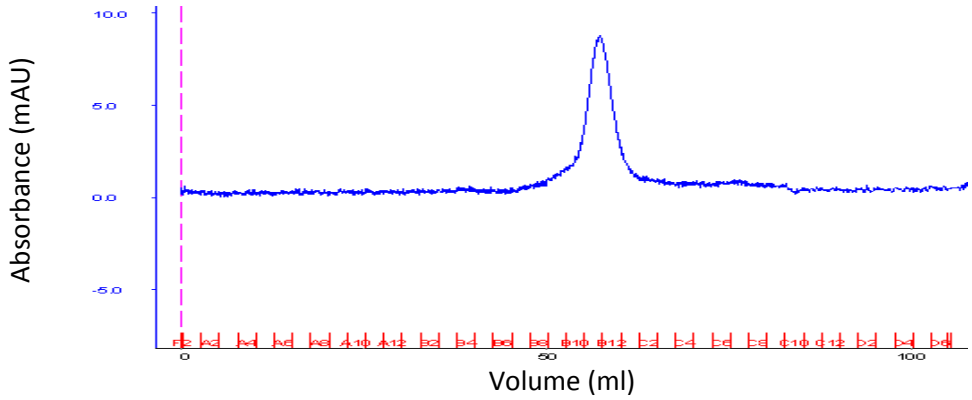


Figure 2.10. Size exclusion chromatogram of sample containing *Limulus* α2m reacted with methylamine. The pink dashed line shows the point of sample injection; The blue line shows the absorbance of the effluent at 280nm thus indicating protein presence. The sample was ran at a flow rate of 0.2ml/min and collected in 2.5ml fractions.

Size exclusion chromatography of the methylamine reacted *Limulus* α2m (SEC4) Figure 2.10, shows a single peak that starts in fraction B7 which occurs 47.5ml after sample injection and fraction C2 which is eluted 65ml after sample injection. The early region of the peak (fractions B7-

B9) shows extremely shallow curvature and given the elution volume is consistent with the void volume and the region that typically contains haemocyanin, which may be present here in fractions B7, B8 and B9 in trace amounts. Whereas absorbance in fractions B10, B11 and B12 and to a lesser extent C1 and C2 fall in the typical elution volume range that α2m falls within, and those here that show the greatest absorbance. These fractions (B10, B11, B12, C1 and C2) were combined and concentrated in a centrifugal filter device before analysis via gel electrophoresis.

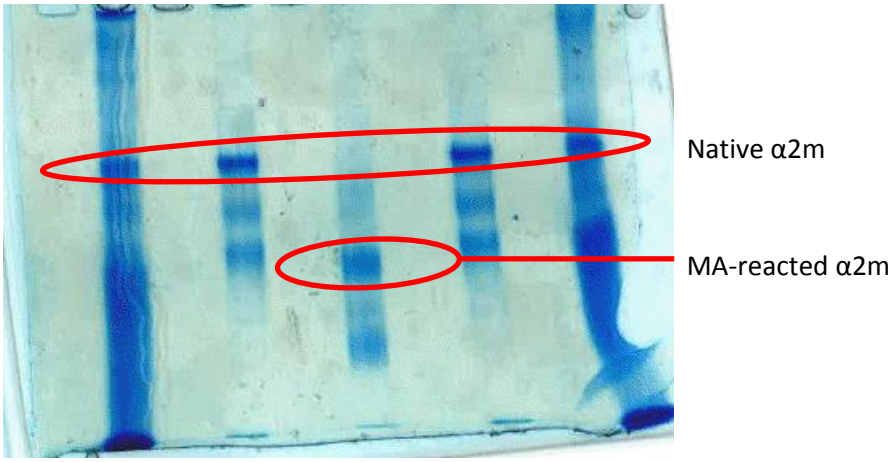


Figure 2.11. Native PAGE gel (4%) analysis of the peak from SEC4 (fractions B10, B11, B12, C1 and C2). Lanes 1 and 5 contain native PEG cut serum at a protein content of 5.6µg. Lanes 2 and 4 both contain purified *Limulus* α2m from SEC3 at a concentration of 8.7µg. Lane 3 contains fractions B10, B11, B12, C1 and C2 which contains *Limulus* α2m that has been incubated methylamine for two hours, at a protein content of 7.7µg.

Analysis of the successful reaction of *Limulus* α2m with methylamine was carried out using a 4% native PAGE gel. This was done as SDS-PAGE analysis would not differentiate between the reacted and unreacted forms of α2m as the cause of the increased electrophoretic mobility is due to the condensing of the quaternary structure, which destroyed during sample preparation in SDS-PAGE and thus no discernible difference can be seen. Rather than use molecular weight markers, PEG cut serum was used to show the migration distance of unreacted α2m. There is a band consistent across Lanes 1, 2, 4 and 5, given that it is known that the only protein present in the samples used in Lanes 2 and 4 is made up of a protein with subunits between 250kDa and 150kDa as shown in Figure 2.11, it can be assumed that the consistent band seen in these lanes is α2m. As the sample used in Lane 3 is taken from the same protein stock used for Lanes 2 and 4, it again can be safe to

assume that the only protein present here is α2m. In this case it is clear that the principle α2m band has travelled further through the gel and thus indicating increased electrophoretic mobility, a defining characteristic of the reacted form of α2m. The other faint bands present in Lanes 2 and 4 that appear to have travelled further than the proposed unreacted α2m band is caused by an unknown protein. Native PAGE gels are notoriously unreliable when it comes to resolving proteins (Shi, & Jackowski, in Hames, 1998). Given that one of the additional bands has a comparable migration distance with that proposed to be the α2m-MA band seen in Lane 3, it is possible that a proportion of the α2m present has become reacted as there are multiple opportunities for reaction to occur. Based on the evidence demonstrated in the chromatograms and gel images shown, the purification from PEG cut serum and the subsequent reaction with methylamine of *Limulus* α2m has been successful. Following this the process was repeated and up-scaled for production of sufficient quantities of pure methylamine-reacted *Limulus* α2-macroglobulin for the crystallographic trials that followed.

Chapter 3 – Crystal Studies of α2-Macroglobulin from *Limulus Polyphemus*

3.1 Introduction to Biomolecular Crystallography

As our curiosity into the fundamental mechanisms of life intensifies, a common theme has become apparent. The key to fully understand the function of anything in biology lies in its structure. This can be said of all structures in biology from the cellular level, all the way down to a molecular level. There are a variety of techniques available to structural biologists, which will be discussed later, the principle method however is x-ray crystallography (Jaskolski, 2010).

3.1.1 Protein Structure

Proteins, often considered the building blocks of life are the products of our genes translated from mRNA by the ribosome. The result is a linear chain of amino acids, a polypeptide which is folded in the Golgi apparatus before being released. Amino acids are chiral in nature, with a central chiral α-carbon atom bound to an amine group, a carboxyl group, a hydrogen atom and a variable side chain group. Shown in Figure 3.1, amino acids as with all chiral molecules there is an L-form and a D-form of almost all amino acids – the non-chiral glycine being the only exception . But in protein only the L-form is seen.

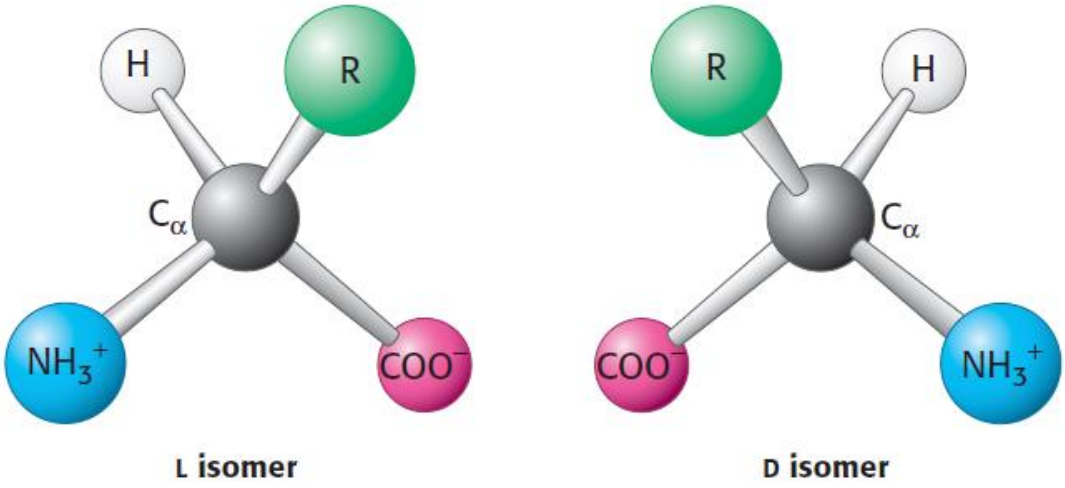


Figure 3.1. Basic amino acid structure of both the L- and D- isomers. The variable R group is what differentiates amino acids from one another and the interactions of these side chains contribute towards the structure of the molecule. Diagram edited from (Berg, *et al*. 2006)

The variable side-chains of amino acids are what differs between them. The side chains can be classified as polar or non-polar based upon the hydrophobicity of the side chain. The polar side chains can then be further classified into acidic (Glutamic acid - Glu/E),basic (Histidine – His/H) or neutral (Threonine – Thr/T). The properties of the side chains are key in the final spatial characteristics of the folded protein. The sequence of these amino acids in the polypeptide chain is known as the primary structure of the protein. Secondary structures within proteins are primarily made up of α-helices and β-sheets, which are held together by inter-side chain hydrogen bonds (Branden, & Tooze, 1999).

In the α-helix hydrogen bonds form between the carboxyl group f residue *n* and the amine group of residue *n+4* of the main chain backbone. The helix has 3.6 residues per full turn In proteins due to the use of the L- form of the amino acids the helix is right-handed as the packing of the side-chains would not allow for a left-handed screw of the helix. The β-strand is made up typically of a fully extended stretch of 5-10 amino acids, where hydrogen bonds between adjacently aligned β-strand carboxyl and amine groups leads to the formation of β-sheets. If the sequence of the β-sheet runs so that the N-Cα-C-N-Cα-C motifs of two aligned, adjacent, and hydrogen bound β-strands run in the same direction a parallel β-sheet is formed, if they run in opposite directions an

anti-parallel β-sheet is formed, with mixed β-sheets also a possibility. Most protein based β-sheets have a twist to their strands, which like the α-helix is always right handed .

Hydrogen bonds occur between oxygen or nitrogen atoms and hydrogen and are relatively weak compared to covalent bonds. Combinations of these secondary structures form motifs such as the pentraxin domain seen in pentraxins (Gewurz, *et al*. 1995), which is made up of a flattened β-jelly-roll with a long α-helix folded on top of the β-sheet. Knowledge of the primary structure and the secondary structures often doesn't reveal the function and mechanisms of a protein. There are a variety of techniques available to learn the primary and secondary structures of a protein. However, in order to fully grasp the function and mechanisms of action we must use techniques that reveal every atom of the structure. Tertiary structures arise from the folding of the peptide chain and the combination of secondary structure motifs. The tertiary structure often represents the individual subunit of a protein molecule. Internal bonds holding together the tertiary structure include: covalent disulphide bridges between cysteine residues, hydrophobic interactions where the hydrophobic side chains of the amino acids 'hide' away from the solvent by remaining buried within the core of the protein, hydrogen bonds, coordination of partially negatively charged side chains by metal or salt ions, and Van de Waals forces. Finally the quaternary structure of the molecule, the arrangement of subunits; stabilised by the same non-covalent and disulphide bonds that hold together the tertiary structure of the protein.

3.1.2. Benefits of Crystallography

X-ray crystallography of macromolecules is the gold standard technique for atomic resolution structure determination. Structures solved by X-ray crystallography have been solved to resolutions as high as 0.5 Å, although structures are more commonly seen at ~2Å. It is this ultra-high resolution that makes crystallography such a valuable technique. It allows us to see the molecular interactions of proteins both natively and with their target ligands at a molecular level; thus revealing the mechanism of action and the residues involved in binding. This is in itself a

valuable piece of information as it can lead to the development of new therapeutic agents in some cases where, drugs can be designed to fit or block binding pockets or proteins can be engineered to bind more effectively. A common argument against crystallography, raises issues the physiological relevance of the structures determined by x-ray crystallography. However crystallographers answer to this, is that the proteins are still biologically active as shown by binding of their natural ligand targets in crystal structures. X-ray crystallography has been used to determine the structure of DNA, small peptides, large macromolecules and even viruses.

It would however be careless to neglect the disadvantages to crystallography. The rate determining step and often the major stumbling block is the production of diffraction quality crystals. Some crystals grow in a matter of weeks and other can take years if they grow at all; and even then they may not be suitable for x-ray diffraction. Hydrogen atoms cannot be seen as they only have one electron to diffract the x-rays; this may seem trivial but as mentioned earlier hydrogen bonds play a key role in inter-residue interactions giving rise to secondary structures.

3.1.2.1. Other Techniques in Structural Biology

Circular Dichroism spectroscopy is a technique that is capable of producing information on the polypeptide backbone conformation of proteins yielding information about the secondary structure and showing the presence of conformational changes by measuring the difference in absorbance of and left- and right- circularly polarised light by a chiral sample (Wallace, & Janes, 2001). Whilst not capable of yielding the super high resolution data of x-ray crystallography, circular dichroism provides us with data such as % α-helix due to the spectral signature shown by helices. Despite the protein sample needing to be of similar purity to that needed for crystallography >95%, the sample does not need to be crystallised to yield data, thus providing a useful complimentary structural tool, that can yield data during the time variable stage of growing crystals while investigating a protein structure (Kelly, *et al*. 2005).

Cryo-electron microscopy (Cryo-EM), is the electron microscopy of an unstained biological specimen in a frozen state; it is capable of yielding structural models to an atomic resolution (Glaeser, & Hall, 2011). Currently cryo-EM has only been able to produce atomic resolution models of large macromolecular proteins >500kDa, and cryo-EM studies have struggled to hit ultra-high resolutions in molecules with multiple conformational states. A relatively new structural biology technique it can be expected that the resolution quality and size range capable with this technique only will improve as advances continue to be made (Glaeser, & Hall, 2011).

Nuclear Magnetic Resonance (NMR) – spectroscopy is also used to study the structures of proteins. It exploits the magnetic properties of certain atomic nuclei and yields structural data that can rival X-ray crystallography (Keeler, 2005).

In the 1960s NMR spectroscopy, a technique that had been predominantly used to study the structure of small organic molecules began to be applied to larger organic molecules; proteins, nucleic acids and carbohydrates (Macomber, 1998). NMR-spectroscopy is a very useful and powerful tool for structural biologists to utilise, as is currently the only method that allows for the production of a high resolution 3-D structure determination of partially or wholly unstructured proteins. NMR-spectroscopy though does have its limits and protein structures solved by NMR-spectroscopy are limited to around 35kDa in size.

X-ray crystallography is still the gold-standard for protein structure determination, it is however not without its challenges it is for this reason that other techniques should be considered and used to compliment crystallographic data as well as providing alternatives for those difficult to crystallise proteins (Sengupta, 2010).

3.2 Protein Preparation and Crystal Growth

In order to determine the structure of a protein molecule using crystallography we first need a crystal. Different methods may be used in the growth and optimisation of the crystals in the search for diffraction quality crystals. Crystals grown for macromolecular crystallography are ideally large in nature (0.1-0.5mm), although smaller crystals may be used, making use of specialised beamlines (I-24 Microfocus – Diamond Light Source, Oxford). Ultimately the quality of the crystal determines the quality of the data, with protein purity being the most vital factor in crystal quality.

3.2.1 Protein Preparation

3.2.1.1 Protein Purification

As stated above protein purity is a key ingredient in the successful growth of diffraction quality crystals. The protein needs to be >95% purity this is to ensure that when the crystal forms it is made a single protein type. Impurities will lead to the presence of multiple asymmetric units being present resulting in the absence of diffraction. Due to α2-macroglobulin possessing two conformational states (fast-form and slow-form, as shown with PAGE analysis (Barrett, *et al*. 1979; Quigley, & Armstrong, 1985) homogeneity of the sample is key. This is but one obstacle in terms of sample preparation for crystallographic studies of α2m, as the slow-form is not as structurally rigid as its fast-form counterpart. The regularity and rigidity of the fast-form molecule makes it a more suitable target for crystallisation. The fast form can be obtained by reaction of α2m with methylamine or any protease that cleaves its bait region triggering the conformational change. There are a wealth of techniques available to produce purified protein in addition to those discussed in the previous chapter, a number of chromatographic techniques are typically used in combination with each other and verified using techniques such as SDS-PAGE and western blots etc.

3.2.1.2 Protein Concentration

As discussed in greater detail earlier, the formation of crystals is dependant of supersaturation of the protein. Too high a protein concentration and supersaturation occurs too fast resulting in the formation of amorphous precipitate, too low and crystallisation will never occur. A number of techniques can be used to concentrate a protein but in the previous chapter I stated that for the experiments in this thesis, ultrafiltration by centrifugation was used.

3.2.2 Crystal Growth: Dynamics & Techniques

Thermodynamics are the driving force behind crystal formation. In forming a crystal rather than staying in solution the proteins are taking the route that is most thermodynamically favourable. This is because the crystalline state represents the greatest number of stable bonds and thus reducing the free energy. The fact that the crystalline form represents the more favourable energy state is shown by, the initial formation of precipitate before it's dissolution, during which crystals form. The reverse cannot be observed.

Crystals grow from supersaturated solutions, that is to say its concentration in solution is greater than the limit of their solubility. This is because in the supersaturated zone the nature of the equilibrium between solid phase and the solution encourages the movement of proteins into the solid phase. Achieving the appropriate level of supersaturation is a fine art; too low and crystals will never form, too high and the protein will fall out of solution as amorphous precipitate rather than beautifully order crystals. To help explain this, the phase diagram was devised to illustrate the relationship between protein concentration and precipitant concentration.

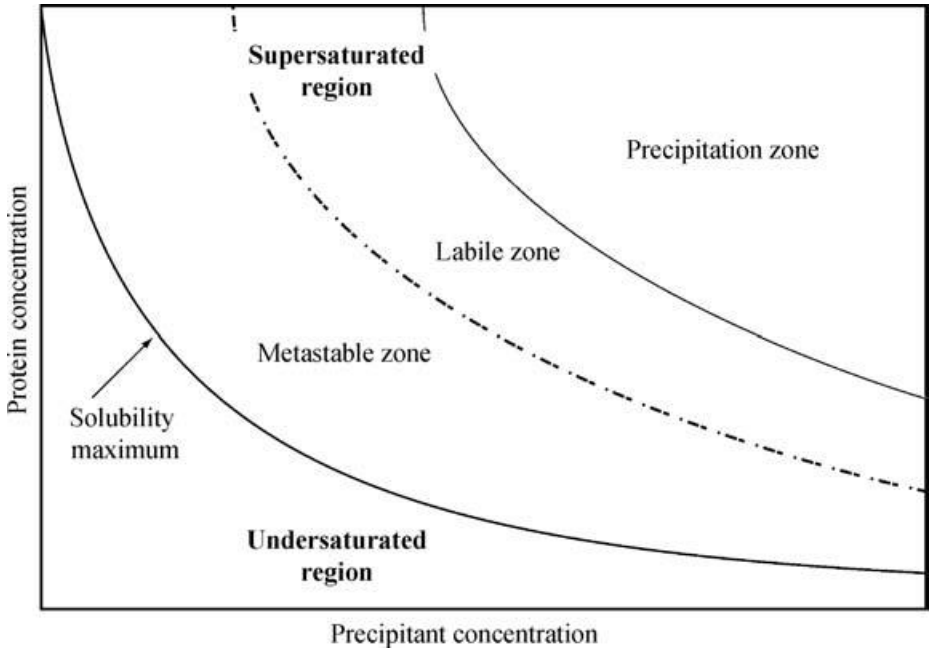


Figure 3.2. The phase diagram for crystallisation. It consists of two primary regions; the undersaturated and the supersaturated. The two regions are divided by the equilibrium line that represents the maximum solubility of the protein. The supersaturated region itself is divided into three zones; the precipitation zone where protein falls out of solution into amorphous precipitate, the labile zone where nucleation and crystal growth occurs, and the metastable zone where crystals do not nucleate but will sustain growth if present (McPherson, 2009).

Once the protein is in the supersaturated region nucleation is required for the crystal to form and grow as opposed to the formation of nonspecific aggregates. The formation of a stable nucleus is key, 'unstable' nuclei form rapidly and spontaneously once the supersaturated state has been achieved but quickly dissolve into solution. A stable nucleus then is one that enlists new molecules onto its growing surfaces at a rate faster than others dissolve back into solution, resulting in a net gain of mass for as long as it remains in the supersaturated condition. Stable nuclei are almost exclusively found in the labile zone of the supersaturated region. The metastable zone does not produce stable nuclei, however if a stable nuclei were to be deposited by some method as in crystal seeding, that nuclei would continue to grow, so long as supersaturation was maintained.

Using the phase diagram in Figure 3.2 that the nucleation zone is between the metastable growth zone of a crystal and the amorphous precipitate zone, which are both defined by the relationship

between protein concentration and the concentration of a precipitant. In terms of crystal growth there are two types of nucleation; homogenous nucleation and heterogeneous nucleation. For homogenous nucleation to occur the [protein]:[ppt] has to be in the nucleation zone also known as the homogenous nucleation zone. Crystals that form from this zone have nuclei made solely of the target protein. When we grow crystals via heterogeneous nucleation we place a preformed nuclei in the metastable growth zone, also known as the heterogeneous nucleation zone, and the target crystal forms around this nuclei of non-target molecules. This process is called crystal seeding and shall be addressed in more detail further on.

3.2.2.1 Thermodynamics of Crystal Nucleation

The birth of a protein crystal from its mother liquor requires a central starting point, a nucleus. These nuclei are made up of growth units that vary from individual monomers to dimers or even tetramers. Growth units by successive aggregation form spherical nuclei known as embryos. Crystal embryos are spherical in shape as a sphere has a low surface area to volume ratio and thus is the typically thermodynamically favoured shape of crystal nuclei for reasons to be explained (Oxtoby, 1998).

The lifetime of these nuclei is dependent upon the forces the growth units exert on one another and on the solution in which they sit. If we imagine that the growth units are spherical themselves and each has six perpendicular forces it can exert upon other growth units or the external solution. The forces they exert on one another are fighting to hold the embryo together whereas the forces they exert on the liquid phase are trying to pull the newly formed cluster apart. The forces that work to maintain the integrity of the embryo is proportional to the number of bonds between growth units and can be considered proportional to the volume of the embryo. The forces working to pull the embryo apart are proportional to the number of free bonds and are proportional to the surface area of the embryo (Garcia-Ruiz, 2003).

Therefore the energy balance can be written as:

$$\Delta G_t = \Delta G_s - \Delta G_v$$

Where ΔG_t is the total Gibbs free energy of an embryo and ΔG_s is the free energy from unsaturated surface bonds and ΔG_v is the free energy contributed by the internal saturated bonds.

As stated previously ΔG_s and ΔG_v are proportional to the surface area and volume of the embryo respectively. As a spherical nucleus is the favoured embryonic form we can then substitute in the volume and surface area of a sphere in to the equation.

$$\Delta G_t = \frac{4\pi r^2}{a^2}\gamma - \frac{4\pi r^3}{3a^2}\Delta\mu$$

Where a is the size of the individual growth unit, r is the radius of the cluster, γ is the free energy of the surface bonds per unit area, also known as surface tension, and Δμ is the difference in chemical potential between the growth units of the solution and those of the embryo.

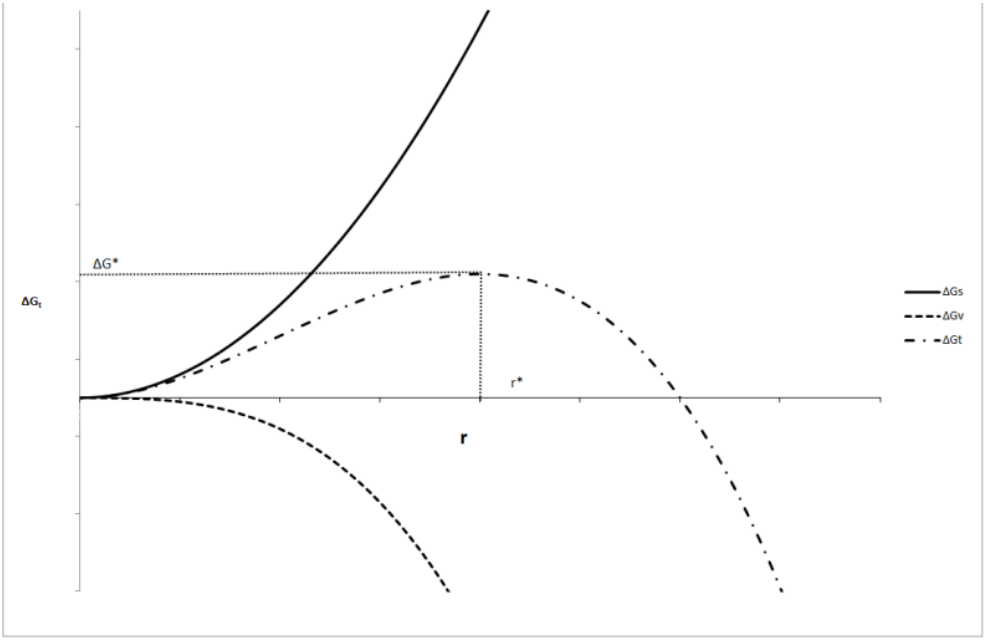


Figure 3.3. Graph depicting the energy balances governing crystal growth. ΔG_t is the total Gibbs free energy, ΔG_s is the free energy from free unsaturated surface bonds and ΔG_v. When depicted like this a clear maximum value for ΔGt is visible ΔG*, which is known as the nucleation barrier (Garcia-Ruiz, 2003)

If we attribute a value to a and plot ΔG_t, ΔG_s and ΔG_v against increasing values of r we can see in Figure 3.3 that ΔG_t, reaches a maximum value at a particle size called the critical size, r^* the energy required to achieve critical radius is known as the nucleation barrier or ΔG^*. When embryos reach the critical size they are now referred to as nuclei; embryos with $r<r^*$ dissolve back into solution, those nuclei with $r>r^*$ are likely to go on to form nuclei and eventually crystals. Once r* has been achieved it is clear that the expansion of r beyond that point lowers the total Gibbs energy further and is thus a spontaneous process (Garcia-Ruiz, 2003). Both the nucleation barrier and the critical radius can be written as a function of the degree of supersaturation:

$$\Delta G^* = \frac{16\pi\gamma^3}{3[kT\ln(S)^2}$$

$$r^* = \frac{2\gamma a}{kT\ln S}$$

Where S is the degree of supersaturation, as a result it is clear that both the critical radius and the nucleation barrier decrease with and increasing supersaturation.

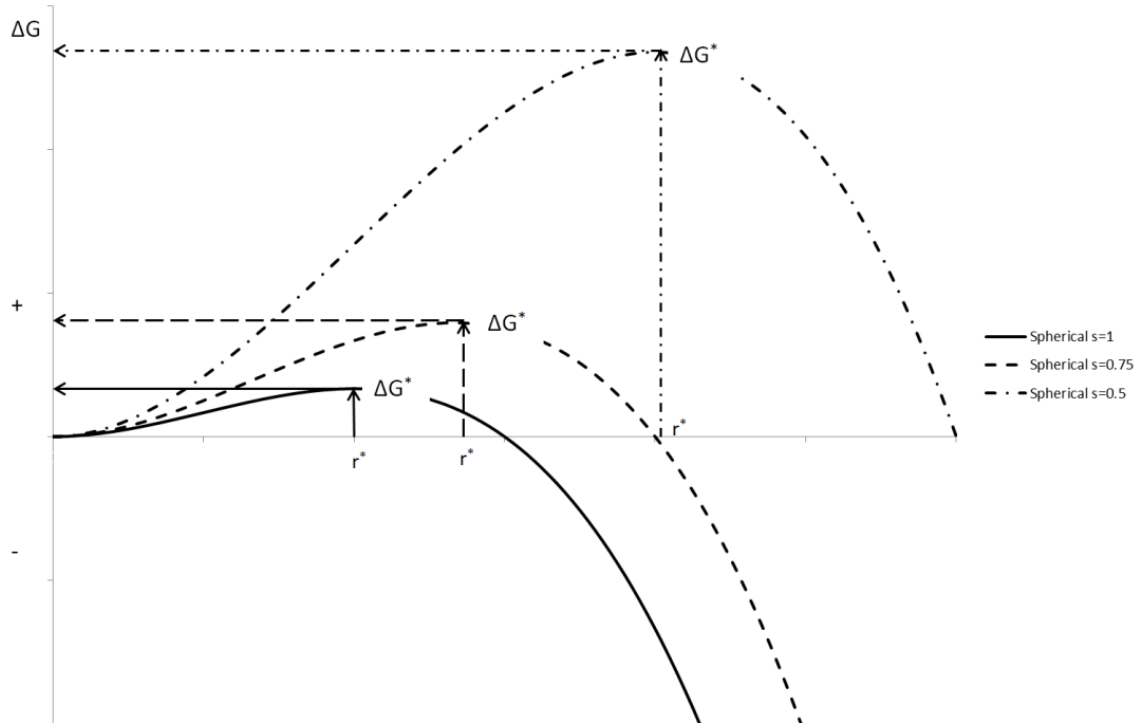


Figure 17 Gibbs energy balance of crystal nucleation. Three examples of varying amounts of supersaturation (S) S=1, S=0.75 and S=0.5, showing that the greater the supersaturation the lower the energy required for successful nuclei formation resulting in a smaller required radius for said crystal nucleus (Garcia-Ruiz, 2003).

By visualising these relationships in Figure 3.4 we can infer that at higher degrees of supersaturation the lower the nucleation barrier and thus the lower the critical radius required for a nucleus to form and continue to grow. At lower levels of supersaturation this energy barrier is higher and as a result the required critical radius is larger. It would seem then that the ideal degree of supersaturation is as high as possible to ensure that nucleation is as energetically favourable as possible. However, if we raise the degree of supersaturation so high that r^* is smaller than the size of an individual growth unit the result is the production of the amorphous phase known as precipitate. Another outcome of having too high a degree of supersaturation is that nucleation occcurs to quickly due to the low energy and small radius required for it to occur and we see a shower of tiny crystals as too many nuclei have formed. In order to grow a small handful of large crystals it is better to have a supersaturation that sits at the lower end of the nucleation zone of the phase diagrams shown earlier in the chapter, whilst energetically less

favourable and thus requiring a larger nuclei to achieve r^*, this will give fewer and ultimately larger crystals which are generally easier to extract data from downstream.

Crystals that form more slowly are generally free of defects and the crystals would move out of the labile zone and into the metastable as the concentration went down as protein was recruited to the nuclei.

3.2.2.2. Crystallisation Experimental Setup

In order to grow crystals proteins are normally dissolved in a buffer containing a precipitant often of the PEG family. Using controlled evaporation, water is slowly moved from the solution, to achieve supersaturation and crystal growth. The most common technique utilising controlled evaporation is – vapour diffusion, of which there several permutations the two most common being the hanging and the sitting drop methods.

The hanging and sitting drop both employ the same theory behind them; however, they differ slightly in experimental setup illustrated in Figure 3.5.

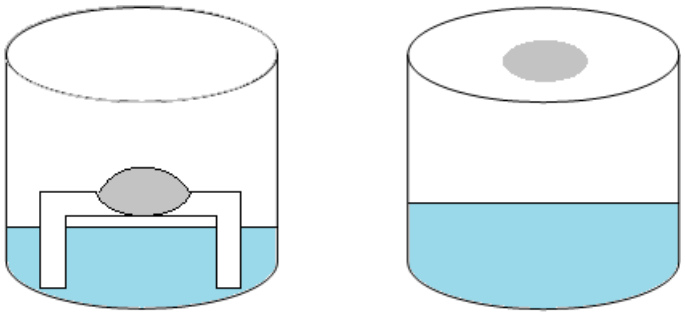


Figure 3.5. Schematic of both sitting and hanging drop experimental setups. The blue area represents the reservoir solution normally containing precipitant, buffer and salt but importantly no protein. The grey droplets represent the protein droplet and ultimately the area where crystal formation will occur. The protein droplet normally consistent of an equal volume of protein solution and of reservoir solution. However this is not always the case. The tops of the wells are sealed with grease, in order to create an air tight isolated system. This is to prevent the reservoir solution from evaporating and drying out as the goal of vapour diffusion is for the droplet to dry out as described below.

Vapour diffusion works on the principle of the concentration difference of the precipitant between the droplet and the reservoir solution will result in the net movement of water from the

droplet into the reservoir. This is because the concentration of the precipitant in the droplet in an equivolume setup is half that of the reservoir solution. Thus water via diffusion leaves the droplet in favour of the reservoir until a state of equilibrium is achieved. This loss of water from the protein droplet not only increases the precipitant concentration but also the protein concentration, thus pushing the protein into the supersaturated state.

Other crystal growth techniques used are liquid-liquid diffusion and dialysis, but they work on the same principle – driving the protein concentration into the supersaturation leading to crystal formation.

3.2.3 Radiation Damage

Due to the photons of X-rays having such high energy levels, when they come into contact with the proteins they can form radicals within the molecules leading to crystal degradation. The oxygen or hydroxyl radicals can diffuse through the crystal causing further damage to the crystal structure that leads to a visible tailing off in the resolution of data being recovered from the crystal. The latest detectors used in x-ray crystallography allow shorter exposure times than their predecessors, reducing the crystals exposure to radiation and subsequently reducing radiation damage. Even if the correct precautions, such as the cryoprotective measures discussed in the next section, are taken radiation damage still occurs but just to a lesser extent. Disulphide bonds in particular are prone to cleavage as a result of radiation damage, the decarboxylation of carboxylic acids are all noted effects of radiation damage as well as an increase in atomic B-factors (Drenth. 2007).

3.2.4 Cryoprotection

Whilst often morphologically indistinguishable small molecule and macromolecular crystals differ greatly in their properties. When compared to crystals formed by small molecules, protein

crystals are generally weaker; this is due to the fewer lattice interactions proportional to molecular mass. Thus the crystals are softer, more flexible and are much more sensitive to a subtle change in conditions. When protein crystals form, there are gaps or channels between molecules in the crystal structure, these gaps are filled with solvent. This solvent which makes up on average 50% of the crystal is free to diffuse in and out of the crystal. These solvent channels are partly responsible for the reduced resolution seen in macromolecular crystals as opposed to small molecule crystals which diffract to much higher resolutions generally. This is because the large gaps between the molecules reduce the likelihood of every molecule lining up in its exact position, thus bringing about slight variation from lattice point to lattice point (McPherson, 2009).

In order to minimise the radiation damage discussed, crystals are cryocooled. Upon their harvesting from the mother liquor, via nylon cryoloops available from companies such a Hampton Research, the crystals are then added to their vial caps and stored in liquid nitrogen until data collection. This freezing of the crystal leads to the immobilisation of the radicals produced by exposure to X-rays and thus minimises the damage they cause. Even during data collection the crystals will be kept under a constant cryostream ensuring they stay frozen. In the process of protecting the crystals from radiation damage another problem has been introduced; water molecules in the above mentioned solvent channels freeze due to the low temperature and form ice crystals. The formation of these ice crystals within the solvent channels of the protein crystal leads to crystal damage as water expands as it freezes. As a result the crystals are introduced to antifreeze solutions that subsequently act as a cryoprotectant preventing the water molecules from forming the ice crystals and leading to the disruption of the crystal. Cryoprotectants are typically molecules such as glycerol or , 2-Methyl-2,4-pentanediol (MPD) and low molecular weight PEG such as PEG 200 MME. Some crystals can be grown already in the presence of a cryoprotectant, and thus don't need to be further cryoprotected as sometimes the addition of a cryoprotectant can lead to the crystal breaking up (Drenth, 2007).

3.3 Protein Crystals

3.3.1 Crystal Properties

Crystals are made up of the periodic assembly of small building blocks, which can be small molecules, nucleic acids, proteins or even massive protein-protein complexes. In protein crystals there is a sparse network of weak intermolecular interactions between protein molecules. Due to the irregularity in shape of proteins as well as the sparsity and lack of strength of the interactions between protein molecules, protein crystals are often extremely weak, fragile and often soft. As a result proteins need to be handled with extreme care to avoid, disintegration, delamination or the breaking off of a desirable chunk.

