Solution Structure of the Central CCP Module Pair of a Poxvirus Complement Control Protein

Colin E. Henderson

A thesis submitted for the degree of Doctor of Philosophy

The University Of Edinburgh

September 2001

Unless otherwise stated, the work described in this thesis is my own work and has not been submitted in whole or in part for a degree or other qualification at this or any other University.

Colin Henderson
September 2001

ABSTRACT

The complement control protein (CP) module (also known as SCR or sushi domain) is prevalent amongst proteins that regulate complement activation. Functional and mutagenesis studies have shown that in most cases two or more neighbouring CP modules comprise specific interaction sites for other molecules. Hence the orientation in space of a CP module with respect to its neighbours and the flexibility of the intermodular junction are likely to be critical for function. The Vaccinia virus complement control protein (VCP) is a complement regulator composed in its entirety of four tandemly arranged CP modules. The central two modules of this protein have been successfully expressed in *Pichia pastoris* and the structure of the modules, numbers 2 and 3 (VCP~2,3), solved using NMR spectrometry. Each module has a typical CP module structure and the two modules do not share an extensive interface. Inspection of fifty structures calculated independently on the basis of the NMR-derived data revealed that the inter-modular orientation does not vary much implying that the 40 linker-to-module NOEs limit the possible movement of the two modules relative to one another during the structure calculation. Module 2 contains a five residue insertion and has a more elongated appearance than module 3. Module 3 appears bulkier due to the laterally protruding hypervariable loop. Module 3 of VCP~2,3 differs a little in structure from module 3 from the VCP~3,4 module pair. The structure of VCP~2,3, allowed reconstruction of VCP~2,3,4 and provides a means of gaining structural information about VCP as a whole using a dissect and rebuild strategy. A recent crystal structure of intact VCP exhibits a somewhat different orientation of the two central modules.

ACKNOWLEDGEMENTS

Completion of this PhD would not have been possible without the assistance and advice of those around me. Nick Mullin was responsible for overseeing much of the molecular biology and protein expression and Dusan Uhrin ran all of the NMR experiments. Brian Smith instructed me on the processing of the spectra and gave good advice and guidance throughout the assignment procedure.

Much of the interpretation of the acquired data was made easier due to Robert Wakefield's 'wonder macros'. Allowing me to use these macros undoubtedly saved me weeks of trial and error programming, so I'm grateful to Rob, who was also a good friend throughout the PhD. Stanislas Blein, when awake, kept me sane at the complement conference in Rhodes. Everyone in 208/209 has helped me at some point and I'm grateful for their time and patience.

Finally, I would like to thank my Mum, who has managed to keep me afloat throughout my university years, my sister for her advice when I was seeking employment and my brother for giving me a place to stay whilst writing my thesis. Paul Barlow gave me the opportunity in the first instance, supervised my work and helped me find the answer to the NMR jigsaw puzzle which became VCP~2.3.

I salute you all.

Colin

Quotes from the wise

Here are some quotes which summarise my PhD research!

"I have achieved nothing that a man of unlimited funds, superb physical endurance, and maximum scientific knowledge could not do"

- Batman

"Inspiration - It's hard to describe. There'll be moments when you get a spark, a gleam of light, and BOOM! you're gone. It seems easy. But then it goes away, and it gets so incredibly hard"

- Robin Williams

"The more you know, the more you don't know.........."

- Aristotle

ABBREVIATIONS

ARIA – Ambiguous restraints used for iterative assignment

B_2GPI – Beta 2 glycoprotein I

BMG – Buffered minimal medium with glucose

BMGY – Buffered minimal medium with glucose and yeast extract

BMM – Buffered minimal medium with methanol

C4BP – C4 binding protein

CCPH – Complement control protein homologue

COSY – Correlated spectroscopy

CP – Complement protein

CCP - Complement control protein

CR1 – Complement receptor 1

DAF – Decay accelerating factor

DNA – Deoxyribonucleic acid

DPFGSE – Double pulsed field gradient spin echo

DQF - Double quantum filtered

EEV – Extracellular enveloped virus

EFG – Epidermal growth factor

FFT – Fast fourier transformation

fH – Factor H

FID – Free induction decay

HAR – Hyper acute rejection

HSQC – Heteronuclear single quantum coherence

IC_{50} – 50% inhibitory concentration

IEV – Intracellular enveloped virus

IgM – Immunoglobulin M

IMP – Inflammation modulatory protein

IMV – Intracellular mature virus

LB – Luria-Bertani broth

LDL – Low density lipoprotein

MAC – Membrane attack complex

MCP – Membrane co-factor protein

MRI – Magnetic resonance imaging

NMR – Nuclear magnetic resonance

NOE – Nuclear overhauser effect

NOESY – Nuclear overhauser enhancement spectroscopy

PCR – Polymerase chain reaction

RAW - Randomisation approach to water suppression

RMSD - Root Mean square deviation

RCA – Regulator of complement activation

SDS PAGE – Sodium dodecyl sulphate polyacrylamide gel electrophoresis

SCR - Short consensus repeat

SCUBA - Stimulated crosspeaks under bleached alpha's

SPICE – Smallpox inhibitor of complement enzymes
TOCSY – Total correlated spectroscopy
VCP – Vaccinia virus complement control protein
VCP~2,3 – VCP modules 2 and 3
VCP~3,4 – VCP modules 3 and 4
VV – Vaccinia virus
YNB – Yeast nitrogen base
YPD – Yeast extract peptone dextrose medium
YPDS – YPD with sorbitol

Chapter 1 - CP modules and the Vaccinia virus complement control protein

1.1 Modular proteins

1.1.1 Introduction

Many proteins are assembled from modules (Baron *et al.*, 1991, Campbell & Downing, 1994, Bork *et al.*, 1996). Such modular or "mosaic" proteins may be large since many modules can be joined together to form the intact protein. This has posed a considerable problem for the study of such proteins by NMR spectroscopy due to the size limitations of the technique.

Modular proteins have been shown to occur in a diverse range of environments, appearing in proteins with a range of functions (Bork *et al.*, 1996). The term 'module' refers to a unit of protein structure, usually less than 200 amino acids in length. Modules are generally encoded by a single exon, have a conserved or recognisable tertiary structure, and tend to occur within proteins that have a diverse range of functions, frequently alongside other modules (Doolittle, 1995, Bork *et al.*, 1996). The overall fold of the module is a result of invariant, or semi-conserved, residues in the protein sequence and it is this "consensus" sequence that determines the module-type (Hegyi & Bork, 1997). Although there are now several instances where intact 3D structures of multi-modular proteins have been solved by X-ray crystallography (e.g. Naismith *et al.*, 1996, Bouma *et al.*, 1999, Schwarzenbacher *et al.*, 1999), crystallisation of modular proteins, which are frequently extended and flexible, has often proved an intractable problem. In most cases, however, it is possible to identify modules and to produce in isolation, single modules or pairs of contiguous modules (Campbell & Downing, 1998), thanks to the rapid advances in genomics and recombinant DNA technology. Several expanding databases that store genomic sequences have been instrumental in allowing the identification of open reading frames that may encode modular proteins. Pattern-searching can indicate the number of repeated, or individual module types that may be present (Ponting *et al.*, 1999). Techniques such as the polymerase chain reaction (Saiki *et al.,* 1985, 1988), which can allow the propagation of a specific piece of DNA conferring a whole, or part of, a gene is now routine. This technique, when combined with the many available expression systems that can utilise plasmid vectors for expression of a protein in a controlled way, has led to the recombinant expression of many different module-types in quantities suitable for structure determination. Recombinant DNA technology has also allowed pairs of contiguous repeats and longer fragments to be structurally analysed. The resulting information can be used in a "dissect and rebuild strategy" to gain an insight into the structure of the intact multiple modular protein.

1.2 The role of modules

Modular proteins may be composed of multiple examples of a single module type, or may contain different module types (mosaic proteins) (Doolittle & Bork, 1993, Herissat & Bork, 1996). The combination of a limited but diverse range of module-types allows a wide-range of possible binding-surfaces. Futhermore, the orientation of one module with respect to its neighbours, whether modules of the same type or not, can vary, thus conferring extra diversity to the shapes and surface features of the parent proteins. Binding sites frequently extend over more than one module.

In this way, a relatively small number of building blocks may be assembled in different ways to provide a wide range of binding activities. An understanding of how the different combinations of modules are arranged, by studying the structures of pairs of modules, could help pave the way for modelling the structures of large, intact proteins.

Some examples of tandemly arranged modules that have been studied include the structure of two immunoglobin-like domains from CD4 (Wang *et al.*, 1992, Brady *et al.*, 1993), fibronectin type 1 domains (Williams *et al.*, 1994), fibronectin type III domains involved in cell adhesion (Leahy *et al.*, 1996), and the fibronectin type I/epidermal growth factor-like domains of tissue type plasminogen activator (Smith *et al.*, 1995). In these examples, the module pairs were determined to have a rigid intermodular orientation. Other examples of module pairs exhibit a significant degree of flexibility between the modules and examples include immunoglobins in the CD2 receptor (Jones *et al.*, 1992), the tandem phosphatase SH-PTP2 phospholipid recognition domains (Eck *et al.*, 1996), adjacent zinc finger motifs (Bruschweiler *et al.*, 1995, Ominchinski *et al.*, 1992) and modules within the cell recognition region of fibronectin (Leahy *et al.*, 1996). Other module pairs studied have shown no apparent interactions at the modular interface (Hansen *et al.*, 1994, Nakaseko *et al.*, 1992).

The orientation of one module with respect to the next, which could be one of the most important considerations with respect to linking structure to function, is defined by the intermodular interface and the extent of the contact between the modules (Barlow *et al.*, 1993, Wiles *et al.*, 1997, Casasnovas *et al.*, 1999, Scwarzenbacher *et al.*, 1999, Spitzfaden *et al.*, 1997). This seems to be dependent on the number of amino acids that link the repeats and the proximity of loops and turns within the neighbouring modules. The extent of the intermodular interface in module-

module interactions studied to date is attributed to the predominantly non-bonding (ionic, hydrophobic, van der Waals) and H-bonding interactions that are present at the interface. In certain cases, however, metal ions (Huber *et al.*, 1994) or disulphide bonds (Eck *et al.*, 1996) have been shown to participate.

1.3 CP Modules

1.3.1 Introduction to CP modules

When considering the contribution of module-module junctions to the overall structure and function of proteins, a well-studied example is the family of modular proteins that contain complement protein (CP) modules. These modules are so called because of their presence in many of the proteins that are part of the complement regulatory system (Reid *et al.*, 1986), a major component of the body's innate immune system (Figure 1 i). Although the CP module has been most commonly associated with proteins from the complement system (Morgan & Harris, 1999), it has been identified in many other animal proteins. These include proteins in many mammalian tissues and some examples are the blood clotting factor XIII beta chain (Day *et al.*, 1989), the GABA$_b$ receptor subtype 1a (Kaupman *et al.*, 1997, Hawrot *et al.*, 1998), the alpha chain of the interleukin 2 receptor, haptoglobin, endothelial leukocyte adhesion molecule 1 and thyroid peroxidase (Day *et al.*, 1989). Prior to the completion of the work described in this thesis, 3D structures have been determined for two individual CP modules (Norman *et al.*, 1991, Barlow *et al.*, 1991, 1992), for three pairs of CP modules

Figure 1 i - CP modules in complement proteins

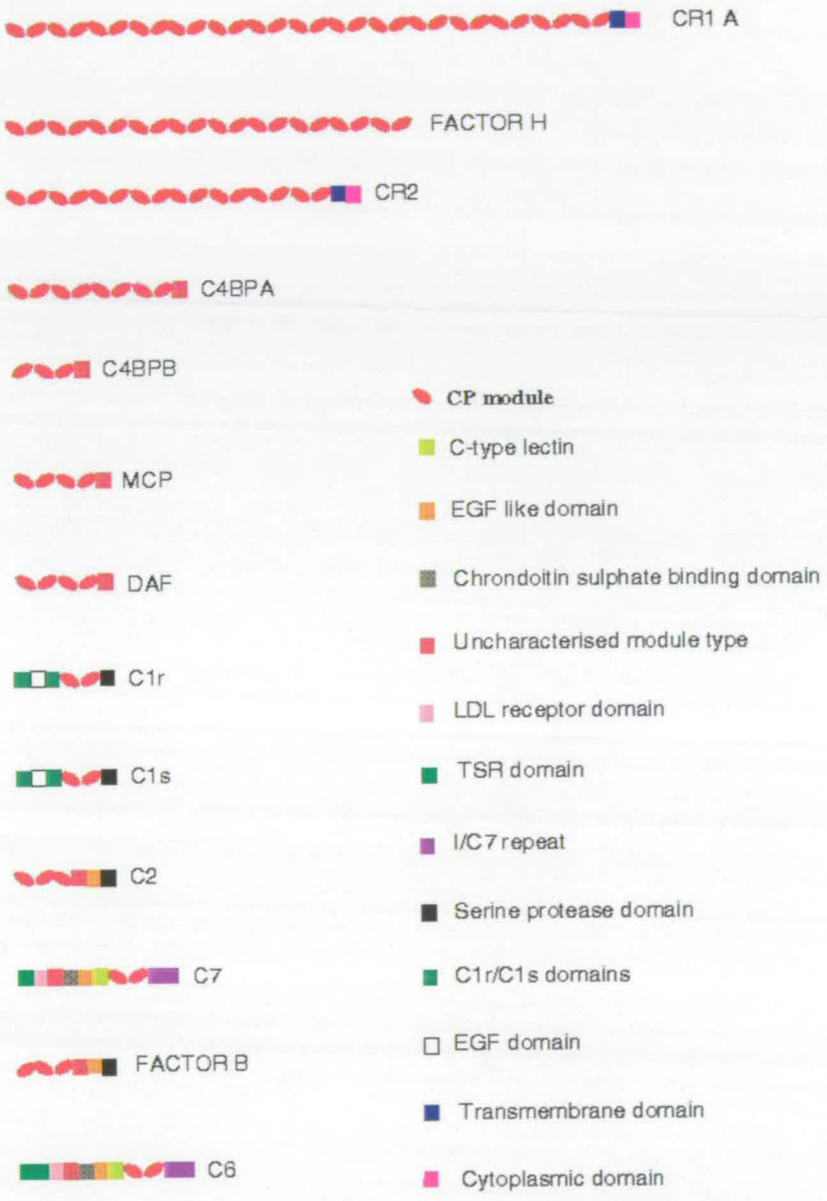

Figure 1 i - Proteins in the complement system that contain CP modules. Many of the proteins contain more than one module type, but have at least two contiguous CP modules.

(Barlow *et al.*, 1993, Wiles *et al.*, 1997, Casasnovas *et al.*, 1999), for a CP-module serine-protease domain pair from C1r (Gaboriaud *et al.*, 2000), and for four CP modules plus one atypical CP-like module from β2 glycoprotein 1 (Bouma *et al.*, 1999, Schwarzenbacher *et al.*, 1999).

The CP module is characterised by a unique consensus sequence that includes glycines, prolines and hydrophobic and aromatic residues (Reid & Day 1989). Most characteristically, each CP module also contains four highly conserved cysteine residues that form two disulphide bonds close in space to the N- and C- termini of the CP module. These disulphide bonds are formed in a I-III, II-IV manner and underpin a structural scaffold that is likely common to all CP modules. This consists of several (normally between five and eight) β-strands that surround a hydrophobic core. The β-strands are linked by H-bonds to form anti-parallel β-sheets. A virtually invariant tryptophan residue lies in the N-terminal half of the module, next to the (I-III) disulphide bond. The CP modules are joined to each other in a C – N (head-to-tail) fashion.

1.3.2 Occurrence of CP modules and structural studies

The 3D structure of a CP module is shown in figure 1 ii. The structures of those CP modules studied to date reveal a high degree of similarity although sequence homology may be low (20% in some cases). The proteins in which CP modules occur, however, have different binding properties. This suggests that non-conserved residues at the intermodular interface or linker, and variations in the distribution of charged residues around the protein, are likely responsible for functional diversity

Two CP modules have been identified in each of C1r and C1s, proteins involved in the initiation of the classical complement activation pathway.

Figure 1ii - Structure of a typical CP module

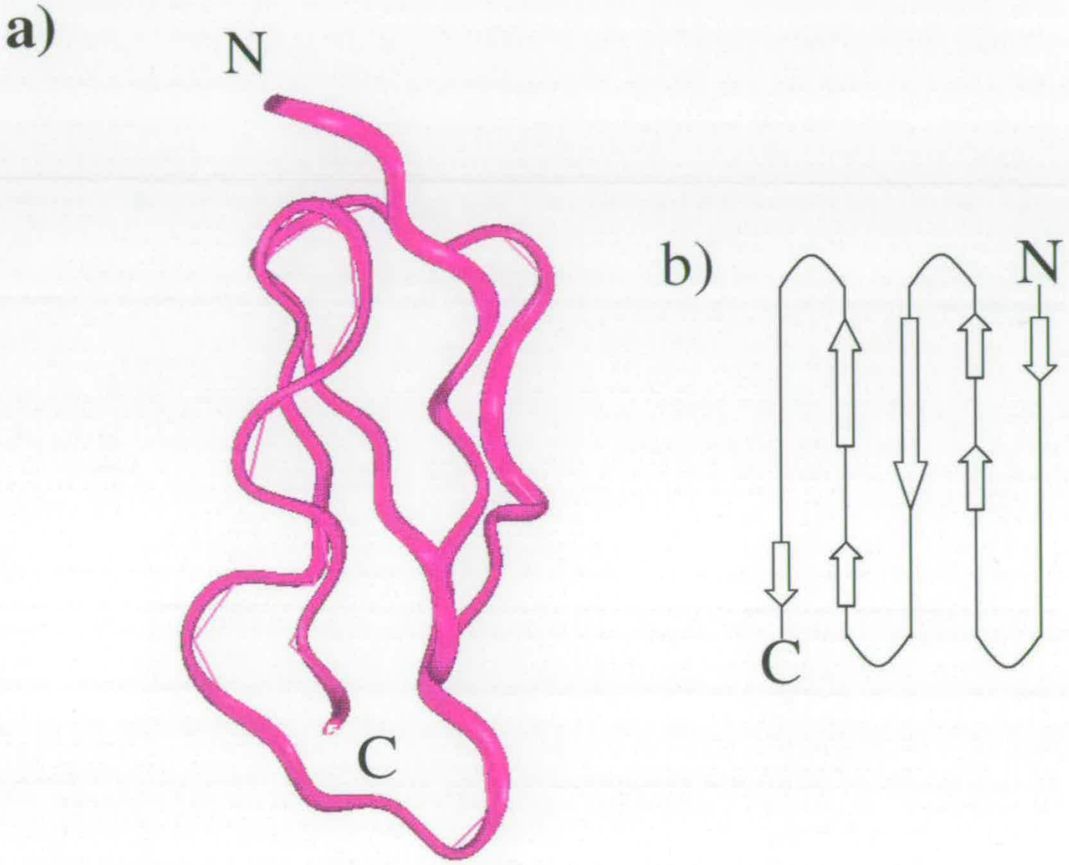

Figure 1 ii - (a) Fold of a typical CP module. (b) arrangement of β-strands that form anti-parallel β-sheets that contribute to maintaining the overall fold.

Pairs of CP modules are also found in other complement proteins such as C7 and C6, which are involved in formation of the membrane attack complex (MAC), the terminal complex in the complement cascade. C2 and factor B, proteins involved in formation of key convertases in the complement cascade contain three contiguous CP modules, as does the β-chain of C4b binding protein, which regulates complement following classical activation of the cascade. Other complement regulators such as decay accelerating factor (DAF) and membrane co-factor protein (MCP) contain four consecutive modules, whereas C4b binding protein's α-chain has eight consecutive modules. Two complement receptors, complement receptor type 1 (CR1) and complement receptor type 2 (CR2) have 30 and 15 (or 16) modules respectively. Factor H, another complement regulator, contains 20 tandemly arranged CP modules.

The diversity in binding of these proteins may be attributed to the variations in the sequence of each module due to local insertions, deletions and substitution of residues in loops and turns. The first example of a CP module to be structurally characterised was the 16th module of human factor H (Barlow *et al.*, 1991), a protein involved in the inhibition of the complement cascade, using solution state NMR techniques. This structure determination was closely followed by the solving of the structure of another module from the same protein (module 5) (Barlow *et al.*, 1992), before the structure of the first CP module pair was published, namely factor H modules 15 and 16 (fH~15,16) (Barlow *et al.*, 1993). These two modules were found not to have an extensive interface between them, since there was a lack of contacts between amino acids in different modules whose close proximity would have been readily detectable as nuclear overhauser effects (NOEs) by NMR. The next module pair to be structurally characterised was that containing modules 3 and 4, i.e. the C-terminal half, of the four CP-module vaccinia virus complement control protein

(VCP) (Wiles *et al.*, 1997). In VCP, which is another complement inhibitor, module 3 has 19% sequence identity compared to fH~15, and module 4 has 17% sequence homology to fH~16. In the case of VCP modules 3 and 4, a well-defined intermodular interface was observed. The intermodular orientations were described mathematically by creating x, y and z axes for each module based on its inertia tensor, and calculating the angles through which one module would have to rotate to be aligned with the other. The z axis in each case was the long axis of the module, and the (mutually orthogonal) x and y axes were defined on the position of the α-carbon of the conserved tryptophan that was judged to occupy an equivalent position in each module. The angles were then described as the 'tilt' (θ), the angles through which one module has to move with respect to the other to align the z-axes, (figure 1 iii) 'twist' (ω) and 'skew' (δ), the angles through which the long axis would have to rotate to align the x and y axes. Angles were calculated for each member of the NMR-derived ensembles and averaged (± standard deviation). In the case of the pair from factor H, fH~15,16, the angles were θ, 50 ± 13 $^{\circ}$; δ, 155 ± 23 $^{\circ}$; ω, 230 ± 17 $^{\circ}$; these may be compared with the angles for VCP~3,4 - θ, 59 ± 4 $^{\circ}$; δ, 324 ± 5 $^{\circ}$; ω, 22 ± 6 $^{\circ}$. Thus as well as being more flexible fH~15,16 appears to have different intermodular angles to VCP~3,4. These findings gave the first clear indication that modelling pairs of CP modules by homology with known structures would not be straightforward and would require more examples to establish patterns that could potentially correlate the residues close to an intermodular interface to a particular orientation.

The structure of the first two CP modules of membrane cofactor protein (MCP, CD46) have been solved by X-ray crystallography (Casasnovas *et al.*, 1999).

Figure 1iii - Angles used to describe different orientations of CP-modules with respect to one another.

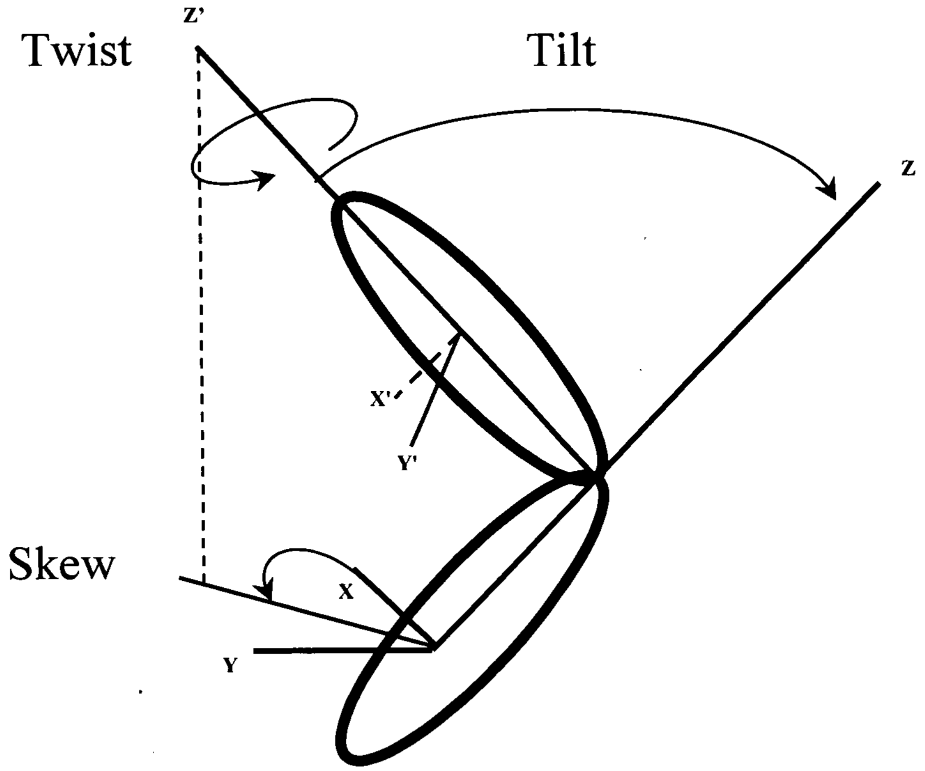

Figure 1 iii - The angles used to describe intermodular orientation are shown. The z-axis of each module is defined as the principal inertia tensor. The x and y axes are defined by the position of the conserved tryptophan residue. The relationship of the second module to the first module is described by; Tilt - the angles through which the second module has to move to align with the z axis of the reference module; Twist - the angle through which the second module would have to turn to align its x and y axes with the x and y axes of the reference module; Skew - the angle at which the original tilt was made with respect to the x-axis. The x axis is defined as the CHα bond of the conserved tryptophan, with the y axis 90^0 to x and perpindicular to the z axis.

Each of the two modules at the N-terminus of MCP, which collectively comprise the Epstein Barr virus-binding portion of MCP, showed characteristics typical of the other CP modules that had been characterised, with β-sheets surrounding a predominantly hydrophobic core. The crystal form of MCP modules 1 and 2 (MCP~1,2) was based on six independent module pairs which form a hexamer in the unit cell, although no biological relevance has been attached to the formation of the oligomeric unit. The two modules, MCP~1 and MCP~2 are joined by a four residue linker sequence. A calcium ion was present at the interface in the crystal structure and has been implicated as a potential stabilising factor at the interdomain interface in this particular case. The calcium ion, however, is not thought to be of any physiological importance. Very few interdomain contacts were apparent in the module pairs as they appeared in the crystal structure. Inspection of the interface of each MCP~1,2 pair revealed that there was a certain degree of flexibility, since tilt angles varied by ~15 °. Further movement might have been restricted by contacts associated with the hexameric oligomer formed within the crystal. These results implied that intermodular movement is a real phenomenon, as opposed to a result of a lack of distance restraints in NMR derived structures.

One of the most interesting CP module-containing structures to emerge recently was the crystal structure of Beta 2-Glycoprotein I, which contains four CP modules together with a fifth atypical example (Bouma *et al.*, 1999, Schwarzenbacher *et al.*, 1999). The protein is present in human plasma and is involved in phospholipid binding. This crystal structure again shows that the CP module interfaces can differ drastically although in this case there was no evidence of flexibility. The CP1-CP2 interface shows ~0 ° tilt angle, but has a twist angle of 115 °. The interface is defined by four H-bonds and several hydrophobic interactions. The CP2-CP3 pair have a tilt angle of ~30 °, a twist angle of ~110 °, and a rather low buried interface surface area (270 Å^2). Three H-bonds stabilise the interface and eleven residues are involved in van der Waals interactions, which contribute to the solvent inaccessible nature of the interface. CP3-CP4 have a tilt angle of ~55 °, a twist of ~40 ° and a higher buried surface area of 318 Å^2, which is defined by hydrophobic interactions between eleven side chains – there are no H-bonds at this interface.

Another recent CP module to be crystallised is in the CP module-serine protease domain pair from C1s, a protein involved in the events following the classical pathway of complement activation (Gaboriaud *et al.*, 2000). The structures of several CP modules are shown in figure 1 iv. It can be seen that the modules are similar in structure, making it very interesting when considering the relationship between sequence and function. By looking at some of the proteins in the complement system, this can be better illustrated.

1.4 Complement and the Vaccinia virus complement control protein

1.4.1 The complement system

The complement system is part of the mammalian innate immune system and serves as one of the first lines of defence against pathogenic attack and entry of foreign material (Law & Reid (eds), 1998). The complement cascade is summarised in Figure 1 v. It is divided into two main pathways, the classical and alternative routes of activation and it serves to target any foreign material for destruction (an example would be the cell surfaces of microorganisms).

Figure 1 iv - Examples of CP module structures

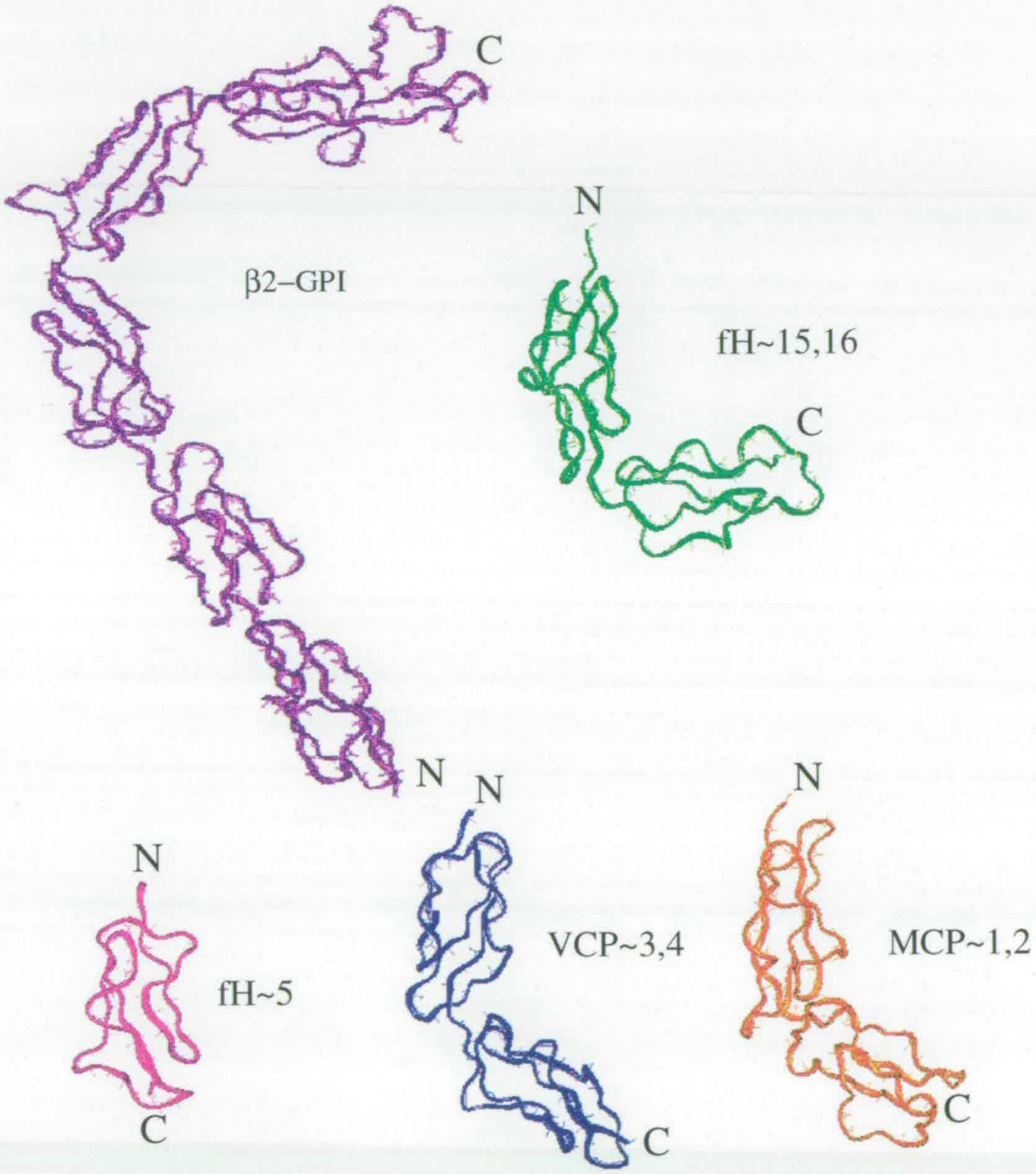

Figure 1 iv - Examples of known CP module structures for β2-glycoprotein I, factor H modules 15 and 16, factor H module 5, Vaccinia virus complement control protein modules 2 and 3 and membrane cofactor protein modules 1 and 2. The individual modules have a similar overall fold, attributed to the homology between the protein sequences that characterise the CP module family. β2-glycoprotein I has 4 CP modules and 1 CP-like module at the C terminus.

Figure 1v - The complement cascade

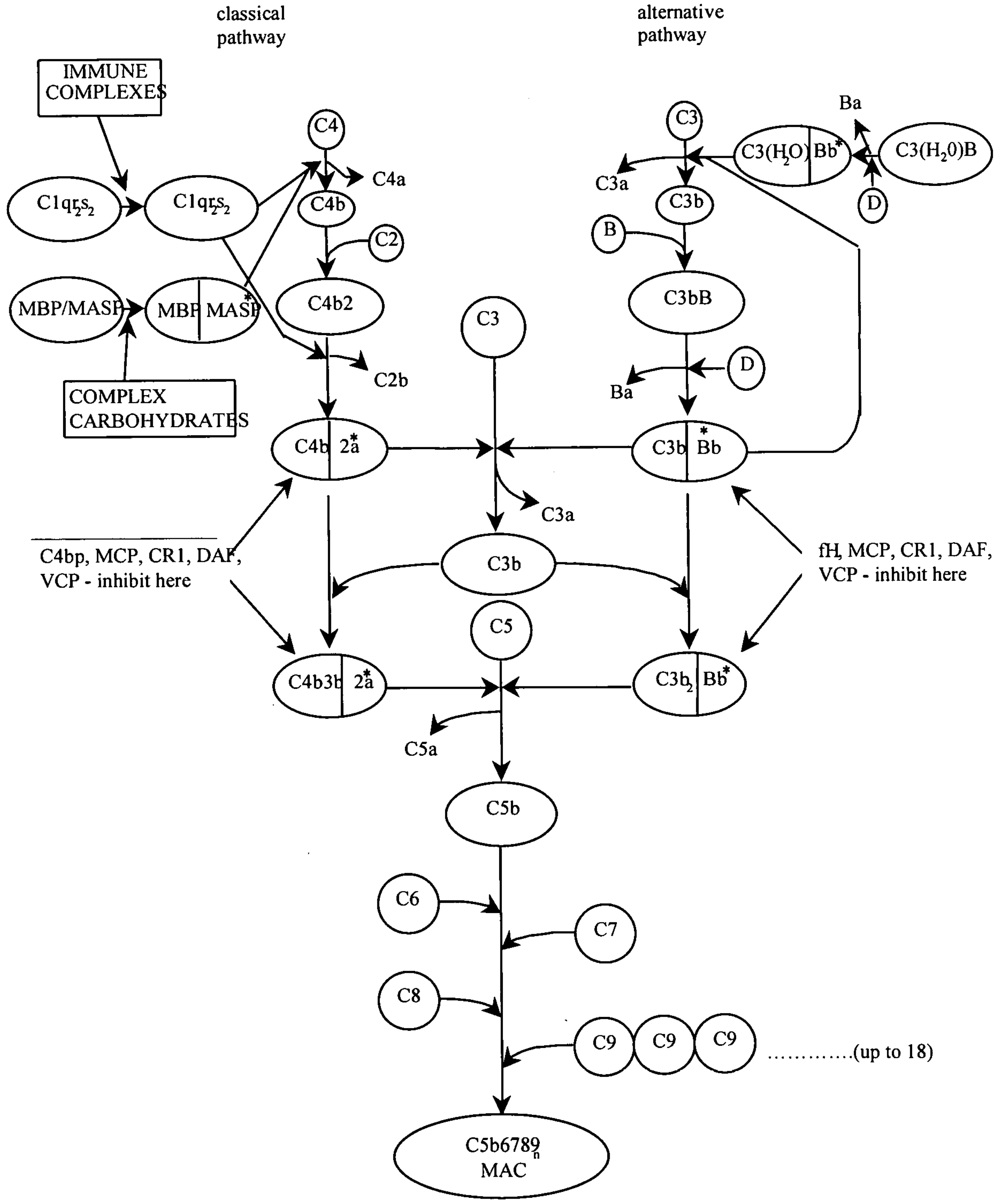

Figure 1 v - Shown are the major reactions in the activation pathways of complement. The MAC (membrane attack complex) is the terminal component of the complement cascade. Anaphylatoxins C3a, C4a and C5a are potent inflammatory mediators. The regulatory proteins (RCA proteins) are shown

The activation mechanisms for the two pathways differ in the initial stages, but involve complement protein 3 (C3) at a later, pivotal point in the cascade. Activation of C3 triggers a positive feedback loop resulting in more activated C3 being produced (Dodds & Sim, 1997). The positive feedback loop allows the complement system to carry out an immune response that may be described as explosive. At several points in the cascade, anaphylatoxins are released that are powerful mediators of the inflammatory response. The entire system is a series of enzymatic interactions which create active fragments that go on to trigger the next stage until the membrane attack complex (MAC) is formed. The MAC is a multiple protein complex that integrates into target cell membranes and causes cell lysis by 'punching' holes in the membranes. C3 is considered activated when it is cleaved to form C3a and C3b. This exposes an internal thiolester bond that can bind to hydroxyl or acceptor groups on cell surfaces and target them for destruction via the MAC (Dodds *et al.*, 1996), immune clearance or for an immune response via antibodies. C4 is also a key protein in the cascade, which possesses a similar thiolester bond. Activation can be antibody-mediated (classical) or non-antibody-mediated (alternative). Neutral sugars on bacterial cell surfaces can also trigger a third pathway, known as the lectin pathway, which operates in much the same way as the classical pathway, but does not require antibody (see Figure 1 v).