The weak forces holding the protein molecules in the crystal lattice include: dipole-dipole interactions, hydrogen bonds and van der Waals forces. The gaps between molecules, the afore mentioned solvent channels, are filled with the mother liquor solution from which the crystal was grown, resulting in crystal solvent content of around 50%.

As proteins are irregular in shape and contain multiple sites for these weak protein-protein interactions, the same protein may form more than one crystal morphology dependent upon the crystallisation conditions. This phenomenon may be helpful in trials where ligand soaking is the target and one crystalline morphology favours a larger ligand and allows for the successful soaking.

3.3.2 The Asymmetric Unit, Space Group Symmetry Operations and the Unit Cell

The asymmetric unit of a crystal is the smallest structure that lacks any inherent symmetry; typically in protein crystallography the asymmetric unit is an individual protein molecule or subunit, on occasion it can be a half or quarter subunit depending upon self-symmetry within the molecule. In the next step of molecular organisation within the crystal, the asymmetric unit has a variety of symmetry operations applied to it; the result is a closely packed additional set of asymmetric units. Multiple symmetry operations may be applied to the asymmetric unit, these range from the simple; translation and rotation, which can combine to give: centres of symmetry and screw axes. These combinations of symmetry operations are known as the space group. There are limits however to which symmetry elements are allowed within a crystal, a comprehensive list is given in the International Tables for X-ray Crystallography Vol. 2 (Hahn, 2005), combinations of all of the possible symmetry operations gives rise to 230 unique three-dimensional space groups. The space group of a crystal might best be described as how using symmetry operations, a set of asymmetric units are related. Due to the stereoisomerism present in amino acids, and proteins only using the L-isomer, the space groups that require inversion symmetry (the changing of hands of the molecule), centre of symmetry, mirror plane and glide planes are all eliminated for selection by proteins reducing the available number of space groups for the molecules to assume within the crystal to 65 (Rupp, 2010).

Screw axes a combination of a N-fold rotation followed by a successive translation, and hence the previous name in earlier literature of a roto-translation. With rotational symmetry the asymmetric unit with N fold rotational symmetry was returned to its starting position; this is not the case for screw axes. In the screw axes with N-fold rotational symmetry the translational vector t, has a magnitude of s/N, where s has values 1-N-1 meaning that as rotational symmetry increases the number of rotational axes also increases (Rupp, 2010). To explain space group nomenclature I will use $P2_1$ as an example. The capital letter at the beginning of the space group denotes the crystal

system in place – in this case primitive, with a symmetry operation of 2_1 (Rhodes, 2000).What this means in reality is that the asymmetric unit undergoes two fold- rotational symmetry, before undergoes a 1/2 or a c/2 translation along the c axis.

The third stage in constructing the building blocks of a protein crystal is to construct the smallest possible parallelepiped that encloses the full set of symmetry operated molecules. The nature of the unit cell is in fact determined by the symmetry operations performed on the unit cell shown below in Figure 3.6.

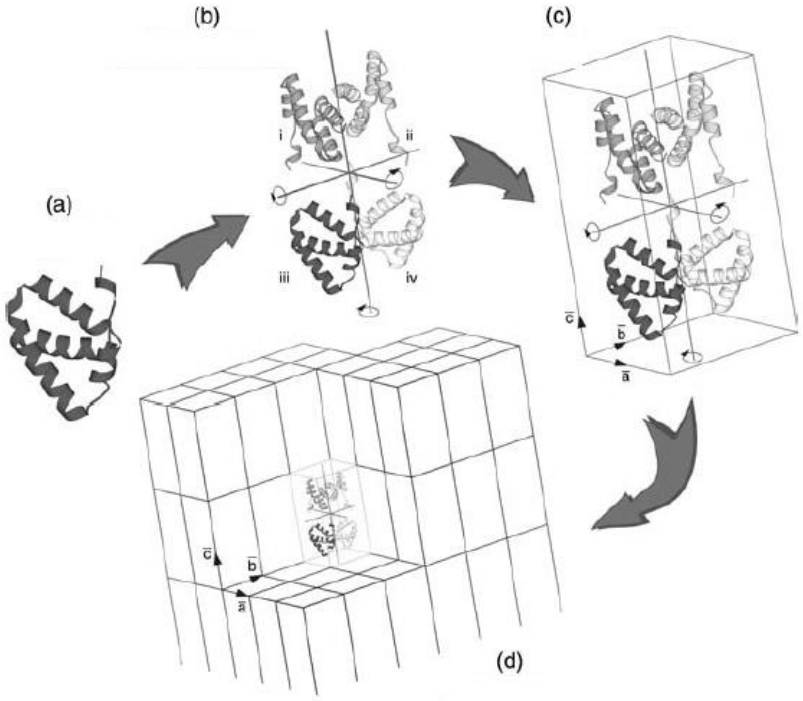


Figure 3.6. The arrangement of the various building blocks of a protein crystal. A) The asymmetric unit, the smallest element of the crystal building block lacking any symmetry. B) The space group related symmetry operations upon the asymmetric unit. In this example the original asymmetric unit labelled i, undergoes two, twofold rotations. The first along the vertical axis to produce asymmetric unit ii and the second along the horizontal axis to produce asymmetric units iii and iv. C) demonstrates the imaginary boundary of the unit cell enclosing the complete collection of symmetry operated asymmetric units, and mimics the symmetry operations of the space group. D) the periodic packing of the unit cells along the unit cell axis forming he crystal lattice (McPherson, 2009).

As the space group determines the characteristics of the unit cell, a nomenclature needs to be used when talking about the properties of the unit cell. The basis vectors of the unit cell are a, b and c all emanating from the origin 0. The angles between the basis vectors of the unit cell are

also determined by the symmetry functions of the space group. The angle between vectors a and b is γ, the vector between a and c is β, and the angle between b and c is α. The planes between vectors are also labelled so that the plane spanned by a and b is C, a and c is B and b and c is plane A.

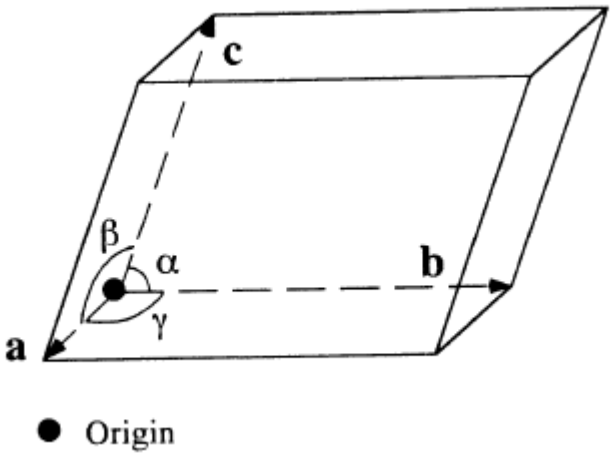


Figure 3.7. The dimensions, angles and planes of the unit cell of a crystal. Understanding these is key for the understanding of crystal systems and lattices. Diagram edited from (Drenth, 2007)

These axes dimensions and the angles, shown above in Figure 3.7, help to classify the crystal system and eliminates certain space groups, and so the unit cell data is extremely valuable.

3.3.3 Crystal Systems, Bravais Lattices and Point Groups

There are seven crystal systems, as defined by the relationship between the axes and their enclosed angles. These crystal systems are described in Table 3.1 along with the types of lattices/Bravais lattices they can accommodate. The unit cell generally only contains one full complement of the symmetry operated asymmetric units, these crystal systems are known as primitive lattices, denoted P (McPherson, 2009). However in higher order crystals exhibiting more symmetry, there may be more than one full complement of symmetry operated asymmetric units. These make up the centred lattice systems that

may either be face centred, F/C where F has all planes centered as opposed to just plane C, body centred I, or rhombohedral centred R. The centred crystal systems make up the remaining 8 Bravais lattice systems to a total of 14.

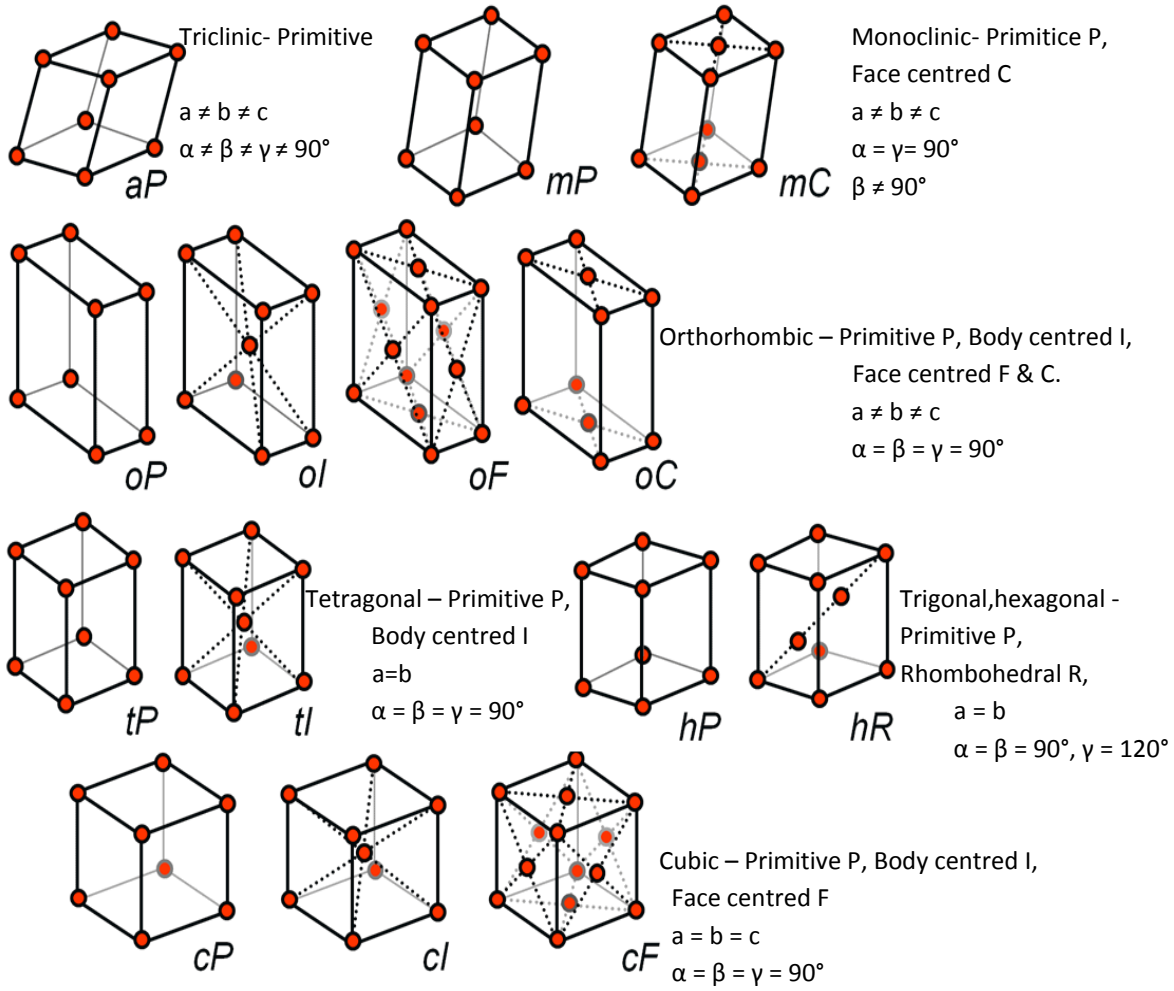


Figure 3.8. The 14 Bravais lattices organised by their crystal systems. Lattice aP is the primitive triclinic crystal system where the a stands for anorthic meaning of oblique axes. Lattices mP and mC are the monoclinic primitive and face centred, respectively. Note that mC has only the faces on the C planes centred and as such is not an mF system. Lattices oP, oI, oF and oC are the orthorhombic primitive, body centred, face centred (all faces and C-face only) respectively. Lattices tP and tI are the tetragonal primitive and body centred lattice systems respectively. Lattices hP and hR are the trigonal/hexagonal primitive and rhombohedral systems respectively. Finally the cP, cI and cF lattices are the cubic primitive, body centred and face centred (all faces) lattice systems. Diagram edited from Rupp 2010 (Rupp, 2010).

The application of symmetry operations to the Bravais lattices, in Figure 3.8, leads to the organisation of 65 chiral space groups, in Table 3.1.

Table 3.1. Crystal systems and their possible lattice types and space groups as well as their unit cells parameters. All 65 possible space groups for biological macromolecules are assorted by their crystal systems (Rupp, 2010; Drenth, 2007; McPherson, 2009).

Crystal System	Lattice Types	Unit Cell Dimensions and Angles	Unit Cell Symmetry	Permissible Space Groups
Triclinic	P	$a \neq b \neq c$, $\alpha \neq \beta \neq \gamma \neq 90°$	None	P1
Monoclinic	P	$a \neq b \neq c$, $\alpha = \gamma = 90°$, $\beta \neq 90°$	A single 2 fold axis parallel to b.	$P2$, $P2_1$
	C			C2
Orthorhombic	P	$a \neq b \neq c$, $\alpha = \beta = \gamma = 90°$	Three mutually perpendicular 2 fold axes	P222, $P222_1$, $P2_12_12$, $P2_12_12_1$
	I			I222, $I2_12_12_1$
	C			$C222_1$, C222
	F			F222
Tetragonal	P	$a = b \neq c$, $\alpha = \beta = \gamma = 90°$	4 fold rotation axis parallel to c	P4, $P4_1$, $P4_2$, $P4_3$, P422, $P42_12$, $P4_122$, $P4_12_12$, $P4_222$, $P4_22_12$, $P4_322$, $P4_32_12$
	I			I4, $I4_1$, I422, $I4_122$
Trigonal	P	$a = b \neq c$, $\alpha \neq \beta = 90°$, $\gamma = 120°$	3 fold rotation axis parallel to c	P3, $P3_1$, $P3_2$, P312, P321, $P3_112$, $P3_121$, $P3_212$, $P3_221$
	R			R3, R32
Hexagonal	P		6 fold rotation axis parallel to c	P6, $P6_1$, $P6_2$ $P6_3$, $P6_4$, $P6_5$,P622, $P6_122$, $P6_222$, $P6_322$, $P6_422$, $P6_522$,
Cubic	P	$a = b = c$, $\alpha = \beta = \gamma = 90°$	Four 3 fold axes along diagonals	P23, $P2_13$, P432, $P4_232$, $P4_332$, $P4_132$
	I			I23, $I2_13$, I432, $I4_132$
	F			F23, F432, $F4_132$

This information regarding the alignment of molecules within the crystal, and the extent to which it is symmetrical, can greatly inform crystallographers on the data collection strategies required.

Space in crystallographic terms is described in one of two concepts; real space, R and reciprocal space R*. As with the real lattice, a reciprocal lattice can be constructed which shares it's symmetry with the real crystal lattice but its dimensions are inversely related. In order to construct the reciprocal lattice, the first step is the assignment of lattice planes.

Lattice planes are effectively slices through the crystal that intersect two lattice points as shown in Figure 3.9. A lattice of a given type will have a variety of lattice planes within its structure. The

nomenclature for these planes, is based on a simple system of counting the number of unit cells the plane has to traverse until it meets a lattice point, this is done for all three axes.

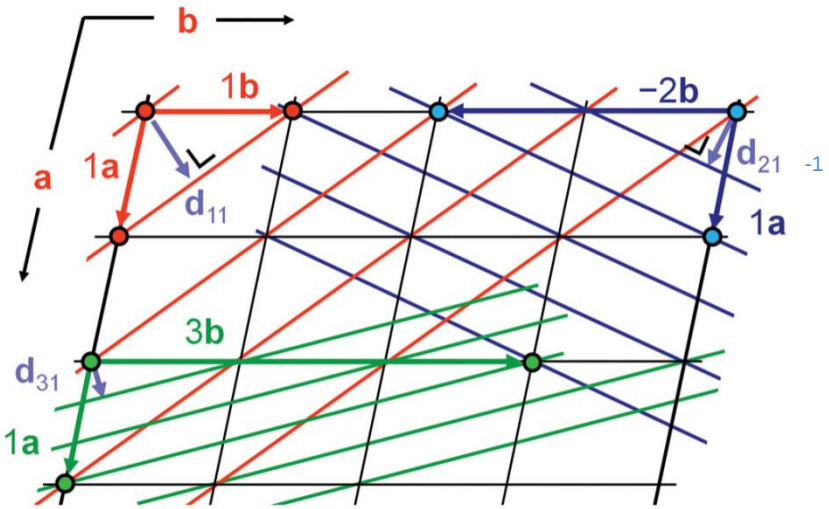


Figure 3.9. A two dimensional representation of three different planes intersecting a lattice d_{11}, $d_{2\text{-}1}$ and d_{31}, using the Miller indexing nomenclature. Coloured in red the d_{11} lattice planes clearly span one unit cell in the a-axis and one in the b axis giving rise to d_{11}. The blue lattice planes, has two planes that intersect the unit cell in the a axes, whereas in the b axes each plane intersects at a lattice point, all be it in the negative, giving rise to the $d_{2\text{-}1}$. The green lattice planes intersect the a axis 3 times and just once in the b axis and thus d_{31}. Edited from (Rupp, 2010).

If using lattice planes in real space, planes that are parallel to the lattice vectors intercept at the infinite, which is why crystallographers use Miller indices, where the reciprocal of these values is taken giving rise to values of 0 for planes parallel to the lattice vectors. Miller indices define the parallel and equidistant lattice planes (h, k, and l), the higher the indices the much closer packed the planes and the more tightly the unit cell is sampled providing greater detail about the sample structure (Rupp, 2010). The real lattice of [0, a, b, c], shares its origin with that of the reciprocal lattice [0, a*,b*, c*], as the planes of real space *hkl* are equivalent to those of reciprocal space $\overline{hkl}$ – the reciprocal lattice is centrosymmetric and retains its lattice type, this relationship is demonstrated in Figure 3.10.

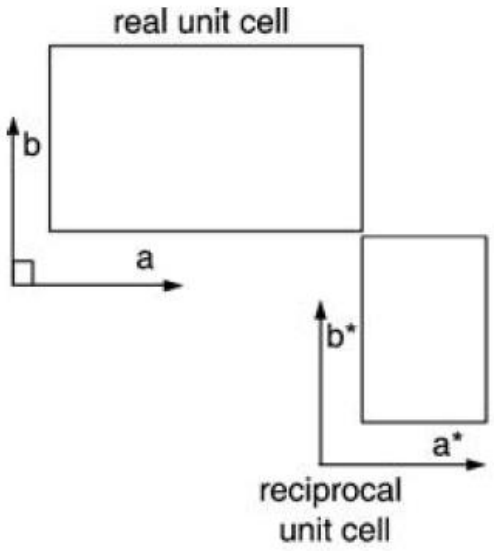


Figure 3.10. The relationship between the real unit cell and the reciprocal unit cell of a orthorhombic crystal system. As all dimensions are inversed the relationship between them stays the same resulting in the retention of lattice type (McPherson, 2009).

A working understanding of the relationships between real space and reciprocal space is vital during the data processing phase, in order to produce an electron density map from the images of scattered x-rays.

3.4 Data Collection

In order to understand how we obtain data, from the crystals we grow we must understand the nature of X-rays and their behaviour when they come into contact with the crystal.

3.4.1 X-ray Sources and Detectors

X-rays are a part of the electromagnetic spectrum that spans the wavelengths of 0.01-10nm or 0.1 to 100Å; and due to the small wavelength these waves are very high energy 100eV to 100keV, with x-rays of around 10keV typically used for x-ray crystallography (Rhodes, 2000). X-rays can be produced by bombarding a metal target plate with an electron stream produced a heated filament, before it accelerated by an electronic field. When a high energy electron collides with

the metal plate the result is the displacement of the electron from a low-lying orbital in the metal atom, to fill the gap an electron from a higher valency orbital drops, in order to do this it needs to liberate itself of its excess energy and does so in the form of an X-ray photon. X-ray sources normally produce more than one wavelength of X-ray as the wavelength is proportional to the energy that needs to be lost and that depends on the orbital of the replacing electron The ideal source for X-rays is one that is monochromatic, that produces X-rays of a single wavelength thus resulting in a single set of reflections.

The three most common X-ray sources seen are the X-ray cathode tube, the rotating anode tube and the particle storage ring, which are capable of producing synchrotron radiation as seen at Diamond Light Source, Oxford, UK. The X-ray cathode tube, as utilised by our in house Oxford Excalibur PX Ultra diffraction instrument, electrons from a hot cathode filament, are accelerated into a water cooled anode of target metal by electrically charged plates. Then ten times more powerful rotating anode tube setups use a rapidly rotating target which thus gives a larger target surface area, and better heat dissipation, the limiting factor of X-ray cathode tube setups. The most powerful X-ray sources, the particle storage rings, have the electrons stored in the giant rings are travelling at near to the speed of light. When a charged body such as an electron moves in a circular motion as seen in these accelerators radiation is given off – synchrotron radiation, in the form of X-rays. Tangential setups to the storage ring, known as beamlines use a system of mirrors, lenses and monochromators provide the experimenter with powerful monochromatic X-rays at a wavelength that can be altered.

Reflections of the X-rays are counted using scintillation counters, that count the photons and give accurate intensity readings, when the scintillator material absorbs an X-ray photon a flash of light is given off and a photocell counts these flashes. Charge-coupled devices (CCDs) are a common detector used in both the small scale Oxford Excalibur setup and large scale Diamond Light Source synchrotron setups. These detectors effectively count the photons that accumulate charge that is

directly proportional to the light that hits them. The CCD detector is coated with phosphors that emit visible light in response to X-ray photons. Specially designed for synchrotron applications the PILATUS3 detector, a hybrid-pixel X-ray detector, is a high performance detector that has improved count rates, which allows for shorter exposure times leading to an overall reduction in the radiation damage conferred to the crystal. The PILATUS3 detector is becoming the gold standard at synchrotron facilities replacing the older CCD detectors (Rhodes, 2000; Wright, *et al*. 2012).

3.4.2 The Behaviour of Waves

Electromagnetic waves such as X-rays, where amplitude of these waves is the electromagnetic field strength and its frequency can be determined as a function of the relationship between the speed of light and its wavelength. When thinking of X-rays it sometimes helps to think of them as photons that the internal resonance frequency that classes them as X-rays. When these photons encounter a crystal, 99% absolutely nothing happens; the remaining 1% of the time sees the photons induce oscillations in the electrons. The electrons then emit a wave of identical frequency, which is a combination of all the scattered waves that have interfered both constructively and destructively. Electromagnetic waves can be thought of as two sine waves that are perpendicular to the wave vector, s_0; which the vector product of the electric field vector E and the magnetic field vector H, depicted in detail in Figure 3.11.

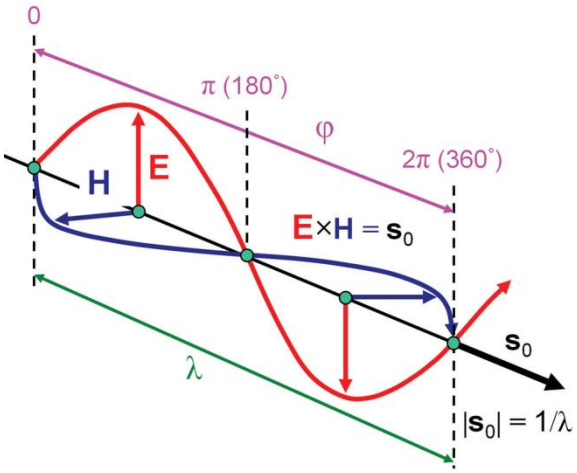


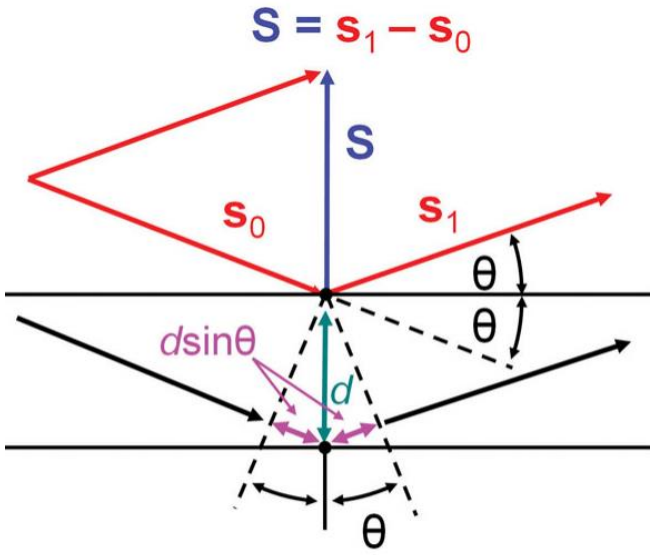
Figure 3.11. The wave or propagation vector s_0, with its two perpendicular vector waves for the electronic field (E) and the magnetic field (H). The wavelength of the vectors is shown in green and is labelled λ. One full wavelength takes 2π radians or 360°, thus any phase variance between two waves is expressed in radians. The amplitude of the resultant wave vector s_0 is defines as $|s_0|=1/\lambda$ (Rupp, 2010).

When the X-rays interact with the electrons of the crystallised molecule, it is only the electric field vector E that stimulates the scattering response. As a result the electromagnetic radiation is described as a classical planar sinusoidal wave. When two waves are combined with one another the result can either be destructive or constructive. For example if two identical waves are out of phase with one another by π or 180°, their peaks and troughs are aligned so that they cancel each other out with a net amplitude of zero. If they are in phase with one another where the phase shift (ϕ) is either 0 or a multiple of 2π the result is that the peaks and troughs combine to form a wave with twice the original amplitude. This is important when we come to look at intensities later.

3.4.3 Crystal Diffraction

3.4.3.1 Bragg's Law

In protein crystallography the key to yielding structural data is the repeating lattice; it acts as a diffraction grating resulting in the scattering of photons in a variety of directions known as orders of diffraction. The father-son combination of W. Lawrence Bragg and W. Henry Bragg proposed a law that describes the reflection from sets of planes within the crystal.



$$2d\,sin\theta = n\lambda$$

Figure 3.12. Bragg's Law describes the reflections of waves from sets of parallel planes within the crystalline structure. S_0 and S_1 are the incident and reflected waves respectively, the difference between which gives the scattering vector S, d is the distance separation between planes, λ is the wavelength of the x-rays which as stated above is the same in both incident and reflected waves, and n is an integer equivalent to the order of diffraction. Diagram edited from (Rupp, 2010).

Because the angle of incidence is the same as the angle of reflection in both planes, two waves that arrive at the point of reflection in phase then become out of phase once reflected because the lower wave of the two has to travel an additional d sinθ in distance either side of the point of reflection, Figure 3.12. For constructive interference to occur n has to be an integer where it isn't destructive interference will occur as the reflected waves will be out of phase (Rupp, 2010). Bragg's Law gives the predicted position of any diffracted X-ray, it does not however yield any information about the intensity or the phase.

3.4.4 Data Collection and Analysis

Getting the data collection parameters correct in x-ray crystallography is key; a crystal may be a non-reproducible once in a lifetime opportunity to extract structural information about an important protein, the difference between 3Å data, where side chain conformations become visible, and 1.5Å data where the central pores on cyclic structures is visible, is stark. The further the diffracted waves are from their origin the greater the number and variety of phases are

present thus weakening the high resolution data. A more intense incident beam leads to a higher achievable resolution as does increasing the exposure time. Both of these factors can lead to increased radiation damage if not carefully monitored and adjusted as well as the increase in the low resolution diffraction spots intensities growing to the point that they become overloaded and thus unusable. In order to collect the full complement of unique reflections the crystal is turned through the phi (ϕ) axis, crystals with high orders of symmetry need to be rotated through a smaller phi angle than those with low levels of symmetry.

The sinusoidal waves used in x-ray crystallography may be analysed as a series of waves that are the combined integral sums of a fundamental frequency known as a Fourier series. A direct Fourier transform from the variation of electron density leads to the calculation of the amplitudes and phases of all its Fourier components. As a result the reverse is also true, reverse Fourier transforms of the amplitudes and phases of the diffracted waves can give rise to the generation of a three dimensional electron density map that the protein molecule fits within and that leads to production of those diffracted waves. Figure 3.13 demonstrates the process a crystallographer must follow from data collection to producing an acceptable model of a protein and any potentially bound ligands or cofactors.

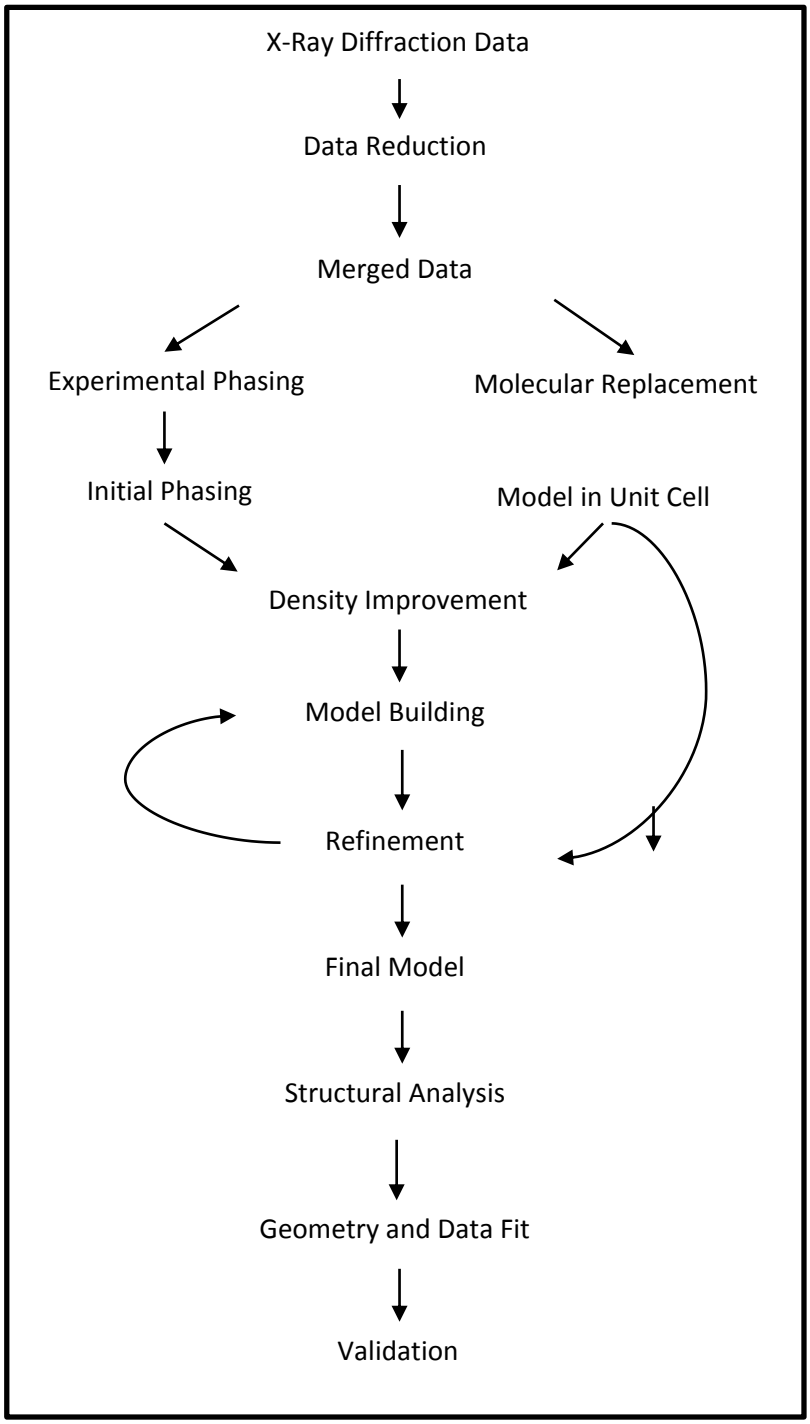


Figure 3.13. Schematic representation of the data processing process carried out by crystallographers to elucidate a structural model from their data

Each diffracted photon that the detector intercepts can be described in terms of the contributions of all of the scattering agents within the unit cell. The sum of these reflections *hkl*s is the structure

factor - F_{hkl}. This is important as the number of photons that terminate at a given position on the detector is dependent upon the scattering probability in that given direction, and to determine this the structure factor for each reflection needs to be calculated. Knowing the number of photons that accumulated provides an accurate measure of the proportionality to the structure factor intensity, and represents the squared structure factor amplitude and as a result can be accessed directly from the data. The result is that only the amplitudes of the reflections are measured and as such phase information is lost upon detection. This is known as the phase problem (Rupp, 2010).

3.4.4.1 The Phase Problem

The missing phase information that is lost upon diffracted photon detection results in the construction of an unusable electron density map as demonstrated by Figure 3.14 below.

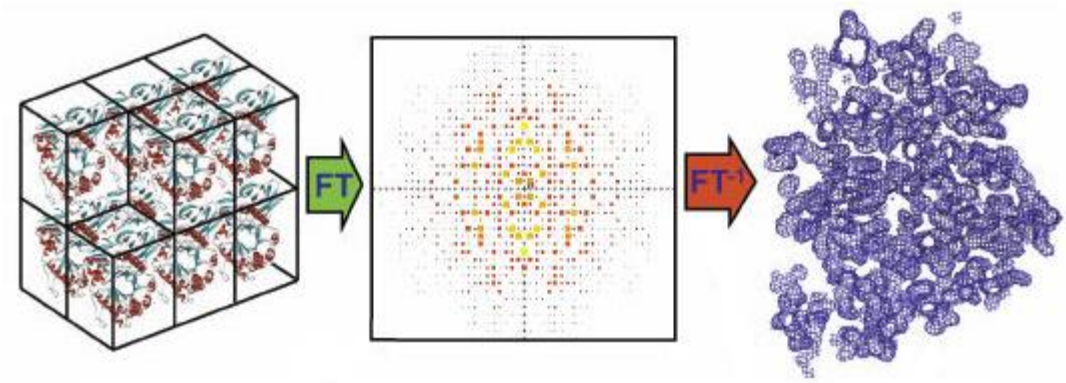


Figure 3.14. How Fourier transforms and reverse Fourier transforms can transform the physical repeating blocks of the unit cell into the diffraction patterns seen upon exposure of the crystals to x-rays and then from the relative positions and intensities of those diffraction spots to construction of the electron density map (Rupp, 2010).

As phase is not discernible from the diffraction patterns produced it must be calculated by other means. If only a small fraction of the diffraction phases can be correctly deduced via the correct location of one or several atoms in their relative positions, this is often enough to commence the

structure solution. All existing phasing techniques initially provide a set of approximate phases that are derived from additional experimental data.

Phasing methods depend upon the determination of marker atom substructure, atoms within the protein superstructure that provide a source of anomalous differences in the diffraction data - typically heavy metal ions that are either native to the protein structure or have been soaked in experimentally. This is technique is often referred to as isomorphous replacement.

The result is a difference in atomic scattering factors ultimately resulting in a variation in the structure factor intensities relative to the reference structure. The variation in these data is then used to locate the source marker atoms within the structure. The creation of variation in intensities between data sets that lack and contain the marker atoms allows the problem to be reduced to solving a structure of fewer atoms compared to those of the whole protein (Perutz, 1956). Dispersive and anomalous scattering factor contributions cause this intensity variation; by the variations that exist between the diffracted intensities of pairs of the same reflection recorded at the same wavelength and difference that arise between intensity/structure factor amplitudes between members of a *Bijovoet pair* (Perutz 1956).

Shown in the Harker diagram in Figure 3.15 below (Rupp, 2010), the combination of structure factor amplitudes for the heavy atom structure F_A, and the native protein F_P is equal to the structure factor for the derivative protein, F_{PA}.

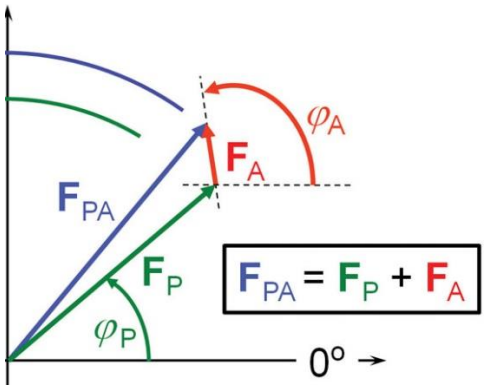


Figure 3.15. The complex structure factors for protein (F_P) in green, derivative (F_{PA}) in blue and heavy atom (F_A) in red. Intensity measurements only reveal the magnitude of the stricture factors for F_{PA} and native F_P, but from the difference data positions of the heavy atoms can be determined. Therefore, the entire complex structure factor for heavy atom F_A is known, and two possible solutions for the phase angle can be deduced.