The potential for damage to tissue is high with such an explosive inflammatory response, so it is important that the main convertases are tightly regulated in order to prevent host tissue being attacked. To limit the amount of destruction to self-tissue, a series of proteins, the regulators of complement activation (RCA) are involved. These proteins include DAF (CD55), which increases the rate of dissociation of the important convertases in the pathway (Fujita *et al.*, 1987), MCP

(CD46), which acts as a cofactor for proteolytic cleavage of C3b and C4b by factor I (Seya *et al.*, 1986), CR1 (CD35), a large complement inhibitor with decay accelerating and co-factor activities for both the classical and alternative pathways (Fearon, 1979, Iada & Nussenzweig, 1981, Medof *et al.*, 1982), C4b binding protein (C4bBP) which has decay accelerating and co-factor activities for the classical pathway (Ogata *et al.*, 1993) and factor H which confers decay accelerating and co-factor activities in the alternative pathway (Reid & Day, 1989). As mentioned previously these RCA proteins all consist in part, or completely, of tandemly repeated CP modules (figure 1 i). Binding studies of the complement regulators have shown that two, and more commonly three or four, contiguous CP modules are required to confer functional activity (Klickstein *et al.*, 1988, Martin *et al.*, 1991, Kalli *et al.*, 1991, Krych *et al.*, 1991, 1998, Clarkson *et al.*, 1995, Gordon *et al.*, 1995, Iwata *et al.*, 1995, Manchester *et al.*, 1995, Brodbeck *et al.*, 1996, Sharma & Pangburn, 1996, Kuhn & Zipfel, 1996, Smith *et al.*, 2000, Rosengard *et al.*, 1999). CR1 has a broad range of complement regulatory activity, with different binding properties and specificities being associated with different triplets within the 30 CP modules, which make up its entire extracellular portion.

Many viruses have evolved proteins able to bind to, or inhibit, key components of the complement system as a way to sustain pathogenicity (Rosengard & Ahearn, 1999). One of the most interesting examples of such a virus, is the Vaccinia virus, which secretes a protein that mimics the human complement regulators, thus defending the protein from the complement cascade.

1.4.2 The Vaccinia Virus

Vaccinia virus is a member of the orthopoxviridae and, like other poxviruses such as the Variola virus and Cowpox virus, replicates in the cell cytoplasm. The virus exists in two forms, the intracellular mature virus (IMV), which is the most abundant of the infectious progeny, and the extracellular enveloped form (EEV), which is formed when the IMV becomes wrapped in membranes to produce an intracellular enveloped virus (IEV) that subsequently fuses with the plasma membrane to form the EEV. The different forms of the virus have different cell surface proteins and this helps to give the virus biological and immunological diversity and contributes to virulence. Different strategies arise from these different forms with regard to immune system evasion although all poxviruses release cytokines, chemokines and interferons from the infected cells to prevent neutralisation (Kotwal, 2000). The EEV is resistant to neutralisation by antibody and is probably the key to dissemination of the virus throughout the host. Historically, the immune-evasive nature of Vaccinia was fundamental to the eradication of smallpox. Smallpox is caused by the Variola virus, which is antigenically related to the Vaccinia virus. Although the Vaccinia virus can escape neutralisation by antibodies, the immune response persists after vaccination and will attack any antigenically related virus.

It was mentioned previously that one of the earliest lines of immune defence the body has is the complement system, which will target all pathogenic and foreign material for destruction. The EEV has been shown to incorporate some of the host's complement regulatory proteins into its viral coat as a way of protecting itself. The IMV also possesses its own complement regulatory activity, through production of a protein that is secreted from cells infected with the virus, which inhibits complement activity by mimicking host RCA proteins (Smith *et al.*, 1997,1999). This protein is the Vaccinia virus complement control protein.

1.4.3 Vaccinia virus complement control protein (sp35)

Vaccinia virus complement control protein, or VCP, was initially discovered after polypeptides were identified in the medium surrounding cells infected with the virus (Kotwal & Moss, 1988). When an attenuated form of the virus, designated the 6/2 mutant, was spontaneously formed these polypeptides were no longer present in the surrounding medium. This mutant had a segment of DNA near the left-hand-side of the genome missing. One of the proteins, of apparent molecular mass of 35 kDa on an SDS polyacrylamide gel, was purified and N-terminal sequence determined. This sequence was then compared to the 17 open reading frames within the deleted segment of the attenuated virus. One of the open reading frames, corresponding to the gene C21L, matched the protein sequence exactly and would encode an open reading frame of 263 amino acids. Comparison with other proteins showed that this protein could be predicted to form four repeating units, which each shared a consensus structure with a module-type (the CP module) known to be common in proteins involved with the regulation of complement activation. The viral protein showed highest sequence homology with the N-terminal half of C4bBP α–chain. This implied that the viral protein may also be involved in complement regulatory activity.

1.5 VCP as a complement regulator

1.5.1 Inhibition of the classical complement pathway

Several experiments were used to test the complement regulatory activity of VCP (Kotwal *et al.*, 1990). The medium from cells infected with the 6/2 attenuated mutant, which lacked the open reading frame for VCP, did not inhibit hemolysis of IgM-sensitised sheep red blood cells. Hemolysis, an effect of the terminal components of complement, was inhibited when the wild type virus was used to infect the cells. Other mutant strains, which were engineered to encode selectable markers instead of VCP were also created (vSIGK1 and vSIGK3). When medium from cells infected with the two mutant strains was used along with human serum and IgM-sensitised sheep red blood cells, hemolysis occurred. This effect was not observed when the wild type virus was used to infect the cells. To prove that VCP was responsible, all the proteins in the medium surrounding the cells infected with the virus were purified using a DEAE "Biogel" column, but only the fraction that eluted as VCP was shown to have complement inhibiting activity. Further analysis subsequently revealed that the protein could bind to C4b, or could promote decay of the C4b2a convertase, two effects that would inhibit the complement cascade (McKenzie *et al.,* 1992).

1.5.2 Inhibition of the alternative pathway

Two pathways exist in the complement system, so it was important to test whether the virus could neutralise both pathways, and if the effect could be attributed to VCP. The serum from guinea pigs, deficient in C4, an essential protein in the classical pathway, was shown to neutralise the attenuated form of the virus, but not the wild type (Isaacs *et al.*, 1992). This suggested that VCP could be affecting the alternative pathway of complement. C3b- and C4b-Sepharose columns were prepared and the contents of the medium surrounding cells infected with the virus were passed down these columns. VCP was shown to stick to both of the columns, and protein subsequently eluted from the C4b-sepharose column was passed down the C3b-sepharose column and retained 75% of the original binding efficiency. Binding to either component of the complement system could explain the complement regulatory activity of VCP (McKenzie *et al.*, 1992).

1.5.3 VCP and other complement regulators

VCP was compared with the other known complement regulators to examine its efficiency at inhibition. When examining the IC_{50} values of lysis of rabbit and sheep erythrocytes, recombinantly generated VCP was shown to be less potent than CR1 and Factor H at inhibiting the alternative pathway, and was less effective at inhibiting the classical pathway than CR1. It was, however, four-times more potent at inhibiting the classical cascade than factor H. VCP also possesses co-factor activity, although its mechanism is slightly different from that of factor H and CR1. Both of these proteins act as co-factors for the cleavage of C3b by factor I at three different sites, termed site 1 (Arg^{1281}-Ser^{1282}), site 2 (Arg^{1298}-Ser^{1299}) and site 3 (Arg^{932}-Glu^{933}). VCP, however, only assists cleavage at site 1 (Sahu *et al.*, 1998).

VCP, therefore, seems to represent a very compact and effective complement inhibitor that can inhibit both the classical and alternative pathways, displaying decay-accelerating, C3b- and C4b-binding, and co-factor activity. CR1 and factor H, have respectively 30 and 20 CP modules in their structures but the viral genome would have no need to be burdened with long stretches of DNA since an efficient and more compact protein has been evolved. The sequence of VCP, which is not glycosylated, is shown (figure 1 vi) and compared with C4b-binding protein, the closest human homologue, and the variola virus smallpox inhibitor of complement enzymes (SPICE), which is four times more potent than VCP at inhibiting complement, yet has 95% sequence similarity in its four CP modules (Rosengard & Ahearn, 1998). Other viral complement regulators, which also contain four tandemly arranged CP modules and are likely similar in structure are the cowpox virus inflammatory modulatory protein (IMP) (Miller *et al.*, 1997) and the herpesvirus saimiri complement control protein homologue (CCPH) (Albrecht & Fleckenstein, 1992).

1.5.4 Therapeutic potential of VCP

Because of its potent functional and compact structural properties, VCP is an excellent candidate to study as an anti-inflammatory therapeutic. The complement system has been implicated in many diseases, and is thought to be the principal cause of the debilitating symptoms of arthritis. The complement system is also responsible for the tissue rejection associated with transplants. It can trigger hyperacute rejection (HAR), an acute immune response that can cause the very rapid (minutes) rejection of transplanted tissue in for example xenotransplantation. The potential to engineer porcine organs to express VCP on their surface is an interesting potential use of this protein and might, in combination with other drugs, allow xenotransplantation to take place effectively, masking the foreign organ from the host's complement system (Al-Mohanna *et al.*, 2001). However, because of the potential risks posed by dormant retroviruses which may exist in the porcine organs, xenotransplantation in this way is

Figure 1 vi - Sequence of VCP and homologues C4b binding protein and SPICE

```
V1   L S C C T I P S R P I N M K F K N S V T E D A N A N Y N I G D T I E Y L C L P G Y R K Q K M G P I   Y A K C T G               T G   W T L F N     Q C I K
C1      N C G P P P   T L S F A A P M D I T   L T E T R F K T G T T L K Y T C L P G Y V R S H S   T Q T L T C N S D         V G E W   V Y N   T F C I
S1   L S C C T I P S R P I N M K F K N S V T E D A N A N Y N I G D T I E Y L C L P G Y R K Q K M G P I   Y A K C T G               T G   W T L F N     Q C I K
V2     R R C P S   P R D I D N G Q L D I         G G   V D   F G S S I T Y S C N S G Y H L I G     E   S K S Y C E L G S T G S M V W N P E A P   I C E S
C2   Y K R C R H   P G E L R N G Q V E I     K T D   L S   F G S Q I E F S C S E G F F L I G S       T T S R C E V Q D R G I G   W S H P L P   Q C E
S2     R R C P S   P R D I D N G _ L D I         G G   V D   F G S S I T Y S C N S G Y Y L I G     E   Y K S Y C _ L G S T G S M V W N P _ A P   I C E S
V3     V K C Q S P P   S I S N G R H N G   Y   E D   F Y T D   G S V V T Y S C N S G Y S L I G N     S G V L C S G                 G E W N D P   P   T C Q I
C3   I V K C K P P P   D I R N G R H S G   E   E N   F Y A   Y G F S V T Y S C D P R F S L L G H     A S I S C T V E N E T   G V W R P S P P   T C E V
S3     V K C Q L P P   S I S N G R H N G   Y   E D   F Y T D   G S V V T Y S C N S G Y S L I G N     S G V L C S G                 G E W N N P   P   T C Q I
V4     V K C P   H P   T I L N G Y L S S   F G K R   S Y S   Y N D N V D F N C K Y G Y V L S G S     S S S T C S P G             N T W Q P E L P   K C V R
C4   K I T C R   K P   D V S H G E M V S   F G G P   I Y N   Y K D T I V F K C Q K G F K L R G S     S V I H C D A D             S K W N P S P P   A C E
S4     V K C P   H P   T I S N G Y L S S   F G K R   S Y S   Y N D N V D F K C K Y G Y V L S G S     S S S T C S P G             N T W K P E L P   K C V R
```

Figure 1 vi - Sequence of VCP modules 1 (V1) to 4 (V4). The sequence of two homologues of VCP are also shown; C4b binding protein (C1-C4); Smallpox inhibitor of complement enzymes (SPICE, S1-S4). The consensus residues are highlighted in bold. SPICE is 95% homologous to VCP. The residues in SPICE that differ from VCP are underlined in bold.

strongly opposed. Other potential therapeutic effects would be in Alzheimers Disease (Daly & Kotwal, 1998). The structure of VCP will provide an excellent basis for elucidating details of how the CP modules act to inhibit the complement system, providing information about exposed surface residues and charge distribution. While the individual CP modules could be modelled by homology, module-module interactions and angles are difficult to model at present. Nonetheless modelling studies have been used to generate structures of VCP (Smith *et al.*, 2000), MCP (Listewski *et al.*, 2000*)* and several other RCAs. The NMR structure of the C-terminal half of VCP has been solved previously and it was shown to be composed of two typical CP modules, with (as mentioned earlier) an extensive interface between the modules such that they have a well defined orientation to one another. Both of these modules are (as observed in other CCP modules, see above) composed of eight β-strands, which form H-bonded β–sheets and they are well-defined in general, with little deviation between the structures. The area of highest deviation was the "hypervariable loop", a region that shows little sequence conservation when compared to other CP motifs and did not adopt a set structure in module 3 (Wiles *et al.*, 1997). After the work described in this thesis had been completed, a crystal structure of intact VCP was solved (Murthy *et al.*, 2001). This is discussed, and compared with the NMR results, in Chapter 6.

Solving the structure of the intact protein by NMR would be difficult due to the heavy overlap of cross-peaks and the line-broadening associated with large proteins. Even if the protein were C^{13} labelled, it would still be badly overlapped in most regions due to the high degree of conservation in chemical shifts for equivalent residues in the respective CP modules. To make progress with understanding the structure and function relationships of CP module groups, this thesis presents the structure of VCP module 2 and 3 (VCP~2,3), the central two modules of the protein. The intention was to provide another CP module structure and another example of a CP module pair with its associated intermodular orientation. The fact that module 3 is common to both VCP2,3 and VCP3,4, allowed, for the first time, a reconstruction of what the intact VCP2,3,4 may look like. This work will aid our understanding of this interesting viral mimic.

References

Albrecht, J. C. & Fleckenstein, B. (1992). *J. Virology*, **66**, 3937-3940

Al-Mohanna, F., Parhar, R. & Kotwal, G. J. (2001). *Transplantation* **71**, 796-801.

Barlow, P. N., Baron, M., Norman, D.G. , Day, A. J. , Willis, A. C. , Sim, R.B. & Campbell I. D. (1991). *Biochemistry.* **30,** 997-1004.

Barlow, P. N., Norman, D. G., Steinkasserer, A., Horne, T. J., Pearce, J., Sim, R. B. & Campbell I. D (1992). *Biochemistry.* **31,** 3626-3630

Barlow, P. N., Steinkasserer, A., Norman, D. G., Kieffer, B., Wiles, A. P., Sim R. B. & Campbell, I. D (1993). *J. Mol. Biol* **232** 268-284.

Baron, M., Norman, D. & Campbell, I. D. (1991). *Trends Biol. Sci.* **16**, 13-17

Bork, P., Downing, A. K., Keiffer, B. & Campbell, I. D. (1996). *Quart. Rev. Biophys.* **29**, 119-167

Bouma B., De Groot P. G.,.Van Den Elsen J. M. H, Ravelli R. B. G., Schouten A., Brady, R. L., Dodson, E.J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). *Science*, **260**, 979-983

Brodbeck, W. G., Liu, D. C., Sperry, J., Mold, C. & Medof, M. E. (1996). *J. Immunol.* **156**, 2528-2533

Bruschweiler, R., Liao, X. B. & Wright, P. E. (1995). *Science*, **268**, 886-889

Campbell, I. D. & Downing, A. K. (1994). *Trends Biotech.* **12**, 168-172

Campbell, I. D. & Downing, A. K. (1998). *Nature Struct. Biol* **5 (suppl)**, 476-479

Casasnovas, J. M., Larvie, M. & Stehle, T. (1999). *EMBO J*, **18,** 2911-2922.

Clarkson, N. A., Kaufman, R, Lublin, D. M., Ward, T., Pipkin, P. A., Minor, P. D., Evans, D. J. & Almond, J. W. (1995). *J. Virol.* **69**, 5497-5501

Daly, J. & Kotwal, G. J. (1998). *Neurobiol. Aging.* **19,** 619-627

Day, A. J., Campbell, R. D. & Reid, K. B. M. (1989). (F. Melchers. *et al.*, eds), **vol 7**, 209-212, Springer-Verlag, Heidelberg

Dodds, A. W. & Sim, R. B. (1997). *Complement: A practical approach.* Oxford University Press, Oxford

Dodds, A. W., Ren, X. D., Willis, A. C. & Law, S. K. A. (1996).

Doolittle, R. F. & Bork, P. (1993). *Sci. Am.* **269**, 50-56

Doolittle, R.F. (1995). *Annu. Rev. Biochem.* **236**, 1079-1092

Eck, M. J., Pluskey, S., Trub, T., Harrison, S. C. & Shoelson, S. E. (1996). *Nature*, **379**, 277-280

Fearon, D. T. (1979). *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5867-5871

Fujita, R., Inoue, T., Ogawa, K., Iada, K. & Tamura, N. (1987). *J. Exp. Med.* **166**, 1221-1228

Gaboriaud, C., Rossi, V., Bally, I., Arlaud, G. J. & Camps, J. C. F. (2000). *EMBO J.* **19**, 1755-1765

Gordon, D. L., Kaufman, R. M., Blackmore, T. K., Kwong, J. & Lublin, D. M. (1995). *J. Immunol.* **155**, 348-356

Hansen, A. P., Petros, A. M., Meadows, R. P. & Fesik, S. W. (1994). *Biochemistry*, **33**, 15418-15424

Hawrot, E., Xiao,Y., Shi, Q., Norman, D., Kirkitadze, M. & Barlow, P. N. (1998). *FEBS Letts,* **432,** 103-108.

Hegyi, H. & Bork, P. (1997). *J. Protein Chem.* **16**, 545-551

Herissat B. & Bork P. (1996). *Protein Eng.* **9**, 725-726

Huber, A. H., Wang, Y. M. E., Bieber, A. J. & Bjorkman, P. J. (1994). *Neuron,* **12**, 717-731

Iada, K. & Nussenzweig, V. (1981). *J. Exp. Med.* **153**, 1138-1150

Isaacs, S. N., Kotwal, G. J. & Moss, B. (1992). *Proc.Natl.Acad.Sci. USA*, **89**, 628-632.

Iwata, K., Seya, T., Yanagi, Y., Pesando, J. M., Johnson, P. M., Okabe, M., Ueda, S., Ariga, H. & Nagasawa, S. (1995). *J. Biol. Chem.* **270**, 15148-15152

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. J. (1992). *Nature*, **360**, 232-239

Kalli, K. R., Hsu, P. H., Bartow, P. J., Ahearn, J. M., Matsumoto, A. K., Klickstein, L. B. & Fearon, D. T. (1991). *J. Expt. Med.* **171**, 1451-1460

Kaupmann, K., Huggel, K., Heid, J., Flor, P., Bischoff, S., Mickel, S. J., McMaster, G., Angst, C., Bittiger, H., Froestl, W. & Bettler, B. (1997). *Nature*, **386**, 239-246

Klickstein, L. B., Bartow, T. J., Miletic, V., Rabson, L. D., Smith, J. A. & Fearon, D. T. (1988). *J. Expt. Med.* **168**, 1699-1717

Kotwal, G. J. & Moss, B. (1988). *Nature,* **335**, 176-178

Kotwal, G. J. (2000) *Immunol. Today*, **21**, 242-248

Kotwal, G. J., Isaacs, S. T., McKenzie, R., Frank, M. M. & Moss, B. (1990). *Science,* **250**, 827-830.

Krych, M., Hauhart, R., & Atkinson, J. P. (1991). *Annu. Rev. Immunol.* **9**, 431-455

Krych, M., Hourcade, D. & Atkinson, P. P. (1991). *Proc. Natl. Acad. Sci. U.S.A.* **88**, 4353-4357

Kuhn, S. & Zipfel, P. F. (1996). *Eur. J. Immunol.* **26**, 2383-2387

Law, S. K. A. & Reid K. B. M. (1990) *Complement*, 2nd Edition, IRL Press

Leahy, D. J., Aukhil, I. & Erickson, H. P. (1996). *Cell*, **84**, 155-164

Manchester, M., Valsamakis, A., Kaufman, R., Liszewski, M. K., Alvarez, J., Atkinson, J. P., Lublin, D. M. & Oldstone, M. B. A. (1995). *Proc. Natl. Acad. Sci. U.S.A.,* **92**, 2303-2307

Martin, D. R., Kalli, K. R., Yuryev, A., Fearon, D. T. & Ahearn, J. M. (1991). *J. Expt. Med.* **174**, 1299-1331

McKenzie, R., Kotwal, G.J., Moss, B., Hammer, C. H. & Frank, M. M. (1992). *J. Infect. Dis.* **166**, 1245-1250

Medof, C. G., Iada, K., Mold, C. & Nussenzweig, V. (1982). *J. Exp. Med.* **156**, 1739-1754

Miller, C. G., Shchelkunov, S. N. & Kotwal, G. J. (1997). *Virology,* **229**, 126-133

Morgan, B. P. & Harris, C. L. (1999). *Complement regulatory proteins,* Academic Press, San Diego

Murthy, K. H., Smith, S. A., Ganesh, V. K, Judge, K. W., Mullin, N., Barlow, P. N, Ogata, C. M and Kotwal, G. J (2001). *Cell* **104**, 301-311

Naismith J. H., Devine T. Q., Kohno T. & Sprang S. R (1996) *Structure*, **4**, 1251-1262.

Nakaseko, Y., Neuhaus, D., Klug, A. & Rhodes, D. (1992). *J. Mol. Biol.* **228**, 619-657

Norman, D.G., Barlow, P. N., Baron, M., Day, A.J., Sim, R.B. & Campbell, I. D. (1991). *J. Mol. Biol.* **219,** 717-725.

Ogata, R. T., Mathais, P., Bradt, B. M. & Cooper, N. R. (1993). *J. Immunol.* **150**, 2273-2280

Omichinski, J. G., Clore, G. M., Robien, M., Sakaguchi, K., Appella, E. & Gronenborn, A. M. (1992). *Biochemistry,* **31**, 3907-3917

Ponting, C. P., Schult, J., Milpetz, F. & Bork, P. (1999) *Nucleic Acid Res.* **27**, 229-232

Reid, K. B. M., & Day A. J. (1989). *Immunol Today.* **10,** 177-180.

Reid, K. B. M., Bentley, D. R., Campbell, R. D., Chung, L. P., Sim, R. B., Kristtensen, T. & Tack, B. F. (1986). *Immun. Today,* **7**, 230-234

Rosengard, A. M, Alonso, L. C, Korb, L. C., Baldwin III, W. M, Sanfilippo, F., Turka, L. A., Ahearn, J. M. (1999) *Molecular Immunology,* **36**, 685-697

Rosengard, A. M. & Ahearn, J. M. (1998). *Mol. Immunol.* **35**, 397, Abstr

Rosengard, A. M. & Ahearn, J. M. (1999). *Immunopharm.* **42**, 199-106

Sahu, A., Isaacs, S. N., Soulida, A.M. & Lambris, J. D. (1998). *J. Immunol.* **160**, 5596-5604

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B & Erlich, H. A. (1988) *Science* **239**,487-491

Saiki, R. K., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230**, 1350-1355

Schwarzenbacher, R., Zeth, K., Diederichs, K., Gries, A., Kostner, G. M, Laggner, P & Prassi, R. (1999). *EMBO J.*, **18**, 6228-6239.

Seya, T. J., Turner, J. R. & Atkinson, J. P. (1986). *J. Exp. Med.* **163**, 837-855

Sharma, A. K. & Pangburn, M. K. (1996). *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10996-11001

Simmelink M. J. A, Derksen R. H. W. M., Kroon, J. & Gros P. (1999) *EMBO J.* **18** 5166-5173.

Smith, B. O., Downing, A. K., Driscoll, P. C., Dudgeon, T. J. & Campbell, I. D. (1995). *Structure,* **3**, 823-833

Smith, G. L (1999). *Immunol. Lett.* **65**, 55-62

Smith, G. L., Symons, J. A, Khanna, A., Vanderplasschen, A. & Alcami, A. (1997). *Immunol. Rev.* **159**, 137-154

Smith, S. A, Mullin, N. P., Parkinson, J., Shchelkunov, S. N., Totmenin, A. V., Loparev, V. N., Srisatjaluk, R., Reynolds, D. N., Keeling, K. L., Justus, D. E., Barlow, P. N. & Kotwal, G. J. (2000). *J. Virol.* **74**, 5659-5666

Spitzfadden C., Grant, R. P., Mardon, H. J. & Campbell, I. D.(1997). *J.Mol.Biol*, **265**, 565-579.

Wang, J. H., Yan, Y. W., Garrett, T. P. J., Liu, J. H., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E. L. & Harrison, S. C. (1990). *Nature*, **348**, 411-418

Wiles, A. P., Shaw, G., Bright, J., Perczel, A., Campbell, I. D. and Barlow, P. N. (1997). *J. Mol. Biol.* **272**, 253-265.

Williams, M. J., Phan, I., Harvey, T. S., Rostagno, A., Gold, L. I. & Campbell, I. D. (1994). *J. Mol. Biol.* **235**, 1302-1311

Chapter 2 - Protein production and purification

2.1 *Pichia pastoris* as an expression system

Choosing an effective protein expression system is an important consideration in a project of this nature due to the requirement for milligram quantities of pure, isotopically labelled material. In this respect, biological expression systems are more suitable than chemical synthesis since isotopic labels can be easily incorporated into the expressed protein provided the recombinant organisms are able to survive on minimal media.

Although bacterial expression systems have been used widely (Smith & Johnson, 1988, Matthews *et al.*, 1994, Itakura *et al.*, 1977), a eukaryotic expression system is a logical choice for proteins that contain disulphide bonds. Using a eukaryotic expression system has several advantages; it can properly process the protein, fold it and secrete it. *Escherichia coli* expression, on the other hand, often requires additional refolding steps and more extensive purification protocols.

The *Pichia pastoris* expression system is similar to that of *Saccharomyces cerevisiae,* and is capable of producing large quantities of protein in a cost-effective manner. *P. pastoris* has many advantages over the *S. cerevisiae* system. It has been shown, in general, to produce better yields of protein, with one copy of the gene in *P. pastoris* expressing comparable amounts of protein to multi-copy insertions in *S. cerevisiae*. *P. pastoris* plasmids are also more stable than those from *S. cerevisiae* and the expressed protein will more often undergo hyperglycosylation in *S. cerevisiae,* which can lead to problems with purification and activity of the expressed protein (Cereghino & Cregg, 2000). Very high levels of protein production are possible with the *P. pastoris* expression system - 1 – 10% of all successful transformants result in a multiple gene-insertion event, where the gene of interest is spontaneously inserted several times at a specific locus through genetic recombination between the yeast genome and several successfully transformed plasmids.

2.2 AOX genes involved in protein transcription

The expression system based on *P. pastoris* utilises the methylotropic nature of the yeast, which allows it to utilise methanol as its sole carbon-source. In order to do this effectively, without accumulation of toxic by-products such as hydrogen peroxide, *P. pastoris* contains two genes involved in controlling the transcription of alcohol oxidase (AOX). AOX activity is expressed when methanol is the sole carbon source available. The enzyme acts as a catalyst in the oxidation of methanol to formaldehyde and hydrogen peroxide (Buckholx & Gleeson, 1991, Egli *et al.*, 1980). The AOX activity is controlled by two genes that encode alcohol oxidases, *AOX1* and *AOX2*. These genes are 97% homologous but have different activities. The presence of only one gene instead of two alters the phenotype of the yeast. If *AOX1* alone is transcribed, the resulting phenotype is known as mut$^+$ (wild type methanol utilisation). This is the phenotype of GS115 *P. pastoris* strains, which can utilise methanol very efficiently due to the high levels of alcohol oxidase that *AOX1* encodes. The *AOX1* gene is responsible for most of the alcohol oxidase activity in the cell (Veehhuis *et al.*, 1983, Ellis *et al.*, 1985, Cregg *et al.*, 1989). Substitution of the *AOX1* gene with *ARG4* from *S. cerevisiae* results in the expression of the *AOX2* gene, which can also be used to metabolise methanol, but results in a mut^s phenotype (for methanol utilisation slow; strain KM71 is an example). This has a much slower rate of utilisation of methanol due to the decrease in alcohol oxidase levels in the absence of the *AOX1* gene. Both *AOX1* and *AOX2* genes are under the control of the single AOX promoter (Tschopp *et al.*, 1987, Chiruvolu *et al.*, 1997, Cregg & Madden, 1987). The gene encoding the protein of interest is inserted into a plasmid that will confer one of the phenotypes, mut^s or mut$^+$, depending on the strain selected. The gene encoding the target protein is subsequently incorporated into the yeast genome due to a single crossover event that can occur at any one of three different loci in the genome. In successful transformants, the gene of interest will lie downstream of the AOX promoter, which will promote transcription of the protein when methanol is the only carbon source available. Protein production can therefore be induced by supplying minimal media with methanol as the sole carbon source (Cereghino & Cregg, 2000). The *P. pastoris* system can be engineered to incorporate a gene that encodes the *S. cerevisiae* α-factor pre-propeptide which acts as a signal sequence for secretion of the protein (Kurjan & Herkowitz, 1982). The resulting protein contains 19 extra amino acids (the signal-sequence). These amino acids are normally cleaved from the expressed protein in two stages. Initially, the product of the *KEX2* gene cleaves at a dibasic (Lys-Arg) site, whilst the *STE13* cleavage product acts at one of two possible cleavage sites consisting of Glu-Ala repeats (discussed later).

2.3 Plasmids in Pichia Pastoris

Two plasmids were employed in VCP~2,3 expression. The first plasmid used was pPICZαA (Figure 2 i), which confers a phenotype of *P. pastoris* that requires histidine to be supplied to the growth media. pPICZαA has a zeocin-resistance gene that can be used as a selection marker for positive transformants carrying the plasmid.

The plasmid pPIC9K was also used; it carries ampicillin and kanamycin-resistance and contains the *HIS4* gene, allowing the cells to metabolise histidine intracellularly. Both plasmids were manufactured by Invitrogen (San Diego, CA, U.S.A). The plasmids are inserted into the yeast by transformation and the plasmid DNA is then incorporated into the yeast genome by genetic recombination (figure 2 ii).

Figure 2.i – Plasmids used in P. pastoris expression system

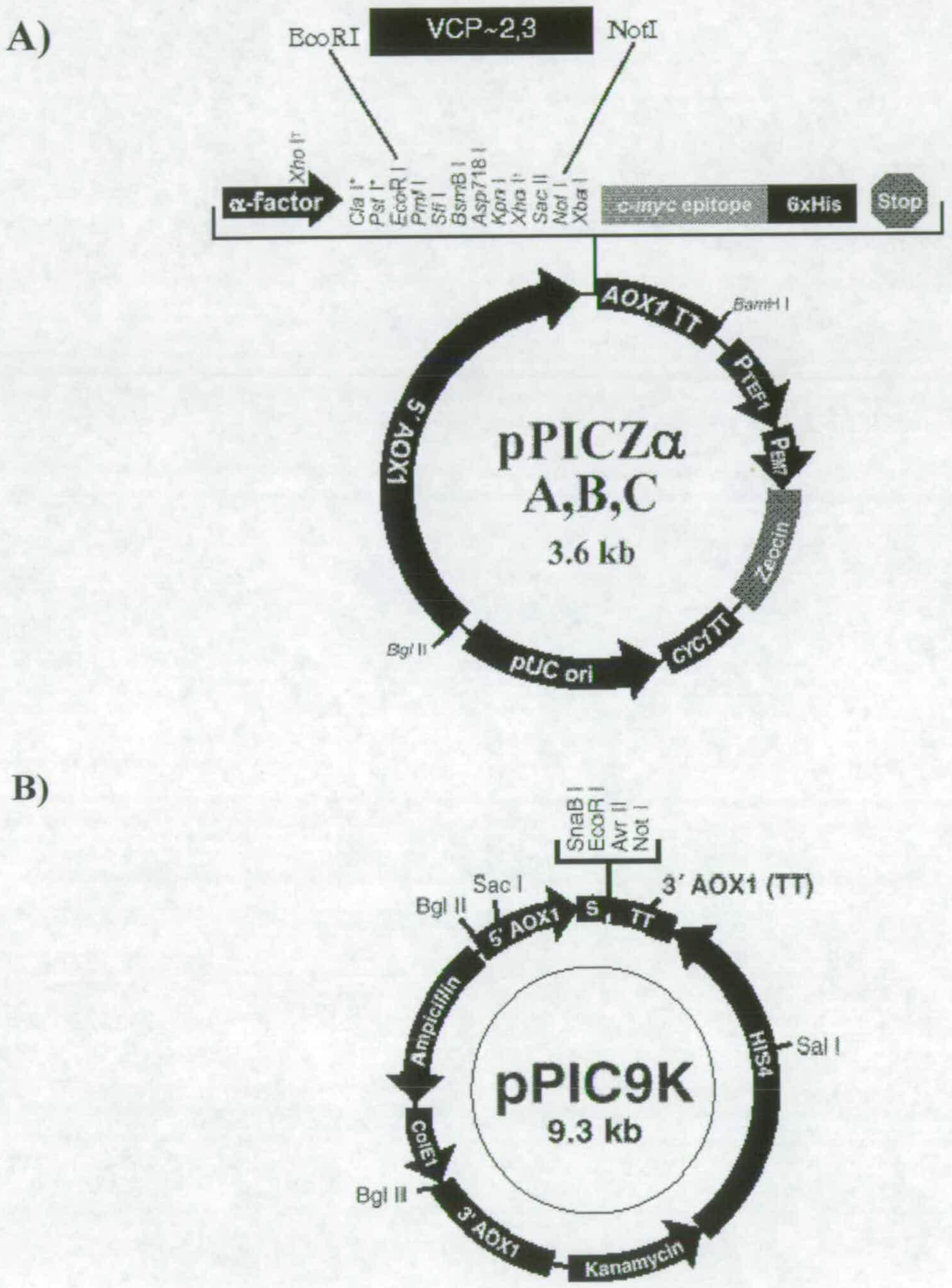

Figure 2 i - Maps of plasmids A) pPICZαA and B) pPIC9K. Each plasmid contains genes that confer antibiotic resistance. The plasmids also contain AOX1 promoter regions and restriction sites. pPIC9K also carries the HIS4 gene. This allows cells to metabolise histidine intracellularly. The gene of interest is inserted into the plasmid at the appropriate restriction sites, as shown (adapted from Invirogen manual).

Figure 2 ii – Genetic recombination event involving P. pastoris genome and pPICZαA

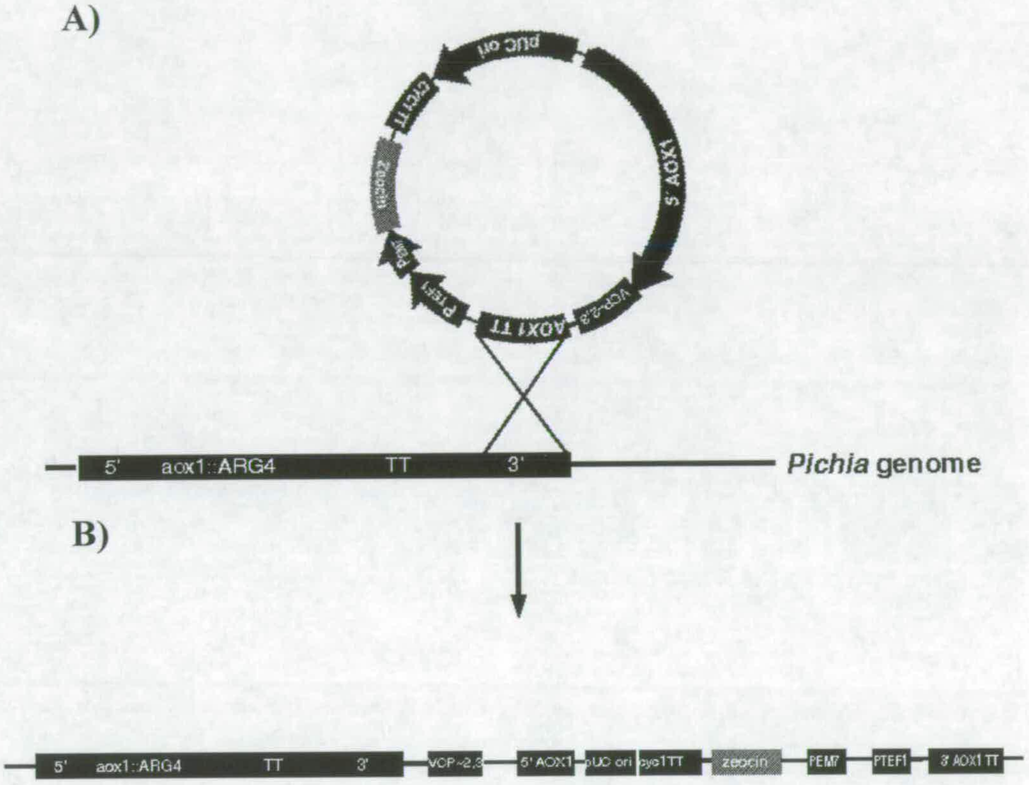

Figure 2 ii – A) The pPICZαA plasmid undergoes a genetic recombination event, as shown, resulting in the plasmid being incorporated into the yeast's genome shown in B. (figure adapted from Invitrogen manual).