Anomalous scattering is a second means of obtaining phases from heavy-atom derivatives which utilises the capability of the heavy atoms to absorb x-rays of a specific wavelength. Due to this absorption Friedel's law does not apply and the result is that reflections hkl and -h-k-l are no longer of equal intensity. It is well established that atoms as well as diffracting x-rays also absorb them, absorption of x-rays however drops significantly at wavelengths just below the characteristic emission wavelength and this drop-off point is referred to as an absorption edge, Figure 3.16.

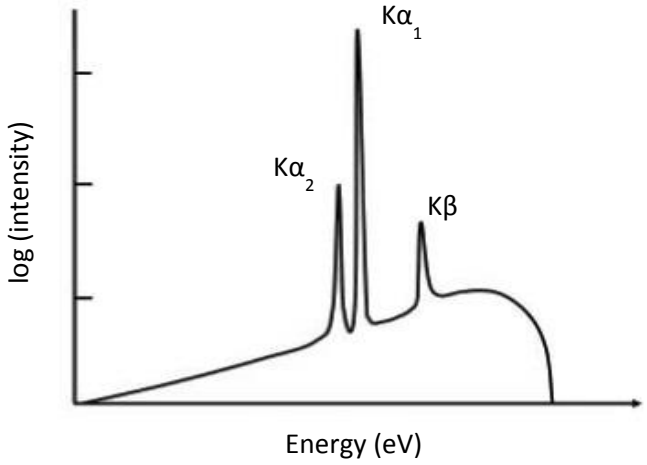


Figure 3.16. An x-ray emission spectrum. When an anode material, such as copper, is bombarded with high-energy electrons, a characteristic emission is observed. Characteristic radiation emanates when holes in core shells generated by the bombardment are filled with electrons from upper shells.

Elements exhibit anomalous scattering when the x-ray wavelength nears this edge. As the wavelength of emissions are longer for elements with lower atomic numbers, light atoms such as carbon, nitrogen and oxygen do not contribute to anomalous scattering, but absorption edges of heavier atoms in this range do. Synchrotron sourced x-rays are tuneable to a chosen wavelength, and as such data can be collected under conditions that maximise anomalous scattering by heavy atoms. As with isomorphous displacement, anomalous scattering offers a direct way of estimating phase through the measurement of differences between normal and anomalous scattering; but rather than comparing the differences between structure factors of native and heavy atom-soaked proteins, comparisons are made between intensities of (unequal) Friedel pairs. When the electron transition occurs in heavy atoms, the corresponding atomic scattering factor, f, behaves as a complex number, appearing with two correction terms: a real number (f') and an imaginary one (f''), whose total value is dependent on the frequency of the incident photon. In order to interpret how the correction factors modify the scattering factor, it is necessary to understand that that f' is 180° out of phase with normally scattered waves, and f'' is 90° out of phase.

The equation for the atomic scattering factor, f, during anomalous dispersion can be written in terms of f_0, the normal atomic scattering factor:

$$f = f_0 + f' + if''$$

If all the atoms in a structure are heavy atoms and susceptible to anomalous dispersion, Friedel's law is broken due to changes in their phases, although their intensities are still equal:

$$I_A(h,k,l) = I_A(-h-k-l)$$

And as such the structure factor of the reflection has a magnitude proportional to the square root of the observed intensity: $F_A(h,k,l) = F_A(-h-k-l)$.

However, not all the atoms in the structure show anomalous dispersion of x-rays, and these differ from heavy atoms in both intensity and phase. As with isomorphous replacement, exacting phases using anomalous dispersion requires comparing three sets of data: one with native crystals to establish amplitudes for each reflection, F_P; one for a heavy atom derivative to establish amplitude for F_A; and finally one with a different wavelength to maximise anomalous scattering by the heavy atoms. The non-equivalence of Friedel pairs enable the crystallographer to gather the phases of the heavy atoms data, which is then used to derive structure factors of the native data.

Similar to anomalous scattering, multiwave anomalous dispersion (MAD) exploits the effects of anomalous contributions by using Friedel pairs, but also utilises dispersive differences (that is, differences that exist between the diffracted intensities of pairs of the same reflection) between data sets collected at different wavelengths. To optimise the differences, an x-ray excitation scan for the phasing element must be recorded to define the absorption edge, and at least two MAD wavelengths are selected: one that maximises the anomalous signal, and a second that optimises the dispersive differences between wavelengths. This will give a peak data set and an inflection data set, corresponding to the positive and negative peaks of the absorption edge respectively.

Relative intensity differences exist between data recorded at different wavelengths due to anomalous contributions in the structure affecting the atomic scattering factors. In addition to wavelength-dependent differences between Friedel pairs, individual reflection intensities vary slightly with wavelength (dispersive differences), which also contain phase information that can be extracted by solving equations much like those for anomalous diffraction (Hendrickson, 1985).

Another method bypasses the phase problem by using structures for which the phases are already known. Molecules that share homologies in domains and folds types that are often family members allow the matching of homologous regions leading to the estimation of the phases (Rossmann, & Blow, 1962; Rossmann, 1990). Using the suite of programs – CCP4 (Collaborative Computational Project No. 4), electron density maps are produced and then undergo several rounds of refinement before being ready to publish

3.5. Crystallisation of α2-macroglobulin from *Limulus Polyphemus*

The path to ascertaining successful crystallisation conditions for a protein is often a long one, for example the first published studies on the human α2m molecules were seen in 1991 (Andersen, *et al*. 1991) but not until 2012 was the structure solved (Marrero, *et al*. 2012) and at what was at the lower end of the scale. Between these years numerous other papers were published with a variety of conditions until ultimately the successful conditions were published. As there are no known conditions for the crystallisation of *Limulus* α2m, a combination of strategies was used. Of the many commercially available kits Molecular Dimensions Ltd.'s Structure Screens 1 and 2 as well as cacodylate free Structure Screens 1 and 2, were used to find initial starting conditions for crystallisation. Structure Screens contain 50 crystallisation conditions each, that were derived from known to result in the crystallisation of proteins they use a broad range of buffer, pH, precipitants, salts and additives (Jancarik, & Kim, 1991). In addition the crystallisation conditions

for the human α2m were used too as a potential starting point (Marrero, *et al*. 2012). To these conditions the pre-concentrated and purified *Limulus* α2m-MA was added in equivolumetric drops with the crystallisation conditions, at a starting concentration between 5 and 8 mg/ml set up in a sitting drop method.

If a condition produced crystals the crystals were then stained to ascertain the likelihood of them being protein or small molecule. Staining the crystals utilises the ability of dyes to move into the solvent channels that run through protein crystals. The majority of protein crystals take up the dyes and as a result assume the deep violet colour. The commercially available stain Izit^{TM} and Crystal violet were used in assessing the potential of crystals grown. Staining protein crystals isn't fool proof however as some protein crystals will not take up the dye as seen in work done by this research group with rfhSP-D. In experiments when the protein does take up the dye the cationic components of the dye once inside the solvent channels bind to the negatively charged amino acids on the proteins surface.

Table 3.2. Conditions for the successful growth of crystals. Crystals of some description (protein/small molecule) were grown in the following wells, with all conditions being from or a derivative of the Molecular Dimensions Ltd Structure Screens. SS1-indicates structure screen 1, SS2-indicates structure screen 2, and SS1CD-indicates the cacodylate-free structure screen 1. Full details of components and tray composition is laid out in the appendices.

Crystal Tray	Well	Conditions	Outcome
MN01	A3	SS1-3	Small Mol.
MN01	C2	SS1-14	Dissolved
MN01	D3	SS1-21	Small Mol.
MN02	A3	SS1-27	
MN02	B4	SS1-34	Small Mol.
MN07	A2	Based on SS1-14: 0.1M Na Cacodylate pH 6.5, 20% (w/v) PEG8K, 0.2M $(NH_4)_2SO_4$	Small Mol.
MN07	C1	Based on SS1-14: 0.1M Na Cacodylate pH 6.5, 40% (w/v)	Small Mol.

		PEG8K, 0.05M $(NH_4)_2SO_4$	
MN07	C3	Based on SS1-14: 0.1M Na Cacodylate pH 6.5, 40% (w/v) PEG8K, 0.5M $(NH_4)_2SO_4$	Small Mol.
MN19	C2	SS1CF-14	Small Mol.
MN20	A4	SS1CF-28	No Diffraction
MN20	B1	SS1CF-31	
MN20	C5	SS1CF-41	Dissolved
MN20	C6	SS1CF-42	Small Mol.
MN20	D1	SS1CF-43	
MN21	B4	SS2-8	
MN21	C1	SS2-11	
MN21	D5	SS2-21	Small Mol.
MN22	C5	SS2-39	

Using the sparse matrix approach of the structure screens from Molecular Dimensions Ltd yielded a number of hits for potentially suitable conditions for producing crystals. However the crystals that were produced from the structure screens and their derivatives that survived cryoprotection and were tested upon either turned out to be small molecule or showed no diffraction at all.

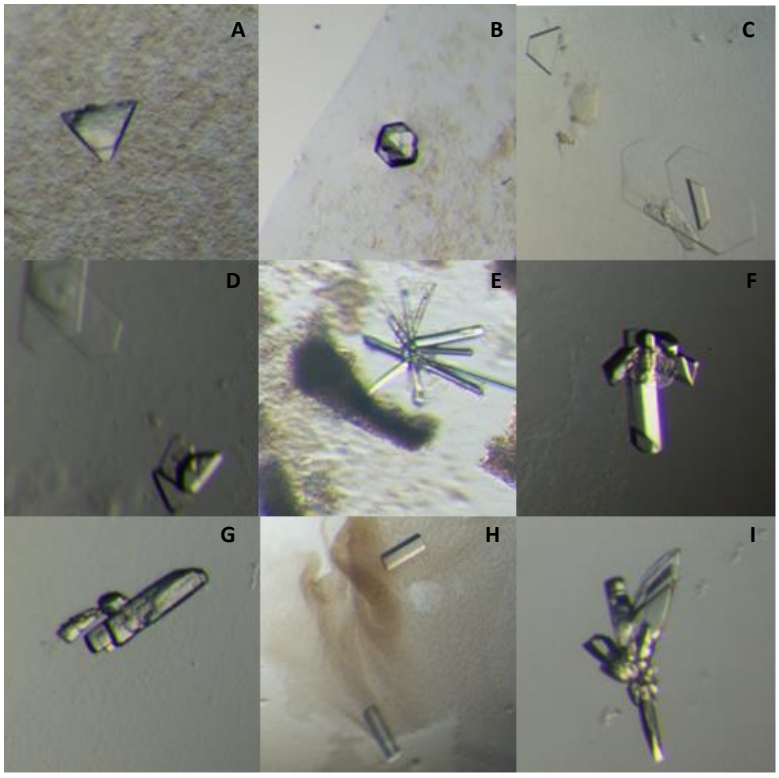


Figure 3.17. A selection of crystal images from trials using the Molecular Dimensions Ltd Structure Screens 1 & 2 as well as Structure Screen 1 Cacodylate-free. A – Small flat pyramidal crystal grown in MN01A3 (SS1-3), tested at Diamond Light Source (DLS), was determined to be small molecule. B – Small hexagonal crystal grown in MN01C2 (SS1-14), dissolved upon addition of cryoprotectant. C – Thin trapezoid morphology of crystals grown in MN01D4 (SS1-21), tested at DLS, proved to be small molecule; and trapezoid morphology found in other wells containing tri-sodium citrate. D – Thicker trapezoid crystals found in MN02B4 (SS1-34), which also contains tri-sodium citrate. E – Typical morphology of ammonium sulphate crystals found in MN07C3, which had high levels of $(NH_4)_2$ SO4 (0.5M). F – Multiple crystal from MN20A4 (SS1CF-28), tested at DLS but no diffraction. G – A poorly formed crystal in MN20C5 (SS1CF-41), that dissolved upon contact with cryoprotectant. H – Another example of ammonium sulphate crystals grown in MN20C6 (SS1CF-42), untested. I – An unusual multiple crystal grown in MN21D5 (SS2-21), tested at DLS, proved to be small molecule.

Whilst a broad range of morphologies of crystals were grown utilising the structure screen conditions, demonstrated in Figure 3.17, not one of the screened conditions when tested showed protein spots. Fortunately the work that was based on the human conditions was more successful in this regard (Marrero, *et al.* 2012).

Table 3.3. Partially successful crystallisation conditions for the growth of crystals using conditions derived from those used in the crystallisation of human α2-macroglobulin (Marrero, *et al*. 2012). For full crystal tray composition sees the appendices.

Crystal Tray	Well	Conditions

MN04	A2	0.2M Ammonium Citrate, 15% (w/v) PEG 2000, 50mM NaF
MN04	A3	0.2M Ammonium Citrate, 15% (w/v) PEG 2000, 75mM NaF
MN04	B5	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF
MN04	C4	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF
MN04	D3	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF
MN09	C1	0.2M Ammonium Citrate, 25% (w/v) PEG 2000, 50mM NaF
MN09	C2	0.2M Ammonium Citrate, 25% (w/v) PEG 2000, 75mM NaF
MN09	C3	0.2M Ammonium Citrate, 25% (w/v) PEG 2000, 100mM NaF
MN11	A4	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF
MN11	C1	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF
MN11	D1	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF
MN11	D3	0.2M Ammonium Citrate, 15% (w/v) PEG 3350, 50mM NaF

Initial trials based on the human conditions of 0.2M Ammonium Citrate pH 6.4, 15% (w/v) PEG 3350, 50mM NaF (Marrero, *et al.* 2012), that varied molecular weight PEGs from 2000-4000 as well as varying the concentration of NaF from 25mM to 75mM. Following the appearance of crystals in some wells, additional trays and wells were added to screen around those conditions.

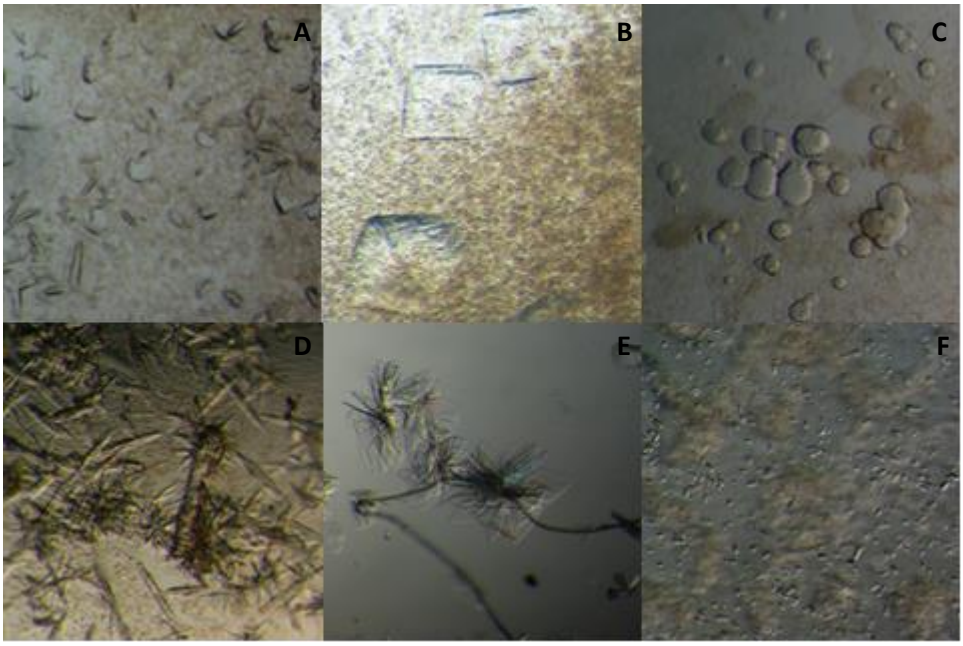


Figure 3.18. A selection of different crystal morphologies from crystal well conditions based on the successful crystallisation of the human α2m (Marrero, *et al*. 2012). A – Irregular shaped crystals grown in MN04A2, tested at DLS, protein spots to low resolution (20Å). B – Large rectangular crystals grown in MN04A3, tested at DLS, protein spots to 8-9Å. C - Irregular round crystalline objects grown in MN04B5, these were tested with stain, which they did not take up and so were untested. D - Many irregular needles with poorly defined edges grown in MN09C3. E – Extremely thin kite shaped crystals that fan out around a single point of growth, grown in MN11A4 and tested at DLS with no diffraction visible. F – Small 'rugby ball' shaped crystals grown in MN13A2 tested at DLS with no diffraction visible. Full details of the well conditions available in the appendices.

The crystals in Figure 3.18 represent some of the less successful crystal morphologies grown using the human conditions as a basis (Marrero, *et al*. 2012). Some proved to be protein via the presence of diffraction spots, whereas some showed no diffraction at all. Those conditions that grew crystals that proved to be protein were screened around further. The most successful well that was laid down was MN04D3. The conditions for this well were 0.2M Ammonium Citrate pH 6.4, 15% PEG 3350, 50mM NaF, with protein stocks at 8.319mg/ml used. Following the initial success seen in MN04A3, which showed protein spots to 8-9Å, additional wells were laid down varying the percentage weight to volume of the PEG in row D of the crystal tray. Six weeks after the wells were laid down crystal growth was discovered in MN04D3.

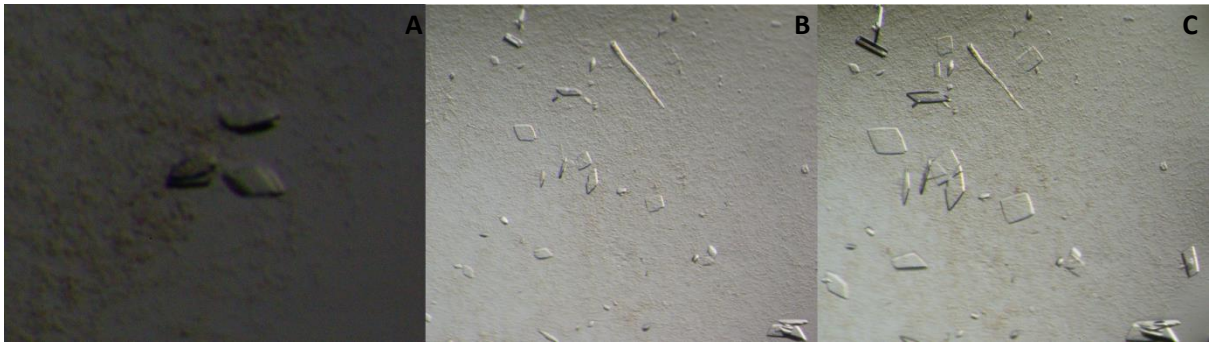


Figure 3.19. Well progression of crystals found in MN04D3. A – The first crystals detected were poorly ordered and diamond shaped. B – Four days later other crystals had begun to form the majority of which fit the diamond shape morphology but were well ordered with clean edges. C – Another four days later and the crystals had grown further with one or two really promising crystals present.

Six weeks after the wells were set up the beginnings of the diamond shaped crystals visible above were present. As time progressed the crystals continued to grow and improve in quality until the appeared to grow no more (Figure 3.19 –C and Figure 3.20).

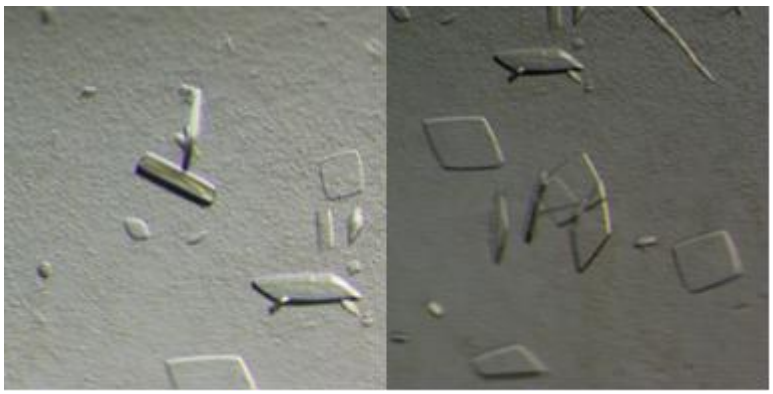


Figure 3.20. Crystals from MN04D3 prior to cryoprotection and testing at DLS. Crystals are better ordered than those seen in MN04A3 which shared a similar morphology.

The well was selected for testing at Diamond Light Source, prior to data collection and freezing the crystals needed to be cryoprotected. 2- Methyl, 2,4-pentanediol (MPD) was selected as a cryoprotectant due to trials carried out with crystals from similar conditions such as MN04A2 and A3. MPD was added by 5% incremental increases until 20% MPD/well solution was exchanged in the droplet (10µl taken out and 10µl added. Once the exchange had taken place the crystals were allowed to soak in the cryoprotectant with a view to as much as possible moving into the solvent channels thus offering protection during the freezing process.

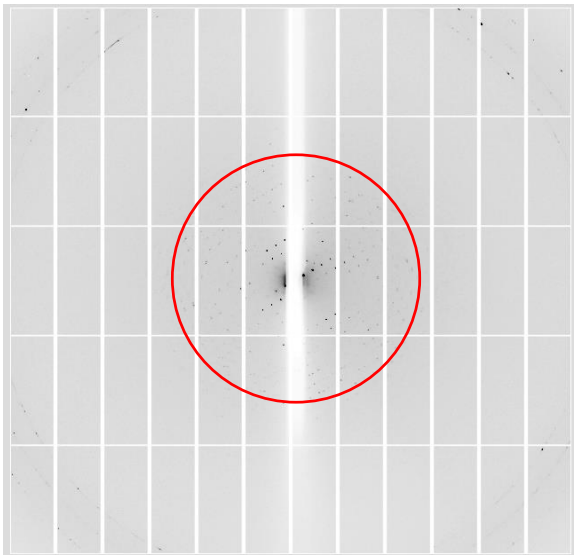


Figure 3.21. Diffraction image from the data collection of MN04D32 with diffraction spots initially to 6Å, on beamline I02 at Diamond Light Source, Oxford, UK. The red circle indicates the extent of the low resolution diffraction spots. Higher resolution spots would be expected outside the red circle.

Crystals were frozen and taken to DLS on beamline I02 where data was collected to 6Å (Figure 3.21) on MN04D32 (the second crystal looped out from MN04D3) as well as two other data sets of slightly lower resolution.

3.5.1. Discussion

Unfortunately, as you will see if you tried to search the Protein Databank for a structure at 6Å, this data lacks the resolution required to build a model using the data processing software, this was however sufficient enough for EDNA (**E**nhanced automate**D** collectio**N** of dat**A**) a program available at both DLS and the European Synchrotron Radiation Facility (ESRF), that proposed a strategy for data collection based on the radiation damage and symmetry information obtained from the initial few images (Incardona, *et al*. 2009). EDNA suggested that the crystals were in the P222 space group and had unit cell dimensions of a = 115Å, b = 141Å and c = 338Å. Table 3.1 shows that P222 is a primitive orthorhombic space group which has unit cell dimensions of $a \neq b \neq c$ and $\alpha = \beta = \gamma = 90°$. Interestingly the human model (Marrero, *et al*. 2012) has an orthorhombic space group of $P2_12_12_1$ and unit cell dimensions of a = 130.7Å, b = 260.3Å, and c = 281.8Å. Whilst both the human and the *Limulus* asymmetric units undergo two symmetry operations in each of

the three axes the human asymmetric unit's symmetry operations are screw axes as opposed to simple rotations around a point as seen in *Limulus* based on EDNA's prediction.

Future experiments should aim to build on the work carried out here; firstly by replicating the crystals grown in MN04D3 and then looking at how the conditions can be adjusted to improve the resolution to at least 4Å but preferably higher. Key to this is the homogeneity and purity of the protein sample being used. Not only is the α2m purified from a serum that contains two significantly more abundant contaminants (haemocyanin and the pentraxins) but given the nature of the structural differences between reacted and unreacted, simply having purified α2m is not enough, as a result every effort should be made to ensure its homogeneity (Armstrong, & Quigley, 1999). Secondly there are a number of known conditions that were published that showed diffraction with the human α2m with a variety of ligands.

Table 3.4. Known conditions to provide protein crystals of the human α2m with a variety of ligands to low resolution when exposed to synchrotron X-rays.

Reference	Protein	Conditions	Resolution
Andersen, *et al*. 1991	α2m-MA	20-30% $(NH_4)_2SO_4$ pH 7.3-7.8, 0.15%-1.5% β-octyl-glucoside, Tris-HCl	9Å
	α2m-Trypsin	20-25% $(NH_4)_2SO_4$ pH 7.0-8.0, OR 1.5M Mg SO_4 pH 6.5-7.5, Tris-HCl	10
	α2m-Plasmin	1.3-1.5M MgSO4 pH 6.75-7.25, Tris-HCl MOPS	11
Andersen, *et al*. 1994a	α2m-MA	10% MPD, 0.5mM $ZnCl_2$, 1M NaCl, pH6-7, at 4°C	8.5
Andersen, *et al*. 1994b	α2m-MA	20mM Tris-HCl pH 7.7, 1.0M NaCl, 21mM MES, 0.5mM ZnSO4, 13% MPD, 2:1 protein to reservoir ratio, at 4°C	9
	α2m-MA	20mM Tris-HCl pH 7.7, 0.5M NaCl, 21mM MES,	15

		0.5mM $ZnSO_4$, 13% MPD, 2:1 protein to reservoir ratio, at 4°C	

This in addition to the existing condition (0.2M Ammonium Citrate pH 6.4, 15% PEG3350, 50mM NaF) that diffracted to 6Å in this body of work means there are a number of potential avenues to be explored during the continued pursuit of the crystal structure of *Limulus* α2m.

Crystallography is a discipline that truly tests the patience of its practitioners; in 1991 the first published data of the crystallisation of human α2-macroglobulin (Andersen, *et al*. 1991) with crystals being diffracted to 9Å, the culmination of years of work was ultimately published by the same research group in 2012 (Marrero, *et al*. 2012) to a resolution of 4.45Å. Using this for perspective when looking at the work carried out in this thesis it is clear that whilst there is still work to be done to reach a resolution that will provide a workable dataset and ultimately an structural model, it is clear that a solid starting point has been found for the continued research into the structure of *Limulus* α2-macroglobulin.

Chapter 4 - Bioinformatic Analysis and Structure prediction of α2-Macroglobulin from *Limulus Polyphemus*.

4.1. Introduction

Bioinformatics provide powerful tools for the analysis of biological molecules, utilising multiple computing techniques with common uses including: proteomics, phylogeny studies, genomics, and sequence analysis of genes and the proteins coded by them. ExPASy is an online hub of bioinformatics resources providing bioinformatic data for the above mentioned studies and providing a great number of tools for this study.

An ever blossoming field of bioinformatics lies in protein structure prediction. Due to the challenging nature of structure solution by x-ray crystallography, tools have been developed to provide highly accurate means of predicting the structure of a protein of known amino acid sequence, as an intermediate until, if possible, the structure is solved by experimental means. There are currently a wide range of programs available for the prediction of protein structure and knowing which one to use and whether its results are to be trusted or not is a difficult decision to make. However this is aided by The Critical Assessment of protein Structure Prediction (CASP). CASP can best be thought of as a competition and a quality assurance method for the structure prediction suites available. Running biannually since 1994 CASP invites the programs to take part in a blind assessment of their structure predicting abilities. Now in its 10^{th} iteration with the results of the 11^{th} pending it allows interested parties to assess the field and select the method of structure prediction that best suits their requirements. For the upcoming structure prediction work done, the Zhang labs I-TASSER server was used. More of an amalgamated suite of programs than one single program, it has been ranked 1^{st} place in every CASP since its inception CASP7-10. For this reason as well as its ability to produce models for extremely large sequences such as that of α2m it was the obvious choice.

4.2. Sequence homology in the α2-Macroglobulin Superfamily

Homology between amino acid sequences can infer a great deal of information about the molecules of interest, highlighting similar regions between molecules of the same family within the same organism and those family members from other organisms. If pre-existing structural data exists for one of the aligned sequences it can be used to deduce potential structural and functional features of the sequences that lack high resolution structural data.

The sequence of α2-macroglobulin can be found using the ExPASy program UniProtKB under the code of O01717. This yields a 1507 residue long amino acid sequence that has since been proven experimentally to be 1482 with a 25 amino acid signal peptide (Iwaki, *et al.* 1996). This sequence was then fed into the ExPASy program BLAST, which searches through the UniProtKB database seeking out homologous sequences and ranking them by percentage homology.

Table 4. The various members of the α2m superfamily, from various species showing their sequence length and their percentage homology to the α2m of *Limulus polyphemus.*

UniProtKB code	Organism	Protein	Sequence length (aa)	% Sequence Homology
O01717	*Limulus polyphemus* (Horseshoe crab)	α2-macroglobulin	1482	100
B8R3M2	*Ixodes ricinus* (Common tick)	α2-macroglobulin splicing variant	1486	43
E2AC15	*Camponotus floridanus* (Florida Carpenter Ant)	α2-macroglobulin-like protein	1762	37
G9BIX6	*Pacifastacus leniusculus* (Signal Crayfish)	α2-macroglobulin 2 isoform 3	1598	33

K9J6H8	*Sus scrofa* (Pig)	α2-macroglobulin	1478	33
Q2VJB3	*Xenopus laevis* (African clawed frog)	α2-macroglobulin	1474	33
Q61838	*Mus musculus* (Mouse)	α2-macroglobulin	1495	33
Q5R4N8	*Pongo abelii* (Sumatran Orangutan)	α2-macroglobulin	1474	32
P01023	*Homo sapien* (Human)	α2-macroglobulin	1474	32
P20742	*Homo sapien* (Human)	Pregnancy Zone Protein (PZP)	1482	31
Q6YHK3	*Homo sapien* (Human)	CD109	1445	30
W5J4Q1	*Anopheles gambiae*	Thiol-ester containing protein	1397	28
P01024	*Homo sapien* (Human)	Complement component C3	1663	20
P01031	*Homo sapien* (Human)	Complement component C5	1676	19

Table 4.1 clearly shows that α2-macroglobulin is found in a wide variety of species from various phyla and is found in both vertebrates and invertebrates. The high homology levels between species not only indicates a high level of importance for the molecule but is also highly suggestive

of the molecule being evolutionarily ancient, further evidenced by its presence in the phylogenetically ancient *Limulus polyphemus*.

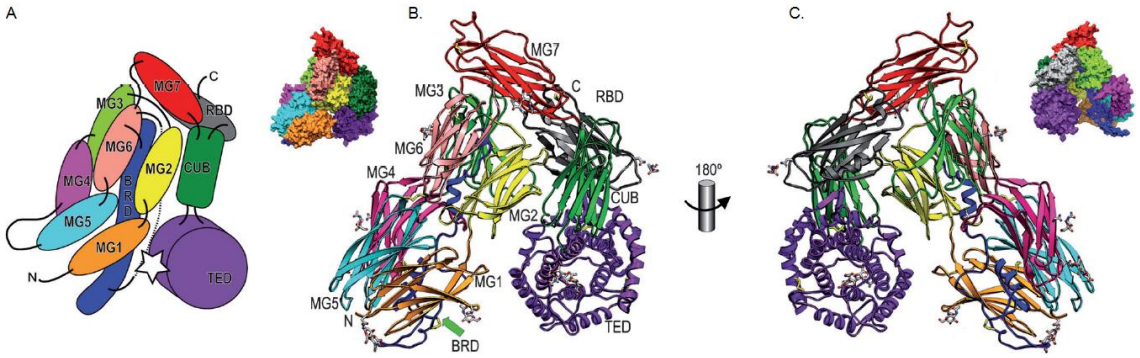


Figure 1.6. A) A schematic representation of the domain organisation/tertiary structure of a subunit of human α2m; with the bait region represented as a dashed line and the star demonstrating a potential protease cleavage site. B) Ribbon and space fill representations of the tertiary structure of an α2m subunit (convex face) with the green arrow indicating the location of the bait region. C) The back/ concave face of the α2m subunit from humans demonstrating the bait region to the rear of the molecule. Diagrams edited from (Marrero, *et al*. 2012).

In order to infer some of the possible structural characteristics of *Limulus* α2m more detailed comparisons need to be made with those family members whose structures are known and currently are available in the PDB. Initially a sequence alignment with the human homologue α2m, the sequence of which was obtained from the UniProtKB database on the ExPASy server, was performed using the ExPASy program CLUSTAL O. The aligned sequences were then broken down according to the domain structure of the human molecule: macroglobulin domains (MG) 1-7, the bait region domain (BRD), the CUB domain, thiol-ester domain (TED) and the receptor binding domain (RBD), as per Figure 1.6, repeated above for clarity, to be able to properly compare both secondary and tertiary structure of the proteins. In all of the alignments described below residues in red indicate their presence in a helix, whereas blue residues are found in beta strands in accordance with the findings of Marrero et al. (Marrero, *et al*. 2012) for the *Limulus* α2m molecule the predicted secondary structure from the I-TASSER suite was used (Zhang, 2008; Roy, *et al*. 2010).

4.3. Structure Prediction of *Limulus* α2-Macroglobulin

4.3.1. Introduction to the Protein Structure Prediction Server I-TASSER

I-TASSER is an open source suite of programs available online via a server rather than direct download (http://zhanglab.ccmb.med.umich.edu/I-TASSER). It is capable of processing sequences up to 1500 residues in length, following submission of the sequence the I-TASSER server will email to inform the user that their query is complete. The completed query yields a mass of data, from the predicted secondary structures and solvent accessibility of the residues to up to five potential models with assigned confidence scores. In addition to this the I-TASSER results also inform the user of the templates it used from the PDB during the model building. This information can often yield clues to the structure, function and family of proteins to which the query sequence belongs. Querying of I-TASSER with the sequence from *Limulus* α2m, showed the top structural homologues from the PDB of which there were nine were with their respective PDB codes: 1 - TEP1r from *Anopheles gambiae* (PDB code - 2PN5), 2 - Complement C5 - in complex with SSL7 from *S. aureus* (3KLS), 3 - Bovine Complement C3 (2B39), 4 - Complement C5b6 (4A5W), 5 - Human α2m 4ACQ, 6 - Complement C5 complexed with cobra venom factor (3PVM), 7 - Complement C4 in complex with MASP-2 (4FXG), 8 - Mammalian (*Sus scrofa*) Fatty acid synthase (2VZ9), 9 - *Streptomyces avermitilis* α-L-rhamnosidase. Of these only fatty acid synthase and α-L-rhamnosidase are not part of the α2m superfamily of thiol-ester containing proteins.

The first stage of I-TASSER processing is threading. Here template proteins are identified from the solved structures in the PDB. The first stage, involves the query sequence passing through a non-redundant sequence database PSI-BLAST which identifies evolutionary relatives of the sequence and thus those most likely to share structural motifs. The result is the creation of a sequence profile, based on the multiple alignment of all the sequence homologues. This sequence profile is then utilised to produce the secondary structure prediction using PSIPRED. With the aid of both the sequence profile and the predicted secondary structure the query sequence is then threaded through a representative PDB structure library using the I-TASSER suite LOMETS, a locally installed

meta-threading server which is made up of 7 state-of-the-art threading programs (FUGUE, HHSEARCH, MUSTER, PROSPECT, PPA SP3, and SPARKS). Each individual program within the LOMETS server ranks the templates by a variety of sequence and structure based scores. The templates with the highest scores from each program are then selected for further consideration where the quality of their alignment is judged based on the statistical significance of the best threading alignment (Roy, *et al*. 2010). Stage two of I-TASSER processing results in the structural assembly of the query sequence model. Continuous fragments from the template structures in threading alignments are used to construct structural conformations of the well aligned regions. The unaligned sections, predominantly loop regions and tails, are built via *ab initio* modelling. A modified replica-exchange Monte Carlo simulation is used for fragment assembly which utilises; statistical data from the PDB, spatial restraints from threading templates and sequence based contact predictions from SVMSEQ. Conformations that were generated in the low-temperature replicas during the refinement simulation are clustered using SPICKER, to identify low-free energy states. The average of the 3D coordinates of all the clustered structural decoys is then to obtain cluster centroids. In the third stage, model selection and refinement, the cluster centroids produced earlier are fed back into the fragment assembly simulation, this removes steric clashes in addition to refining the global topology of the cluster centroids. The output is a second generation of decoys that are once more clustered before the lowest energy structures are selected to be input into REMO which produces the final structural models by building the full model from C_α traces from the hydrogen-bonding networks. The fourth stage of I-TASSER gives an inference to the function of the query molecule by structurally matching it to the same existing PDB structures that provided templates in the earlier stages. In addition to using those with similar global folds, the software also highlights those that differ in global fold but that do/may share a conserved active/binding site. Functional analogues are ranked using a number of criteria such as: TM-score (template modelling score, which is a measure of similarity of the two structure with values that range from 0 to 1), RMSD, and sequence identity. The TM-score is potentially the

best indicator in terms of modelling accuracy where a TM-score < 0.17 indicates a randomly selected protein, whereas a TM-score > 0.5 identifies the query protein and the template as sharing similar folds. The main assessment of prediction accuracy however is given as a confidence or C-score for each model produced by I-TASSER. The C-score is based on the quality of the threading alignments coupled with the convergence of the I-TASSER's structural assembly refinement simulations. The C-score has been extensively tested in large-scale bench marking tests. This showed that the Pearson correlation between the C-score and the TM-score, which is equal to the absolute difference between the model and the native structure, was equal to 0.91 which is a highly significant value when we consider that Pearson values range from 0 for random variables to 1 for identical variables. If a cut off score for the C-score of -1.5 is used the result is that 90% of quality predictions are correct.

It should be noted that during its development I-TASSER has been optimised for modelling single domain proteins, however there is still a built in process for the prediction of multi-domain models. LOMETS defines the domain boundaries. It does so by assessing a segment of query sequence, if >80 residues gave no alignment with template proteins in the top two threading hits, it is identified as a multiple domain protein, and the domain boundaries are defined by the boundaries of aligned/unaligned regions. Following this, two types of assembly simulations are run, one that aims to model the whole chain with a view to guide the domain orientations in the tertiary structure of the molecule and the other that models each domain individually. Then to generate the full length model the models of the individual domains generated previously are docked together using the data from the whole chain simulation as a template. This docking simulation is performed to produce a model that has similar domain organisation whilst producing the minimal amount of steric clashes. However this process is only carried out in instances that are multi-domain but that only have partial alignment with the top-scoring templates. If the top scoring templates are multi-domain as well and all of the domains from the query sequence align, the whole chain is modelled in I-TASSER using the full chain technique.

I-TASSER allows its users to specify whether or not they want their query structure to be based on a particular template. For example it may in this case have been advantageous to model the *Limulus* α2m on the human structure, by specifying that I-TASSER should use PDB code 4ACQ (Marrero, *et al*. 2012). However, given that the human structure shows α2m that has had its thiol-ester reacted with methylamine, and thus represents the activated form of the molecule, using this as the I-TASSER template would most likely yield a model that represents the activated form of *Limulus* α2m rather than the native molecule.

4.3.2. Protein Structure Prediction Results and Discussion

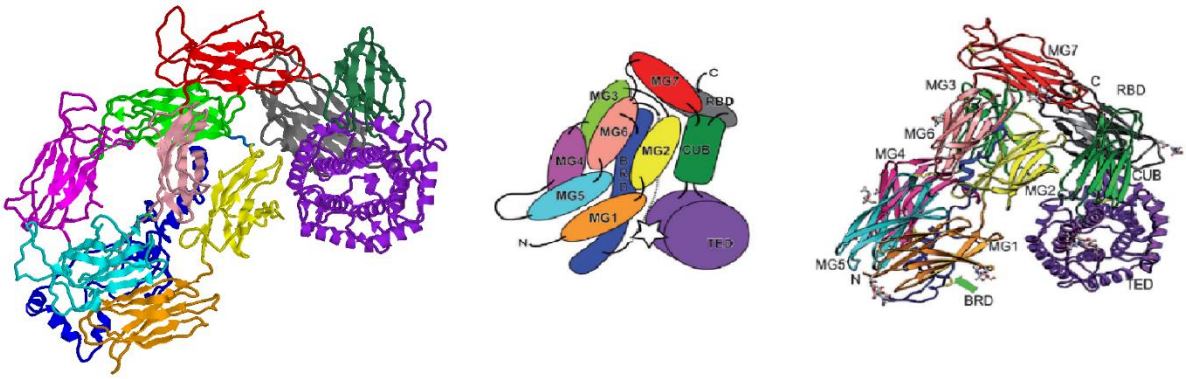


Figure 4.1. From left to right: ribbon diagram of an individual subunit of Limulus α2m produced using I-TASSER (Zhang, 2008; Roy, *et al*. 2010), a schematic diagram of the human subunit to show domain organisation and ribbon diagram of a single subunit of methylamine activated human α2m (Marrero, et al. 2012). Domains are coloured according to the central schematic from Marrero et al. (Marrero, *et al*. 2012); MG1 - orange, MG2 - yellow, MG3 - light green, MG4 - magenta, MG5 - cyan, MG6 - pink, BRD - blue, MG7 - red, CUB - green, TED - purple, RBD - grey.

Figure. 4.1 shows the I-TASSER model of the human subunit in comparison to its activated human counterpart. It is important to note that the human model is the methylamine activated structure given the structural reorientation that occurs during activation. It is however clear from the two diagrams that the general domain arrangement and topology remains conserved between species. The right-handed one-and-a-half turn ellipsoidal super helical structure formed by MG1-6 (Marrero, *et al*. 2012) is visible with the BRD lying behind the coordinating MG domains. It should be noted that the *Limulus* CUB and TED domains both sit a little higher in relation to MG6 than they do in the human protein, which may be a functional outcome of the difference between native and activated molecules. It is known that upon activation the RBD becomes revealed as

part of the molecular reorganisation process, a process which may involve the descent of the CUB and TED domains to allow the exposure of the BRD.

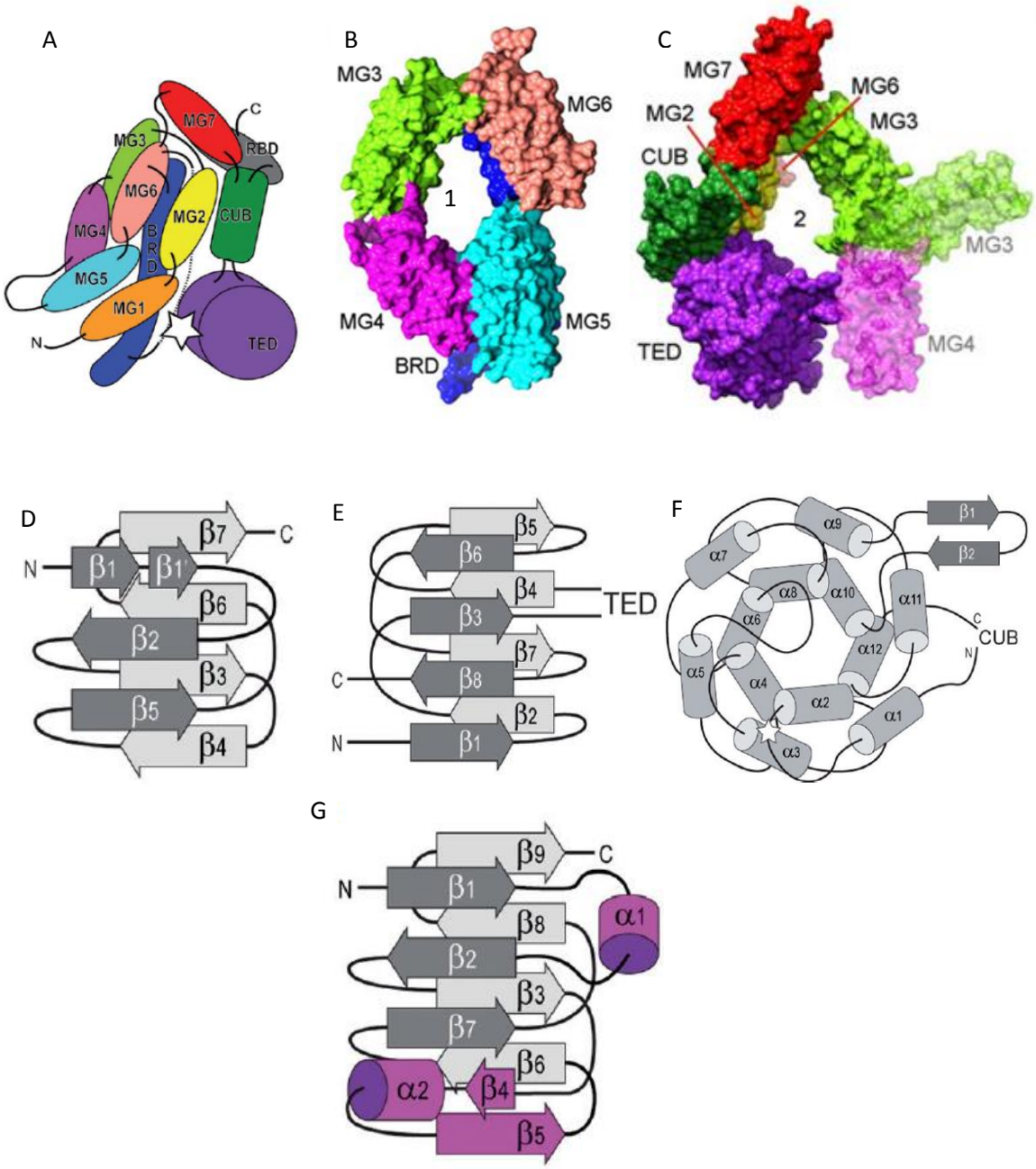


Figure 4.2. Overall domain organisation (A), formation of entrances one (B) and two (C) as well as schematic representation of domain architecture for the MG macroglobulin domains (D), the CUB domain (E), the TED domain (F) and the RBD (G) all seen in the methylamine activated human α2m structure (Marrero, *et al*. 2012). The colour scheme for A is retained for B and C, whilst in G the secondary structures shown in magenta highlight the difference between the RBD (G) and the MG domains (D).

The respective domains of human and *Limulus* α2m have been aligned by sequence in Fig. 4.2 below. In addition to the sequence alignment the confidence vales for the predicted secondary

structure are also included with a scale of 0-9 for low to high confidence respectively. Residues shown in red indicate their presence/predicted presence in an α-helix; blue residues are present/predicted present in β-strands, whereas the black residues indicate no secondary structure present and represent loop regions. The alignment also shows the predicted solvent accessibility from the I-TASSER modelling, with values that range from 0 for a buried residue to 9 for a highly exposed residue. The nature of these exposures may provide key insights into the functional roles played by certain residues. Residues in bold depict those that are N-glycosylated (Iwaki, *et al*. 1996). Residue numbers for the human sequence are included above the alignment for ease of navigation. Work of this nature often presents as many questions as it does answers but it does present significant and reliable new clues about the structure of *Limulus* α2m.

4.3.2.1. Macroglobulin Domain 1

MG1 has 9.1% sequence homology between the human and *Limulus* molecules. They are 98aa and 102aa long and span, Gly^{4} – Glu^{102} and Lys^{1} – Asp^{102}, for human and *Limulus* α2m respectively.

```
                  10|        20|        30|         40|         50|
HUMAN α2M    ---GKPQYMVLVPSLLHTETTEKGCVLLSYLNE--TVTVSASL-ESVRGNRSLFTDLEAE
LIMULUS α2M  ----KSGFILTAPKSLTPGKSNI--LNLHLFDIKTNGFLRIGVKDQDDGNVVAETEVSFN
                 *  ::: .*. *   .::   : *  ::   .  :  .: :.  **    *::. :
CONF. SCORE  ----977999999869679986--99999669998889999998899979877899971
SOLVENT ACC  ----631000000120105340 1--00000032644160201022376332224442434

             60|        70|        80|        90|       100|
HUMAN α2M    N-DVLHCVAFAVPKSSSNE-EVMFLTVQV-KGPTQEFKKRTTVMVKNE
LIMULUS α2M  KDNPSSSIQLTIPSGVEVKRPKLYANGSYSSPSSNDFFFEKDINMHKD
             : :   .: :::*.. . :   :: . .  .  :::*  .. : ::::
CONF. SCORE  798517999983788766613999999984156651787788998725
SOLVENT ACC  464332203040144444641312031222333334143434031444
```

Figure 4.3. Sequence alignment diagram of the MG1 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Residues shown in bold are glycosylated.

Based on the human structure (Marrero, *et al*. 2012) it is known that MG1 does not play a key role in quaternary structure of the α2m molecule but does play a role in coordinating the BRD along

with MG2 and 3. This may suggest why the sequence homology is lower than average as its role structurally is minimal. Despite a sequence homology of just 9.1% the secondary structures line up exceptionally well and with an average confidence score for the predicted secondary structure of the domain of 7.96 on a scale of 0-9 in Figure 4.3. This suggests a high level of structural homology even if the sequences lack that same level of homology. The low sequence homology seen in MG1 and to a lesser extent the other MG domains is not overly surprising, considering the evolutionary theory about the development of the α2m superfamily described previously (Janssen, *et al*. 2005; Sahu, & Lambris, 2001). It is known that Asn^{32} and Asn^{47} within MG1 in the human α2m are N-linked glycosylation sites (Sottrup-Jensen, *et al*. 1984). Of these two only Asn^{47} is conserved in the *Limulus* homologue in Asn^{44} but the residue is not glycosylated (Iwaki, *et al*. 1996). The lack of glycosylation of this residue may be due to the low predicted solvent accessibility of which it scores 3; if the residue is buried within the tertiary structure then it is easily conceivable that it is not glycosylated. There is however one glycosylation site present in MG1 and is found on residue Asn^{80}, this residue also has a predicted solvent accessibility of three however but as this is a known and experimentally proven glycosylation site (Iwaki, *et al*. 1996), the low score may prove irrelevant.

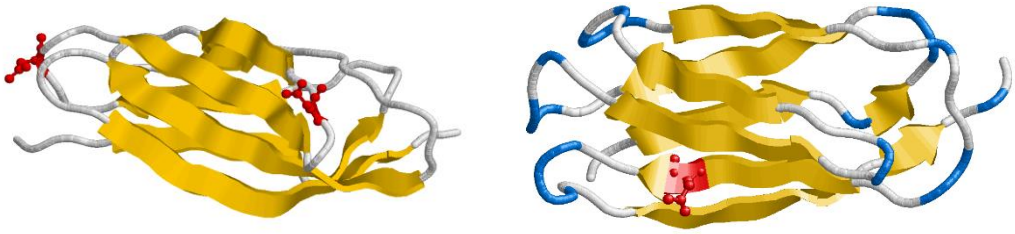


Figure 4.4. From left to right: the known human structure for MG1 of α2m (Marrero, *et al*. 2012); the predicted model for the *Limulus* MG1 domain. B-strands are depicted in yellow, loop regions are shown in white, turns in blue and the glycosylation sites shown with the residues in red ball and stick. For ease of viewing the domains are both orientated with the N-terminal to the left and their C-terminal to the right. Figure generated using RASMOL (Sayle, & Milner-White, 1995).

Figure 4.4 shows the domain topologies of human and *Limulus* MG1 from α2m. By analysing the 3-D structure using RASMOL (Sayle, & Milner-White, 1995) it is clear that the 8-stranded β-sandwich is conserved. I-TASSER however did not show Ile^{97} of the *Limulus* α2m as part of β8 this despite of the secondary structure prediction stating it should be part of β8 which was predicted

with a confidence score of 9. The reasoning for this gap in β8 is unknown. Whilst it was already known, that glycosylation of this domain varies between species, the production of the I-TASSER model for MG1 shows the variation in location of these glycosylation sites. The loop regions of the two models, whilst having to be similarly positioned to maintain the β-sandwich motif, show minimal structural homology which is a likely product of the low sequence identity of 9.1% for the domain.

4.3.2.2. Macroglobulin Domain 2

MG2 has 32.7% sequence homology between the human and *Limulus* molecules. They are both 98aa long and span, Asp^{103} – Phe^{201} and Lys^{103} – Tyr^{201}, for human and *Limulus* α2m respectively.

```
                                                                  110|
HUMAN α2M                                                        DSLVFVQTDKSI
LIMULUS α2M                                                      KLIVFVQTDKPL
                                                                 . :******* :
CONF. SCORE                                                      403999928256
SOLVENT ACC                                                      311000101221

              120|       130|       140|       150|       160|       170|
HUMAN α2M    YKPGQTVKFRVVSMDENFHPL-NELIPLVYIQDPKGNRIAQWQSFQLEGGLKQFSFPLSS
LIMULUS α2M YKPGQTVKVRILPTTPDLKLVPKETIGSFQIENPDGIVLGYWPMLSFAEGIAQFELALPD
            ********.*::    ::: : :* *  . *::*.*  :. *  :.:  *: **.: * .
CONF. SCORE 22787767999999688773667872699998799909999998537751999999899 8
SOLVENT ACC 13242310000000134131234522020213324333034243343430002030302 6

               180|       190|       200|
HUMAN α2M    EPFQGSYKVVVQKKSGGRTEHPFTVEEF
LIMULUS α2M EPTYGMWRIKGNIE-DTEIYENFEVKEY
            **  * :::  : :   .  . * *:*:
CONF. SCORE 76663599999928-9569999999705
SOLVENT ACC 41440301010326-6443334030440
```

Figure 4.5. Sequence alignment diagram of the MG2 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Those shown in red indicate present or predicted α-helices. Residues shown in bold are glycosylated.

MG2 in human and *Limulus* α2m have a significantly higher sequence homology than MG1. The human structure shows that MG2 plays a much more important role in the quaternary structure as well as the tertiary interaction with the BRD as seen in MG1. MG2, contributes towards the formation of entrance 2 (Figure 4.2) to the prey chamber of the molecule. The secondary

structures almost align almost as well, as shown in Figure 4.5, as those in MG1 apart from where the second β-strand lies in the human (Ser^{113} – Lys^{116}) I-TASSER has predicted a small helical structure (Pro^{113} – Tyr^{115}). This is highly unlikely as it disagrees with the architecture of the MG domains which is a fibronectin type III domain made up of 8 β-strands arranged in a β-sandwich. Further questions are raised about the validity of the prediction by the program itself with an average confidence score for the helical structure of only 4.3, compared to the average confidence of all the remaining β-strands of the domain which is 7.95. This in turn brings the overall confidence score for predicted secondary structures of the domain to an average of 7.5 inclusive of what appears to be an error in the helical assignment. Of the three amino acid residues encompassed in this proposed helical structure of the Limulus model (Pro^{113}, Leu^{114}, and Tyr^{115}) it is visible that Leu^{114} has a strong property correlation (:) with the Ile^{114} of the human structure and Tyr^{115} has direct identity (*), further strengthening the theory that this prediction is likely to be in error.

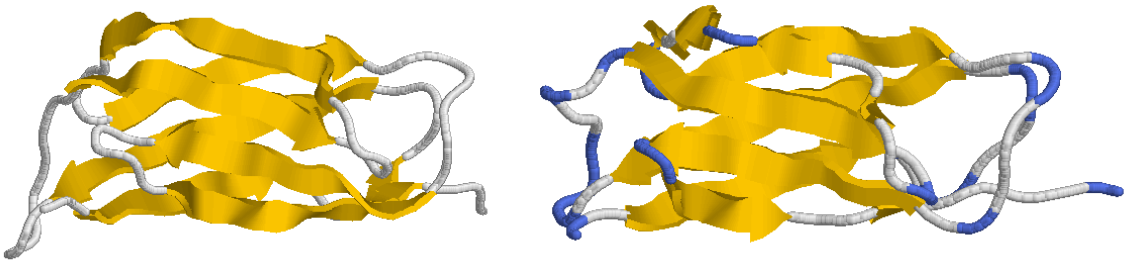


Figure 4.6. From left to right the human and Limulus MG2 domains respectively from α2m (Marrero, *et al*. 2012; Zhang, 2008; Roy, *et al*. 2010). Loop regions are labelled in white with turns in the *Limulus* model depicted in blue. The β-strands are indicated in yellow. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

It is clear from comparing the two 3-D structures of MG2 (Figure 4.6) that this domain's highly conserved sequence (32.7%) is reflected in the (predicted) structure of the domain. In addition to the conserved 8-strand β-sandwich motif, the highly flexible loop regions also show high levels of structural homology.

4.3.2.3. Macroglobulin Domain 3

MG3 has 20.9% sequence homology between the human and *Limulus* molecules. They are 115aa and 106aa long and span, Phe^{206} – Thr^{321} and Phe^{206} – Tyr^{313}, for human and *Limulus* α2m respectively.

```
                                                 210|       220|       230|
HUMAN α2M                                       FEVQVTVPKIITILEEEMNVSVCGLYTY
LIMULUS α2M                                     FEVKITPPSYLLTNADSITWKICAQYTY
                                                ***::* *. :     :.:. .:*. ***
CONF. SCORE                                     79998357247515981899999985 47
SOLVENT ACC                                     03030424331114444030301011 3

                240|       250|       260|       270|       280|       290|
HUMAN α2M    GKPVPGHVTVSICR-KYSDASDCHGEDSQAFCEKFSGQLNSHGCFYQQVKTKVFQLKRKE
LIMULUS α2M GQPVEGTFVAETNVVKYNWEKE--GV--P----VIHKEGLIDGCLDVTVNSSALGFNEQR
            *:** * ....    **.  .:  *         :  :   .**:   *::..: ::.:.
CONF. SCORE 872178579999965588554 4--31--0----00107720771678753343144544 5
SOLVENT ACC 122043211000001333243 3--21--3----30233122313021314243042322 2

                 300|      310|      320|
HUMAN α2M    YEM-KLHTEAQIQEEGTVVELTGRQSSEIT
LIMULUS α2M LSYRAVNMFAEVTEKGTGIKMNATDS--IY
             .   :.  *:: *:** :::.. :*  *
CONF. SCORE 67615999999999606279987788--99
SOLVENT ACC 23112010102021423333343433--03
```

Figure 4.7. Sequence alignment diagram of the MG3 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus.* Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Those shown in red indicate present or predicted α-helices. Residues shown in bold are glycosylated.

MG3 in human and *Limulus* α2m have a sequence homology of 20.9% which reflects its role in coordinating both tertiary and quaternary structure (Figure 4.7). In tertiary terms it acts in much the same way as the previous two MG domains in coordinating the BRD. It's participation in the right-handed one-and-a-half turn ellipsoidal super-helix leads to it forming part of the boundary of entrance 1, as shown in Figure 1.5 (Marrero, *et al*. 2012). Perhaps most significant about this domain is its role in entrance 2 (Figure 4.2). As a contributing domain to the formation of entrance 2 it is di-sulphide bound to the MG4 of the neighbouring di-sulphide linked subunit. Interestingly Cys^{278} of this domain in the human structure does not share sequence identity with

the *Limulus* α2m. There is sequence identity with the two intra-domain disulphide linked residues Cys^{251} - Cys^{299} and Cys^{228} - Cys^{269} in human and *Limulus* respectively (Figure 4.7). This supports the work by Husted *et al*. (Husted, *et al*. 2002) as it was shown that the disulphide responsible for the dimerization in the *Limulus* α2m is between Cys^{719} of the *Limulus* BRD of each subunit. MG3 throws up some interesting discrepancies when comparing the known secondary structure of human α2m to the predicted data of its *Limulus* counterpart. Similarly to MG2 the addition of a helix has occurred (Ser^{276}, Ser^{277} and Ala^{278}), this time not as a substitute for an expected β-strand but in addition to the 8 strands expected. This was placed by I-TASSER with an exceptionally low average confidence score of 3.33. Whilst no direct identity was present for these residues Se^{276}, Ser^{277}, Ala^{278} from the *Limulus* protein have strong property, weak property, and weak property similarity to the Thr^{283}, Lys^{284} and Val^{285} respectively of human α2m. The overall confidence score for the domain is 6.26, which is brought down from 6.6 if we exclude the helix as an error. The lower overall confidence score for predicted secondary structures here is due to low confidence in some of the β-strands in spite of them being well aligned with their human equivalent, which is likely due to the lower sequence identity seen. The human α2m contains a single glycosylation site Asn^{224}, whereas the *Limulus* α2m is glycosylated at both Asn^{275} and Asn^{307} (Iwaki, *et al*. 1996). Human Asn^{224} is not conserved with its aligned residue in the sequence being Thr^{224} which shares some similarities with them both having polar uncharged side chains. *Limulus*' glycosylated residue Asn^{275} in the human molecule is substituted for Lys^{284}. Both these residues have amine

groups in their R-groups with the Asn^{275} being bound to the sugar, with Lys^{284} the amine group is positively charged and so will look to make polar interactions. Asn^{307} aligns with Thr^{313}, both, as mentioned above, loosely similar residues in terms of their properties.

Upon analysis of the MG3 domains from both human and *Limulus* α2m, one noticeable feature is that the *Limulus* MG3 model lacks the α-helix present in the human α2m RASMOL rendition spanning the amino acids Lys^{282} - Phe^{286}. The model is taken from the PDB structure of Marrero *et al*. (Marrero, *et al*. 2012), with the supplementary materials for this paper providing the secondary structure data used above in the alignment (Figure 4.7). In this published data there is no secondary structure feature shown in this region, so the appearance of a helix structure in this region is a likely product of RASMOL believing there is a helix present as it uses different parameters to those used by other programs. Counter to this the predicted secondary structure as discussed above for the *Limulus* MG3 domain states an α-helix exists through Ser^{276} - Ala^{278}, however in the RASMOL rendition no helix appears during this region. Both the present and absent helices of MG3 from human and *Limulus* respectively are in alignment with one another. There is however clear structural homology between these regions as the *Limulus* MG3 render (Figure 4.8) shows a near helical structure in that loop region so the difference between the human structure and the *Limulus* models for MG3 with regard to this helix is likely negligible.

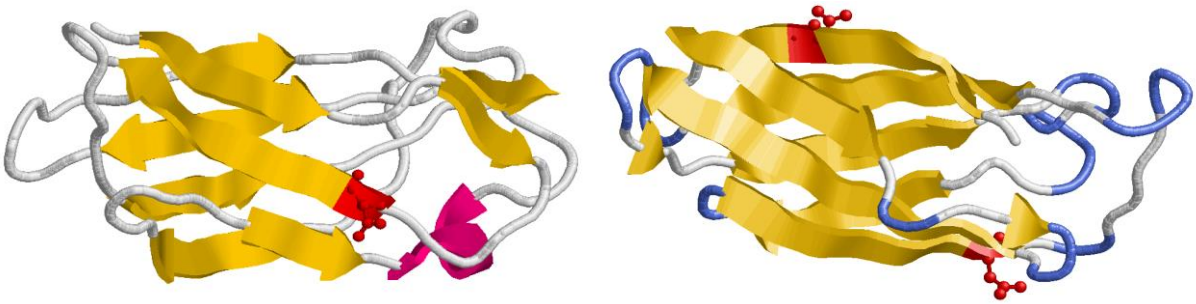


Figure 4.8. From left to right: The human MG3 domain (Marrero, *et al*. 2012) and the Limulus MG3 as predicted by I-TASSER (Zhang, 2008; Roy, *et al*. 2010). B-strands are depicted in yellow with α-helices are shown in magenta. Loop regions are shown in white with the turns in the Limulus model shown in blue. Residues that are glycosylation sites are shown in ball and stick display mode and in red. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

Interestingly the positioning of the glycosylated Asn^{224} and Asn^{275} of human and *Limulus* α2m respectively is very similar with both sitting on the same face of the β-sandwich motif. Whilst not

conserved the fact that their relative positions are conserved is certainly significant. The glycosylated Asn^{307} of *Limulus* is however on the opposing face of the MG3 domain to Asn^{275}. The placement of these residues and ultimately their bound sugars may prove to be physiologically significant.

4.3.2.4. Macroglobulin Domain 4

MG4 has 20.8% sequence homology between the human and *Limulus* molecules. They are 106aa and 124aa long and span, Arg^{322} – Val^{428} and Arg^{314} – Trp^{437}, for human and *Limulus* α2m respectively.

```
                                                      330|      340|
HUMAN α2M                                       RTITK-----LSFVKVDSHFRQGIPFFGQV
LIMULUS α2M                                     RTSNPLNIMYLEPTSGKGYLKPGLPFYGKL
                                                * .       *. .. ..::: *:**:*::
CONF. SCORE                                     984032478732667755032899525789
SOLVENT ACC                                     113320201003244444103221201000

            350|      360|                    370|      380|      390|
HUMAN α2M   RLVDGKGVPIPNKVIFIRGNE--------------ANYYSNATTDEHGLVQFSINTTNVM
LIMULUS α2M KVEKPDGTPAPGEQIELCRFADRERWNRKRWLEEKIRACKEFTSDEAGIIKFTVPPQTPD
              :: . .*.* * :* :                  .  .: *:** *:::*::   .
CONF. SCORE 997589984578549999985244211223445541002467606753899997488777
SOLVENT ACC 001113332043230201122231212233234443443442324441203020214433

                400|      410|        420|
HUMAN α2M   GTSLTVRVNYKDRSPCYGYQWVSEEHE--EAHHTAYLV
LIMULUS α2M ITS----FRFKAKALQYGKKDGDNKLNQPQHSFTVSSW
             **    ..:* ::  ** :  .:: :  :  .*.
CONF. SCORE 259----99999996520013631012213578985555
SOLVENT ACC 223----02010311323232111111111101110221
```

Figure 4.9. Sequence alignment diagram of the MG4 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Those shown in red indicate present or predicted α-helices. Residues shown in bold are glycosylated.

The fourth MG domain in human and Limulus α2m also share a good level of sequence identity, at 20.8%. This domain plays a key role in the dimerization as well as the formation of entrance 2 (Figure 4.2). It forms symmetrical contacts with the TED of its disulphide linked subunit and forms

part of the boundary of entrance 2 in this capacity as well as forming the boundary of entrance 1 (Figure 4.2) as part of the right-handed one-and-half-turn ellipsoidal super-helix structure. Whilst the human MG4 has a disulphide bond as mentioned above between the Cys^{431} and the Cys^{278} of the disulphide linked subunit. In the case of *Limulus* these cysteine residues are not present and the Cys^{431} of MG4 has been substituted in *Limulus* to Gln^{440} (Marrero, *et al*. 2012; Husted, *et al*. 2002). Areas of uncertainty lie within the predicted structure of this domain as a number of the predicted structural features have low confidence levels but those with even low confidence levels align well with the known human structure (Figure 4.9). In particular there is the two

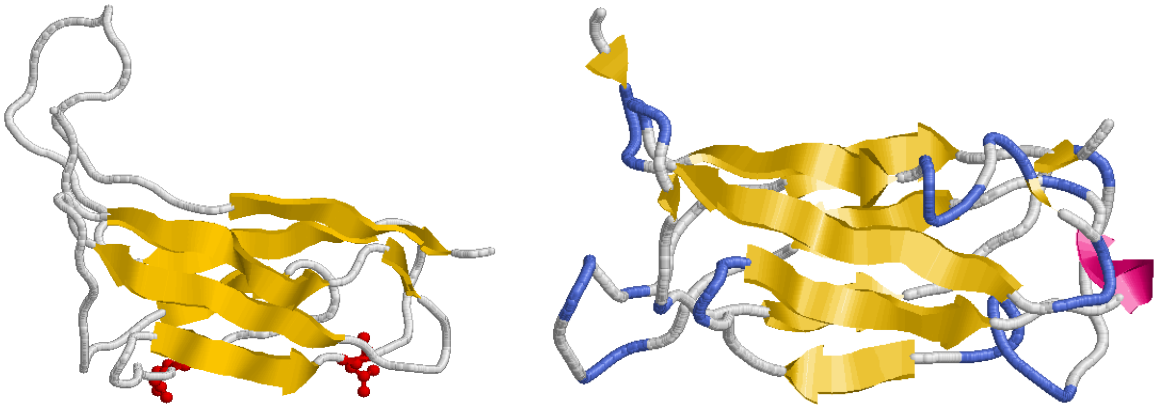


Figure 4.10. From left to right the human and Limulus MG4 domains (Marrero, *et al.* 2012; Zhang, 2008; Roy, *et al*. 2010). B-strands are shown in yellow, α-helices are shown in yellow, loop regions are white and turns are in blue on the Limulus model only. Glycosylation sites are shown in ball and stick and coloured red. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

residue long strand structure (Tyr^{332} and Leu^{333}) which has an average confidence of 1.5 over the two residues yet lines up reasonably well with an equally short region in the human sequence (Phe^{336} and Arg^{337}). The overall sequence identity for the predicted features is 6.09 which isn't particularly high but when looking at the alignment the fit is reasonably good with a few strands larger in one species and a few loops between strands larger in other areas. Of the two only MG4 of human α2m is glycosylated with glycosylation occurring at Asn^{373} and Asn^{387}.Neither of these residues are conserved with their aligned *Limulus* equivalent being Glu^{384} and Pro^{398} respectively.

Apart from the general topology being conserved, the most noticeable distinction between the two models is the appearance of an α-helix in the RASMOL rendition of *Limulus* α2m from Thr^{401} - Asp^{403} (Figure 4.10). This helix does not feature in the predicted secondary structure yielded by the I-TASSER results using the program PSIPRED. This may be due to LOMETS using a different model than the human α2m (Marrero, *et al*. 2012) being the primary template for this region or it maybe as suggested earlier an artefact of RASMOL rendering due to the similarities between helices and turns. The I-TASSER results ranked the templates used with the highest ranked template coming from TEP1r of *Anopheles gambiae* (Baxter, *et al*. 2007). This sequence does contain an α-helix in the equivalent position between residues Thr^{398} - Val^{399} and it has an additional helix between Ser^{348} - Val^{350}. The Thr^{398} – Val^{399} helix is shared with the human thiol-ester containing protein complement component C3. Whilst bovine C3 is not one of the ranked templates used by I-TASSER, it is ranked 3rd (Fredslund, *et al*. 2006). This shows that the inclusion of α-helices into the macroglobulin domains by I-TASSER may not be as erroneous as first though as both Human C3, and TEP1r of *Anopheles gambiae* (Baxter, *et al*. 2007) contain α-helices in the majority of their MG domains, without disrupting the conserved β-sandwich motif.

4.3.2.5. Macroglobulin Domain 5

MG5 has 33.0% sequence homology between the human and *Limulus* molecules. They are 109aa and 141aa long and span, Phe^{431} – Cys^{540} and Phe^{440} – Cys^{579}, for human and *Limulus* α2m respectively.

```
                                                                  440|      450|
HUMAN α2M                                                        PSKSFVHLEPMSHELPCGHT
LIMULUS α2M                                                      PSGSHLQLEPITEEIECGKP
                                                                 ** *.::***::.*: **:
CONF. SCORE                                                      5785489970455423797 6
SOLVENT ACC                                                      3330203032454404244 4

                460|      470|      480|      490|
HUMAN α2M   QTVQAHYILNGGTLLGLKKLSFYYLIMAKGGIVRTGTHGLL--VKQE-------------
LIMULUS α2M LTVKFKYTTGE-----EKKQKFYYQIMARNFIVDTGSFEHEFLLSEDKSGLTDETYLPID
             **: :*          ** .*** ***:  ** **:.     :.::
CONF. SCORE 89999854188-----652179999960873898505998775100111211048851 00
SOLVENT ACC 13030312044-----332220000011322000123232423234443323321112 13

                                500|      510|      520|      530|
HUMAN α2M   ----------------------DMKGHFSISIPVKSDIAPVARLLIYAVLPTGDVIGDSAK
LIMULUS α2M VTALSLNPPNEPEWENNVIVPPHIGETSLTLIPSFEMNPSAKILVFYVREDGETVADSTK
                                  *. *:::  . :: * *::*:: *   *:.:.**:*
CONF. SCORE 00122013442111332001466660677874133568726999999926983999999 9
SOLVENT ACC 12313133243232333221334433222303123220110000000003321000220 3

              540|
HUMAN α2M   YDVENC
LIMULUS α2M ITVKKC
              *::*
CONF. SCORE 998404
SOLVENT ACC 030441
```

Figure 4.11. Sequence alignment diagram of the MG5 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Those shown in red indicate present or predicted α-helices. Residues shown in bold are glycosylated.