2.4 Engineering *Pichia pastoris* to synthesise VCP~2,3

2.4.1 Preparation of the VCP~2,3/pPICZαA

The Vaccinia virus (VV) DNA was a kind gift from Dr. Geoffrey Smith, Department of Pathology, Oxford University, Oxford, UK. The VV DNA was used to amplify the DNA encoding VCP~2,3 using primers that incorporated *EcoR*I and *Not*I restriction sites at the 5' and 3' ends of the VCP~2,3 coding sequence respectively.

***EcoR*I primer 5'-3':**

CCGGAATTCATTAAACGGAGATCGCCATCG

***Not*I primer 5'-3':**

ATAGTTTAGCGGCCGCCTATTAAACAATCTGACACGTGGGTGGATCGGA

The primers were used to amplify the coding sequence forVCP~2,3 (Kowtal & Moss, 1988) using the polymerase chain reaction (PCR) (Saiki *et al.*, 1985, 1988). The resulting PCR products were then loaded onto an agarose gel and fragments of ~400 bp were cut out and purified using a QIAGEN gel extraction kit (Qiagen, Hilden, Germany). The fragments were then incubated at 37 °C overnight in the presence of *Eco*RI and *Not*I restriction enzymes and purified using a QIAGEN mini-prep kit. The plasmid pPICZαA was also incubated overnight in the presence of the same restriction enzymes, *Eco*RI and *Not*I, and the linearised product was purified using the QIAGEN system. The VCP~2,3 DNA was ligated into the plasmid at room temperature in the presence of T4 DNA ligase and left overnight (Weiss *et al.*, 1968, Cohen *et al.*, 1973). The resulting mixtures were purified using the QIAGEN method, as before. The reaction solutions and conditions are outlined in protocol 2 i.

2.4.2 Transformation into *E.Coli*

Escherischia coli was used to propagate the pPICZαA/VCP~2,3 construct. Competent *E. coli* Top10 F' cells were prepared to enable uptake of the plasmid. Transformation of the construct into the cells was via the calcium chloride (Cohen *et al.*, 1972) method (protocol 2 ii) and the resulting cells were plated on Luria-Bertani (LB) plates containing 50 μg/ml zeocin to select for positive transformants. Control plates with competent cells (untransformed) were also prepared and incubated overnight at 37 °C to allow growth. Growth was observed on the plate containing the cells transformed with the plasmid and eight colonies from the pPICZαA/VCP~2,3 cells were used to inoculate 10 ml of low-salt LB containing 25 μg/ml of zeocin and grown overnight at 37 °C in a shaking incubator. The resulting cells were harvested by centrifugation (7 mins at 4000 g; supernatant discarded) and the plasmid DNA was extracted using a QIAGEN mini-prep kit. A sample containing 17 μl of the extracted plasmid DNA from each of the eight initial colonies was then left overnight in the presence of *Eco*RI and *Not*I restriction enzymes. Analysis on an agarose gel showed that transformantion had been successful, as indicated by the presence of two bands representing the linearised plasmid pPICZαA and the ~400bp VCP~2,3 fragment (Figure 2 iii). Three of the colonies that contained the successful constructs were, again, grown in 10 ml low-salt LB containing 25 μg/ml zeocin, the plasmid DNA extracted as before and sequenced using the dye termination method (Amersham Pharmacia, Uppsala, Sweden). The DNA sequencing showed that one colony harboured a plasmid that contained the correct DNA sequence for VCP~2,3. This

Protocol 2 i - PCR, restriction digest and ligation conditions

a) PCR Conditions

Reactant	Volume (μl)
Thermopol buffer	10
dNTP's	10
5' primer	1
3' primer	1
VV genome	1
Taq DNA polymerase (deep vent)	1
H_2O	1

Cycle Number	Time (s)	Temp (°C)
1	300	95 (melting)
2-36	30	95 (annealing)
	30	55 (annealing)
	60	72 (annealing)
37	300	72 (complete elongation)

b) Restriction digest mixture

Contents for PCR digest	Vol. used (μl)
QIAGEN purified PCR product	5
10x *EcoR*I buffer	2
*EcoR*I	0.5
*Not*I (kept on ice until used)	0.5
H_2O	13

Contents for pPICZαA digest	Vol. used (μl)
QIAGEN purified pPICZαA	5
10x *EcoR*I buffer	2
*EcoR*I	0.5
*Not*I (kept on ice until used)	0.5
H_2O	13

c) Ligation reaction mixture

Contents for ligation reaction	Vol (μl)
pPICZαA cut *EcoR*I - *Not*I	1
PCR amplified fragment cut *EcoR*I - *Not*I	7
T4 DNA ligase	1
10x ligation buffer	1

Protocol 2 i cont'd

d) PCR conditions for dye terminations sequencing

Reactant	Volume (μl)
PCR rection mix	4
QIAGEN purified DNA	3.5
5' primer	0.5
H_2O	2

Cycle Number	Time (s)	Temp (°C)
1	300	95 (melting)
2-36	30	96 (annealing)
	15	50 (annealing)
	240	60 (annealing)
37	300	72 (complete elongation)

colony was grown overnight in low-salt LB zeocin to generate more plasmid DNA, and plated on low-salt LB plates containing 25 μg/ml of zeocin to create stocks. The plasmid DNA was purified using the Qiagen mini-prep system, as before. Details of the transformation methods are outlined in protocol 2 ii.

2.4.3 Transformation of VCP~2,3/pPICZαA into *Pichia pastoris*

To linearise the VCP~2,3/pPICZαA construct, the plasmid was left to incubate overnight in the presence of the *Sac*I restriction enzyme, since the pPICZαA contains this restriction site. The linearised DNA was separated from the mixture by phenol/chloroform-extraction and applied to an agarose gel to confirm the presence of the linearised product. Although the linearised plasmid is the same molecular weight as the circular plasmid, the intact plasmid will supercoil, which provides a simple means of assessing whether the linearisation has been successful or not. Analysis of the agarose gel showed that the plasmid had been linearised successfully (figure 2 iv). Electroporation was used to transform the linearised plasmid DNA into the yeast cells. Application of a 1500 V pulse on a Bio-Rad pulser (resistance 200 Ω, capacitance 25 μF) causes the yeast cell wall to rupture temporarily, at which time surrounding DNA may be taken in. Both strains of the yeast, GS115 (Cregg *et al.*, 1985) and KM71 (Cregg & Madden, 1987), were prepared for electroporation as described in the Invitrogen manual.

Figure 2 iii - Agarose gel of linearisation products

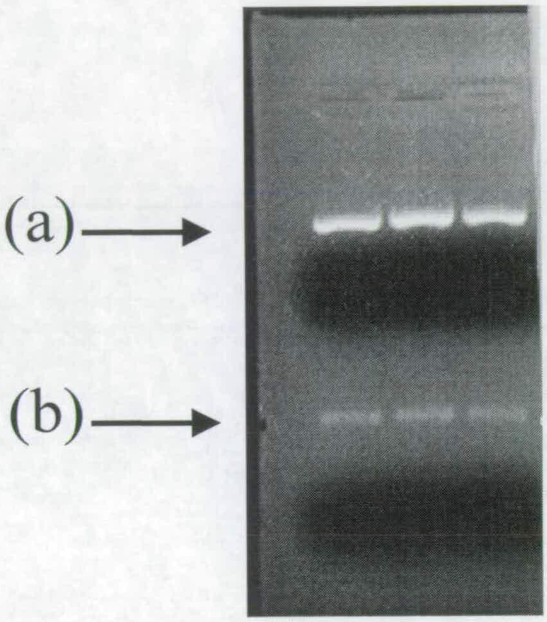

Figure 2 iii- Agarose gel showing restriction digest products after incubation of plasmid from transformed cells with *Eco*RI and *Not*I restriction enzymes. (a) shows the linearised plasmid pPICZαA. (b) shows the 400bp fragment of VCP~2,3.

Protocol 2 ii - Reaction conditions for creating pPICZαA/VCP~2,3 construct

a) Preparation of competent cells and insertion of plasmid DNA

A single colony of *E. coli* top 10F' cells was added to 100 ml LB (Luria-Bertani (LB) broth containing 1% Tryptone, 0.5% yeast extract (YE), 1% NaCl, pH 7.0) in a 1 litre flask and incubated in a shaking incubator at 37 °C for 3 hours at 300 r.p.m. The cells were then transferred to sterile 50 ml Falcon tubes using aseptic techniques and stored on ice for 10 mins. Cells were harvested by centrifugation (10 mins at 4000 rpm, 4° C) and the tubes inverted to drain traces of the media. Each pellet was resuspended in 10 ml ice cold $CaCl_2$ (0.1 M) and stored on ice. The cells were then centrifuged and the media drained, as before. The pellet was then resuspended in 2 ml ice cold $CaCl_2$ and 200 μl of each suspension was transferred to a sterile microfuge tube, and 5 μl of pPICZαA/VCP~2,3 DNA was added to each. The contents were mixed and left on ice for 30 mins. The tubes were then placed in a water bath which had been pre-heated to 42 °C and left to incubate for 90 s before tranferring the tubes to an ice bath and leaving to chill for 2 mins. LB media (1.5 ml) containing 50 μl/ml zeocin was added and left to incubate for 45 mins at 37 °C. Top agar (1 ml) was then added to the solution and left to set. The plates were inverted and left to incubate overnight at 37 °C.

b) Restriction digest reaction mixture

Contents of mixture	Volume (μl)
QIAGEN purified DNA	17
*Eco*RI	0.5
*Not*I	0.5
*Eco*RI buffer	2

Both strains were plated on yeast peptone dextrose (YPD) plates containing 1 M sorbitol and 1 % histidine, which helps the yeast recover after electroporation) and left to incubate overnight. Several colonies were then re-plated on YPD plates containing 25 μg/ml zeocin and 1% histidine to select for the positively transformed cells that contained the construct. A total of nine KM71 colonies and four GS115 colonies grew successfully when zeocin was present. The details of the methods are outlined in protocols 2 iii.

2.5 Protein production and purification

2.5.1 Screening *Pichia pastoris* for VCP~2,3 production

Initially, BMGY (buffered minimal medium with glucose and yeast extract) solutions had been used to screen for protein, but this method proved ineffective due to infection by *E. coli*. *E. coli* do not grow as efficiently in the absence of yeast extract, hence BMG (buffered minimal medium containing glucose) was used subsequently.

Protocol 2 iv lists the ingredients for BMG and BMM (buffered minimal media + methanol). Each of the nine KM71 and four GS115 colonies were used to inoculate 15 ml of BMG solution, containing 1% histidine. The cells were allowed to grow in a glucose-rich medium until an optical density at 600 nm of between 2 and 6 had been attained. At this point the cells are in the log-phase of growth and will produce most protein when methanol replaces glucose as the sole carbon-source.

Figure 2 iv - Agarose gel showing linearised plasmid

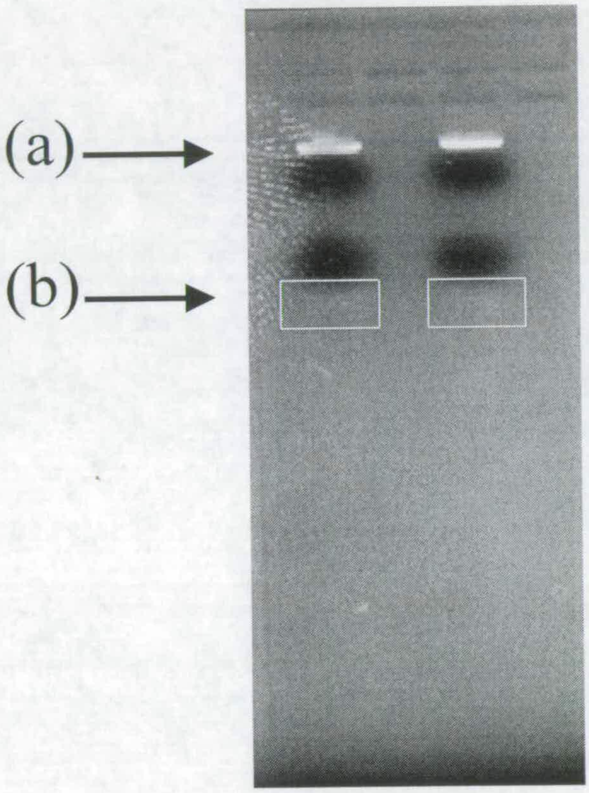

Figure 2 iv - Restriction digest using *Sac*I revealed two bands on agarose gel. a) Linearised plasmid DNA, as expected. b) Feint band corresponding to supercoiled plasmid DNA.

Protocol 2 iii

a) Linearising Plasmid DNA

Contents of mixture for Linearisation	Volume (μl)
BSA (bovine serum albumin)	2
QIAGEN purified pPICZαA/VCP~2,3 plasmid DNA	176
NEB (New England Biolabs) buffer	20
*Sac*I restriction enzyme	4

b) Phenol/Chloroform extraction

Phenylchloroform (200 μl of 1:1) was added to the reaction mixture containing the linearised plasmid DNA. The resulting solution was mixed, centrifuged at 12000 g for 30 s in a microcentrifuge and the top layer (~100 μl) removed. The DNA was recovered by ethanol precipitation. This involved adding 10 μl of 3 M sodium acetate (pH 5.2) to the extracted layer to give a concentration of 0.3 M. Ethanol (250 μl of 70% v/v in water) was then added, the mixture centrifuged for 20 mins at 13000 g and the supernatant carefully discarded. The DNA pellet was the washed using 1 ml 70% ethanol, centrifuged for 1 min at 13000 g and the supernatant discarded. The DNA pellet was then vacuum-dried and resuspended in 10 μl water.

c) Electroporation and Transformation

Pichia pastoris was grown overnight at 30 °C in a 50 ml conical flask containing 5 ml YPD (yeast extract peptone dextrose medium, 1% yeast extract, 2% peptone, 2% dextrose). The overnight growth (0.5 ml) was added to 500 ml of fresh YPD in a 2 L flask and grown until an OD_{600} of 1.5 had been attained (~ 16 hours). The resulting cells were harvested (5 mins at 1500 g, 4 °C) and the pellet was re-suspended in ice-cold sterile water. The cells were the harvested, as before, and re-suspended in 250 ml ice-cold sterile water. The cells were harvested again and re-suspended in 20 ml ice-cold sorbitol (1 M). Finally, the cells were harvested and re-suspended in 1 ml ice-cold sorbitol (1 M). This created the cells that were competent for electroporation. For transformation, 80 μl of the cells were mixed with 2 μl of the ethanol-precipitated plasmid DNA and transferred to an ice-cold 0.2 cm^3 cuvette. The cells were left to incubate for 5 min before applying the 1500 V pulse (25 μF, 200 Ω) for 10 ms. Ice cold sorbitol (1 ml of 1 M) was immediately added to the cuvette and the contents transferred to an ice-cold, 15 ml sterile Falcon tube. The cells were then left to incubate at 30 °C for 2 hours. Cells (~100 μl) were spread on separate YPDS plates plates (as with YPD with addition of 1 M sorbitol). The plates were incubated at 30 °C.

The *P. pastoris* cultures were grown overnight in a shaking incubator at 30 ^{o}C and the OD_{600nm} was checked frequently to detect for log-phase growth. Once the desired cell density had been attained, the cells were harvested by centrifugation (5 mins at 1500 g) and resuspended in 15 ml BMM /1% histidine solution (which is equivalent to BMG with methanol replacing the glucose as the carbon source). This turns on the AOX promoter and induces protein production. Methanol induction levels were maintained by adding 1% methanol and 1% histidine to each inoculated vial every 24 hours. After two days of induction, samples of 1.0 ml were removed daily and analysed by SDS polyacrylamide gel electrophoresis (PAGE).

Several of the colonies tested, both KM71 and GS115 strains, showed the presence of an SDS-PAGE band (~20 kDa) that could indicate expression of VCP~2,3. This was ~6 kDa higher than the theoretical molecular weight of VCP~2,3, although this may be expected due to the anomalous running on an SDS-PAGE gel of intact VCP observed by Kotwal (Kotwal *et al.*, 1988). The band was particularly prominent in one of the KM71 colonies, so a scale-up procedure was carried out. This involved growing an initial 10 ml culture of transformed KM71 in BMG overnight, and using this to inoculate 100 ml of BMG solution. This was grown until an OD_{600nm}

Protocol 2 iv

a) Preparation of BMG and BMM media

One litre of BMG medium contains 100 ml of 1 M potassium phosphate buffer, pH 6.0, 100 ml of 10x YNB (134 g yeast nitrogen base dissolved in 1000 ml water), 2 ml 500 x biotin (20 mg biotin dissolved in 100 ml water), 100 ml 10x GY (100ml glycerol added to 900 ml water), made up to 1 litre with water. Each solution was filter-sterilised through a 0.4 µm membrane, or autoclaved before use. Preparation of BMM used the same recipe with 10 x M (5 ml methanol mixed with 95 ml water) replacing the 10 x GY.

of 2.5 had been reached and the cells were then harvested and re-suspended in 20 ml BMM (because of KM71's slow rate of utilisation of methanol, the BMM volume was 20% of the BMG volume; GS115 strains are resuspended in an equivalent amount of BMM to BMG because of the faster rate of utilisation of the methanol). The culture of KM71 cells in BMM was grown overnight in a baffled flask in a shaking incubator (~250 rpm at 30°C); the baffles were important for effective oxygenation of the cell culture during expression. Methanol and histidine were again added to a final concentration of 1% daily to maintain induction levels, and the level of protein secretion was assessed by SDS-PAGE using 1 ml samples removed from day 2 onwards (figure 2 v). The induction was stopped after five days, as recommended by the Invitrogen manual.

Inspection of the SDS-PAGE gels revealed high levels of protein production for the KM71 strain that had been subjected to the scale-up procedure. This indicated that a "jackpot" clone had been created due to a multiple-copy insertion event. To test that the protein being secreted was indeed VCP~2,3, a 1.0 ml sample of the supernatant from five days post-induction was subjected to SDS-PAGE and subsequently blotted onto nitrocellulose for N-terminal amino acid sequencing. The sequencing confirmed the presence of EAEFIKRRCP and EFIKRRCPSP corresponding to the N-terminus of VCP~2,3, preceded by an additional six residues (EAEFIK) for one species, or four residues (EFIK) for the second species. The two species were present in an apparent ratio of 85:15 in the sequenced protein sample. The additional residues were artefacts of incomplete cleavage of the signal sequence (EA) and of the cloning procedure (EFIK) (figure 2 vi).