MG5, a region of high sequence homology (33.0%) in human and Limulus α2m considering the evolutionary distance between the two species, is active in both the boundaries of entrance 1 and 2 in the human structure (Marrero, *et al*., 2012). Despite the high sequence homology there are

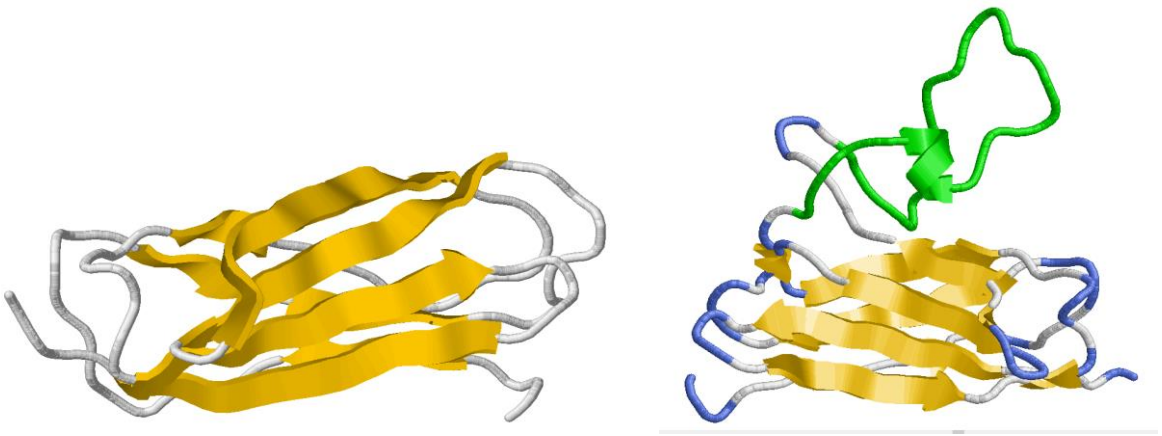


Figure 4.12. From left to right: The human MG5 domain from α2m (Marrero, *et al*. 2012), and the Limulus MG5 domain from α2m as predicted by I-TASSER (Zhang, 2008; Roy, *et al*. 2010). As with previous domain comparisons β-strands are shown in yellow, loop regions in white and turns in the Limulus model are shown in blue. The Limulus model also has a region highlighted in green that shows an inserted region spanning Lys^{502} - Pro^{535}. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

multiple misaligned regions with very low confidence levels (5.22) in the predicted structure (Figure 4.11). The 5th strand of the *Limulus* MG5 (Asp^{507} – Ile^{513}) as well as the 6th and 7th strands (Thr^{516} – Leu^{520} and Val^{532} – Ile^{533} respectively) lies within an insertion compared to the human sequence. The respective confidences of these predicted strands are 3.7, 1 and 0. This could be due to the fact that they are found within a 34 residue long insertion (Lys^{502} – Phe^{535}), compared to the human. As a result the software lacked a direct structural model in this region and as a result produced a low confidence region. If this region is excluded the remaining structure is a 7 stranded domain with an average confidence level for the strands of 6.79. This is one strand less than the human structure, and when analysing the human sequence it is clear that at the 2nd strand (Leu^{445} – Phe^{446}) there is an absence of a strand in the *Limulus* domain. Looking at the directly aligned residues Ile^{454} and Glu^{455} these residues have been designated C by I-TASSER for main chain but were done so with confidence scores of 2 and 3 respectively. It may be that because the β-sheet is so small, the program overlooked the possibility as it perhaps had a much lower confidence value when designated as a strand.

In MG5 it appears that the MG motif of a β-sandwich is conserved, but there is one glaringly noticeable difference between this domain in human and *Limulus* α2m. This difference is an inserted loop region of 33 residues spanning Lys^{502} – Pro^{535} visible in the sequence alignment and highlighted above in green in Figure 4.12 for ease of visualisation. This insertion in *Limulus* α2m relative to the human protein appears between strands 5 and 6, and has a predicted secondary structure (PSIPRED) containing three β-strands Asp^{507} - Ile^{513}, Thr^{516} - Leu^{520} and Val^{532} - Ile^{533} as mentioned earlier. Upon evaluation of this region of the I-TASSER predicted model for *Limulus* MG5 it is clear that no β-sheets were in fact incorporated into the model, rather a short α-helix was built in at residues Glu^{527}, Trp^{528}, Glu^{529} and Asn^{530}. Following the appearance of this unexpected helix, which may be a result of the top ranked template from *Anopheles gambiae* TEP1r, a sequence alignment using CLUSTAL was carried out. The results showed that compared to TEP1r the insertion is longer still at 41 residues, spanning His^{494} - Ser^{542} with an aligned region

sitting between Pro[522] and Glu[529]. It may then be possible to infer that this inserted loop region plays a significant physiological role in *Limulus* α2m as the insertion of this region is distinct from the sequences of its human homologue (Marrero, *et al*. 2012), C3 and the TEP1r protein from *Anopheles gambiae* (Baxter, *et al*. 2007).

4.3.2.6. Macroglobulin Domain 6

MG6 has 44.6% sequence homology between the human and *Limulus* molecules. They are both 98aa long and span, Leu[541] – Asp[578] and Glu[706] – Ala[765] for human α2m and Leu[581] – Asp[618] and Val[757] – Gly[816] *Limulus* α2m, sequence gap is shown where BRD is inserted.

```
                  550|       560|       570|
HUMAN α2M    LANKVDLSFSPSQSLPASHAHLRVTAAPQSVCALRAVD
LIMULUS α2M  LRNKVGLKFGEEKVLPGASSTLQLTASPYSICGIGAVD  |→BRD
             * *** *.*. .: **.: : *::**:* *:*.: ***
CONF. SCORE  67615999766665899858999994799889999952
SOLVENT ACC  13330303144443424440404004230100000001

                 580|       590|       600|       610|       620|       630|
HUMAN α2M    BRD→|ETVRKYFPETWIWDLVVVNSAGVAEVGVTVPDTITEWKAGAFCLSEDAGLGIS
LIMULUS α2M      |VEIRTYFPETWLWELQNIGATGELSLKRDIPHTITEWVGSAICISEETGLGVS
                   :*.******:*:*  : ::*  .:   :*.***** ..*:*:**::***:*
CONF. SCORE        22103574124772002355744899997799876699999999588763223
SOLVENT ACC        33224202300123313334423231303103211100000000124101011

HUMAN α2M    STASLRA
LIMULUS α2M  EAATVKG
             .:*:::.
CONF. SCORE  5359999
SOLVENT ACC  3323130
```

Figure 4.13. Sequence alignment diagram of the MG6 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus* protein. Residues shown in bold are glycosylated.

With the highest sequence identity between human and *Limulus* α2m of any domain MG6 contains a break that accommodates the BRD as well as contributing to the boundary of both

entrance 1 and 2 (Marrero, *et al*. 2012) it also has a confidence rating of 6.55 which is lowered once again by a low confidence short β at Glu^{758} and Ile^{759} without when excluded result in a confidence score of 7.27 (Figure 4.13). A question should be asked however as to whether the residues following that short predicted strand (Arg^{760} – Thr^{766}) should in fact be part of the strand and bridge to the next β-strand in the sequence (Trp^{766} – Leu^{771}). The high sequence identity between human and *Limulus* α2m for this region suggests that this is indeed the case. Of the seven aligned amino acids in those strands, six share sequence identity, this coupled with the fact that the human equivalent β-strand is much longer (11 residues as opposed to 5) and have an average confidence of 3.14 might suggest that these residues should in fact form part of the strand.

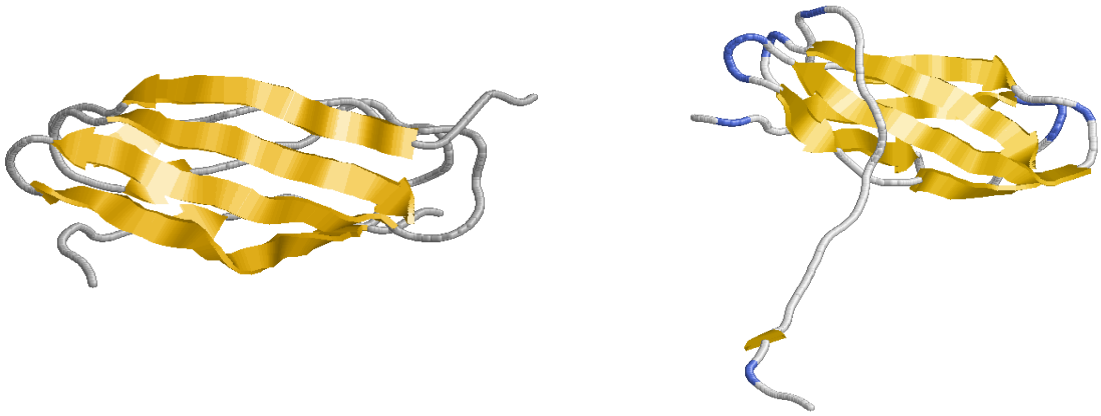


Figure 4.14. From left to right: Human MG6 from α2m excluding the BRD that is inserted between Asp^{578} - Glu^{706} (Marrero, *et al*. 2012) and the MG6 domain of Limulus α2m as predicted by I-TASSER excluding the BRD inserted between Asp^{618} - Va^{757} (Zhang, 2008; Roy, *et al.* 2010). Figure generated in RASMOL (Sayle, & Milner-White, 1995).

The comparison of the two domains shows the region of the *Limulus* domain with a large loop region contains a short helix Glu^{758} and Ile^{759} (Figure 4.14). Upon comparison with the top ranked template TEP1r from *A. gambiae* in this protein there is a large loop region present which may account for the loop in the predicted model of the *Limulus* α2m MG6 domain, but there is no β-strand present in the aligned region. Given the confidence scores for the predicted secondary

structure for Glu[758] and Ile[759], 2 and 1 respectively, it may be that this is an area that requires further examination when the crystal structure for *Limulus* α2m is solved.

4.3.2.7. The Bait Region Domain

The BRD has 18.3% sequence homology between the human and *Limulus* molecules. They are both 126aa long and span, Gln[579] – Thr[705] and Lys[619] – Val[756], for human and *Limulus* α2m respectively.

```
                                                          580|      590|
HUMAN α2M                                                   QSVLLMKPDAELSASS
LIMULUS α2M                                                 KSVHILSSDNRITEEE
                                                            :** ::. * .:: ..
CONF. SCORE                                                 2113316567689668
SOLVENT ACC                                                 0000103363403444

              600|      610|      620|          630|      640|      650|
HUMAN α2M   VYNLLPEKDLTGFPGPLNDQDNEDCINRHNVYIN----GITYTPVSSTNEKDMYSFLEDM
LIMULUS α2M VFNKLGGHDYYWPKQAT--SDYKYCEDYKFKQTEGEHEGSFSSGFTSTNYLDSITAFDEA
            *:* *  :*          .* : * : :    :    *   : .:***  *  : :::
CONF. SCORE 42210035445676656--64554333222346665465046764024664120011235
SOLVENT ACC 13330322111112221--32233033110233422231100101113323200200230

                  660|                 670|      680|       690|
HUMAN α2M   GLKAFTNSKIR-----------KPKMCPQLQQYEMHGPEGLRVGF-YESDVMGRGHARLV
LIMULUS α2M GLVVISDMELETRPCKPSGFEDGGRPCPQYDVAFAAPQAANRIGGGGEAGGFGGG---IR
            ** .::: ::.           : *** :        . *:*   *:  :* *   :
CONF. SCORE 703530211114565410003344453210122022011110234443334322 2---22
SOLVENT ACC 000000101122100100001111120111100011122222121333323224 4---24

          700|
HUMAN α2M   HVEEPHT
LIMULUS α2M KKTNKPV
            :  :  .
CONF. SCORE 1234564
SOLVENT ACC 4444433
```

Figure 4.15. Sequence alignment diagram of the BRD domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus* protein. Those shown in red indicate present or predicted α-helices.

The bait region domain of *Limulus* α2m, one of the key characteristic features of α2m, actually has much lower sequence identity with the human protein than might be expected – 18.3%. However

upon application of context it is clear that the BRD of *Limulus* α2m should differ greatly in sequence to that of human α2m, as the two immune systems have evolved to combat and target different proteases. Of the total predicted secondary structures for *Limulus* α2m the average confidence in the predictions is just a mere 2.39 (Figure 4.15). As can be seen from the human secondary structure as well as the tertiary folding of the domain, the BRD is a largely disordered region to allow maximum accessibility to the cleavage sites of proteases. One possible reason for the low confidence scores, even for those secondary structures that align with their human counterparts, may be because other than human α2m the remaining structures in the PDB that I-TASSER used lack bait regions and thus a BRD. In the other α2m family members different functionally defining domain is inserted into MG6 instead of the BRD, such as the linker (LNK) domain and anaphylatoxin (ANA) domain present in the complement proteins C3, C4 and C5 (Janssen, *et al*. 2005).

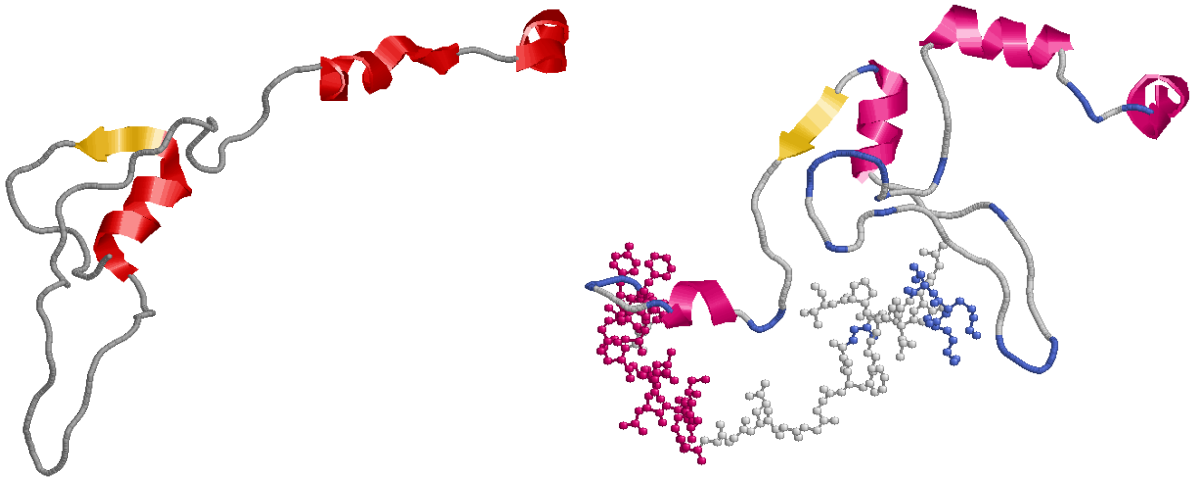


Figure 4.16. From left to right: The human BRD excluding the bait-region - Pro^{667} - Thr^{705} as the region is highly flexible and thus isn't built into the model (Marrero, *et al*. 2012) and the predicted Limulus BRD of α2m (Zhang, 2008; Roy, *et al.* 2010), with the bait region itself shown in ball and stick mode to make it distinguishable. Colour schemes are the same as previous domains. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

In the human structure and the *Limulus* α2m BRD models it is clear that the general topology is the same with conserved structures clearly visible (Figure 4.16). The addition of the bait-region itself in the *Limulus* model by I-TASSER is unlikely to yield potentially significant information

despite its absence from the human structure. There is an α-helix shown in the bait region of the *Limulus* model, Gln^{730} - Asn^{733}, this conformation is unlikely as it is potentially inhibitive of protease cleavage, the inherent flexibility of the bait-region is why it is absent from the human structure and it is this property, in conjunction with its sequence, that allows it to be the target of multiple proteases.

4.3.2.8. Macroglobulin Domain 7

MG7 has 40.3% sequence homology between the human and *Limulus* molecules. They are 119aa and 116aa long and span, Phe^{766} – Pro^{885} and Phe^{817} – Pro^{932}, for human and *Limulus* α2m respectively.

```
                    770|      780|      790|      800|      810|
HUMAN α2M           FQPFFVELTMPYSVIRGEAFTLKATVLNYLPKCIRVSVQLEASPAFLAVPVEK
LIMULUS α2M         FQPFFVSFTLPYSVIRGEKVPIIVTVFNYLSECLPIKLSLEQSDKFEM-QNDT
                    ******.:*:******** . : .**:*** :*: :.:.** *  *     :.
CONF. SCORE         867999824882220498899999999643566205999994587774-0687
SOLVENT ACC         211010102002101341303020101001132130203033144042-4446

          820|      830|      840|      850|      860|      870|
HUMAN α2M   EQAPHCICANGRQTVSWAVTPKSLGNVNFTVSAEALESQELCGTEVPSVPEHGRKDTVIK
LIMULUS α2M NSYTSCVCGGKSDTTRWMIKPRSLGQVNLTVYGASLPNEAICGNQDYST--VTARDAATR
            :.   *:*.   :*. * :.*:***:**:** . :* .: :**.:  *.     :*:. :
CONF. SCORE 505788669973899999832765358999987506886310157741 5--566531389
SOLVENT ACC 443334144643220201020341231302010331644423342413 3--332331232

          880|
HUMAN α2M   PLLVEP
LIMULUS α2M QLLVEP
             *****
CONF. SCORE 988606
SOLVENT ACC 303033
```

Figure 4.17. Sequence alignment diagram of the MG7 domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus* protein. Those shown in red indicate present or predicted α-helices. Residues shown in bold are glycosylated.

MG7 in the human structure does not play a role in entrance 1 like many of the preceding MG domains (Marrero, *et al*. 2012) but it does play a role in the formation of entrance 2 (Figure 4.2).

MG7 can be found sat between MG6, which has the BRD inserted, and the CUB domain, which has the TED inserted. The fact that MG7 lies between these two domains that contain functionally and family defining domains inserted might account for its high sequence identity of 40.3% as it will play a key role in the placement of these domains spatially. The domains predicted secondary structure has an average confidence of 6.39 (Figure 4.17). The first predicted β-strand of the domain, Phe^{817} – Ser^{823}, does in fact begin in the previous MG6 domain with the MG6 portion spanning Ala^{811} - Gly^{816}. When compared to the secondary structure of the known human structure it could be suggested that that potentially the first three residues of MG7 (Phe^{817}, Gln^{818}, Pro^{819}) and the last three residues of MG6 (Val^{814}, Lys^{815}, Gly^{816}), should not in fact be part of a β-strand to bring the structure more in line with the human model. The relatively high confidence for Phe^{817}, Gln^{818}, and Pro^{819} of 8, 6 and 7 respectively suggest that the predicted secondary structure is in fact correct. The second strand which is only a short one Ser^{829} – Ile^{831}, has a very low confidence of 2, 2 and 0 respectively despite the 100% sequence identity for that strand. This seems to be a common problem that I-TASSER has when predicting secondary structures. It appears to correctly predict the structure for that region but the low confidence levels are highly typical of short length strands. The predicted secondary structure for the domains assigns ten β-strands to MG7; this is as previously stated out of keeping with the architecture and general motif of the MG domain which is typically a seven/eight stranded β-sandwich. Judging from the sequence alignment it is clear that the eight strands seen in the human structure align well with eight *Limulus* counterparts. There are two *Limulus* strands however that lack a human partner (Phe^{862} – Met^{864} and Cys^{910} – Tyr^{915}). The respective confidences of 6 and 4, are not as low as expected from a potentially erroneous result especially as Ser^{829} – Ile^{831} appears to be a correct prediction in spite of the average confidence of 1.3, but the evidence does suggest that these represent a potential error in the predicted secondary structure of the MG7 domain.

Human α2m has one glycosylation site in its MG7 domain at Asn^{846}, which is conserved in the

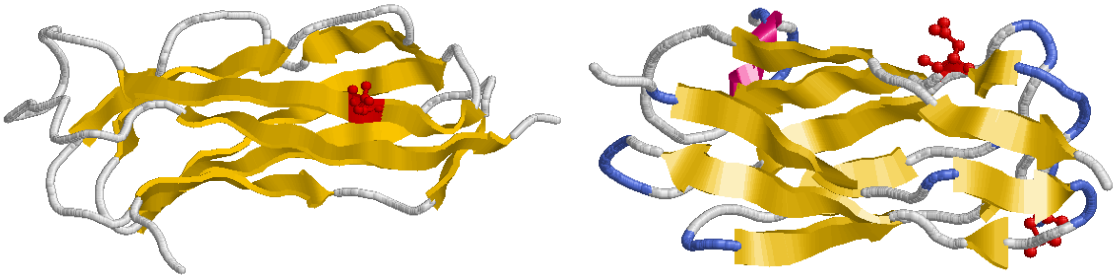


Figure 4.18. From left to right: the human MG7 domain of α2m (Marrero, *et al*. 2012) and the I-TASSER predicted structure of the MG7 domain of Limulus α2m (Zhang, 2008; Roy, *et al*. 2010). Colour schemes as on previous domains with glycosylation sites shown in ball and stick configuration and in red. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

Limulus homologue at residue Asn^{896}, in addition to this glycosylation site *Limulus* α2m MG7 is also glycosylated at Asn^{866}.

Perhaps the most significant observation noted when comparing these two domains is the appearance of an α-helix in the *Limulus* model between residues Leu^{904} - Glu^{907} (Figure 4.18). In addition to this the predicted secondary structure, by the program PSIPRED, states a β-strand is present between Cys^{910} and Tyr^{915}. In spite of this following template search and model refinement the actual *Limulus* α2m model produced shows no such β-strand through these residues.

4.3.2.9. The CUB Domain

The CUB domain has 38.6% sequence homology between the human and *Limulus* molecules. They are both 114aa long and span, Glu^{886} – Ile^{931} and Thr^{1247} – Phe^{1316} for human α2m and Glu^{933} – Leu^{981} and Lys^{1279} – Lys^{1346} *Limulus* α2m, sequence gap is shown where the TED domain is inserted.

```
                    890|       900|        910|       920|       930|
HUMAN α2M           EGLEKETTFNSLLCPSG--GEVSEELSLKLPPNVVEESARASVSVLGDI
LIMULUS α2M         EGFPKEDTWSTFACPKDQNGKFTATSDLLLPEDLVEDSARGYVSITGDL|→CUB
                    **: ** *:.:: **.   *:.:   .* ** ::**:***. **: **:
CONF. SCORE         881389999999860576764058883357766778837999814775
SOLVENT ACC         422334312110012334434334414142144214321301010111

                          1250|     1260|     1270|     1280|     1290|
HUMAN α2M                 TRTGKAAQVTIQSSGTFSSKFQVDNNNRLLLQQVSLPELPGEYSMK
LIMULUS α2M         CUB →|KDELD-LEVGVE-SSGFEKKIMLTKDNSILMQTFRLQTVPSPVDFE
                          .   .  :* :: *. *..*: : ::* :*:* . *  :*.  .::
CONF. SCORE               76556-306782-27862589983787204445404678804999
SOLVENT ACC               34333-130203-35434330303442232334341463434040 4

                     1300|     1310|
HUMAN α2M          VTGEGCVYLQTSLKYNILPEKEEF
LIMULUS α2M        ATGSGCGLVQTSLRYNVNTPPPRK
                   .**.**  :****:**:     .
CONF. SCORE        963235999876780246887665
SOLVENT ACC        042101000000010214346644
```

Figure 4.19. Sequence alignment diagram of the CUB domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus.* Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Those shown in red indicate present or predicted α-helices.

The CUB domain that contains an inserted family defining thiol-ester domain is highly conserved between the two species in question. Not only does it have the same number of amino acids in length but at 38.6% sequence identity it is also one of the domains which have the greatest sequence homology. In humans the domain is arranged as two four-stranded antiparallel β-sheets with the TED inserted between strands 3 and 4. The predicted secondary structure suggest that the TED in *Limulus* mimics the humanoid protein in that it lies between the 3rd and 4th β-strands it does however have a proposed seven strand conformation with the 7th and 8th strands effectively joined into one large strand according to the prediction data. This predicted large 7th strand

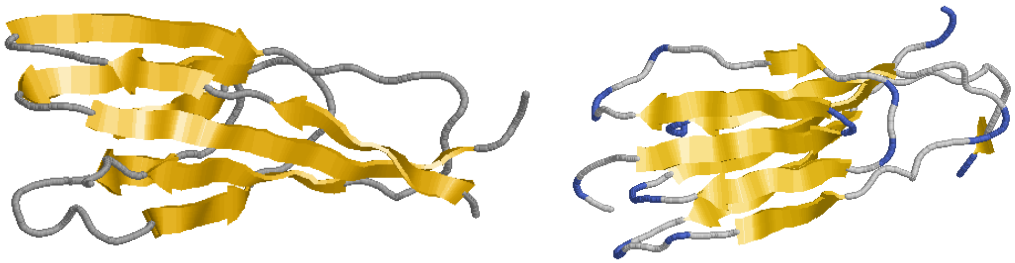


Figure 4.20. From left to right: The CUB domains from human α2m and Limulus α2m as shown by Marrero et al. and I-TASSERs prediction respectively (Marrero, *et al*. 2012; Zhang, 2008; Roy, *et al*. 2010). Colour schemes are as per previous figures. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

(Pro^{1318} - Tyr^{1337}) however can be queried due to Ser^{1326}, Gly^{1327}, Cys^{1328}, having respective confidence scores of 3, 2 and 3 (Figure 4.19). With two of these three residues aligning with the gap between the 7th and 8th strands in the human protein the case for this argument is strong especially as the CUB domain motif is well conserved in the α2m superfamily.

The unglycosylated CUB domains of both human and *Limulus* α2m Figure 4.20 both show the conserved all-β motif made up of two four stranded antiparallel β-sheets. Interestingly, although the LOMETS secondary structure predicts that residues Thr^{1324}, Gly^{1325}, Ser^{1326}, Gly^{1327} and Cys^{1328} are all part of the proposed extended β-strand the reality is that the I-TASSER model shows these residues as a turn between β-strands 7 and 8. As a result the model coincides well with the human structure of the CUB domain (Marrero, *et al*. 2012).

4.3.2.10. The Thiol Ester Domain

The TED domain has 40.8% sequence homology between the human and *Limulus* molecules. They are 314aa and 296aa long and span, Leu^{932} – Phe^{1246} and Met^{982} – Tyr^{1278}, for human and *Limulus* α2m respectively. The thiol-ester bond participating residues have been highlighted in yellow.

Figure 4.21. Sequence alignment diagram of the Thiol Ester domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus* protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus* protein. Those shown in red indicate present or predicted α-helices. Residues in bold are glycosylated.

The α-helix dominated, α- α toroid TED that defines the α2m superfamily of proteins is highly conserved in all species due to its key functional roll in addition to its family defining characteristics. Despite being slightly shorter in length the *Limulus* TED still share 40.8% sequence homology with the human protein. The yellow highlighted residues of Figure 4.21 are involved in the thiol-ester bond, Cys^{949} – Gln^{952} and Cys^{999} – Gln^{1003} from human and *Limulus* respectively, are

all conserved as this is a key functional region that sees high conservation across all family members (Armstrong, & Quigley, 1999). From the alignment of the secondary structures the majority of the helices align with great accuracy however there are a few areas that I-TASSER may have gotten wrong. Tyr^{1052} and Ser^{1053} have been designated β-strands and it has done so with high confidence, both scoring seven. These two residues are conserved between both species and in human α2m do not feature in a β-strand. The α-helix predicted from Pro^{1180} – Thr^{1195} for the *Limulus* α2m may contain an error towards its back end with residues Ile^{1194}, Thr^{1195} as well as the following residue Glu^{1196} predicted to be not included in the helix, having confidence scores of 3, 3 and 0 respectively. These three residues are aligned with a short β-strand in the human molecule in Val^{1152}, Lys^{1153}, and Lys1154. These three amino acids form the first of the β-sheets of the domain seen in the human structure (Marrero, *et al*. 2012). The human structure has shown us that these two β-sheets are linked by a β-hairpin that may play a key role in the reorientation of the TED following thiol-ester cleavage. Both human and *Limulus* α2m thiol-ester domains are glycosylated. The human domain is glycosylated at Asn^{968} which is not conserved in the *Limulus* homologue, the *Limulus* domain shows glycosylation at two sites - Asn^{1089} and Asn^{1145}.

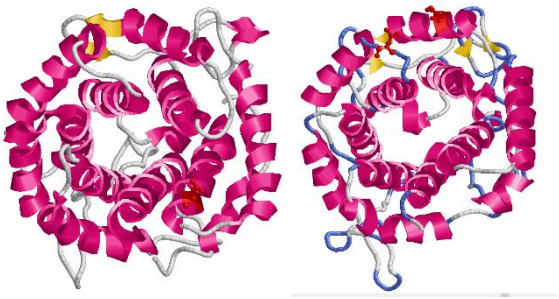


Figure 4.22. From left to right the human α2m TED domain (Marrero, *et al*. 2012), and the I-TASSER predicted structure for the Limulus α2m TED domain (Zhang, 2008; Roy, *et al*. 2010). Colour scheme is as on other domain comparison figures, with the glycosylated residues shown in red and ball and stick configuration for ease of identification. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

The above comparison (Figure 4.22) of the two family defining domains clearly shows a high level of conservation in this domain. Plainly visible is the α/α-toroid topology in the arrangement of six concentric α-hairpins as an α-propeller with a central axis.This gives rise in both models to a thick two-faced disc. The glycosylation sites on both human and *Limulus* sit on the same face of the domain.

4.3.2.11. The Receptor Binding Domain

The RBD domain has 31.3% sequence homology between the human and *Limulus* molecules. They are 128aa and 129aa long and span, Pro[1317] – Ser[1445] and Gly[1347] – Glu[1476], for human and *Limulus* α2m respectively.

```
                                    1320|     1330|     1340|     1350|
HUMAN α2M                            PFALGVQTLPQTCDEPKAHTSFQISLSVSYTGSRSA
LIMULUS α2M                          GFHLEVTVKRGL---YRDCINAHIATCVKYDGKGGV
                                      * * * .         :    . :*: .*.* *. ..
CONF. SCORE                          663079997402---466644999999996167787
SOLVENT ACC                          433030302223---344433130110021334442

               1360|     1370|       1380|     1390|     1400|     1410|
HUMAN α2M   SNMAIVDVKMVSGFIPLKPTVKMLERS--NHVSRTEVSSNHVLIYLDKVSNQTLSLFFTV
LIMULUS α2M SNMAVLEMKMVSGWIPDEESIKNIVDREELNLRRYEVDGNQLNLYFSELTDQNLCFNFWL
            ****:::*****:** : ::* :      .: * **..*:: :*:.::::*.*.: * :
CONF. SCORE 763699986375303143034453115764158987039979999965789868999889
SOLVENT ACC 010000001000211014420430343442303313145310000013146442102020

                  1420|     1430|     1440|
HUMAN α2M   LQDVPVRDLKPAIVKVYDYYETDEFAIAEYNAP--CS
LIMULUS α2M EQDIEVQETKPATIRLYDYYELEQEVVTSYSIDENCE
             **: *:: *** :::***** :: .::.*.    *.
CONF. SCORE 999982331214999999048975379998899876 2
SOLVENT ACC 213031432120101000113254303420324463 6
```

Figure 4.23. Sequence alignment diagram of the receptor binding domains of human and *Limulus* α2m, with confidence score for the predicted secondary structures and solvent accessibility for the *Limulus*

protein. Residues highlighted in blue indicate a β-strand in humans or a predicted β-strand in the *Limulus*. Those shown in red indicate present or predicted α-helices. Residues in bold are glycosylated.

The RBD is a key functional domain that allows the α2m-protease complex to be endocytosed by the immune system for processing and clearance. LRP-1 the human receptor for α2m is part of an evolutionarily ancient family of proteins and a homologue found in *Limulus* LRP – receptor associated protein, does bind the human protein but its binding of the *Limulus* α2m has yet to be

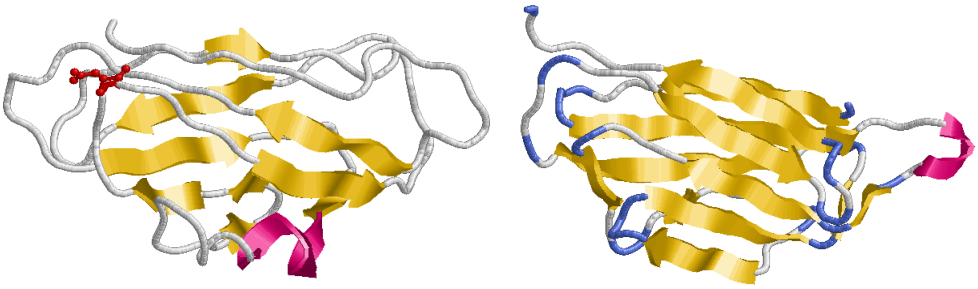


Figure 4.24. From left to right, the human RBD of α2m (Marrero, *et al*. 2012), and the I-TASSER predicted structure of the Limulus α2m RBD (Zhang, 2008; Roy, *et al*. 2010). The colour scheme is as before, with β-strands in yellow, α-helices in magenta, loop regions in white and turns in the loops shown in blue. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

assessed. This cross species recognition may be due to the 31.3% sequence homology between the two receptor binding domains which also show a very good secondary structure fit. The predicted *Limulus* model (Figure 4.23) lacks the first α-helix seen in the human structure (Lys^{1333}, Ala^{1334}, His^{1335}, and Thr^{1336}) which correspond to Arg^{1360}, Asp^{1361}, Cys^{1362} and Ile^{1363} in *Limulus*. These residues were predicted C for main chain with respective confidences of 6, 6, 6, and 4. Although Lys^{1333} in human and Arg^{1360} in *Limulus* are both positively charged and have similar properties, the remaining three residues are entirely dissimilar. Ala^{1334}, in human is a very passive amino acid whereas Asp^{1361} with its interactive carboxylic acid is far more likely to interact with other molecules. His^{1335} and Cys^{1362} in human and *Limulus* respectively, are reactive but in different ways Cys^{1362} forming an intrasubunit disulphide bond with Cys^{1475}, the penultimate residue of the domain. Whether or not the absence of this helix in *Limulus* is correct remains to be seen, it may in fact be correct if that region is the part that interacts with its receptor. Only the human RBD has any glycosylation which takes place at Asn^{1435}.

Interestingly upon observing and comparing the two RBD models above, it was clear that an α-helix that was predicted by the secondary structure prediction is missing. The predicted helix that runs from Glu^{1397} - Val^{1404} with only an average confidence of 3.0, is in fact replaced with a loop region and a small strand running from Ile^{1400} - Lys^{1401}. Another interesting discrepancy is seen in the helix of the human α2m RBD which the supplementary materials to Marrero *et al*. (Marrero, *et al*. 2012), state runs from Lys^{1369} - Ser^{1379}. The model in fact however shows that the α-helix runs only from Lys1374 - Ser^{1379}. Judging however from the backbone depiction of the RBD it is clear that all the residues, Lys^{1369} - Ser^{1379} are arranged in a helical manner but one that lacks uniformity in diameter of the central cavity of the helix that may be why it has not shown all of the residues as part of an α-helix. This may be once again due to the way in which RASMOL renders the coordinates into a three dimensional model (Figure 2.24).

4.3.2.12. Tertiary and Quaternary Structure

Sequence homology varies greatly between the various domains with the lowest homology seen in MG1 (9.1%) and the highest seen in MG6 (44.6%) whilst the functionally significant and family defining TED also has a high sequence homology for such an evolutionarily distant species (40.8%). An encouraging sign for the competence of the software is the alignment of known secondary structures with those predicted by the software. One could argue that this is to be expected as the human structure was indeed one of the PDB models used to build the theoretical one and thus determine the proposed secondary structure of *Limulus* α2m ranking 5^{th} in the structure homology for known structures. This means a difficult question must be asked, and one that currently there is no answer for: Does *Limulus* α2m genuinely have greater structural homology with TEP1r from *Anopheles gambiae* than it does with its human homologue? The only way to truly answer this question lies in the solution of the crystal structure of the molecule and then making a direct comparison of the structures in question.

The human structure is currently the best functional homologue within the PDB to *Limulus* α2m. Given that the model above produced by I-TASSER represents a non-activated form compared to the activated human structure, it was decided to run a second I-TASSER simulation with the human model defined as the primary template in order to produce a model with similar domain positioning to the human model and thus yielding potential information about the rearrangement of the domains upon activation.