Figure 2 v - SDS gel analysis of 1 ml samples to monitor protein expression

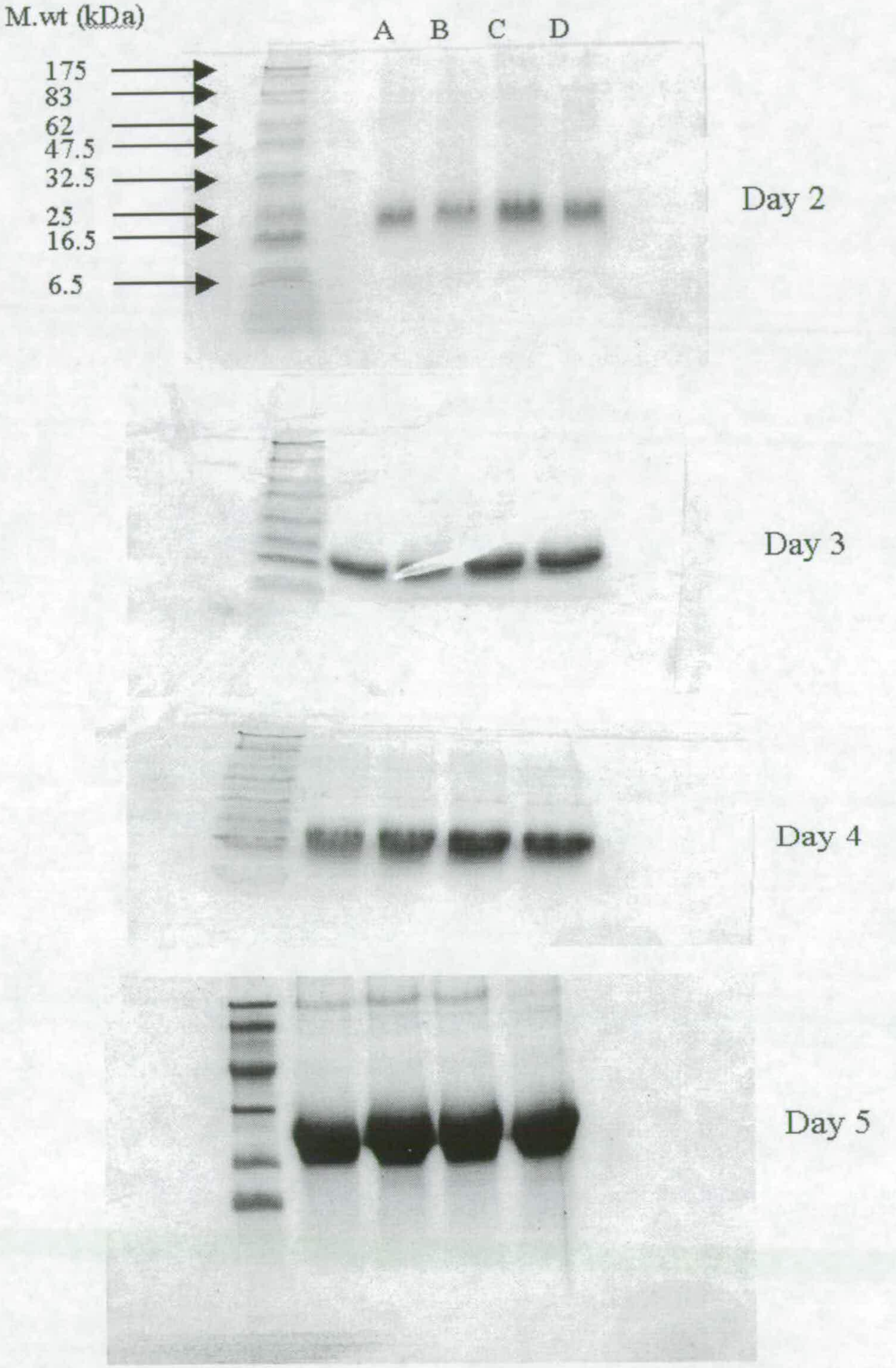

Figure 2 v - SDS gels of 0.5 ml samples from days 2 - 5 of protein induction. Each of the four lanes A-D represents the protein from one of the four growth flasks.

Figure 2 vi - cloning artefacts

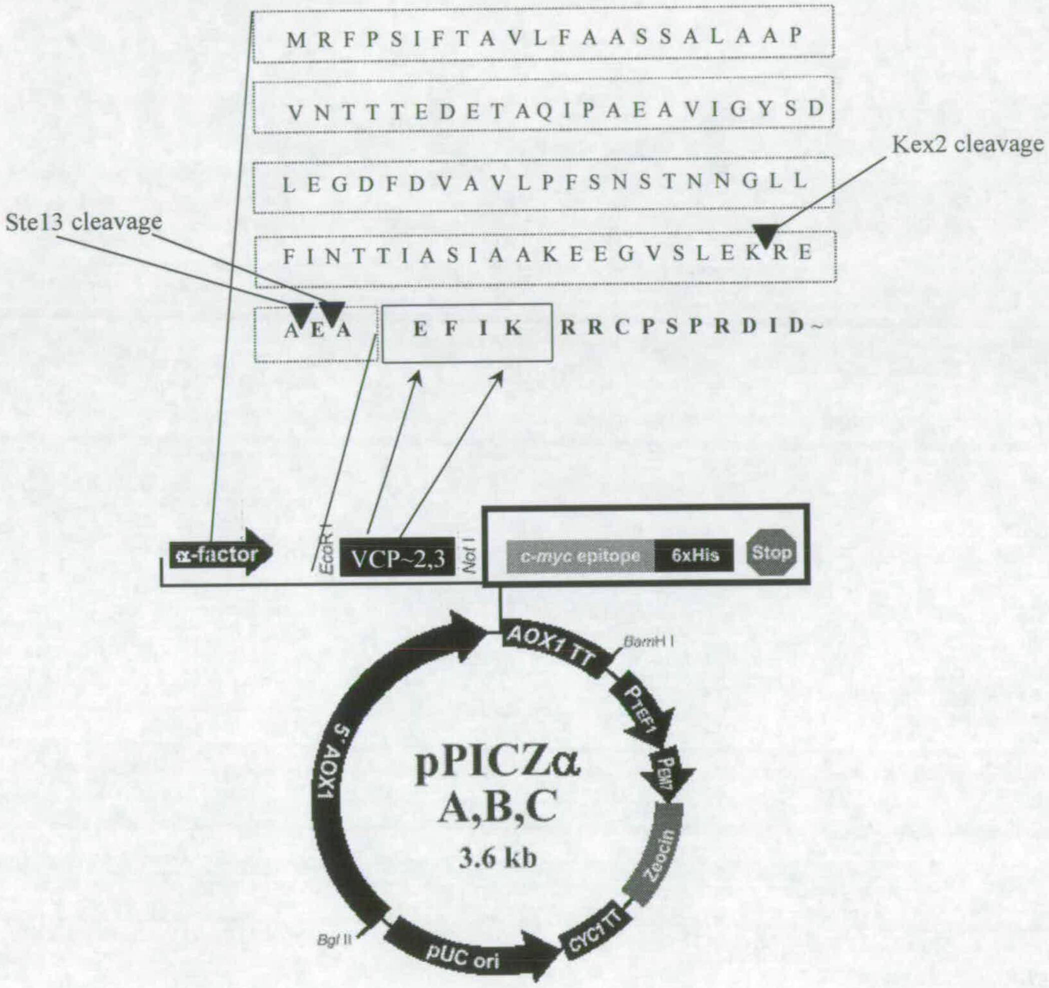

Figure 2 vi - Residues within the dotted box are part of α-factor signal sequence. Initially, the *Kex2* gene product cleaves at the site indicated by the black triangle. The Ste13 gene product then cleaves at one of two positions indicated by the other two black triangles. Residues in bold are part of expressed VCP~2,3 + cloning artifacts. Residues surrounded by the thin box are due to incorporation of *Eco*RI restriction site. The thick black box shows part of the expression system not included in VCP~2,3 due to cloning of a stop codon after the *Not*I restriction site.

2.5.2 Large scale growth and protein purification

For large-scale preparation of VCP~2,3, undertaken in order to produce the NMR samples, a 10 ml culture of the "jackpot" clone in BMG was prepared and allowed to grow overnight in a shaking incubator at 30 °C. This 'starter' culture was then used to inoculate 2 l of BMG, and cultured in a shaking incubator until an OD_{600} of 2.5 had been attained. The pellet was then harvested by centrifugation (5000 g for 10 mins) and re-suspended in 400 ml of BMM, to induce expression of VCP~2,3. Aliquots of 1.0 ml of the suspension were removed and analysed by SDS-PAGE, which revealed the presence of protein at the expected molecular weight. Methanol and histidine were both added to a final concentration of 1% every 24 hours. After five days of induction, the supernatant was collected (10000 g for one hour) and concentrated to ~50 ml using an Amicon pressure cell with 3 kDa molecular weight cut off membrane. The retained concentrate was then centrifuged at 20000 g for 1 hour and the pellet discarded. The protein was then applied to a PD-10 (Amersham Pharmacia, Uppsala, Sweden) gel filtration column, which removed 95% of the salt in the concentrate. The concentrate was eluted from the column in a volume of 3.5 ml and applied to a separate PD-10 column that had been previously equilibrated with 5 mM sodium acetate, pH 4.0, and was eluted in a volume of 3.5 ml. This provided an effective means of exchanging the concentrate into the desired buffer for cation-exchange chromatography. The eluate was filtered through a 0.2 µm membrane and applied to a mono-S cation-exchange column, also equilibrated with 5 mM sodium acetate pH 4.0. A linear gradient of 0 - 1 M NaCl (1 ml/min over 20 mins) was applied to the column and the protein was eluted at 0.6 M NaCl. The cation-exchange purification technique was chosen since pure protein was eluted from the column over a short time period (confirmed by SDS-PAGE analysis of the eluent (figure 2 vii)). This proved to be a better purification system than a previously tested anion-exchange protocol that used a mono-Q column, equilibrated with 15 mM Tris buffer, pH 9.0 and a 0 - 1 M NaCl gradient over 20 mins to elute any bound protein. Using mono-Q, the protein was eluted immediately (figure 2 viii); inspection by SDS-PAGE indicated that the protein was pure, but it was subsequently shown to co-elute with a non-covalently bound carbohydrate that gave strong peaks in 1D NMR analysis. In all ion-exchange protocols, the absorbance was monitored at 280 nm throughout the elution profile to check for those fractions that contained protein. All fractions corresponding to absorbance peaks at 280 nm were analysed by SDS-PAGE for the presence of VCP~2,3. For mass spectrometric analysis, the fractions that contained VCP~2,3 were desalted using a C4 reverse phase column, (eluting with an acetonitrile gradient, figure 2 ix), freeze-dried, dissolved in 50% acetonitrile in aqueous solution containing 0.1% formic acid and analysed by electrospray-ionisation mass spectrometry (figure 2 x). The mass obtained was 13582 Da, which was within one mass unit of the expected molecular weight of VCP~2,3 plus the cloning artefacts (EAEFK) at the N-terminus. Two separate peaks were also present, one at 13381 Da, corresponding to the alternative form of the protein without the E and A residues at the N-terminus (resulting from the Ste13 cleavage at a different position to the site of cleavage in the 'major' species), and the other at 13620 Da (the $+ K^+$ adduct).

Figure 2 vii - sds gel mono s column

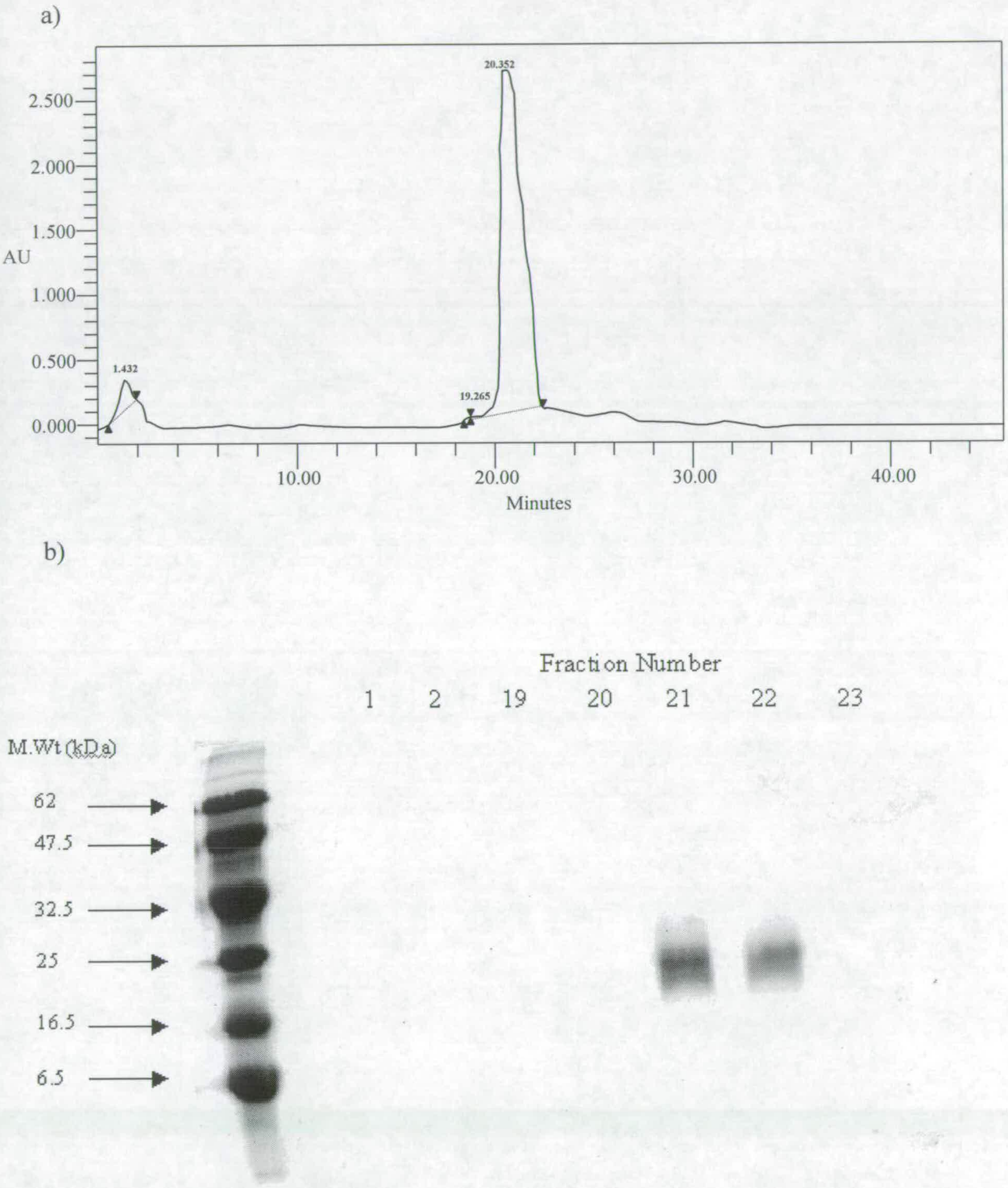

Figure 2 vii - i (a) Elution profile from mono-S cation exchange column; 5 mM sodium acetate buffer was used (pH 4.0) with a linear salt gradient of 0 - 1 M NaCl applied after 10 mins. Gradient was applied over 20 min period. Fractions were collected at 1.0 ml/min. (b) SDS analysis of fractions which gave absorption peaks at 280 nm. 50 μl of each 1.0 ml sample collected was used for SDS polyacrylamide analysis.

Figure 2 viii - Elution profile and SDS gels from Mono-Q anion exchange column

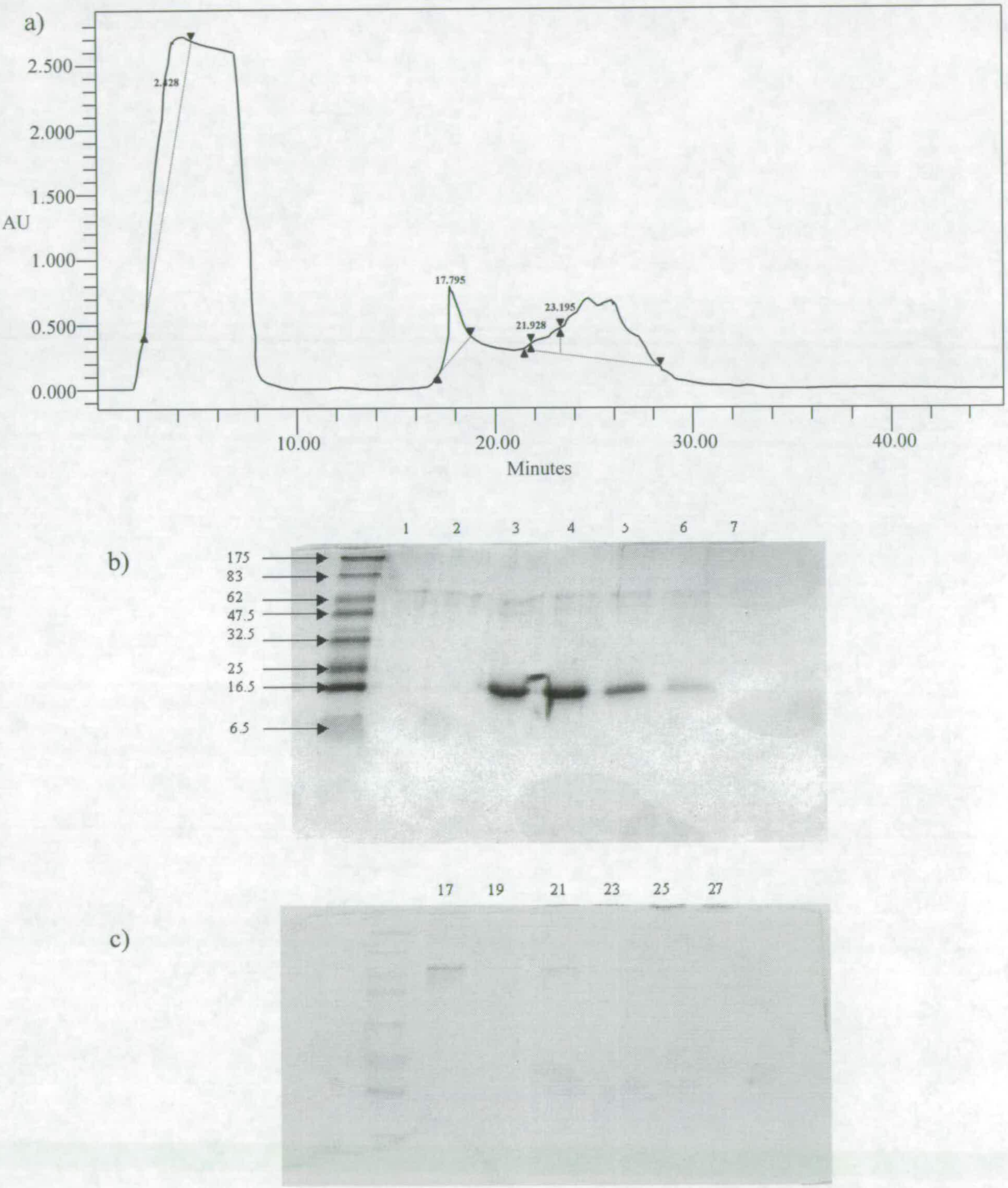

Figure 2 viii - a) Elution profile from mono-Q cation exchange column; 50 mM Tris phosphate buffer was used (pH 9.0) with a linear salt gradient of 0 - 1 M NaCl applied after 10 mins. Gradient was applied over a 20 min period b) & c) Corresponding SDS gels from fractions corresponding to absorbance peaks on elution profile.