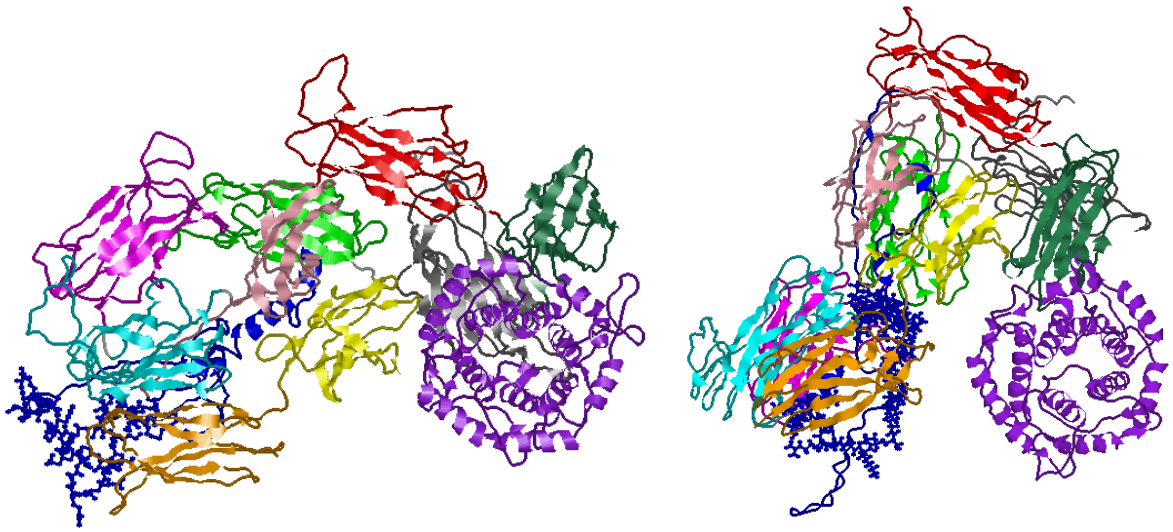


Figure 4.25. To the left is the "native" model of Limulus α2m and to the right is the "activated" model. The "native" model is based on an open search with I-TASSER without specifying a template to use, whereas the "active" model was built with the human structure specified as the template. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

As shown in Figure 4.25 the 'native' model to the left is the original I-TASSER prediction with no template specified, the 'activated' model to the right was built by I-TASSER when the human model (PDB code – 4ACQ) (Marrero, *et al.* 2012) was specified as the template. The result unsurprisingly, is an activated structure that far greater resembles that of the human model in terms of domain organisation. Upon closer inspection using RASMOL it is clear that there are a few major differences in terms of the relative positions of the domains between the native and activated *Limulus* models. MG1-6 and the BRD in the native *Limulus* α2m model seem to have

twisted ~90° with respect to MG6. It is known that these domains form the right handed one-and-a-half turn ellipsoidal super-helix that forms entrance 1 (Marrero, *et al*. 2012). This twisting of the super-helix may inhibit access through entrance 1 to the central prey chamber where the bait region and thiol-ester lie. In addition the thiol ester domain of the activated model seems to have 'dropped' with respect to MG6 which could be described as the apex of the subunit. In the 'native' *Limulus* α2m model the higher position of the TED, means that the CUB domain limits the accessibility of the RBD. However in the 'activated' model the TED drops and the result is that it pulls the CUB domain down further alongside the RBD causing more of it to be revealed to the solvent. During the 'dropping of the TED' the TED also appears to undergo a massive internal shift. In the 'native ' *Limulus* α2m model the residues that form the thiol ester Cys^{999} and Gln^{1003} are positioned up by the RBD, in the 'activated' *Limulus* α2m model however they appear much lower down relative to the RBD due to the angle of α-helix-2 having shifted by ~15°.

It may then be proposed that upon activation of α2m the ellipsoidal super-helix twists, thus closing off entrance 1, and simultaneously the thiol ester domain drops, pulling with it the CUB and exposing the RBD. The result is a more compact structure, with entrance 1 closed off, the thiol ester residues moving more centrally and the RBD exposed which allows the activated *Limulus* α2m to bind its receptor for clearance. The inherent flexibility of the native α2m molecule has been known and demonstrated for some time (Armstrong, *et al*. 1991). Once activated the TED domain shifts as a result trapping the protease within it.

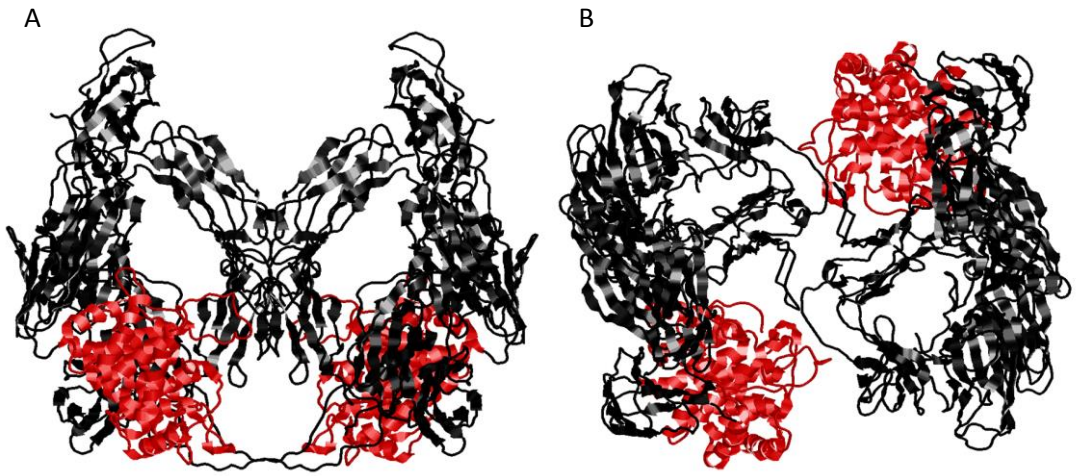


Figure 4.26. The activated human molecule shown as a dimer thus representing the activated *Limulus* dimer, the TED domains have been highlighted in red to highlight their relative positions in both a side on (A) and top view (B) of the molecule (Marrero, *et al*. 2012). Figure generated in RASMOL (Sayle, & Milner-White, 1995).

Figure 4.26 shows the human α2m structure manipulated into a dimer, to highlight the shift expected to be seen in *Limulus*. It depicts a rather "demonic" looking molecule with the TED domains extended and thus inhibiting the 'escape' of a potential prey molecule. In the native structure the TED domains would be withdrawn higher up, thus restricting access to the RBD. This allows greater flexibility around the MG1-6 and MG7-CUB-TED interface allowing access to the BRD by proteases, and upon cleavage of the bait region the thiol ester bond is broken as a result, thus triggering the drop of the TED and entrapment of the protease in question. However the mechanics of this model don't appear to work for the dimerisation, certainly based on the human structure. The final positions of the TED domains demonstrated in red in Figure 4.24 represent the positions after activation, it is difficult to imagine how when moving from a higher position how exactly they would trap a protease. To further understand this the interface between the two subunits of the human dimer was investigated. Disulphide bonds hold the human dimer in position occurring between Cys^{255}A (MG3) and Cys^{408}B (MG4) and Cys^{408}A and Cys^{255}B where, as described earlier A represents one subunit and B represents the other disulphide bonded subunit (Marrero, A., *et al*. 2012). The *Limulus* dimer is formed due to disulphides forming between Cys^{719} of each subunit, which resides in the bait region of the molecule (Iwaki, *et al*. 1996). Using this as the anchor point for the two subunits, the result is something quite different.

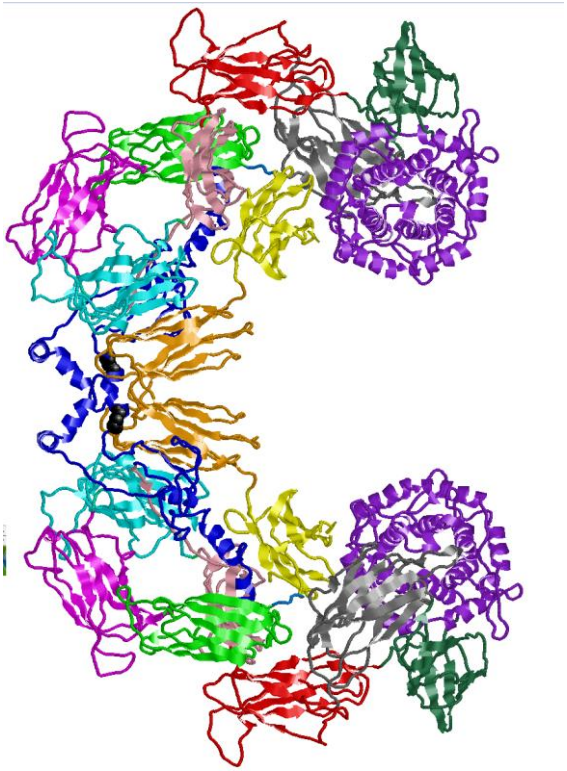


Figure 4.27. The proposed structure for the native *Limulus* α2m dimer. MG1-6 are coloured, orange, yellow, light green, magenta, cyan , and pink respectively. The BRD is shown in blue, MG 7 in red, the Cub domain is dark green whilst the TED and RBD are coloured purple and grey respectively as per the colour scheme used in the domain illustrations in Marrero et al. (Marrero, *et al*. 2012). Cys^{719} of each subunit I shown in black and in spacefill display to show them more clearly. The subunits were manipulated into position so that their relative distances were close to disulphide linkages. The result is a large dimer with flexibility around the disulphide that links the two subunits. Figure generated in RASMOL (Sayle, & Milner-White, 1995).

The disulphide linked dimer created and shown in Figure 4.27 creates a 'Pacman' shaped molecule opposed to the 'Donnie Darko' shaped molecule proposed in Figure 4.26, which allows access to its bait region through the mouth of the 'Pacman'. The 'mouth' of this model measures ~35Å across which is large enough to allow the passage of a 20-25kDa protease into the prey chamber and is consistent with what is seen in the human structure (Marrero, *et al*. 2012). Cleavage of the bait region results in the cleavage of the thiol ester bond between Cys^{999} and Gln^{1003}, by a mechanism in *Limulus* yet to be revealed.

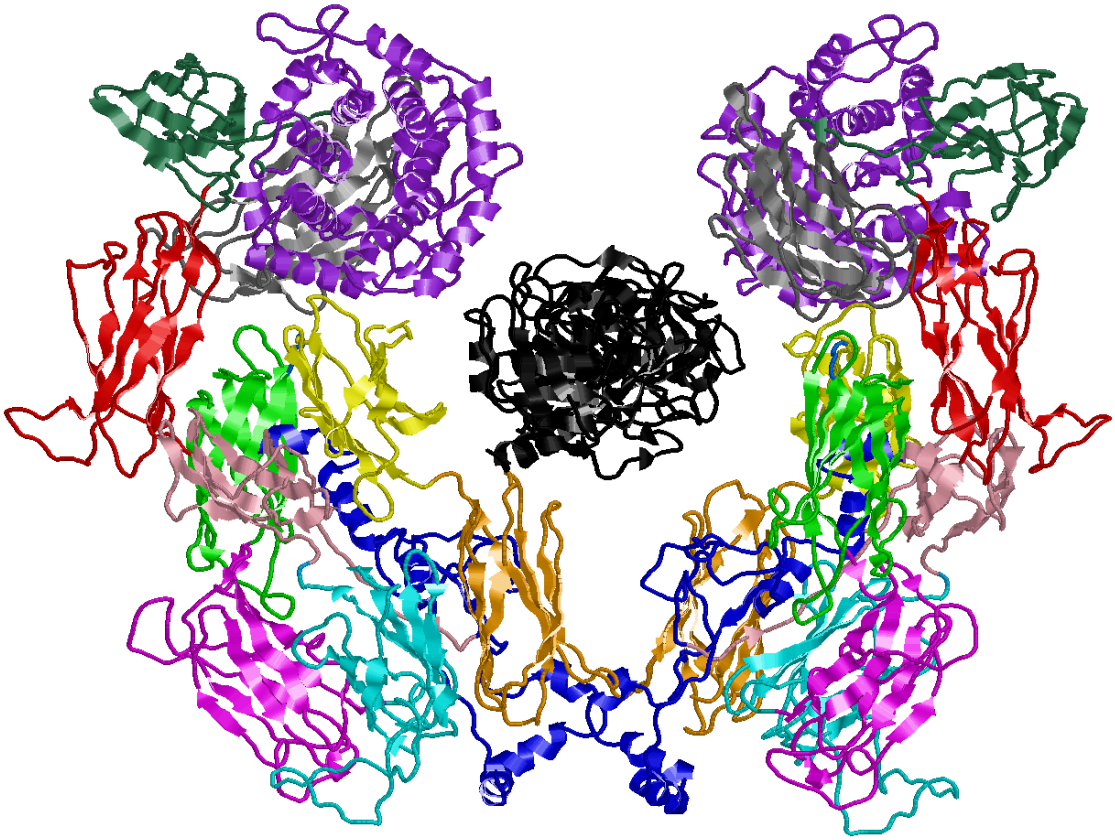


Figure 4.28. Native Limulus α2m dimer model produced by I-TASSER, maintaining the colour scheme in Figure 4.27, with trypsin (shown in black) in the prey chamber. Upon activation the purple TED domains will move centrally thus closing off the 35Å wide mouth behind the trypsin rendering it trapped. The resultant exposure of the RBD then allows for this protease-inhibitor complex to be cleared from circulation. Figure generated by COOT and RASMOL (Emsley, *et al*. 2010; Sayle, & Milner-White, 1995).

This then allows for the TED domain to shift as is the norm for α2m family members upon activation (Janssen, *et al*. 2005; Baxter, *et al*. 2007), this shift in the TED domain results in the entrapment of the protease as 'Pacman' effectively closes his jaws around the molecule and revealing the RBD previously partially covered by the CUB domain which has also descended with the tethered TED.

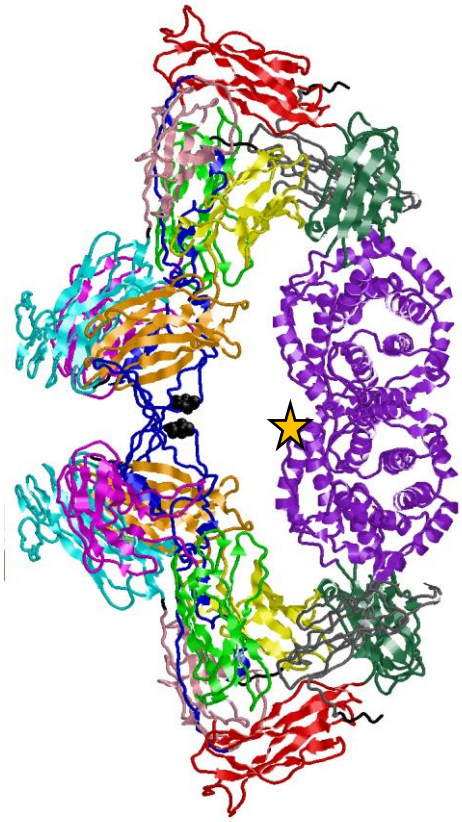


Figure 4.29. The proposed activated *Limulus* dimer based on the I-TASSER results using human α2m-MA activated structure as its principle template. Colour schemes as in Figure 4.26. The two subunits are bound as before by disulphide bonds between the Cys^{719} of their respective bait regions. As with all family members upon activation of *Limulus* α2m the CUB-TED complex has dropped relative to MG7, thus closing the binding pocket behind the protease. The Cys^{999} of one of the TED domains is shown by a gold star to show it's position.

Due to activation the thiol ester of the TED domains are now accessible. In the human structure the broken thiol ester then forms and anchor to the protease covalently binding it, while in *Limulus* this is not the case (Armstrong, & Quigley, 1999). This is likely to occur so that once the thiol ester has been cleaved *Limulus* Cys^{999} is then free to bind to the cytolytic protein limulin (Swarnakar, *et al*. 2000). As described earlier in part 1.2.2.3 - Limulin is a sialic acid binding pentraxin that causes cell lysis by binding surface sialic acid and then inserting itself into the membrane forming a pore. Activated but not native Limulus α2m inhibits this by binding limulin via Cys^{999}. The model activated dimer depicted in Figure 4.29 does not initially appear much smaller than the native dimer shown in Figure 4.27 and 4.28, which appears to contradict the gel electrophoretic evidence shown in Figure 2.11. This may be due to the fact that the native molecule shown in Figure 4.27 is highly flexible and not in one set conformation unlike that seen

in Figure 4.29, and whilst the depiction in 4.27 is accurate as seen in transmission electron microscopy studies, the native form actually takes on multiple 'conformations' (Armstrong, *et al*. 1991).

In summary the bioinfomatic analysis and structure prediction presented here provides structural models of various kinds as well as insights into the mechanism that *Limulus* α2m undergoes upon activation and how this might relate to its functions downstream.

Chapter 5 – General Discussion: Conclusions and Future Work

In invertebrates which lack an adaptive immune system, the importance of the innate immune system cannot be overstated. The roles performed by the innate immune system have allowed phylogenetically ancient species such as the horseshoe crab, *Limulus polyphemus* to remain unchanged evolutionarily for 400-500 million years. Whilst as the scientific world has seen, the adaptive immune system is a marvelously specific and powerful weapon in the immunological arsenal, its initial response time is slow. As a result, to stop initial infection running rampant a background system must stem the tide of the rate of infection. The innate immune system does this with a combination of barriers and protein molecules. The innate immune family members are broad and interestingly have homologues in a wide-range of species. The existence of homologous proteins in animals that lack an adaptive immune system is further evidence of the importance of the innate immune system. This is further emphasized by the fact that for many innate immune proteins; no absence has ever been reported, suggesting that their absence is incompatible with life and further highlighting their significance in the struggle against infection. Thiol ester containing proteins such as α2-macroglobulin can be found in a variety of species, with humans having multiple family members as well as the α2m protease inhibitor itself. In *Limulus polyphemus* the α2-macroglobulin homologue is the only known circulating protease inhibitor and thus its role in protection from proteolytic attack is potentially more important than that of the human α2m which also has a battery of specific active-site inhibitors to aid it in this role (Quigley, & Armstrong, 1983). In addition to its pan-protease inhibitor role *Limulus* α2m has also been shown to be a regulator of the *Limulus* cytosolic system, where activated α2m binds and negates the activity of limulin the cytosolic pentraxin (Swarnakar, *et al*. 1995). The aim of the research presented in this thesis was to shed light on the structure of α2-macroglobulin from *Limulus polyphemus* using crystallographic and bioinformatic techniques. This thesis represents the first ever reported instance of *Limulus* α2m being crystallised. Of the crystals grown those that grew in crystallization well MN04D3 were tested at a synchrotron light source, Diamond Light Source in

Oxford, and they diffracted to a resolution of 6Å. Whilst not a high enough resolution for the production of a model this does provide a starting point for future crystallographic study. The range of possibilities for future crystallographic studies is broad, but firstly the methodology of the production of α2m should be assessed. Here the protein was purified from PEG cut serum using multiple stages of affinity and size exclusion chromatography to remove the contaminants haemocyanin and the pentraxins. PEG cut serum was used due to the reduced levels of haemocyanin, in theory making the purification of α2m easier. As shown in Figure 2.2 however PEG cut serum still contains very high quantities of haemocyanin and this was further evidenced in SEC1 (Figure 2.4) and the SDS-PAGE gel of the products of SEC1 (Figure 2.5). Haemocyanin persistence is but one obstacle as the addition of PEG cut serum to a phosphoethanolamine linked agarose column results in the precipitation of some of the pentraxins (Shrive, *et al*. 2009). One possible method to avoid excess contaminants would be to treat the α2m with trypsin or another protease. Whilst methylamine treatment results in the cleavage of the thiol ester bond, it will not impact greatly upon any other contaminant proteins. Treatment of a 'semi-pure' stock of α2m with trypsin would render the bait region cleaved and thus α2m activated and complexed with trypsin, as well as cleaving any remaining protein contaminants. This requires available proteolytic sites and whilst all of the subunits of haemocyanin, CRP and SAP contain multiple trypsin cleavage sites (as shown in Table 5.1), many of those sites may be inaccessible to the protease due to tertiary and quaternary folding of the protein.

Table 5.1. The subunits of the key serum proteins of *Limulus polyphemus*, showing the number of proposed trypsin cleavage sites assessed by peptide cutter of the ExPASy server. (Gasteiger, *et al*. 2005).

Uniprot Ascension Code	Protein	Length (aa)	Proposed Trypsin Cleavage sites
P06205	Limulus CRP 1.1	242	20
P06206	Limulus CRP 1.4	242	16
P06207	Limulus CRP 3.3	242	19

Q8WQK3	Limulus SAP-like pentraxin	234	14
A2AX56	Limulus Haemocyanin subunit II	629	64
A2Ax57	Limulus Haemocyanin subunit IIIa	627	66
A2AX58	Limulus Haemocyanin subunit IV	624	62
A2AX59	Limulus Haemocyanin subunit VI	638	58
G8YZR0	Limulus Haemocyanin subunit IIIb	628	59

The feasibility of such an experiment would be easily tested as during the purification process pure stocks of haemocyanin, CRP and SAP were produced. These stocks could be incubated with trypsin prior to SDS-PAGE and possibly size exclusion chromatographic analysis to assess the success of the digest. Should the digest prove to be successful then the stocks of α2m that have been purified already could then be treated with trypsin to digest any remaining contaminants. This would then be followed by a SEC run to isolate the reacted α2m with trypsin trapped within from the digested contaminants and any free trypsin not trapped by α2m. This of course would now differ from the α2m stocks that were used in this thesis where α2-macroglobulin was reacted with methylamine, but as the mechanism of structural change has been shown to be the same between methylamine and protease activated α2m, this may have little impact on the crystallisation of the complex. The example above uses only trypsin as an example protease but such work could be carried out with any protease that will sufficiently digest the contaminant proteins to an extent that would ease purification.

Given the importance of the innate immune system in invertebrates such as *Limulus polyphemus*, future work with *Limulus* α2m should focus on its ability to block the cytolytic activity of limulin the sialic acid binding pentraxin. Only the activated form of *Limulus* α2m has been shown to have this effect (Swarnakar, *et al*., 1996), the reaction of α2m in *Limulus* generating a free thiol at Cys^{999}. This free thiol is thought to be crucial to the inhibition of limulin mediated cytolysis as to

treatment of methylamine reacted α2m with the alkylating agent iodoacetamide results in the alkylation of the thiol group and removes the inhibitive ability of the protein on cytolysis (Swanakar, *et al*. 2000). Coupled with the fact that none of the sugars on the surface of the glycosylated *Limulus* α2m are sialic acids (Iwaki, *et al*. 1996) this suggests that this free thiol at Cys^{999} is the primary means of interaction between *Limulus* α2m and limulin. In humans the complement based system of cytolysis is extensive and requires many components, making the simple model in *Limulus* utilising the sialic acid binding of limulin of great interest. Crystallographic studies of the interaction between activated α2m and limulin could provide key functional insights into this cytolytic mechanism as well as shedding light on the evolution of such mechanisms. Isolation and purification of limulin from PEG cut serum has already been performed by the Armstrong group (Swarnakar, *et al*. 2000), suggesting that following the same protocols followed by crystal trials should be a relatively, simple process and an avenue well worth investigating.

In this thesis the structure of the native model of *Limulus* α2m has been proposed based on the structure prediction by the I-TASSER server (Zhang, 2008; Roy, *et al*. 2010). The thesis also leads to the proposal that upon activation *Limulus* α2m undergoes a conformational change that results in the twisting of MG1-6 by approximately 90° as well as the 'dropping' of the thiol ester domain with respect to the CUB domain. This proposal is supported by the structural evidence of the mechanism of C3 activation into C3b (Janssen, *et al*. 2005), where, upon cleavage into C3b, the TED drops a distance of 85Å as shown in Figure 5.1.

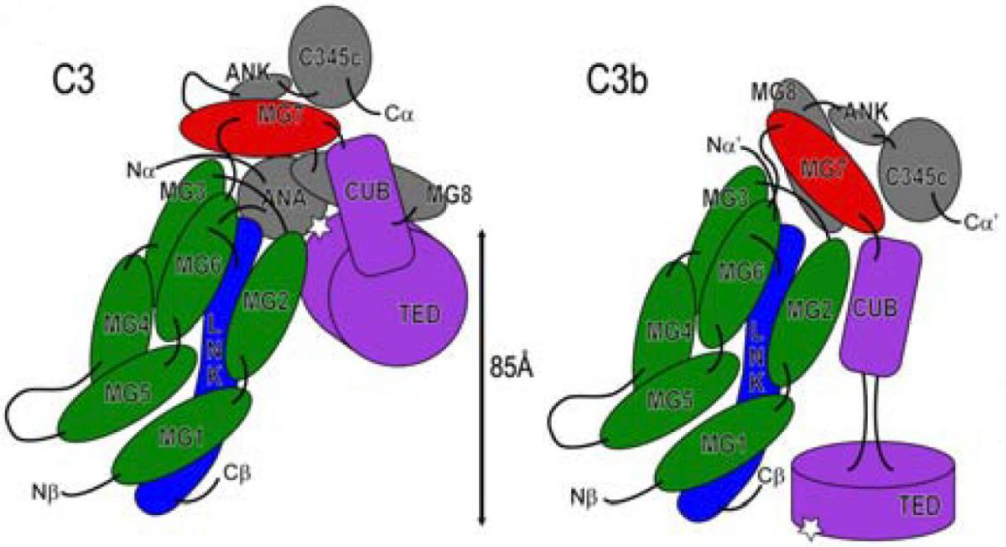


Figure 5.1. Domain representation of the structures of human C3 and the activated form C3b (Janssen, *et al*. 2005; Marrero, *et al*. 2012). MG1-6 are depicted in green, the LNK domain is coloured blue, MG7 is shown in red, the CUB and TED domains in pink and the MG8, ANK and C345c domains are all shown in grey as well as the ANA domain of C3 which is released when C3 is activated. Upon activation the TED domain drops 85Å relative to MG 7.

Analysis of the structures that exist of the α2m family members show that the MG1-6 and MG7-CUB-TED act almost as two independent bodies which twist around one another upon activation and as a result leads to the dropping of the TED. This argument is further strengthened by the results of the initial I-TASSER run used TEP1 of the native form of *Anopheles gambiae* as its principle template, which showed its TED domain tucked up in its inactivated starting position (Baxter, R. *et al*. 2007). The 'activated' *Limulus* α2m model produced by I-TASSER used the human α2m-MA (thus activated) as its principle template and so best represents the activated form of the molecule as shown in Figure 5.2A. Figure 5.2B shows the comparison between the activated and native forms of *Limulus* α2m showing the dropping of the TED domain.

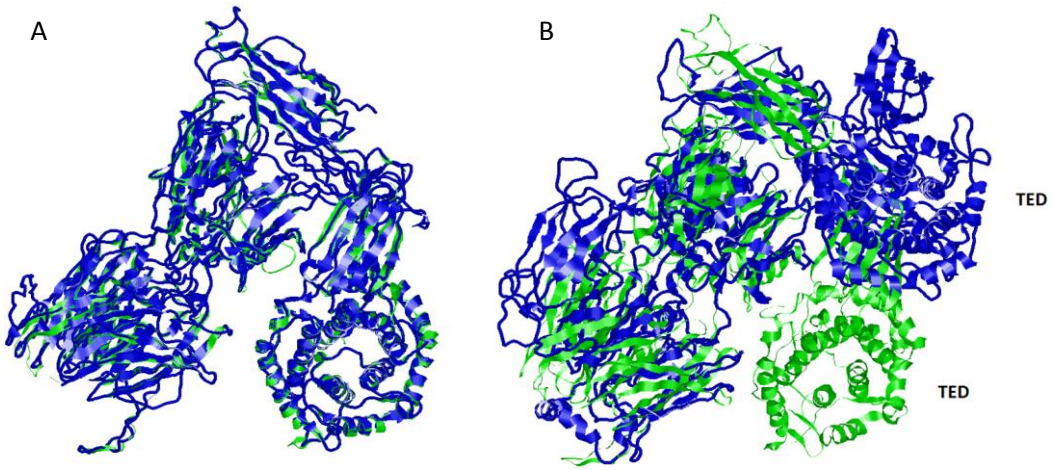


Figure 5.2. A. Overlay of the models for the predicted structure of activated *Limulus* α2m, shown in blue, and the structure for the activated human α2m (Marrero, *et al*. 2012), shown in green. B. Overlay of the models for the predicted structure of native *Limulus* α2m, shown in blue and the model for the predicted structure of the activated *Limulus* α2m shown in green, with the TED domains labelled to highlight the shift upon activation.

The mechanism for activation of *Limulus* α2m proposed here may thus shed further light upon the mechanisms at play once the α2m molecule has been activated such as the inhibition of the limulin mediated cytolytic system (Swarnakar, *et al*. 2000). One experiment that would yield results to support this would be to perform Small Angle X-ray Scattering (SAXS) which provides very low resolution structural data but can clearly show the shape of the molecule. Whilst the lack of structural homogeneity with regard to conformations in the native might confound this process it may provide confirmation of the conformation of the activated *Limulus* α2m dimer proposed in this thesis.

Further afield, the human α2m structure was recently published in its methylamine reacted form (Marrero, *et al*., 2012) so this too requires investigation in terms of its native form, as well as exploration of the structural differences, if any occur, between the methylamine reacted and protease reacted forms. Additionally human α2m provides further avenues for exploration due to its interactions with h-SPD, and various cytokines (Craig-Barnes, 2010; Rehman, *et al*. 2013). H-SPD has been shown to bind to a wide range of sugars and it is known that it will bind to the sugars on the surface of α2m. Unfortunately as fascinating as it would be especially to those

whose research focuses on innate immune proteins there are one or two key obstacles that may prove insurmountable in this line of research. The first problem would lie in ensuring homogeneity between h-SPD binding sites on α2m. Each subunit of the human α2m molecule has eight glycosylation sites, and ensuring homogeneity of binding in each α2m-h-SPD complex may prove a barrier to success. Secondly long chain sugars are highly flexible mainly around their glycosidic bonds greatly reducing the possibility for regular ordered asymmetric units to form and thus further limiting the chances of crystal growth. The cytokine binding activity of α2m however may be a more tractable area for investigation with some factors binding with native α2m and others binding only with activated α2m (Rehman, *et al.*, 2013). Of interest are those cytokine that bind via the Zn^{2+} of free sulfhydryl groups such as IL-1β a key immune mediator. Finally, interactions of α2m with its receptor could be investigated in both species. This work would have to involve the use of reacted α2m as only upon activation does α2m expose its previously concealed receptor binding domain.

This work represents the first chapter in crystallography based structural studies of *Limulus* α2-macroglobulin with a broad horizon for future research opportunities that are set to reveal fascinating secrets, as well as the first proposed activation and molecular reorganisation mechanism consistent with the early work in the literature.

References

Aimes, R. T., Quigley, J. P., Swarnakar, S., Strickland, D., & Armstrong, P. B. 1995. "Preliminary investigations on the scavenger receptors of the amebocyte of the American horseshoe crab, *Limulus polyphemus*", *The Biological Bulletin*, vol. 189, pp. 225-226.

Andersen, G.R., Jacobsen, L., Thirup, S., Nyborg, J., & Sottrup-Jensen, L. 1991. "Crystallisation and preliminary X-ray analysis of methylamine-treated α_2-macroglobulin and 3 α_2-macroglobulin-protease complexes", *FEBS Letters,* vol. 292., No. 1,2, pp. 267-270.

Andersen, G. R., Koch, T., Sørensen, A. H., Thirup, S., Nyborg, J., Dolmer, K., Jacobsen, L., & Sottrup-Jensen, L. 1994a. "Crystallisation of Proteins of the α2-Macroglobulin Superfamily", Annals of the Academy of Sciences, vol. 737, pp. 444-446.

Andersen, G. R., Thirup, S., Nyborg, J., Dolmer, K., Jacobsen, L., & Sottrup-Jensen, L. 1994b. "Low-Resolution X-ray Diffraction Data Obtained from Hexagonal Crystals of Methylamine-Treated α_2-Macroglobulin", *Acta Crystallographica D*, vol. 50, pp. 298-301.

Andersen, G. R., Koch, T. J., Dolmer, K., & Sottrup-Jensen, L. 1995. "Low Resolution X-ray Structure of Human Methylamine-treated α2-Macroglobulin", *Journal of Biological Chemistry*, vol. 270, no. 2, pp. 25113-25141.

Andrews, A. T. 1986. "Electrophoresis: theory, techniques, and biochemical and clinical applications", 2nd Edition, Clarenden Press, New York.

Annapoorani, P., Dhandapany, P. S., Sadayappan, S., Ramasamy, S., Rathinavel, A., & Selvam, G. D. 2006. "Cardiac isoform of alpha-2 macroglobulin – A new biomarker for myocardial infracted diabetic patients", *Atherosclerosis*, vol. 186, pp. 173-176.

Armstrong, P. B., & Quigley, J. P. 1987. "*Limulus* α2-macroglobulin: First evidence in an invertebrate for a protein containing an internal thiol ester bond", *Journal of Biochemistry*, vol. 248, pp. 703-707.

Armstrong, P. B., Quigley, J. P., & Rickles, F. R. 1990. "The *Limulus* Blood Cell Secretes α2-Macroglobulin When Activated", *The Biological Bulletin*, vol. 178, pp. 137-143.

Armstrong, P. B., Mangel, W. F., Wall, J. S., Hainfield, J. F., Van Holde, K. E., Ikai, A., & Quigley, J. P. 1991. "Structure of the α2-Macroglobulin from the Arthropod *Limulus polyphemus*", *Journal of Biological Chemistry*, vol. 266, no. 5, pp. 2526-2530.

Armstrong, P. B., Swarnakar, S., Srimal, S., Misquith, S., Hahn, E. A., AImes, R. T., & Quigley, J. P. 1996. " A cytolytic function for a sialic-acid binding lectin that is a member of the pentraxin family of proteins", *Journal of Biological Chemistry*, vol. 271, pp. 14717-14721.

Armstrong, P.B., & Quigley, J.P. 1999. "α2-macroglobulin: an evolutionarily conserved arm of the innate immune system", *Developmental and Comparative Immunology*, vol. 23, pp. 375-390.

Armstong, P. B. 2006. "Proteases and protease inhibitors: a balance of activites in host-pathogen interaction", *Immunobiology*, vol. 211, pp. 263-281.

Armstrong, P. B., & Conrad, M. 2008. "Blood Collection from the American Horseshoe Crab, *Limulus Polyphemus*", *Journal of Visualised Experiments*, vol. 20, pp. 958.

Armstrong, P.B., 2010. "Role of α_2-macroglobulin in the immune response of invertebrates", *Invertebrate Survival Journal*, vol. 7, pp. 165-180.

Arnold, J.N., Wallis, R., Willis, A.C., Harvey, D.J., Royle, L., Dwek, R.A., Rudd, P.M., & Sim, R.B. 2006. "Interaction of Mannan Binding Lectin with α_2 Macroglobulin via Exposed

Oligomannose Glycans: A Conserved Feature of The Thiol Ester Protein Family?", *The Journal of Biological Chemistry*, vol. 281, no. 11, pp. 6955-6963.

Atha, D. H., & Ingham, K. C. 1981. "Mechanism of Precipitation of Proteins by Polyethylene Glycols; Analysis in Terms of Excluded Volume", *Journal of Biological Chemistry*, vol. 256, no. 23, pp. 12108-12117.

Athippozhy, A, Huang, L., Wooton-Kee, C.R., Zhao, T., Jungsuwadee, P., Stromberg, A.J., & Vore, M. 2011. "Differential gene expression in liver and small intestine from lactating rats compared to afge-matched virgin controls detects increased mRNA of cholesterol biosynthetic genes", *BMC Genomics*, vol.12, 95-111.

Barrett, A.J., Brown, M.A., & Sayers, C.A. 1979. "The Electrophoretically 'Slow' and 'Fast' Forms of the α_2-Macroglobulin Molecule", *Journal of Biochemistry*, vol. 181, pp. 401-418

Basu, S., Binder, R.J., Ramalingam, T., & Srivastava, P.K. 2001. "CD91 is a common receptor for heat shock proteins gp96, hsp90, hsp70 and calreticulin", *Immunity*, vol. 14, pp. 303-313.

Baxter, R. H. G., Chang, C. I., Chelliah, Y., Blandin, S., Levashina, E. A., & Deisenhofer, J. 2007. "Structural basis for conserved complement factor-like function in the antimalarial protein TEP1", *Proceedings of the National Academy of Science, USA*, vol. 104, pp. 11615-11620.

Berg, J. M., Tymoczko, J. L., & Stryer, L. 2006 (6th Edition), *Biochemistry*, W. H. Freeman, New York.

Birkenmeier, G., Kämfper, I., Kratzsch, J., & Schellenberger, W. 1998. "Human leptin forms complexes with α2-macroglobulin which are recognised by the α2-macrogloulin receptor/low density lipoprotein receptor-related protein", *European Journal of Endocrinology*, vol. 139, pp. 224-230.

Borth, W. & Luger, T. A. 1989. “Identification of alpha 2-macroglobulin as a cytokine binding plasma protein. Binding of interleukin-1 beta to "F" alpha 2-macroglobulin.”, *Journal of Biological Chemistry*, vol. 264, no. 10, pp. 5818-5825.

Borth, W. 1992. “α2-Macroglobulin, a multifunctional binding protein with targeting characteristics”, *FASEB Journal*, vol.80, pp.3345-3353

Borth, W., Feinman R.D., Gonias, J.P., & Strickland, D.K. (eds). 1994. “Biology of α2-Macroglobulin, its receptor, and related proteins2”, *Annals of the New York Academy of Sciences*, vol. 737, pp

Branden, C. & Tooze, J. 1999, “Introduction to Protein Structure”, 2nd Edn, Garland Publishing, New York.

Breton, C. B., Blisnick, T., Jouin, H., Barale, J. C., Rabilloud, T, & Pereira da Silva, L. H. 1992. “*Plasmodium chabaudi*, p68 serine protease activity required for merozite entry into mouse erythrocytes”, *Proceedings of the National Academy of Science. USA,* vol. 89, pp. 9647-9651.

Brindley, P. J., Gam, A. A., McKerrow, J. H., & Neva, F. A. 1995. “Ss40: the zinc endopeptidase secreted by infective larvae of *Strongyloides stercoralis*” *Experimental Parasitology,* vol. 80, pp. 1-7.

Bursey, C. R. 1997. “Histological response to injury in the horseshoe crab, *Limulus polyphemus*”, *Canadian Journal of Zoology*, vol. 55, pp. 673-676.

Catsimpoolas, N. 1983. “Proteins in Chromatography Part B: Applications”, Heftmann, E. (Eds.), Elsevier, Amsterdam, pp. 53-74.

Chung, H., Brazil, M. I., Soe, T. T., & Maxfield, F. R. 1999. “Uptake, degradation, and release of fibrilar and soluble forms of Alzheimer’s amyloid beta-peptide by microglial cells”, *Journal of Biological Chemistry*, vol. 274, pp. 32301-32308.

Cohen, F. E., Gregoret, L. M., Amiri, P., Aldape, K., Ratley, J., & McKerrow, J. H. 1991. "Arresting tissue invasion of a parasite by protease inhibitors chosen with the aid of computer modelling", *Biochemistry*, vol. 30, pp. 11221-11229.

Craig-Barnes, H. A., Doumouras, B.S., & Palaniyar, N. 2010. "Surfactant Protein D Interacts with α_2 Macroglobulin and Increases Its Innate Immune Potential", *The Journal of Biological Chemistry*, vol. 285, no. 18, pp. 13461-13470.

Crouch, E.C. 2000. "Surfactant Protein-D and Pulmonary Host Defence", *Respiratory Research*, vol. 1, no. 2, pp. 93-108.

Crouch, E., Nikolaidis, N., McCormack, F., McDonald, B., Allen, K., Rynkiewicz, M., Cafarella, T., White, M., Lewnard, K., Leymarie, N., Zala, J., Seaton, B., Hartshorn, K. 2011. "Mutagenesis of SP-D Informed By Evolution and X-ray Crystallography Enhances Defenses Against Influenza A Virus In Vivo", *The Journal of Biological Chemistry*, vol. 286, no. 47, pp. 40681-40692.

Cuatrecases, P., Wilchek, M, & Anfinsen, C. B. 1968. "Selective Enzyme Purification by Affinity Chromatography", *Proceedings of the National Academy of Science, USA*, vol. 61, no. 2, pp. 636-643.

Cutler, C. W., Arnold R. R., & Schenkein, H. A. 1993. "Inhibition of C3 and IgG proteolysis enhances phagocytosis of *Porphyromonas gingivalis*", *Journal of Immunology*, vol. 151, pp. 7016-7029.

Devriendt, K., Van den Berghe, H. Cassiman, J.J., & Marynen, O. 1991. "Primary Structure of pregnancy zone protein: Molecular cloning of a full-length PZP cDNA clone by the polymerase chain reaction", *Biochimica et Biophysica Acta*, vol. 1099, pp. 95-103.

Dodds, A. W., & Law, S. K. A. 1998. “The phylogeny and evolution of the thiolester bond-containing proteins C3, C4 and α2-macroglobulin”, *Immunological Reviews*, vol. 166, pp. 15-26.

Dolmer, K., & Gettins, P.G.W. 2006. “Three Complement-like Repeats Compose the Complete α2-Macroglobulin Binding Site in the Second Ligand Binding Cluster of the Low Density Lipoprotein Receptor-related Protein”, *Journal of Biological Chemistry*, vol. 281, no. 45, pp. 34189-34196.

Drenth, J. 2007. “Principles of Protein X-Ray Crystallography”, Springer, New York.

Du, Y., Bales, K. R., Dodel, R. C., Liu, X., Glinn, M. A., Horn, J. W., Little, S. P., & Paul, S. M. 1998. “α2-macroglobulin attenuates β-amyloid peptide 1-40 fibril formation and associated neurotoxicity of cultured fetal rat cortical neurons”, *Journal of Neurochemistry*, vol. 70, pp. 1182-1188.

Duus, K., Hansen, E.W., Tacnet, P., Frachet, P., Arlaud, G.J., Thielens, N.M., & Houen, G. 2010. “Direct interaction between CD91 and C1q”, *FEBS Journal*, vol. 277, pp. 3526-3527.

Eggleton, P., Lieu, T.S., Zappi, E.G., Sastry, K., Coburn, J., Zaner, K.S., Sontheimer, R.D., Capra, J.D., Ghebrehiwet, B., & Tauber, A.I. 1994. “Calreticulin is released from activated neutrophils and binds to C1q and mannan-binding protein”, *Clinical Immunology and Immunopathology*, vol. 72, pp. 405-409.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. 2010. “Features and development of *Coot*”, *Acta Crystallographica Section D*, vol. 66, pp. 486-501.

Enghild, J.J., Thogersen, I.B., Roche, P.A., & Pizzo, S.V. 1989. “A conserved region in α2-macroglobulins participates in binding to the mammalian α2-macroglobulin receptor”. *Biochemistry*, vol. 28, pp. 1406-1412.

Fabrizi, C., Businaro, R., Lauro, G. M., Starace, G., & Fumagalli, L. 1999. "Activated α2-macroglobulin increases beta-amyloid (25-35)-induced toxicity in LAN5 human neuroblastoma cells", *Experimental Neurology,* vol. 155, pp. 252-259.

Fredslund, F., Jenner, L., Husted, L. B., Nyborg, J., Andersen, G. R., & Sottrup-Jensen, L. 2006. "The structure of bovine complement component 3 reveal the basis for thioester function", *Journal of Molecular Biology*, vol. 361, pp. 115-127.

Fredslund, F., Lauresen, N. S., Roversi, P., Jenner, L., Oliveira, C. L. P., Pedersen, J. S., Nunn, M. A., Lea, S. M., Discipio, R., Sottrup-Jensen, L., & Andersen, G. R. 2008. "Structure of and influence of a tick complement inhibitor onm human complement component 5", *Nature Immunology*, vol. 9, no. 7, pp. 753-760.

Frenoy, J.P., Bourrillon, R., Lippoldt, R., & Edelhoch, H. 1977. "Stability and Submit Structure of Human α_2-Macroglobulin", *The Journal of Biological Chemistry*, vol. 252, No. 4, pp. 1129-1133.

Fuchs, H., Wallich, R., Simon, M. M., & Kramer, M. D. 1994. "The outer surface protein A of the spirochete *Borrellia burgdoferi* is a plasmin(ogen) receptor", *Proceedings of the National Academy of Science, USA*, vol. 91, pp. 12594-12598.

Gaal, O., Medgyesi, G. A., & Vereczkey, L. 1980. "Electrophoresis in the Separation of Biological Macromolecules", Wiley, New York.

Gaddy-Kurten, D. & Richards, J. S. 1991."Regulation of alpha2-macroglobulin by lutenizing hormone and prolactin during cell differentiation in the rat ovary", *Molecular Endocrinology*, vol. 5, pp. 1280-1291.

Garcia-Ruiz, J. M. 2003. "Nucleation of Protein Crystals", *Journal of Structural Biology*, vol. 142, pp. 22-31.

Gardai, S.J., Xiao, Y.Q., Dickinson, M., Nick, J.A., Voelker, D.R., Greene, K.E., Henson, P.E. 2003. "By Binding SIRPα or Calreticulin/CD91 Lung Collectins Act as Dual Function Surveillance Molecules to Suppress or Enhance Inflammation", *Cell,* vol. 115, pp. 13-23.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., Bairoch, A. "Protein Identification and Analysis Tools on ExPASy Server" in Walker, J. M. (Ed). 2005. "The Proteomics Protocols Handbook, Humana Press, New York.

Gewurz, H., Zhang, X.H., Lint, T.F., 1995. "Structure and Function of the Pentraxins", *Current Opinions in Immunology*, vol. 7, pp. 54-64.

Gigli, I., Von Zabern, I, & Porter, R. R. 1977. "The Isolation and Structure of C4, the Fourth Component of Human Complement", *Journal of Biochemistry*, vol. 165, pp. 439-446.

Glaeser, R. M., & Hall, R. J. 2011. "Reaching the Information Limit in Cryo-EM of Biological Macromolecules: Experimental Aspects", *Biophysical Journal,* vol. 100, pp. 2331-2337.

Gourine, A.V., Gourine, V. N., Tesfaigzi, Y., Caluwaerts, N., Van Leuven, F., & Kluger, M. J. 2002. "Role of α2-macroglobulin in fever and cytokine responses induced by lipopolysaccharide in mice", *American Journal of Physiology – Regulatory, Intergrative and Comparative Physiology*, vol. 283, pp. 218-226.

Hahn, T. (Eds). 2005. "International Tables for Crystallography Volume A: Space-Group Symmetry", Springer, New York

Harrington, J. M., Chou, H. T., Gutsmann, T., Gelhaus, C., Stahlberg, H., Leippe, M., & Armstrong, P. B. 2008. "Membrane pore formation by pentraxin proteins from *Limulus,* the American horseshoe crab", *Journal of Biochemistry*, vol. 413, pp. 305-313.

Hartshorn, K., Chang, D., Rust, K., White, M., Heuser, J., & Crouch, E. 1996, "Interactions of recombinant human pulmonary surfactant protein D and SP-D multimers with influenza A",

American Journal of Physiology - Lung Cellular and Molecular Physiology, vol. 271, no. 5, pp. L753-L762.

Hendrickson, W. A. 1985. "Analysis of Protein Structure from Diffraction Measurement at Multiple Wavelengths", *Transactions of the American Crystallographic Association*, 21.

Herz, J., Goldstein, J.L., Strickland, D.K., Ho, Y.K., & Brown, M.S. 1991. "39kDa protein modulates binding of ligands to low density lipoprotein receptor-related protein/ a2-macroglobulin receptors", *Journal of Biological Chemistry*, vol. 266, pp. 21232-21238.

Herz, J. & Strickland D.K., 2001. "LRP: a multifunctional scavenger and signalling receptor", *Journal of Clinical Investigation*, vol. 108, pp. 779-784.

Herzfeld, J. 1996. "Entropically Driven Order in Crowded Solutions: From Liquid Crystals to Cell Biology", *Accounts of Chemical Research*, vol. 29, no. 1, pp. 31-37.

Hiemstra, P. S., 2002, "Novel Roles of protease inhibitors in infection and inflammation", *Biochemical Society Transactions*, vol. 30, part 2, pp. 116-120

Hinze, V., Miron, A., Moeller, S., Schnabelrauch, M., Wiesman, H. P., Worch, H., & Scharnweber, D. 2012. "Sulfated hyaluronan and chondroitin sulphate derivatives interact differently with human transforming growth factor-β1 (TGF-β1)", *Acta Biomaterialia*, vol. 8, pp. 2144-2152.

Holtet, T.L., Nielsen, K.L., Etzerodt, M., Moestrup,S.K., Ghemann, J>, Sottrup-Jensen, L., & Thorgersen H.C. 1994. "Recombinant α2M receptor binding domain binds to the α2M receptor with high affinity. *Annals of the New York Academy of Science*, vol. 737, pp. 480-482.

House, E., Collingwood, J., Khan, A., Korchazkina, O., Berthon, G., & Exley, C. 2004. "Aluminium, iron, zinc and copper influence the in vitro formation of amyloid fibrils of A beta

(42) in a manner which may have consequences for metal chelation therapy in Alzheimer's disease", *Journal of Alzheimer's Disease*, vol. 6, no. 3, pp. 291-301.

Hughes, S. R., Khorkova, O., Goyal, S., Knaeblein, J., Heroux, J., Riedel, N. G., & Sahasrabudhe, S. 1998. "α2-macroglobulin associates with β-amyloid peptide and prevents fibril formation", *Proceedings of the National Academy of Sciences USA*, vol. 95, pp. 3275-3280.

Husted, L. B., Sørensen, E. S., Armstrong, P. B., Quigley, J. P., Kristensen, L., & Sottrup-Jensen, L. 2002. " Localisaion of Carbohydrate Attachment Sites and Disulfide Bridges in *Limulus* α2-Macroglobulin: Evidence for two forms differing primarily in their bait region sequence", *Journal of Biological Chemsitry,* vol. 277, no. 46, pp. 42698-43706.

Incardona, M. F., Bourenkov, G. P., Levik, K., Pieritz R. A., Popov, A. N., & Svensson, O. 2009. "EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis", *Journal of Synchrotron Radiation*, vol. 16, no. 6, pp. 872-879.

Iwaki, D., Kawabata, S., Miura, Y., Kato, A., Armstrong, P.B., Quigley, J.P., Nielsen, K.L., Dolmer, K., Sottrup-Jensen, L., & Iwanaga, S. 1996. "Molecular Cloning of *Limulus* α_2-Macroglobulin", *European Journal of Biochemistry*, vol. 242, pp. 822-831.

Jaskolski, M. 2010, "Personal remarks on the future of protein crystallography and structural biology", *Acta Biochimica Polonica*, vol. 57, no. 3, pp. 261-264

Jancarik, J. & Kim, S. H. 1991. "Sparse matrix sampling: a screening method for crystallization of proteins", *Journal of Applied Crystallography*, vol. 24, pp. 409-411.

Janssen, B. J. C., Huizinga, E. G., Raaijmakers, H. C. A., Roos, A., Daha, M. R., Nilsson-Ekdahl, K., Nilsson, B., & Gros, P. 2005. "Structures of complement component C3 provide insights into the function and evolution of immunity", *Nature*, vol. 437, no. 22, pp. 505-511.

Kang, D.E., Saitoh, T., Chen, X., Xia, Y., Masliah, E., Hansen, L. A., Thomas, R. G., Thal, L. J., Katzman, R. 1997. “Genetic association of the low-density lipoprotein receptor-related protein gene (LRP), an apolipoprotein E receptor, with late-onset Alzheimer’s disease”, *Neurology*, vol. 49, pp. 56-61.

Keeler, J. 2005, "Understanding NMR Spectroscopy", 1st Edn, Wiley Blackwell, New York

Kelly, S. M., Jess, T. J. & Price, N. C. 2005. “How to study proteins by circular dichroism”2, *Biochimica et Biophysica Acta*, vol. 1751, pp. 119-139.

Khan, M. M., Shibuya, Y., Nakagaki, T., Kambara, T., & Yamamoto, T. 1994. “α2-macroglobulin as the major defence in acute pseudomonal septic shock in the guinea-pig model”, *International Journal of Experimental Pathology*, vol. 76, pp. 21-28.

Kidmose, R. T., Laursen, N. S., Dobo, J., Kjaer, T. R., Sirotkina, S., Yatime, L., Sottrup-Jensen, L., Thiel, S., Gal, P., & Andersen, G. R. 2012. ”Structural basis for activation of the complement system by component C4 cleavage”, *Proceedings of the National Academy of Science, USA*, vol. 109, no. 38, pp. 15425-15430.

Kilian, M., & Reinholdt, J. 1986. “Interference with IgA defense mechanisms by extracellular bacterial enzymes”. In: *Medical Microbiology*, Academic Press, London, pp. 173-208.

Kishore, U., Greenhough, T. J., Waters, P., Shrive, A. K., Ghai, R., Kamran, M. F., Bernal, A. L., Reid, K. B. M., Madan, T., & Chakraborty, T. 2006, "Surfactant proteins SP-A and SP-D: Structure, function and receptors", *Molecular Immunology*, vol. 43, no. 9, pp. 1293-1315.

Kovacs, D. M. 2000. “α2-Macroglobulin in late-onset Alzheimer’s disease”, *Experimental Gerontology*, vol. 35, pp. 473-479.

Kristensen, T., Moestrp, S.K., Gliemann, J., Bendtsen, L., Sand, O., & Sottrup-Jensen, L. 1990. "Evidence that the newly cloned low density-lipoprotein receptor related protein (LRP) is the a2-macroglobulin receptor", *FEBS Letters*, vol. 276, pp. 151-155.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, D., Gibson, H., & Higgins, D. G. 2007. "ClustalW and ClustalX version 2 (2007)", *Bioinformatics*, vol. 23, no. 21, pp. 2947-2948.

Leytus, S. P., Bowles, L. K., Konisky, J., & Mangel, W. F. 1981. " Activation of plasminogen to plasmin by a protease associated with the outer membrane of *E. Coli* ", *Proceedings of the National Academy of Science, USA,*vol. 78, pp.1485-1489.

Li, Z.F., Wu, X.H., & Engvall, E. 2004. "Identification and characterisation of CPAMD8, a novel member of the complement3/α2-macroglobulin family with a C-terminal Kazal domain", *Genomics*, vol. 83, pp.1083-1093.

Lin, M., Sutherland, D. R., Horsfall, W., Totty, N., Yeo, E., Nayar, R., Wu, X., & Schuh, A. C. 2002. "Cell surface antigen CD109 is a novel member of the α2-macroglobulin/C3, C4, C5 family of thioester containing proteins", *Blood*, vol. 99, pp. 1683-1691.

Lindsey, M. L. 2004. "MMP Induction and Inhibition in Myocardial Infarction", *Heart Failure Reviews*, vol. 9, pp. 7-19.

Macomber, R. S. 1998. "A Complete Introduction to Modern NMR Spectroscopy", 1st Edn, Wiley, New York.

Maeda, H., Molla, A., Oda, T., & Katsuki, T. 1987. "Internalisation of serratial protease into cells as an enzyme-inhibitor complex with α2-macroglobulin and regeneration of protease activity and cytotoxicity", *Journal of Biological Chemistry*, vol. 262, pp. 10946-10950.

Man, X. Y., Finnson, K. W., Baron, M., & Philip, A. 2012. "CD109. A TGF-β co-receptor, attenuates extracellular matrix production in scleroderma skin fibroblasts", *Arthritis Research and Therapy*, vol. 14, R144.

Mantovani, A., Garlanda, C., Doni, A., Bottazzi, B. 2008. "Pentraxins in Innate Immunity: From C-Reactive Protein to the Long Pentraxin PTX3", *Journal of Clinical Immunology*,vol. 28, pp 1-13

Markiewski, M. M., DeAngelis, R. A., Lambris, J. D., 2006, "Liver inflammation and regeneration: two distinct biological phenomena or parallel pathophysiologic processes?", *Molecular Immunology,* vol. 43, pp. 45-56

Marrero, A., Duquerroy, S., Trapani, S., Goulas, T., Guevara, T., Andersen, G.R., Navaza, J., Sottrup-Jensen, L., & Gomis-Rüth, F.X. 2012. "The Crystal Structure of Human α_2-Macroglobulin Reveals a Unique Molecular Cage", *Angewandte Chemie International Edition,* vol. 51, no. 14, pp. 3340-3344.

Martin, A. G., Depoix, F., Stohr, M., Meissner, U., Hagner-Holler, S., Hammouti, K., Burmester, T., Heyd, J., Wriggers, W., & Markl, J. 2007. "*Limulus Polyphemus* Haemocyanin: 10 Å Cryo-EM Structure, Sequence Analysis, Molecular Modelling and Rigid-body Fitting Reveal the Interfaces Between the Eight Hexamers", *Journal of Molecular Biology*, vol. 336, pp. 1332-1350.

Matthijs, G., Devriendt, K., Cassiman, J. J., Van den Berghe, H. & Marynen, P. 1992. "Structure of the human alpha-2 macroglobulin gene and its promoter", *Biochemical and Biophysical Research Communications*, vol. 184, no. 2, pp. 596-603.

McPherson, A. 2009. "Introduction to Macromolecular Crystallography", 2nd Edition, Wiley-Blackwell, New York

Melchior, R., Quigley, J. P. & Armstrong, P. B. 1995. “α2-Macroglobulin-mediated Clearance of Proteases from the Plasma of the American Horseshoe Crab, *Limulus polyphemus*”, *Journal of Biological Chemistry*, vol. 270, no. 22, pp. 13496, 13502.

Meyers, C. D., 1991. Role of B-Cell Antigen Processing and Presentation in the Humoral Immune Response. *FASEB Journal.* vol. 5 pp. 2547-2553.

Moestrup, S.K., Kaltoft, K., Peterson, C.M., Pedersen, S., Gliemann, J., & Christensen, E.I. 1990. “Immunochemical identification of the human a2-macroglobulin receptor in monocytes and fibroblasts: monoclonal antibodies define the receptor as a monocyte differentiation antigen” *Experimental Cell Research*, vol. 190, pp. 195-203.

Moestrup, S.K., & Hokland, P. 1992. “Surface expression of the a2-macroglobulin receptor on human malignant blood cells”. *Leukemia Research*. vol. 16, pp. 227-234.

Mold, M. J., Shrive, A. K., Exley, C., 2012. “Serum Amyloid P Component Accelerates the Formation and Enhances the Stability of Amyloid Fibrils in a Physiologically Signification Under-Saturated Solution of Amyloid-beta (42), *Journal of Alzheimers Disease*, vol. 29, no. 4, pp. 875-881.

Molla, A., Oda, T., & Maeda, H. 1987. “Different binding kinetics of Serratia 56K protease with plasma α2-macroglobulin and chicken egg white ovomacroglobulin”, *Journal of Biochemistry*, vol. 101, pp. 199-205.

Narita, M., Holtzman, D. M., Schwartz, A. L. , Bu, G. 1997. “α2-macroglobulin complexes with and mediates the endocytosis of β-amyloid peptide via cell surface low-density lipoprotein receptor-related protein”, *Journal of Neurochemistry*, vol. 69, pp. 1904-1911.

Neels, J.G., van Den Berg, B.M., Lookene, A., Olivecrona, G., Pannekoek, H., & van Zonneveld, A.J. 1999. “The second and fourth cluster of class A cysteine-rich repeats of the low density

lipoprotein receptor-related protein share ligand-binding proteins", *Journal of Biological Chemistry*, vol. 29, no. 44, pp. 31305-31311.

Nyberg, P., Rasmussen, M., & Björck, L. 2004. "α2-Macroglobulin-Proteinase Complexes Protect *Streptococcus pyogenes* from Killing by the Antimicrobial Peptide LL-37", *Journal of Biological Chemistry*, vol. 279, no. 51, pp. 52820-52823.

Oberley, R. E. & Snyder, J. M. 2003, "Recombinant human SP-A1 and SP-A2 proteins have different carbohydrate-binding characteristics", *American Journal of Physiology-Lung Cellular and Molecular Physiology*, vol. 284, no. 5, p. L871-L881.

Overall, C. M. & Lopez-Otin, C. 2002. "Strategies for MMP inhibition in cancer: innovations for the post-trial era", *Nature Reviews: Cancer*, vol. 2, pp. 657-672.

Oxtoby, D. W. 1998. "Nucleation of first-order phase transitions", *Accounts of Chemical Research*, vol. 31, pp. 91-97

Persson, A., Chang, D., & Crouch, E. 1990. "Surfactant Protein D Is a Divalent Cation-Dependant Carbohydrate-Binding Protein", *The Journal of Biological Chemistry*, vol. 265, no. 10, pp. 5575-5760.

Perutz, M. F. 1956. "Isomorphous replacement and phase determination in non-centrosymmetric space groups", *Acta Crystallographica*, vol. 9, pp. 867-873.

Pham, C., T., N., 2006. "Neutrophil serine proteases: specific regulators of inflammation", *Nature Reviews Immunology*, vol. 6, pp. 541-550.

Plaut, A. G. 1983. "The IgA1 proteases of pathogenic bacteria", *Annual Reviews in Microbiology*, vol. 37, pp. 603-622.

Poller, W., Faber-J. P., Klobeck, , G., & Olek, K. 1992. “Cloning of the α2-macroglobulin gene and detection of mutations in two function domains: the bait region and the thiol ester site”, *Human Genetics*, vol. 88, pp. 313-319.

Poon-King, R., Bannan, J., Viteri, A., Cu, G., & Zabriskie, J. B. 1993. “Identification of an extracellular plasmin binding protein from nephritogenic streptococci”, *Journal of Experimental Medicine*, vol. 178, pp. 759-763.

Qu, H., Ricklin, D., Lambris, J. D., 2009, “Recent developments in low molecular weight complement inhibitors”, *Molecular Immunology*, vol. 47, pp. 185-195.

Quigley, J. P., & Armstong, P. B. 1983. “An Endopeptidase Inhibitor, Similar to Mammalian α2-Macroglobulin, Detected in the Hemolymph of an Invertebrate, *Limulus polyphemus*”, *Journal of Biological Chemistry*, vol. 258, no. 13, pp. 7903-7906.

Quigley, J. P., & Armstrong, P. B. 1985. “A Homologue of α2-Macroglobulin Purified from the Hemolymph of the Horseshoe Crab *Limulus polyphemus*”, *Journal of Biological Chemistry*, vol. 260, no. 23, pp. 12715-12719.

Quigley, J. P., Ikai, A., Arakawa, H., Osada, T., & Armstrong, P. B. 1991. “Reaction of proteinases with α2-Macroglobulin from the American Horseshoe crab, *Limulus*”, *Journal of Biological Chemistry*, vol. 266, no. 29, pp. 19426-19431.

Ramadori, G., Knittel, T., Bieber, F., Rieder, H., & Meyer zum Buschenfelde, K. H. 1991. “Dexamethasone modulates alpha2-macroglobulin and apolipoprotein E gene expression in cultured rat liver fat-sorting (Ito) cells”, *Hepatology*, vol. 14, pp. 875-882.

Ramasamy, S., Omnath, R., Rathinavel, A., Kannan, P., Dhandapany, P. S., Annapoorani, P., Balakumar, P., Singh, M., Ganesh, R., & Selvam, G. S. 2006. “Cardiac isoform of alpha 2

macroglobulin, an early diagnostic marker for cardiac manifestations in AIDS patients", *AIDS*, vol. 20, no. 15, pp. 1979-1981.

Ramasamy, S., Chengat, V., Clifton, J. D., Rathinavel, A., Bidulescu, A., Tharmaraian, R., & Selvam, G. S. 2010. "Cardiac isoform of alpha 2 macroglobulin and its reliability as a cardiac marker in HIV patients", *Heart, Lung, and Circulation,* vol. 19, no. 2, pp. 93-95.

Rasmussen, M., Miller, H. P., Bjorck, L. 1999. "Protein GRAB of *Streptococcus pyogenes* regulates proteolysis at the bacterial surface by binding α2-macroglobulin", *Journal of Biological Chemistry*, vol. 274, pp. 15336-15344.

Reed, S. L., Keene, W. E., & McKerrow, J. H. 1989a. "Thiol proteinase expression and pathogenicity of *Entamoeba histolytica*", *Journal of Clinical Microbiology*, vol. 27, pp. 2772-2777.

Reed, S. L., Keene, W. E., McKerrow, J. H., & Gigli, L. 1989b. "Cleavage of C3 by a neutral cysteine protease of *Entamoeba histolytica*", *Journal of Immunology*, vol. 143, pp. 189-195.

Reed, S., Bouvier, J., Pollack, A. S., Engel, J. C., Brown, M., Hirata, K., Que, X., Eakin, A., Hagblom, P., & Gillin, F. 1993. "Cloning of a virulence factor of *Entamoeba histolytica*. Pathogenic strains possess a unique cysteine proteinase gene", *Journal of Clinical Investigation*, vol. 82, pp. 1560-1566.

Rehman, A. A., Alisan, H., & Khan, F.H. 2013. "Alpha-2-Macroglobulin: A Physiological Guardian", *Journal of Cellular Physiology*, vol. 228, no. 8, pp. 1665-1675.

Reneker, L.W., Bloch, A., Xie,L., Overbeek, P.A, & Ash, J.D. 2010. " Induction of Corneal Myofibroblasts by Lens-derived Transforming Growth Factor β1 (TGFβ1): A Transgenic Mouse Model", *Brain Research Bulletin*, vol. 81, pp. 287-296.

Rhodes, G. 2000. “Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models”, 2nd Edn, Academic Press, New York.

Ritchie, R. F. Palomaki, G. E., Neveux, L. M., Navolotskaia, O., Ledue, T. B., & Craig, W. Y. 2004. “Reference distributions for alpha2-macroglobulin: a practical, simple and clinically relevant approach in a large cohort”, *Journal of Clinical Laboratory Analysis*, vol. 18, pp. 139- 147.

Robey, F. A., & Liu, T. Y. 1981. “Limulin: a C-reactive protein from *Limulus polyphemus*,” *Journal of Biological Chemistry*, vol. 256, pp. 969-975.

Roe, S., 1989. “Protein Purification Methods. A Practical Approach”, Harris, E. L. V. % Angal, S. (Eds), IRL Press, New York, pp. 175-244.

Rossmann, M. G. 1990. “The molecular replacement method”, *Acta Crystallographica A*, vol. 46, pp. 73-82.

Rossmann, M. G., & Blow, D. M. 1962. “The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, vol. 15, pp. 24-31.

Roy, A., Kucukural, A., & Zhang, Y. 2010. “I-TASSER: a unified platform for automated protein structure and function prediction”, *Nature Protocols*, vol. 5, pp. 725-738.

Rupp, B. 2010. “Biomolecular Crystallography: Principles, Practise and Application to Structural Biology”, 1st Edn, Garland Science, New York.

Sahu, A., & Lambris, J. D. 2001. “Structure and biology of complement protein C3, a connecting link between innate and acquired immunity”. *Immunological Reviews*, vol. 180, pp. 35-48.

Sand, O., Folkersen, J., Westergaard, J.G., & Sottrup-Jensen, L. 1985, “Characterisation of human preganancy zone protein Comparison with human alpha 2-macroglobulin”, *Journal of Biological Chemistry*, vol. 260, pp. 15723-12735.

Sano, H., Sohma, H., Muta, T., Nomura, S. i., Voelker, D. R., & Kuroki, Y. 1999, "Pulmonary Surfactant Protein A Modulates the Cellular Response to Smooth and Rough Lipopolysaccharides by Interaction with CD14", *The Journal of Immunology*, vol. 163, no. 1, pp. 387-395.

Sarcione, E.J., & Biddle, W.C. 2001. “Elevated serum pregnancy zone protein levels in HIV-1 infected men”, *AIDS*, vol. 15, pp. 2467-2469.

Sarma, J. V., & Ward, P. A., 2011, “The Complement System”, *Cell Tissue Research*, vol. 343, pp. 227-235.

Sayle,R., & Milner-White, E. J. 1995. “RasMol: Biomolecular graphics for all”, *Trends in Biochemical Sciences*, vol.20, no. 9, pp.374.

Schenkein, H. A., Fletcher, H. M., Bodnar, M., & Macrina, F. L. 1995. “Increased opsonisation of a prtH-defective mutant of *Porphyromonas ginivalis* W83 is caused by reduced degradation of complement-derived opsonins”, *Journal of Immunology*, vol. 154, pp. 5331-5337.

Schwalbe, R. A., Schwalbe, Dahlbäck, Coe, J. E., & Nelsestuen, G. L. 1992. “Pentraxin family of proteins interact specifically with phophorylcholine and/or phosphorylethanolamine”, *Biochemistry*, vol. 31, no. 20, pp. 4907-4915.

Selkoe, D. J. 2001. “ Alzheimer’s Disease: Genes, Proteins, and Therapy ”, *Physiological Reviews*, vol. 81, no. 2, pp. 741-766.

Sengupta, J. 2010. “Complementarity of structural biology methods: ribosome in the spotlight”, *Current Science*, vol. 98, no. 12, pp. 1584-1591.

Shi, D.L.,Savona, C., Gagnon, J., Cochet, C., Chambaz, E. D., & Feige, J. J. 1990.”Transforming growth factor-β stimulates the expression of α2-macroglobulin by cultured bovine adrenocortical cells”, *Journal of Biological Chemistry*, vol. 265, pp. 2881-2887.

Shi, Q., Jackowski, G. 1998. In “Gel Electrophoresis of Proteins; A Practical Approach”, 2nd Editions, Hames, B. D. (Eds), IRL Press, New York.

Shrive, A. K., Metcalfe, A. M., Cartwright, J. R., & Greenhough, T. J. 1999. “C-reactive protein and SAP-like pentraxins are both present in Limulus polyphemus haemolymph: crystal structure of Limulus SAP”, *Journal of Molecular Biology*, vol. 290, no. 5, pp. 997-1008.

Shrive, A., Burns, I., Chou, H.T., Stahlberg, H., Armstrong, P.B., Greenhough, T.J., 2009. “Crystal Structures of Limulus SAP-Like Pentraxin Reveal Two Molecular Aggregations”. *Journal of Molecular Biology*, vol. 386, no. 5, 1240-1254.

Sloan, D. J. & Hellinga, H. W. 1999. “Dissection of the protein G B1 domain binding site for the human IgG Fc fragment”, *Protein Science*, vol. 8, pp. 1643-1648.

Sodeinde, O. A., Subrahmanyam, Y. V., Stark, K., Quan, T., Bao, Y., & Goguen, J. D. 1992. “A surface protease and the invasive character of plage”, *Science*, vol. 258, 1004-1007.

Springer, T.A. 1998. “An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components”. *Journal of Molecular Biology*, vol. 283, pp. 837-862.

Skornicka, E.L., Kiytakina, N., Weber, M.C., Tykocinski, M.L., & Koo, P. H. 2004. “Pregnancy zone protein is a carrier and modulator of placental protein-14 in T-cell growth and cytokine production”, *Cellular Immunology*, vol. 232, pp. 144-156.

Solchaga, L.A., Kee, C.H., Roach, S., & Snel, L.B. 2012. “Saftey of recombinant human platelet-derived growth factor-BB in Augment Bone Graft”, *Journal of Tissue Engineering*, vol. 3, pp. 1-6.

Solomon, K. R., Sharma, P., Chan, M., Morrison, P. T, & Finber, R. W. 2004. “CD109 represents a noel branch of the α2-macroglobulin/complement gene family”, *Gene*, vol. 327, pp. 171-183.

Sottrup-Jensen, L., Stepanik, T.M., Kristensen, T., Wiezbicki, D.M., Jones, C.M., Lønblad, P.B., Magnusson, S., & Petersen, T.E. 1984. “Primary Structure of Human α2-Macroglobulin. V: The Complete Structure”, *The Journal of Biological Chemistry*, vol. 259, no. 13, pp. 8318-8327#

Sottrup-Jensen, L., Sand, O., Kristensen, L., & Fey, G.H. 1989. “ The α-Macroglobulin Bait Region: Sequence Diversity and Localisation of Cleavage Sites for Proteinases in Five Mammalian α-Macroglobulins”, *The Journal of Biological Chemistry*, vol. 264, no. 27, pp. 15781-15789.

Swarnakar, S., Melchior, R., Quigley, J. P., & Armstrong, P. B. 1995. “Regulation of the plasma cytolytic pathway of *Limulus polyphemus* α2-macroglobulin”, *The Biological Bulletin*, vol. 189, pp. 226-227.

Swarnakar, S., Quigley, J. P., & Armstrong, P. B. 1996. “The Plasma-Based Cytolytic System of the American Horseshoe Crab, *Limulus polyphemus*: Cooperative Interaction of the Sialic Acid-Binding Lectin Limulin and Thiol Ester-Reacted α2-Macroglobulin”, *Biological Bulletin,* vol. 191, pp. 298.

Swarnakar, S., Asokan, R., Quigley, J. P., & Armstrong, P. B. 2000. “Binding of α2-macroglobulin and limulin: regulation of the plasma haemolytic system of the American horseshoe crab, *Limulus*”, *Journal of Biochemistry*, vol. 347, pp. 679-685.

Tayade, C., Esadeg, S., Fang, Y., & Croy, B.A. 2005. “Functions of alpha 2 macroglobulins in pregnancy”, *Molecular and Cellular Endocrinology*, vol. 245, pp. 60-66

Tharia, H. A., Shrive, A. K., Mills, J. D., Arme, C., Williams, G. T., & Greenhough T. J. 2002. “Complete cDNA sequence of SAP-like pentraxins from Limulus Polyphemus: implications for pentraxins evolution”, *Journal of Molecular Biology,* vol. 316, no. 3, pp. 583-597.