Figure 2 ix - elution profile from reverse phase column

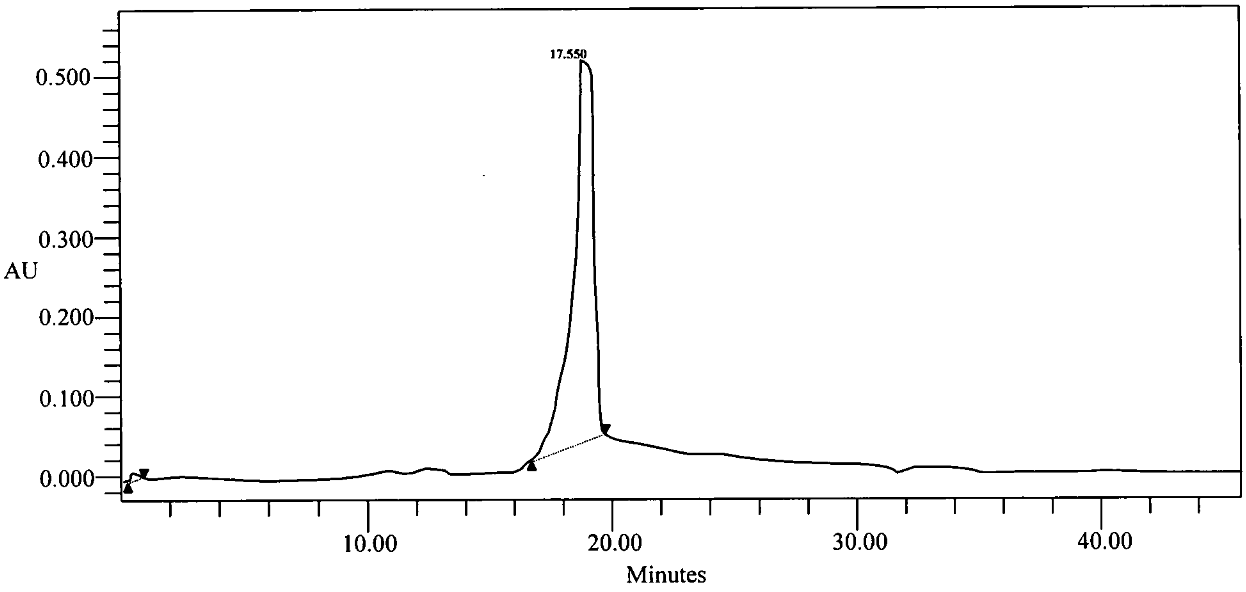

Figure 2 ix - Elution profile from C4 reverse phase column, used to desalt VCP~2,3 before mass spectrometric analysis. An acetonitrile gradient of 80% H_2O (0.1% TFA)/20% AcN (0.1% TFA) to 10% H_2O/90% AcN was applied over a 30 minute period.

Figure 2 x - Electrospray ionisation mass spectrometry of the HPLC-purified protein fraction

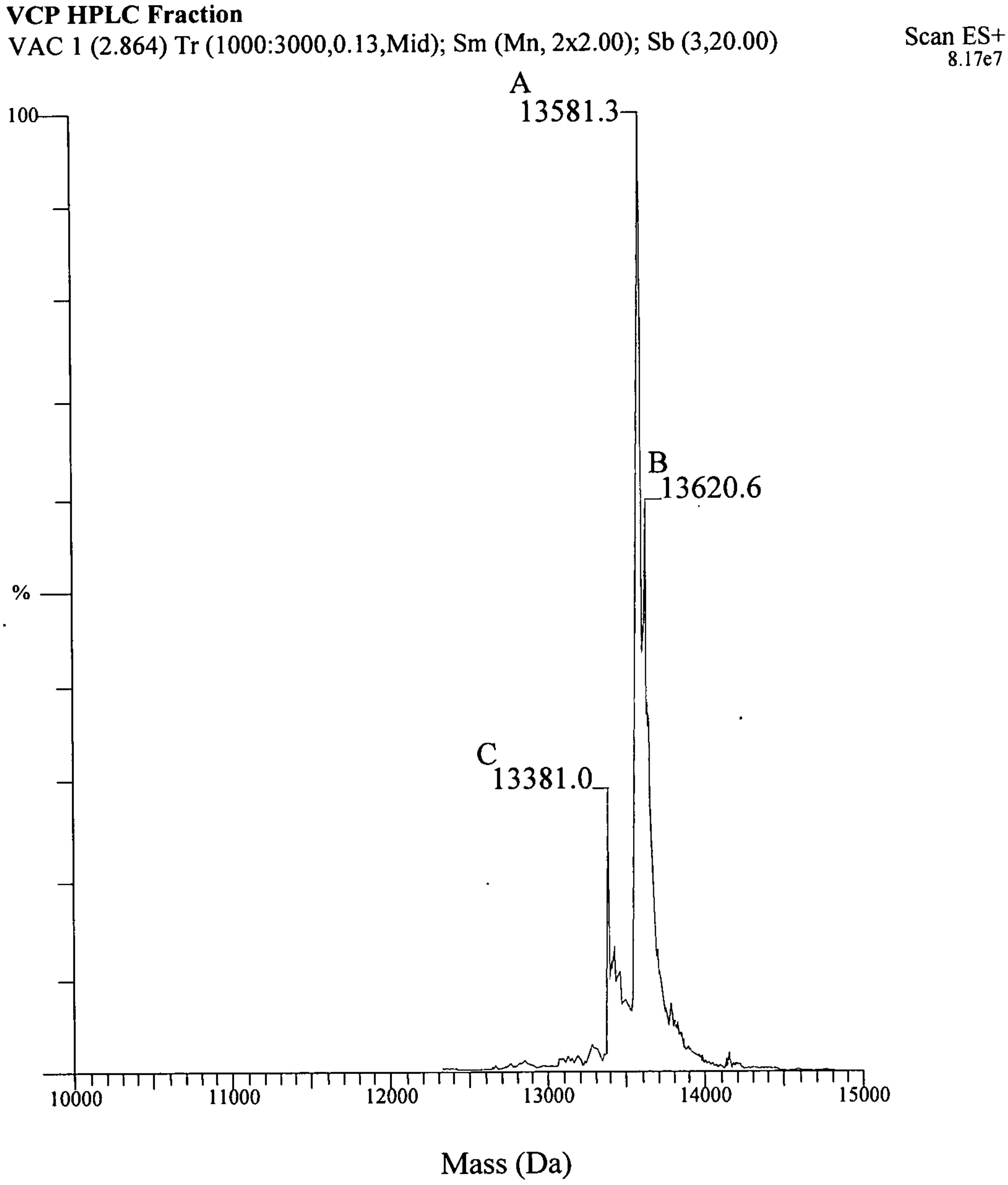

Figure 2 x - Mass spectrometric analysis of protein fragment shows three main bands. Peak A - VCP~2,3 with EAEFIK at N-terminus. B - VCP~2,3 with EAEFIK + K^+. C - VCP~2,3 with EAFIK at N-terminus.

2.6 Preparation of VCP~2,3 for NMR analysis

2.6.1 1D and 2D homonuclear experiments

For preparation of VCP~2,3 for NMR analysis, the freeze-dried protein was dissolved in 1.00 ml H_2O. The absorbance of the protein at wavelengths 280_{nm} (13.95), 320_{nm} (0.1), and 350_{nm} (0.05) were recorded and applied to equation 1, as an estimate of the protein concentration (Mach *et al.*, 1992). The concentration of the protein from the absorbance measurements indicated that the final sample was ~2.5 mM, which was suitable for 2D NMR experiments. The protein was dissolved in 550 µl of 10 mM sodium phosphate, pH 6.0, and 50 µl of ^2H_2O for the 1D and 2D NMR experiments.

Equation 1

$$C = \frac{A_{280} - 10\,(2.5 \log A_{320} - 1.5 \log A_{350})}{5540_{nW} + 1480_{nY} + 134_{nC}}$$

C: Concentration in moles/litre
$_{nW}$: Number of tryptophan residues in VCP~2,3 (2)
$_{nY}$: Number of tyrosine residues in VCP~2,3 (4)
$_{nC}$: Number of cysteine residues in VCP~2,3 (8)
A_{280}: Absorbance at 280_{nm}
A_{320}: Absorbance at 320_{nm}
A_{350}: Absorbance at 350_{nm}

2.6.2 Preparation of ^{15}N - labelled VCP~2,3 for 3D NMR analysis

To prepare the ^{15}N-labelled sample, the nitrogen source for both BMG and BMM media, normal yeast nitrogen base (YNB), was replaced with YNB that contains no amino acids or ammonium sulphate. Normally, ammonium sulphate is present at a concentration of 1% in the YNB but a series of test cultures were prepared in YNB with ammonium sulphate present as the sole nitrogen source at concentrations of 0.2%, 0.4%, 0.6%, 0.8% and 1% (w/v). These test cultures revealed no decrease in protein production (estimated from SDS-PAGE) when 0.2% ^{15}N-labelled ammonium sulphate was used, so all subsequent ^{15}N-labelled growths were carried out using this concentration. However, the KM71 strain used for protein production required a histidine supplement in the growth medium due to the Mut^{s}/His^{-} phenotype of the strain. To ensure that all amino acids, including histidines, had been ^{15}N-labelled, the plasmid pPIC9K was transformed into the 'jackpot' cell line, using identical methods to the initial transformation of pPICZαA/VCP~2,3 into *P. pastoris* (protocol 2.4.2.i). This produced a KM71 strain with phenotype Mut^{s}/His^{+} (as pPIC9K contains the HIS4 gene), allowing the cells to metabolise histidine intracellularly.

The protein concentration was estimated using Equation 1. The absorption values were A_{280}, 13.35, A_{320}, 0.1, A_{350}, 0.05 which gave a sample concentration of ~2.5 mM. The buffer used for 3D experiments was 10 mM sodium phosphate, pH 6.0.

References

Buckholz, R.G & Gleeson, M.A.G (1991) *Biotechnol* **9**, 1067-1072

Cereghino, J. L. & Cregg, J. M. (2000) *FEMS Microbiol.Rev.* **24,** 45-66

Chiruvolu, V., Cregg, J. M. & Meagher, M. M. (1997) *Enzyme Microb Technol* **21**, 277-283

Cohen *et al.* (1972) *Proc.Natl.Acad.Sci. USA* **69**, 2110-2114

Cohen *et al.* (1973) *Proc.Natl.Acad. Sci USA* **70**, 3240-3244

Cregg , J. M., Madden, K. R., Barringer, K. J., Thill, G. P. & Stillman, C. A. (1989) *Mol.Cell.Biol* **9**, 1316-1323

Cregg, J. M & Madden, K.R (1987) *In: Biological Research on industrial yeasts (Stewart et al.)* CRC Press, BOCG Raton, Fl **Vol 2,** 1-18

Cregg, J. M., Barringer, K. J., Hessler, A. Y. & Madden, K, R. (1985) *Mol. Cell. Biol* **5**, 3376-3385

Egli, T., van Dijken, J. P., Veehuis, M., Harder, W. & Feichter, A.. (1980) *Arch. MicroBiol* **124**, 115-121

Ellis, S. B., Brust, P. F. Koutz, P. J., Waters, A. F., Harpold, M. M. & Gingeras, T. R. (1985). *Mol.Cell.Biol* **5**, 1111-1121

Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heynecker, H., Bolivar, F. & Boyer, H. W. (1977) *Science* **198**, 1056-1063

Kotwal, G. J & Moss, B (1988) *Nature* **335**, 176-178

Mach, H., Middaugh, R. & Lewis, R.V. (1992) *Analyt. Biochem.* **200**, 74-80.

Matthews *et al.* (1994) *Nature* **370**, 666-668

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B & Erlich, H. A. (1988) *Science* **239**,487-491

Saiki, R. K., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230**, 1350-1355

Smith, D. B & Johnson, K. S. (1988) *Gene* **67**, 31-40

Tschopp, J. F., Brust, P. F., Cregg, J. M., Stillman, C. A & Gingeras, T. R. (1987) *Nucleic Acids Res* **15**, 3859-3876

Veehhuis, M., van Dijken, J. P. &Harder, W. (1983) *Adv.Microb.Physiol* **24**, 1-82

Weiss, B., Jacquemin-Sablon, A., Live, T. R., Fareed, G. C.& Richardson, C. C. (1968) *J.Biol.Chem* **243,** 4543-4555

Chapter 3 - NMR

3.1 Basic principles of NMR

Nuclear magnetic resonance, or NMR, is an effective tool for gaining information about the chemical composition or structure of large and small molecules. It is used in classifying compounds, determining the chirality of small molecules, solving protein 3D structures and measuring the dynamics of substances in solution.

NMR relies on the fact that some atoms have a spin quantum number that is a multiple of ½ . These nuclei possess a magnetic moment and behave like tiny bar magnets when exposed to a high-strength magnetic field. The orientation of the nuclei is restricted to 2I+1 orientations in space, where I is the magnetic moment of the nuclei in question. The different orientations have different associated energy levels and slightly more nuclei will exist in the lower state than the higher state, according to the Boltzman distribution. Transitions between these states can be induced by applying electromagnetic radiation at an appropriate frequency, which disrupts the population states of the nuclei. When an external magnetic field, B_o is applied, the nuclei do not align exactly with or against the field, but precess around it at the Larmor frequency (ν_0). The Boltzman distribution ensures that when this initial magnetic field is applied, the bulk magnetisation is in the lower energy state. Application of an intense radio-frequency pulse at the Larmor frequency can induce inversion or reduction of the bulk magnetisation. The nuclei then precess around the new direction of the bulk magnetisation. As they precess, the magnetisation varies and can be detected by placing a wire coil near the oscillating nuclei. When the pulse is cut off, the magnetisation decays back to the equilibrium state. During this period, the nuclei, which have different local chemical environments, will resonate at individual frequencies. The oscillating frequency decay, which varies sinusoidally as the nuclei precess, is known as the free induction decay (FID) and can be readily measured.

In essence, the FID encodes the frequency of each nucleus in the sample, shifted upfield or downfield from the reference (Larmor) frequency depending on the magnetic environment of the nucleus. Thus the 'chemical shift' of nuclei in similar environments tend to approximate to characteristic values. The NMR spectrum is created by performing a Fourier transformation (FT) on the FID. This allows the intensity of each chemical shift to be plotted on a continuous frequency scale.

The quality of a NMR spectrum depends on the strength of the magnetic field used. For small molecules, a magnet of 6.4 Tesla (270 MHz) is sufficient, whereas large proteins, which will have nuclei with many overlapping chemical shifts, require a magnet of 14.2 Tesla (600 MHz) or above. As the magnetic field is increased, the ratio of signal:noise improves (due to increases in the population differences between the two energy states), but just as importantly the lines are narrower on the parts per million (ppm) scale that is used to quantify the chemical shift. This gives an increased resolution in the resulting spectrum (Cavanah *et al.* (Eds), 1996, Rattle (Ed), 1994).

For a protein such as VCP~2,3, which has a MW of ~13 500 Da, 1D NMR is useful as an indicator of sample strength and of whether the protein is folded and mono-disperse. For structural analysis of a protein, however, 2D and 3D spectra are necessary to resolve the large number of resonances that are present. The most important nuclei in biomolecular NMR are ^{1}H, ^{13}C, ^{15}N, along with ^{19}F (incorporated artificially) and ^{31}P (in e.g. DNA or phospoproteins). For the study of VCP~2,3, experiments using ^{1}H and ^{15}N nuclei were sufficient for interpretation of the nuclear spins – but for a bigger protein, additional labelling with ^{13}C would have been necessary. The various NMR spectra that were acquired for the structural analysis of VCP~2,3 are described briefly below along with their merits (the pulse sequences shown to illustrate the individual experiments were the ones used to acquire the spectra of VCP~2,3).

3.2 1D and 2D NMR experiments

3.2.1 1D NMR; pulse and acquire

A fully Fourier transformed and processed 1D ^{1}H NMR spectrum will contain a peak for every proton frequency in the sample although in most cases the peaks will be overlapped. A 1D spectrum is obtained by transmitting a radio-frequency (rf) pulse at the Larmor frequency, which rotates the bulk magnetisation into a different direction. After the pulse of rf, the receiver is switched on, which will detect the oscillating frequencies of the nuclei as they relax back to their initial (ground) state. If Cartesian co-ordinates are considered, the pulse is normally applied along the x-axis, causing the bulk magnetisation (aligned along z) to 'flip' into the xy plane. This would be a 90 ^{o} pulse. In fact, the duration and direction of the pulse can be altered to tilt the bulk magnetisation through any angle. The 'pulse sequence' used to obtain a 1D spectrum is shown in figure 3 i.

Figure 3 i - 1D NMR pulse sequence

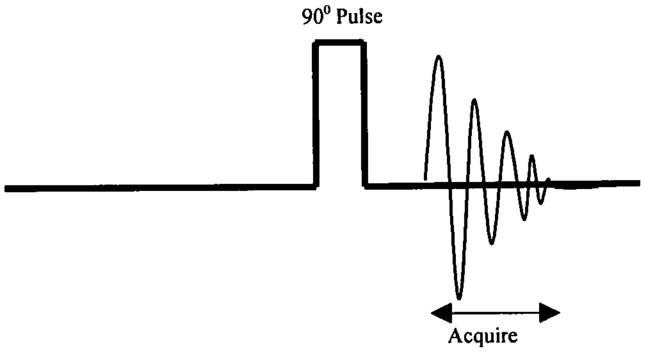

Figure 3 i - The basic pulse and acquire sequence for obtaining a 1D NMR spectrum. The resonances from the precessing nuclei are detected during the acquire stage as they decay back to the ground state.

Figure 3 ii - Four stages in 2D NMR experiment

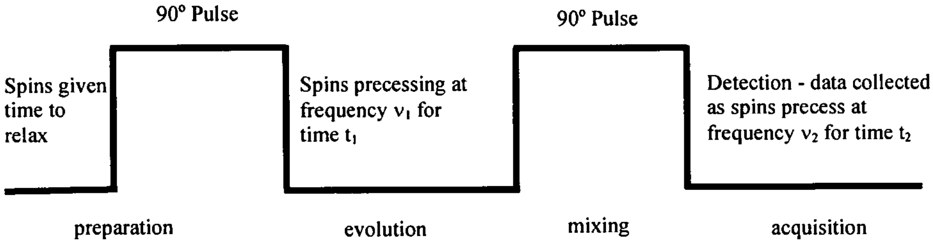
90° Pulse
90° Pulse
Spins given time to relax
Spins precessing at frequency ν_1 for time t_1
Detection - data collected as spins precess at frequency ν_2 for time t_2
preparation
evolution
mixing
acquisition

Figure 3 ii - Basic scheme for a 2D NMR pulse sequences. The first pulse excites the nuclei and they are given time to evolve during t_1. Magnetisation transfer between protons occurs during the mixing period when atoms labelled by their Larmor frequency are coupled to nearby atoms through dipolar or scalar coupling.