Trommsdorff, M., Borg, J.P., Margolis, B., & Herz, J. 1998. “Interaction of cytosolic adaptor proteins with neuronal apolipoprotein E receptors and the amyloid precursor protein”, *Journal of Biological Chemistry*, vol. 273, pp. 33556-33560.

Van den Elsen, J. M. H., Martin, A., Wong, V., Clemenza, L., Rose, D. R., & Isenman, D. E. 2002. “X-ray Crystal Structure of the C4d Fragment of Human Complement Protein C4”, *Journal of Molecular Biology*, vol. 322, pp. 1103-1115.

Van Iwaarden, J. F., Pikaar, J. C., Storm, J., Brouwer, E., Verhoef, J., Oosting, R. S., Vangolde, L. M. G., & Vanstrijp, J. A. G. 1994, "Binding of Surfactant Protein-A to the Lipid-A Moiety of Bacterial Lipopolysaccharides", *Biochemical Journal*, vol. 303, pp. 407-411.

Van Vyve, T., Chanez, P., Bernard, A., Bousquet, J., Godard, P., Lauweijs, R. &, Sibilie, Y. 1995. “Protein Content in Bronchoalveolar Lavage Fluid of Patients With Asthma and Control Subjects”, *Journal of Allergy and Clinical Immunology*, vol. 95, no. 1, pp. 60-68.

Vandivier, R.W., Ogden, C.A., Fadok, V.A., Hoffmann, P.R., Brown, K.K., Botto, M., Walport, M.J., Fisher, J.H., Henson, P.M., & Greene, K.E. 2002. “Role of surfactant proteins A, D and C1q

in the clearance of apoptotic cells in vivo and in vitro: calreticulin and CD91 as a common collectin receptor complex", *Journal of Immunology*, vol. 169, pp. 3978-3986.

Villems, R., & Toomik, P. 1993. "Overview: Handbook of Affinity Chromatography", Kline, T. (Eds.), *Chromatographic Science Series*, vol. 63, pp. 3-60.

Volanakis, J. E., & Kaplan, M. H. 1971. "Specificity of C-reactive protein for choline phosphates residues of Pnemococcal C-Polysaccharide", *Proceedings of the Society for Experimental Biology and Medicine*, vol.236, pp. 612-614.

Wallace, B. A., & Janes, R. W. 2001. "Synchrotron radiation circular dichroism spectroscopy of proteins: secondary structure, fold recognition and structural genomics", *Current Opinion in Chemical Biology*, vol. 5, pp. 567-571.

Wallis, R., 2007, "Interactions between mannose-binding lectin and MASPs during complement activation by the lectin pathway", *Immunobiology*, vol. 212, pp. 289-299.

Wang, H., Head, P., Kosma, P., Brade, H., Muller-Loennies, S., Sheikh, S., McDonald, B., Smith, K., Cafarella, T., Seaton, B and Crouch, E. (2008), "Recognition of Heptoses and the Inner Core of Bacterial Lipopolysaccharides by Surfactant Protein D". *Journal of Biohemistry* vol. 47, pp. 710-720.

Weiser, J. N., Shchepetov, M., & Chong, S. T. 1997, "Decoration of lipopolysaccharide with phosphorylcholine: a phase- variable characteristic of Haemophilus influenzae", *Infection and Immunity*, vol. 65, no. 3, pp. 943-950.

Wood, P.,(eds) 2006. "*Understanding Immunology*", Harlow: Pearson Education Limited.

Wright, G. S. A., Lee, H. C., Schulze-Briese, C., Grossmann, J. G., Strange, R. W., & Hasnain, S. S. 2012. "The application of hybrid pixel detectors for in-house SAXS instrumentation with a

view to combined chromatographic operation", *Journal of Synchrotron Radiation*, vol. 20, pp. 383-385.

Wyatt, A.R., Yerbury, J.J., Dabbs, R.A., & Wilson, M. R. 2012. "Roles of Extracellular Chaperone in Amyloidosis", *Journal of Molecular Biology*, vol. 421, pp. 499-516.

Yamashiro, D.J., Borden, L.A., & Maxfield, F.R. 1989. "Kinetics of a2-macroglobulin endocytosis and degradation in mutant and wild-type chinese hamster ovary cells", *Journal of Cell Physiology*, vol. 139, pp. 377-382.

Zhang, Y. 2008. "I-TASSER server for protein 3D structure prediction", *BMC Bioinformatics*, vol. 9 pp 40.

Zuo, X, & Woo, P. T. 1997. "Natural anti-proteases in rainbow trout, *Oncorhynchus mykiss* and brook charr, *Salvelinus fontinalis*, and the in vitro neutralisation of fish α2-macroglobulins by the metalloproteases from the pathogenic hemoflagellate, *Cryptobia salmositica*", *Parasitology*, vol. 114, pp. 375-381.

Appendix

Appendix 1

Molecular Dimensions Structure Screen Conditions

Structure Screen 1 – Catalogue Number MD1-01

1	0.02M Calcium chloride dihydrate	0.1M Na Acetate trihydrate pH 4.6	30% v/v 2-methyl-2,4 pentanediol
2	0.2M Ammonium acetate	0.1M Na Acetate trihydrate pH 4.6	30% w/v PEG 4000
3	0.2M Ammonium sulphate	0.1M Na acetate trihydrate pH 4.6	25% w/v PEG 4000
4	None	0.1M Na acetate trihydrate pH 4.6	2.0M Sodium formate
5	None	0.1M Na acetate trihydrate pH 4.6	2.0M Ammonium sulphate
6	None	0.1M Na acetate trihydrate pH 4.6	8% w/v PEG 4000
7	0.2M Ammonium acetate	0.1M tri-sodium citrate dihydrate pH 5.6	30% w/v PEG 4000
8	0.2M Ammonium acetate	0.1M tri-sodium citrate dihydrate pH 5.6	30% v/v 2-methyl-2,4-pentanediol
9	None	0.1M tri-Sodium citrate dihydrate pH 5.6	20% w/v 2-propanol, 20% w/v PEG 4000
10	None	0.1M Na Citrate pH 5.6	1.0M Ammonium

			dihydrogen phosphate
11	0.2M Calcium chloride dihydrate	0.1M Na acetate trihydrate pH 4.6	20% v/v 2-propanol
12	None	0.1M Na Cacodylate pH 6.5	1.4M Na acetate trihydrate
13	0.2M tri-sodium citrate dihydrate	0.1M Na Cacodylate pH 6.5	30% v/v 2-propanol
14	0.2M Ammoniium sulphate	0.1M Na Cacodylate pH 6.5	30% w/v PEG 8000
15	0.2M Magnesium acetate tetrahydrate	0.1M Na Cacodylate pH 6.5	20% PEG 8000
16	0.2M Magnesium acetate tetrahydrate	0.1M Na Cacodylate pH 6.5	30% v/v 2-methyl-2,4-pentanediol
17	None	0.1M Imidazole pH 6.5	1.0M Sodium acetate trihydrate
18	0.2M Sodium acetate trihydrate	0.1M Na Cacodylate pH 6.5	30% w/v PEG 8000
19	0.2M Zinc acetate dihydrate	0.1M Na Cacodylate pH 6.5	18% w/v PEG 8000
20	0.2M Calcium acetate hydrate	0.1M Na Cacodylate pH 6.5	18% w/v PEG 8000
21	0.2M tri-sodium citrate dihydrate	0.1M Na Hepes pH 7.5	30% v/v 2-methyl-2,4-pentanediol
22	0.2M Magnesium chloride	0.1M Na Hepes pH 7.5	30% v/v 2-propanol

	hexahydrate		
23	0.2M Calcium chloride dihydrate	0.1M Na Hepes pH 7.5	28% v/v PEG 400
24	0.2M Magnesium chloride hexahydrate	0.1M Na Hepes pH 7.5	30% v/v PEG 400
25	0.2M tri-sodium citrate dihydrate	0.1M Na Hepes pH 7.5	20% v/v 2-propanol
26	None	0.1M Na Hepes pH 7.5	0.8M K, Na tartrate tetrahydrate
27	None	0.1M Na Hepes pH 7.5	1.5M Lithium sulphate monohydrate
28	None	0.1M Na Hepes pH 7.5	0.8M Na dihydrogen phosphate
		0.8M K dihydrogen phosphate monohyd.	
29	None	0.1M Na Hepes pH 7.5	1.4M tri-Sodium citrate dihydrate
30	None	0.1M Na Hepes pH 7.5	2% v/v PEG 400, 2.0M Amm sulphate
31	None	0.1M Na Hepes pH 7.5	10% v/v 2-propanol, 20% w/v PEG 4000
32	,None	0.1M Tris HCl pH 8.5	2.0M Ammoniium sulphate
33	0.2M Magnesium chloride hexahydrate	0.1M Tris HCl pH 8.5	30% w/v PEG 4000

34	0.2M tri-sodum citrate dihydrate	0.1M Tris HCl pH 8.5	30% v/v PEG 400
35	0.2M Lithium sulphate monohydrate	0.1M Tris HCl pH 8.5	30% w/v PEG 4000
36	0.2M Ammonium acetate	0.1M Tris HCl pH 8.5	30% v/v 2-propanol
37	0.2M Sodium acetate trihydrate	0.1M Tris HCl pH 8.5	30% w/v PEG 4000
38	None	0.1M Tris HCl pH 8.5	8% w/v PEG 8000
39	None	0.1M Tris HCl pH 8.5	2.0M Ammonium dihydrogen phosphate
40	None	None	0.4M K, Na Tartrate tetrahydrate
41	None	None	0.4M Ammonium dihydrogen phosphate
42	0.2M Ammonium sulphate	None	30% w/v PEG 8000
43	0.2M Ammonium sulphate	None	30% w/v PEG 4000
44	None	None	2.0M Ammonium sulphate
45	None	None	4.0M Sodium formate
46	0.05M Potassium dihydrogen phosphate	None	20% w/v PEG 8000
47	None	None	30% w/v PEG 1500

48	None	None	0.2M Magnesium formate
49	1.0M Lithium sulphate monohydrate	None	2% w/v PEG 8000
50	0.5M Lithium sulphate monohydrate	None	15% w/v PEG 8000

Structure Screen 2 – Catalogue Number MD1-02

1	0.1M Sodium chloride	0.1M Bicine pH 9.0	30% w/v PEG monomethylether 550
2	None	0.1M Bicine pH 9.0	2.0M Magnesium chloride hexahydrate
3	2% w/v Dioxane	0.1M Bicine pH 9.0	10% w/v PEG 20,000
4	0.2M Magnesium chloride hexahydrate	0.1M Tris pH 8.5	3.4M 1,6 Hexanediol
5	None	0.1M Tris pH 8.5	25% v/v tert-Butanol
6	0.01M Nickel chloride hexahydrate	0.1M Tris pH 8.5	1.0M Lithium sulphate
7	1.5M Ammonium sulphate	0.1M Tris pH 8.5	12% v/v Glycerol
8	0.2M Ammonium phosphate monobasic	0.1M Tris pH 8.5	50% v/v MPD
9	None	0.1M Tris pH 8.5	20% v/v Ethanol
10	0.01M Nickel chloride hexahydrate	0.1M Tris pH 8.5	20% w/v PEG monomethylether 2000

11	0.5M Ammonium sulphate	0.1M Hepes pH 7.5	30% v/v MPD
12	None	0.1M Hepes pH 7.5	10% w/v PEG 6000, 5% v/v MPD
13	None	0.1M Hepes pH 7.5	20% v/v Jeffamine M-600
14	0.1M Sodium chloride	0.1M Hepes pH 7.5	1.6M Ammonium sulphate
15	None	0.1M Hepes pH 7.5	2.0M Ammonium formate
16	0.05M Cadmium sulphate octahydrate	0.1M Hepes pH 7.5	1.0M Sodium acetate
17	None	0.1M Hepes pH 7.5	70% v/v MPD
18	None	0.1M Hepes pH 7.5	4.3M Sodium chloride
19	None 8% v/v Ethylene glycol	0.1M Hepes pH 7.5	10% w/v PEG 8000
20	None	0.1M Mes pH 6.5	1.6M Magnesium sulphate heptahydrate
21	0.1M Na phosphate monobasic 0.1M K phosphate monobasic	0.1M Mes pH 6.5	2.0M Sodium Chloride
22	None	0.1M Mes pH 6.5	12% w/v PEG 20,000
23	1.6M Ammonium sulphate	0.1M Mes pH 6.5	10% v/v Dioxane
24	0.05M Cesium chloride	0.1M Mes pH 6.5	30% v/v Jeffamine M-600
25	0.01M Cobalt chloride	0.1M Mes pH 6.5	1.8M Ammonium sulphate

	hexahydrate		
26	0.2M Ammonium sulphate	0.1M Mes pH 6.5	30% w/v PEG monomethylether 5000
27	0.01M Zinc sulphate heptahydrate	0.1M Mes pH 6.5	25% v/v PEG monomethylether 550
28	None	0.1M Hepes pH 7.5	20% w/v PEG 10,000
29	0.2M K/Na Tartrate	0.1M Sodium citrate pH 5.6	2.0M Ammonium sulphate
30	0.5M Ammonium sulphate	0.1M Sodium citrate pH 5.6	1.0M Lithium sulphate
31	0.5M Sodium chloride	0.1M Sodium citrate pH 5.6	4% v/v polyethyleneimine
32	None	0.1M Sodium citrate pH 5.6	35% v/v tert-butanol
33	0.01M Ferric chloride hexahydrate	0.1M Sodium citrate pH 5.6	10% v/v Jeffamine M-600
34	0.01M Manganese chloride tetrahydrate	0.1M Sodium citrate pH 5.6	2.5M 1,6 Hexanediol
35	None	0.1M Sodium acetate pH 4.6	2.0M Sodium chloride
36	0.2M Sodium Chloride	0.1M Sodium acetate pH 4.6	30% v/v MPD
37	0.01M Cobalt Chloride hexahydrate	0.1M Sodium acetate pH 4.6	1.0M 1,6 Hexanediol

38	0.1M Cadmium chloride	0.1M Sodium acetate pH 4.6	30% v/v PEG 400
39	0.2M Ammonium sulphate	0.1M Sodium acetate pH 4.6	30% w/v PEG monomethylether 2000
40	2.0M Sodium Chloride	None	10% w/v PEG 6000
41	0.01M Cetyl trimethyl ammoniumbromide	None	0.5M Sodium chloride 0.1M Magnesium chloride hexahydrate
42	None	None	25% v/v Ethylene glycol
43	None	None	35% v/v Dioxane
44	2.0M Ammonium Sulphate	None	5% v/v Isopropanol
45	None	None	1.0M Imidazole pH 7.0
46	None	None	10% w/v PEG 1000, 10% w/v PEG 8000
47	1.5M Sodium Chloride	None	10% v/v Ethanol
48	None	None	1.6M Sodium citrate pH 6.5
49	15% w/v Polyvinylpyrolidone		
50	2.0M Urea		

Structure Screen 1 Eco Screen – Catalogue Number MD1-01-ECO

1	0.02M Calcium chloride dihydrate	0.1M Na Acetate trihydrate pH 4.6	30% v/v 2-methyl-2,4 pentanediol
2	0.2M Ammonium acetate	0.1M Na Acetate trihydrate pH 4.6	30% w/v PEG 4000
3	0.2M Ammonium sulphate	0.1M Na acetate trihydrate pH 4.6	25% w/v PEG 4000
4	None	0.1M Na acetate trihydrate pH 4.6	2.0M Sodium formate
5	None	0.1M Na acetate trihydrate pH 4.6	2.0M Ammonium sulphate
6	None	0.1M Na acetate trihydrate pH 4.6	8% w/v PEG 4000
7	0.2M Ammonium acetate	0.1M tri-sodium citrate dihydrate pH 5.6	30% w/v PEG 4000
8	0.2M Ammonium acetate	0.1M tri-sodium citrate dihydrate pH 5.6	30% v/v 2-methyl-2,4-pentanediol
9	None	0.1M tri-Sodium citrate dihydrate pH 5.6	20% w/v 2-propanol, 20% w/v PEG 4000
10	None	0.1M Na Citrate pH 5.6	1.0M Ammonium dihydrogen phosphate
11	0.2M Calcium chloride dihydrate	0.1M Na acetate trihydrate pH 4.6	20% v/v 2-propanol
12	None	0.1M MES pH 6.5	1.4M Na acetate trihydrate
13	0.2M tri-sodium citrate	0.1M MES pH 6.5	30% v/v 2-propanol

	dihydrate		
14	0.2M Ammoniium sulphate	0.1M MES pH 6.5	30% w/v PEG 8000
15	0.2M Magnesium acetate tetrahydrate	0.1M MES pH 6.5	20% PEG 8000
16	0.2M Magnesium acetate tetrahydrate	0.1M MES pH 6.5	30% v/v 2-methyl-2,4-pentanediol
17	None	0.1M Imidazole pH 6.5	1.0M Sodium acetate trihydrate
18	0.2M Sodium acetate trihydrate	0.1M MES pH 6.5	30% w/v PEG 8000
19	0.2M Zinc acetate dihydrate	0.1M MES pH 6.5	18% w/v PEG 8000
20	0.2M Calcium acetate hydrate	0.1M MES pH 6.5	18% w/v PEG 8000
21	0.2M tri-sodium citrate dihydrate	0.1M Na Hepes pH 7.5	30% v/v 2-methyl-2,4-pentanediol
22	0.2M Magnesium chloride hexahydrate	0.1M Na Hepes pH 7.5	30% v/v 2-propanol
23	0.2M Calcium chloride dihydrate	0.1M Na Hepes pH 7.5	28% v/v PEG 400
24	0.2M Magnesium chloride hexahydrate	0.1M Na Hepes pH 7.5	30% v/v PEG 400
25	0.2M tri-sodium citrate	0.1M Na Hepes pH 7.5	20% v/v 2-propanol

	dihydrate		
26	None	0.1M Na Hepes pH 7.5	0.8M K, Na tartrate tetrahydrate
27	None	0.1M Na Hepes pH 7.5	1.5M Lithium sulphate monohydrate
28	None	0.1M Na Hepes pH 7.5 0.8M K dihydrogen phosphate monohyd.	0.8M Na dihydrogen phosphate
29	None	0.1M Na Hepes pH 7.5	1.4M tri-Sodium citrate dihydrate
30	None	0.1M Na Hepes pH 7.5	2% v/v PEG 400, 2.0M Amm sulphate
31	None	0.1M Na Hepes pH 7.5	10% v/v 2-propanol, 20% w/v PEG 4000
32	,None	0.1M Tris HCl pH 8.5	2.0M Ammoniium sulphate
33	0.2M Magnesium chloride hexahydrate	0.1M Tris HCl pH 8.5	30% w/v PEG 4000
34	0.2M tri-sodum citrate dihydrate	0.1M Tris HCl pH 8.5	30% v/v PEG 400
35	0.2M Lithium sulphate monohydrate	0.1M Tris HCl pH 8.5	30% w/v PEG 4000
36	0.2M Ammonium acetate	0.1M Tris HCl pH 8.5	30% v/v 2-propanol

37	0.2M Sodium acetate trihydrate	0.1M Tris HCl pH 8.5	30% w/v PEG 4000
38	None	0.1M Tris HCl pH 8.5	8% w/v PEG 8000
39	None	0.1M Tris HCl pH 8.5	2.0M Ammonium dihydrogen phosphate
40	None	None	0.4M K, Na Tartrate tetrahydrate
41	None	None	0.4M Ammonium dihydrogen phosphate
42	0.2M Ammonium sulphate	None	30% w/v PEG 8000
43	0.2M Ammonium sulphate	None	30% w/v PEG 4000
44	None	None	2.0M Ammonium sulphate
45	None	None	4.0M Sodium formate
46	0.05M Potassium dihydrogen phosphate	None	20% w/v PEG 8000
47	None	None	30% w/v PEG 1500
48	None	None	0.2M Magnesium formate
49	1.0M Lithium sulphate monohydrate	None	2% w/v PEG 8000
50	0.5M Lithium sulphate monohydrate	None	15% w/v PEG 8000

Appendix 2

A2m crystal tray conditions based on human trials.

Crystallisation Tray No: MN04

Set up By: Michael Nicosia

Date of Trial: A1-C3 - 12/6/12, D1-D5 – 12/7/12, A4-C6 – 21/9/12

Protein and Concentration: Limulus α_2-Macroglobulin (8.319mg/ml)

Summary of Protein Separation: PEG cut serum was cleared of pentraxins by affinity chromatography. Prior to the purification of α_2-Macroglobulin by Gel filtration.

Non-variable conditions: 0.2M Ammonium Citrate pH 6.4

Variable Conditions as indicated by table: 15% PEG 2000, 3350, 4000

25, 50 & 75mM NaF

	1	2	3	4	5	6
A	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 25mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 75mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF
B	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 25mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 75mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF
C	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 4000 25mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 4000 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 4000 75mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF
D	0.2M Ammonium Citrate pH 6.4 5% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 10% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 20% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 25% w/v PEG 3350 50mM NaF	

Crystallisation Tray No: MN09

Set up By: Michael Nicosia

Date of Trial: 25/7/12

Protein and Concentration: Limulus α_2-Macroglobulin-Methylamine (7.79mg/ml)

Summary of Protein Separation: Separation: PEG cut serum was cleared of pentraxins by affinity chromatography. Prior to the purification of α2-Macroglobulin by Gel filtration.

Non-variable conditions: 0.2M Ammonium Citrate pH 6.4

Variable Conditions as indicated by table:

	1	2	3	4	5	6
A	0.2M Ammonium Citrate pH 6.4 5% w/v PEG 2000 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 75mM NaF	0.2M Ammonium Citrate pH 6.4 5% w/v PEG 2000 100mM NaF			
B	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 75mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 2000 100mM NaF			
C	0.2M Ammonium Citrate pH 6.4 25% w/v PEG 2000 50mM NaF	0.2M Ammonium Citrate pH 6.4 25% w/v PEG 2000 75mM NaF	0.2M Ammonium Citrate pH 6.4 25% w/v PEG 2000 100mM NaF			
D						

Crystallisation Tray No: MN11

Set up By: Michael Nicosia

Date of Trial: 17/8/12

Protein and Concentration: Limulus α_2-Macroglobulin-Methylamine (7.79mg/ml)

Summary of Protein Separation: Separation: PEG cut serum was cleared of pentraxins by affinity chromatography. Prior to the purification of α2-Macroglobulin by Gel filtration.

Non-variable conditions: 0.2M Ammonium Citrate pH 6.4

Variable Conditions as indicated by table:

	1	2	3	4	5	6
A	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF		
B	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF		
C	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF		
D	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF		

Crystallisation Tray No: MN13

Set up By: Michael Nicosia

Date of Trial: 15/1/13

Protein and Concentration: Limulus α_2-Macroglobulin (7.79mg/ml)

Summary of Protein Separation: PEG cut serum was cleared of pentraxins by affinity chromatography. Prior to the purification of α_2-Macroglobulin by Gel filtration.

Non-variable conditions: 0.2M Ammonium Citrate pH 6.4

Variable Conditions as indicated by table: % PEG 3350, 35, 50 & 65mM NaF

	1	2	3	4	5	6
A	0.2M Ammonium Citrate pH 6.4 12.5% w/v PEG 3350 35mM NaF	0.2M Ammonium Citrate pH 6.4 12.5% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 12.5% w/v PEG 3350 65mM NaF			
B	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 35mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 65mM NaF			
C	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 35mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 50mM NaF	0.2M Ammonium Citrate pH 6.4 15% w/v PEG 3350 65mM NaF			
D						

Appendix 3

List of amino acids.

Amino Acid	Three letter code	Single letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Appendix 4.

I-TASSER output for the 'native' *Limulus* α2m

Details of the three models of 'native' *Limulus* α2m produced by I-TASSER with the description of the significance of these numbers described below.

Name	C-score	Exp. TM-Score	Exp. RMSD	No. of decoys	Cluster Density
Model 1	0.06	0.72 ± 0.11	9.7 ± 4.6	434	0.1251
Model 2	-0.10			308	0.1251
Model 3	-0.62			124	0.1059

C-score is a confidence score for estimating the quality of predicted models by I-TASSER. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5,2], where a C-score of higher value signifies a model with a high confidence and vice-versa.

TM-score and RMSD are known standards for measuring structural similarity between two structures which areusually used to measure the accuracy of structure modeling when the native structure is known. In case where the native structure is not known, it becomes necessary to predict the quality of the modeling prediction, i.e. what is the distance between the predicted model and the native structures? To answer this question, we tried predicted the TM-score and RMSD of the predicted models relative the native structures based on the C-score.

In a benchmark test set of 500 non-homologous proteins, we found that C-score is highly correlated with TM-score and RMSD. Correlation coefficient of C-score of the first model with TM-score to the native structure is 0.91, while the coefficient of C-score with RMSD to the native

structure is 0.75. These data actually lay the base for the reliable prediction of the TM-score and RMSD using C-score. Values reported in Column 3 & 4 are the estimated values of TM-score and RMSD based on their correlation with C-score. Here we only report the quality prediction (TM-score and RMSD) for the first model, because we found that the correlation between C-score and TM-score is weak for lower rank models. However, we list the C-score of all models just for a reference.

What is TM-score?

TM-score is a recently proposed scale for measuring the structural similarity between two structures. The purpose of proposing TM-score is to solve the problem of RMSD which is sensitive to the local error. Because RMSD is an average distance of all residue pairs in two structures, a local error (e.g. a misorientation of the tail) will araise a big RMSD value although the global topology is correct. In TM-score, however, the small distance is weighted stronger than the big distance which makes the score insensitive to the local modeling error. A TM-score >0.5 indicates a model of correct topology and a TM-score<0.17 means a random similarity. These cutoff does not depends on the protein length.

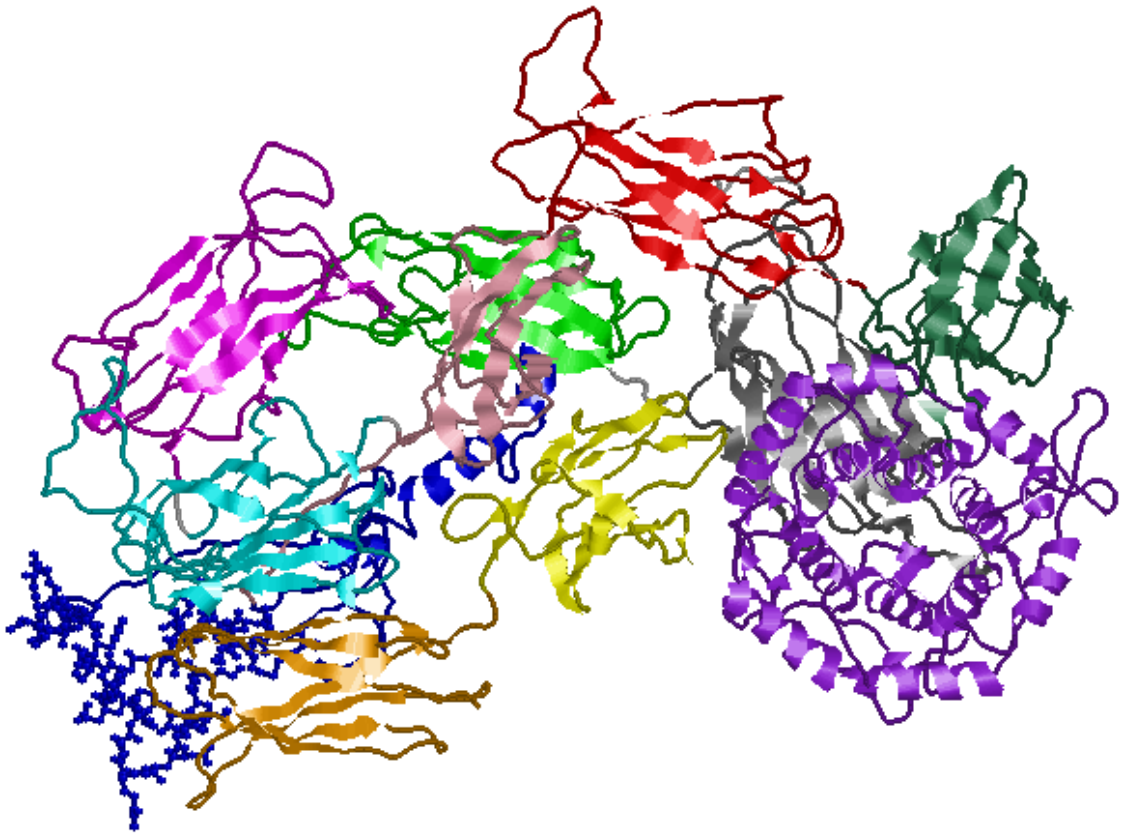
What is Cluster density?

I-TASSER generates full length model of proteins by excising continuous fragments from threading alignments and then reassembling them using replica-exchanged Monte Carlo simulations. Low temperature replicas (decoys) generated during the simulation are clustered by SPICKER and top five cluster centroids are selected for generating full atomic models. The cluster density is defined as the number of structure decoys at an unit of space in the SPICKER cluster. A higher cluster density means the structure occurs more often in the simulation trajectory and therefore signifies

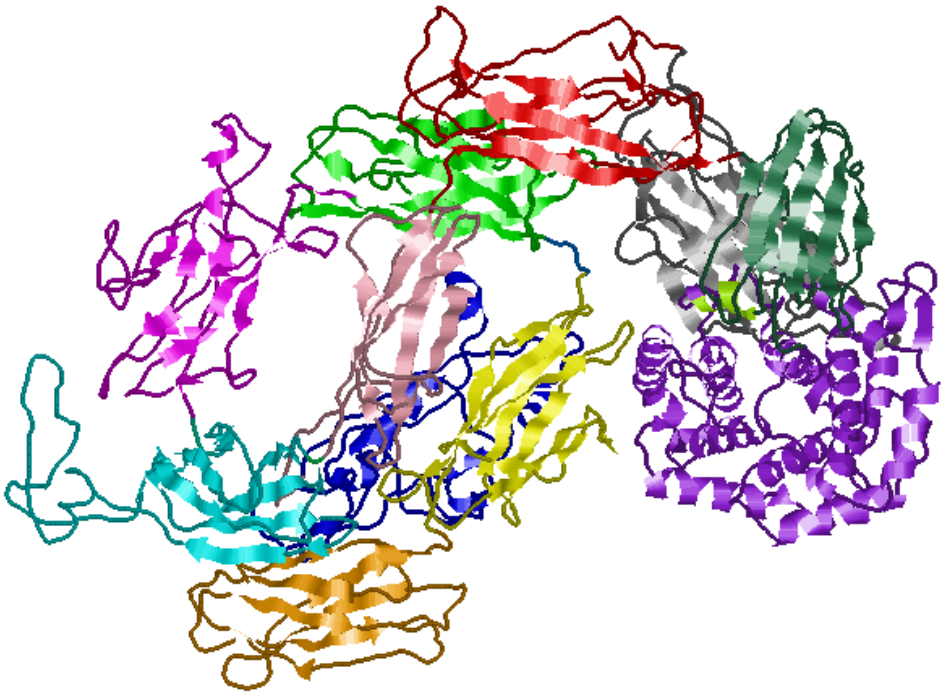
a better quality model. The values in the second last columns of the above mentioned table represents the number of structural decoys that are used in generating each model. The last column represents the density of cluster.

The models below depict the domain in the same colour scheme as seen in the paper for the human structure (Marrero, *et al*. 2012). MG1 - orange, MG2 - yellow, MG3 - light green, MG4 - magenta, MG5 - cyan, MG6 - pink, BRD - blue, MG7 - red, CUB - Green, TED - purple, and RBD - grey.

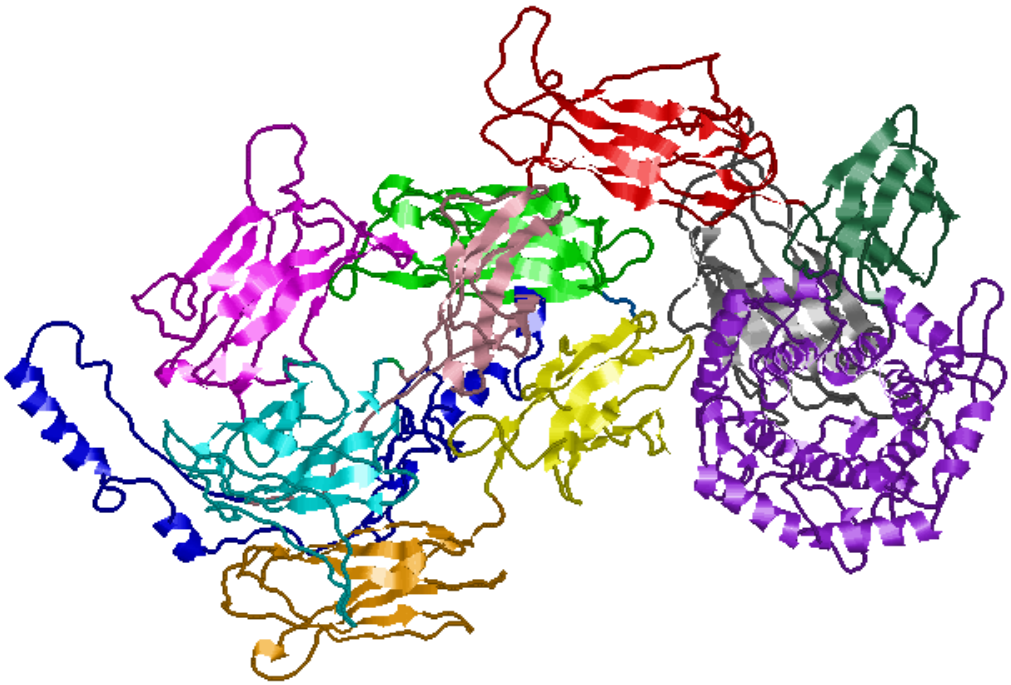
Model 1 of I-TASSER simulation of the 'native' structure of *Limulus* α2m



Model 2 of I-TASSER simulation of the ‘native’ structure of *Limulus* α2m



Model 3 of I-TASSER simulation of the ‘native’ structure of *Limulus* α2m



The Top 10 templates used by I-TASSER in the rendering of model 1.

Rank	PDB - protein	Ident1	Ident 2	Cov	Norm. Z-score
1	2pn5A - TEP1	0.26	0.26	0.85	4.79
2	2pn5A - TEP1	0.25	0.26	0.85	3.49
3	2pn5A - TEP1	0.25	0.26	0.85	10.52
4	2pn5A - TEP1	0.25	0.26	0.85	8.26
5	2pn5A - TEP1	0.25	0.26	0.85	4.30
6	2pn5A - TEP1	0.25	0.26	0.85	4.97
7	2pn5A - TEP1	0.25	0.26	0.85	3.63
8	4acqA - Human α2m	0.32	0.29	0.82	5.23
9	2pn5A - TEP1	0.27	.26	0.66	7.16
10	3cu7A - Human C5	0.21	0.25	0.94	13.56

Where: Ident1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence. Ident2 is the percentage sequence identity of the whole template chains with query sequence. Cov. represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein. Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa. The top 10 alignments reported above (in order of their ranking) are from the following threading programs: 1: MUSTER 2: dPPAS 3: Neff-PPAS 4: PPAS 5: wdPPAS 6: SPARKS-X 7: SP3 8: HHSEARCH2 9: PROSPECT2 10: FFAS03

Top 10 Identified Structural Analogues in PDB, of which there were only 9 available, used by I-TASSER in the prediction of the 'native' model of *Limulus* α2m.

Rank	PDB-protein	TM-score	RMSD	IDEN	COV
1	2pn5A - TEP1	0.841	1.12	0.25	0.847
2	3klsB - C5/SSL7	0.603	5.90	0.167	0.702
3	2bs9A - C3	0.573	5.90	1.66	0.671
4	4a5wA - C5b6	0.370	9.64	0.094	0.544
5	4acqA - Human α2m	0.368	8.94	0.107	0.522
6	3pvmB - C5/CVF	0.363	9.21	0.070	0.527
7	4fxgB - C4/MASP2	0.357	4.50	0.208	0.395
8	2vz9B - fatty acid synthase	0.278	10.42	0.032	0.434
9	3w5mA - α-L-Rhamnosidase	0.273	8.16	0.048	0.368

Ranking of proteins is based on TM-score of the structural alignment between query structure and known structures in the PDB library. RMSD is the RMSD between residues that are structurally aligned by TM-align. IDEN is the percentage sequence identity in the structurally aligned regions. COV. Represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.