3.2.2 2D NMR

All two-dimensional (2D) NMR experiments have the same basic theory and can be broken down into four stages outlined in figure 3 ii. The pulse sequences in 2D NMR are designed to allow detection of magnetisation transfer (coupling) between nuclei. The coupling can be through covalent bonds (scalar coupling) or through space (dipolar coupling). Two pulses can be applied to the bulk magnetisation and the size of the resulting signal will depend on the Larmor frequency of the nuclei and the time (t_1) between applying the first and second pulse. The FID can be collected over another set period of time (t_2). Therefore, alteration of (t_1) will result in different Larmor frequencies (ν_1) having different intensities. Acquiring many FIDs at different t_1 will generate a series of FIDs that can be Fourier transformed with respect to both t_1 and t_2 to give a 2D matrix showing intensity against frequency (ν_1,ν_2). The chemical shifts of the nuclei that are observed in the 1D NMR spectrum remain the same and appear as the diagonal of the 2D spectrum. Signals that arise from coupling are a product of two nuclei, and will therefore have two chemical shifts - the signal is a product of magnetisation transfer between proton I and proton J and will have co-ordinates of I_{ppm},J_{ppm} in the 2D frequency matrix. In a 2D experiment, the various frequencies of the nuclei are 'labelled' by their chemical shifts during t_1 and magnetisation transfer from the frequency-labelled protons occurs during the mixing period. The initial pulse ensures that the nuclei precess coherently, and it is this 'phase coherence' of a group of similarly precessing spins that can be transferred to a different group of protons.

The 2D spectra have more "frequency space" than 1D experiments and therefore there is less overlap between peaks. A combination of different, specifically designed, pulse sequences allows identification of signals that are caused by scalar coupling through no more than three bonds (correlated spectroscopy, COSY) or through up to five or six bonds (total correlation spectroscopy, TOCSY), or by dipolar coupling through space (nuclear Overhauser spectroscopy, NOESY). In combination, these experiments allow one to distinguish between the sets of nuclei in the different side-chains of amino acids and identify pairs of nuclei, perhaps in different amino acid residues, that are close enough to participate in through-space interactions. This provides the information necessary to determine the structure of a protein.

3.2.3 2D correlated spectroscopy (COSY)

A 2D COSY experiment (Jeener, 1971, Aue *et al.*, 1976) is useful for identifying atoms that are no more than three covalent bonds apart. The spin couplings in a COSY experiment are not transmitted through the peptide bond. The experiment produces a spectrum which is particularly useful for amino acid-type assignment, since each ^{1}H nucleus in a side chain will have its own predictable system of coupled spins.

The scalar coupling between two nuclei results in four possible energy levels, due to the different energy states available to each spin (figure 3 iii). The four energy states for the two-spin system are very similar, but have slightly different frequencies, so that transition between the states is possible. In practice, a COSY experiment typically uses the pulse sequence shown in figure 3 iv. The most simple COSY experiment involves a hard 90 ° pulse (prior to t_1) that excites all four transitions at the same time and a subsequent 90 ° pulse causes sharing of the coherence between the

Figure 3 iii - Energy level diagram for a two spin (AX) coupled system

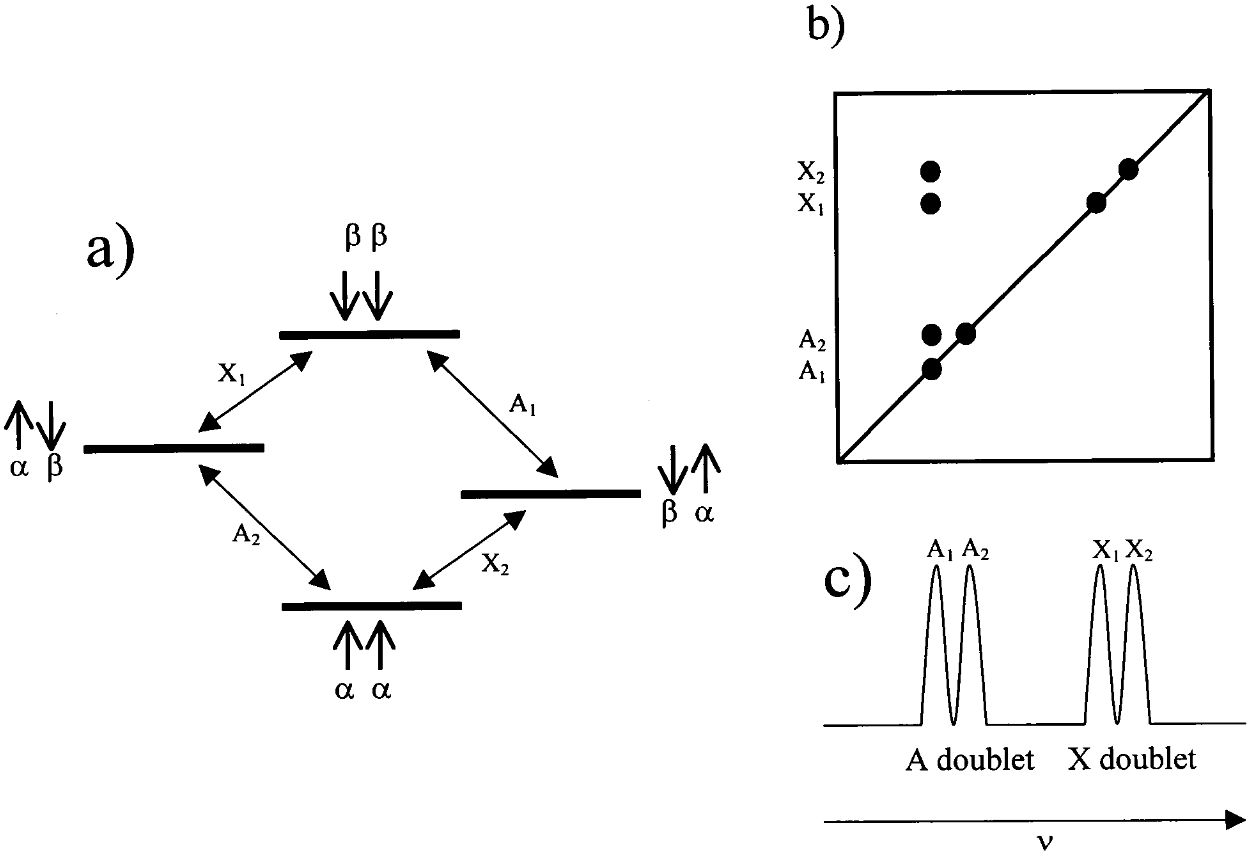

Figure 3 iii: a) Energy level diagram for a two spin (A and X) coupled system. Each of the spins of each nucleus can be aligned with or against M_0 resulting in a doublet for A and X respectively (c). Frequency labelling at ν for A_1 would result in a two-dimensional spectrum shown in (b).

Figure 3 iv - COSY pulse sequence

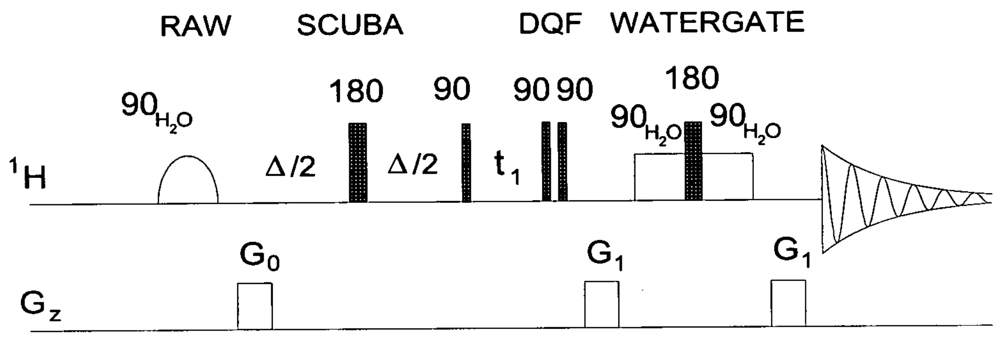

Figure 3 iv – The pulse sequence for a COSY experiment. Techniques such as RAW (randomisation approach to water suppression), SCUBA (stimulated crosspeaks under bleached alphas) and DQF (double quantum filtering) allow protons close to the water resonance or diagonal to be observed. WATERGATE is used to suppress the water signal by applying selective pulses to the water resonance (the gradients serve to dephase and re-phase non-water signals)

Figure 3 v - 2D spectrum of two spin system when all frequencies are labelled

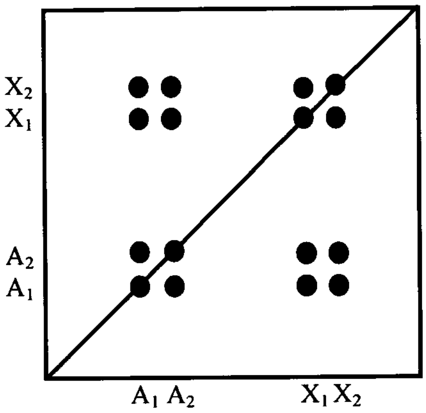

Figure 3 v - A hard 90° pulse excites all four transitions at the same time to produce the sixteen-peak contour map for the two spin system.

transitions. This gives sixteen cross-peaks for the two-spin system, as shown in figure 3 v. This represents an idealised view of the COSY experiment – advances in NMR techniques have resulted in more complex pulse sequences (as shown above) that help isolate the useful signals (Altieri *et al.*, 1995). Artefacts are in fact normally present that have to be eradicated using methods beyond the scope of this thesis, but involving cycling through a number of different pulse and receiver settings for each FID.

3.2.4 Total correlation spectroscopy (TOCSY)

A 2D TOCSY experiment (Braunschweiler & Ernst, 1983), like the COSY experiment, makes use of the fact that spin coupling is not transferred through a peptide bond and thus the spin systems contained within each amino acid residue of a protein are isolated. In a TOCSY experiment, rf is applied throughout the mixing time after a second 90 ° pulse. This provides continuous irradiation and acts as a spin-lock, as the precessing magnetisation vectors are forced to rotate at the frequency of the continuous rf (the Hartmann-Hahn condition) (Hartmann & Hahn, 1962). This makes magnetisation transfer very efficient and the extent of the transfer depends on the length of the mixing time. In a TOCSY pulse sequence (figure 3 vi) (Shaka, *et al.*, 1988, Hwang & Shaka, 1995, Calligihan *et al.*, 1996) a sequence of quickly repeated, 180 ° pulses, causes phase coherence to be transferred throughout the coupled spins (DIPSI-2 stage). The result is a series of signals representing the transfer of magnetisation between all the spins in an amino acid for a chosen mixing time (short mixing times result in less propagation of the coherence throughout the side chain of an amino acid, longer mixing times improve the transfer efficiency).

Examples of the resonance patterns expected for amino acids in a TOCSY experiment are shown in figure 3 vii. It is obvious that the resonances that appear in a COSY spectrum will also appear in a TOCSY spectrum since the latter detects all of the possible transitions within the side chains' spin systems. The sixteen peaks that may result from a two-spin system in a COSY spectrum are not resolved in a TOCSY spectrum due to increased line-width. COSY pulse sequences usually use anti-phase enhancement to allow the actual coupling constants between protons to be measured.

Figure 3 vi - TOCSY pulse sequence

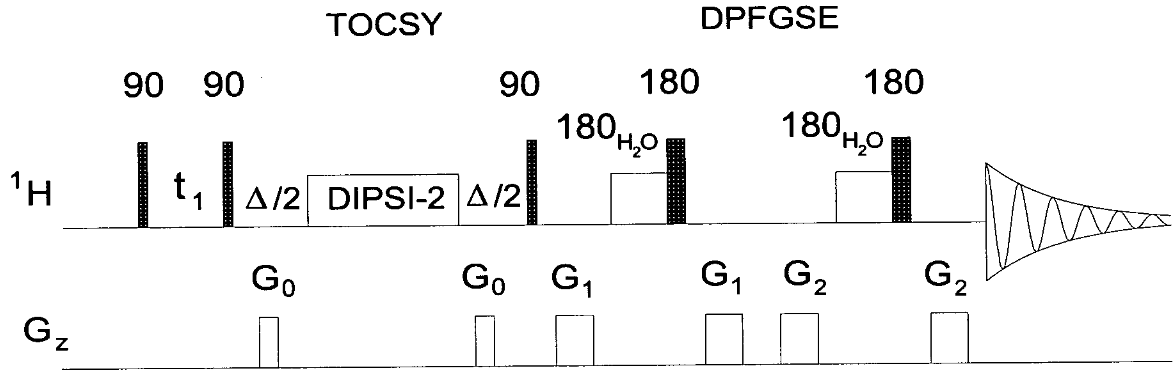

Figure 3 vi – TOCSY pulse sequence. The gradients (G) destroy the xy magnetisation whilst the DIPSI-2 spin-lock period causes polarisation transfer. Water suppression is via the DPFGSE method (double pulsed field gradient spin echo). DPFGSE works with similar principles to WATERGATE, but is three times longer.

Figure 3 vii - 2D contour maps of amino acid patterns from TOCSY experiment

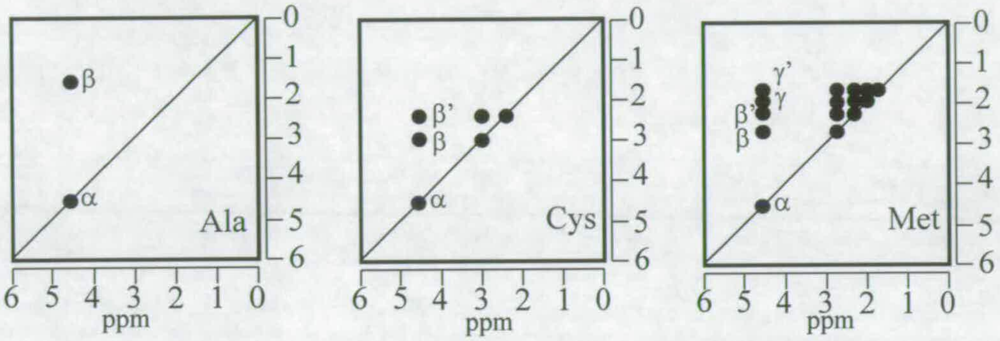

Figure 3 vii– contour maps of TOCSY resonances from side-chains of Alanine, Cysteine and Methionine

3.2.5 Nuclear Overhauser spectroscopy (NOESY)

The NOESY experiment (Jeener *et al.*, 1979) provides information about through-space connectivities between nuclei and can, therefore, be used to map the position of amino acids with respect to one another. A nuclear Overhauser effect (NOE) represents the change in intensity of one resonance when another nearby resonance is perturbed. The affected peak gains or loses intensity due to cross-relaxation, the transfer of magnetisation through space, rather than through scalar coupling. The size of the NOE is dependent on the distance between the two nuclei and decreases as the sixth power of the inter-nuclear distance. Thus, detection of NOEs and measurement of their intensities provide distance restraints and information useful in assignment.

Figure 3 viii – NOESY pulse sequence

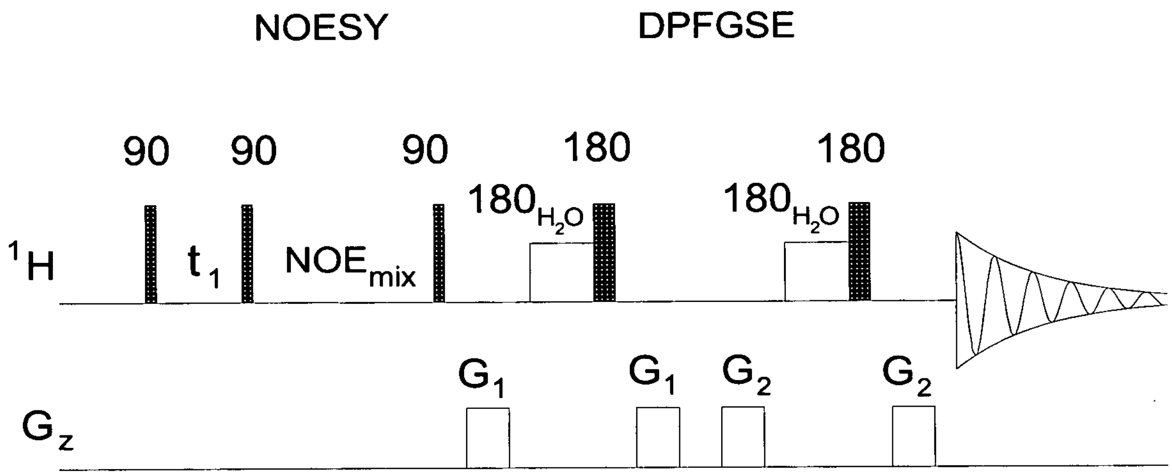

Figure 3 viii – NOESY pulse sequence. NOE_{mix} is the period during which cross-relaxation between protons occurs, resulting in the NOE (detected by applying a third 90 ° pulse). The water suppression, as with the TOCSY experiment, uses DPFGSE.

Callighan *et al.*, 1996). The third pulse in the sequence allows the dipolar coupling between two spins to be observed when a group of spins, precessing coherently during t_1, undergo cross-relaxation during the mixing period and have a different frequency during t_2 due to the effects of nearby spins. This involves applying a 90 ° "saturation pulse" that will perturb the energy state of one group of precessing spins and result in an NOE if cross-relaxation (dipolar coupling) takes place. In practice, this effect can only occur if the two nuclei are <~5.5 Å apart.

Although the 2D NOESY spectrum gives enhanced resolution compared to the 1D version, the spectrum still contains many signals and overlap is inevitable. For larger proteins then, resolution is still a problem, even in 2D experiments. For this reason, use is made of other NMR observable nuclei (^{13}C,^{15}N) and, in the case of VCP~2,3, isotopically labelling with ^{15}N was necessary since this allowed acquisition of 3D experiments.

3.3 3D NMR

3.3.1 Principles

Three dimensional (3D) NMR operates in much the same way as 2D NMR, although the FIDs are now acquired during a third time period, t_3, instead of during t_2. The time periods t_1 and t_2 are still systematically varied and inclusion of a third time period results in a 3D matrix rather than a 2D matrix. The pulse sequences for 3D $^1H^{15}N$ HSQC-NOESY and 3D $^1H^{15}N$ HSQC-TOCSY experiments are shown in figures 3 ix and 3 x (Sklenar *et al.*, 1993, Mori *et al.*, 1995). During a 3D experiment,

Figure 3 ix – ^{1}H-^{15}N HSQC-NOESY pulse sequence.

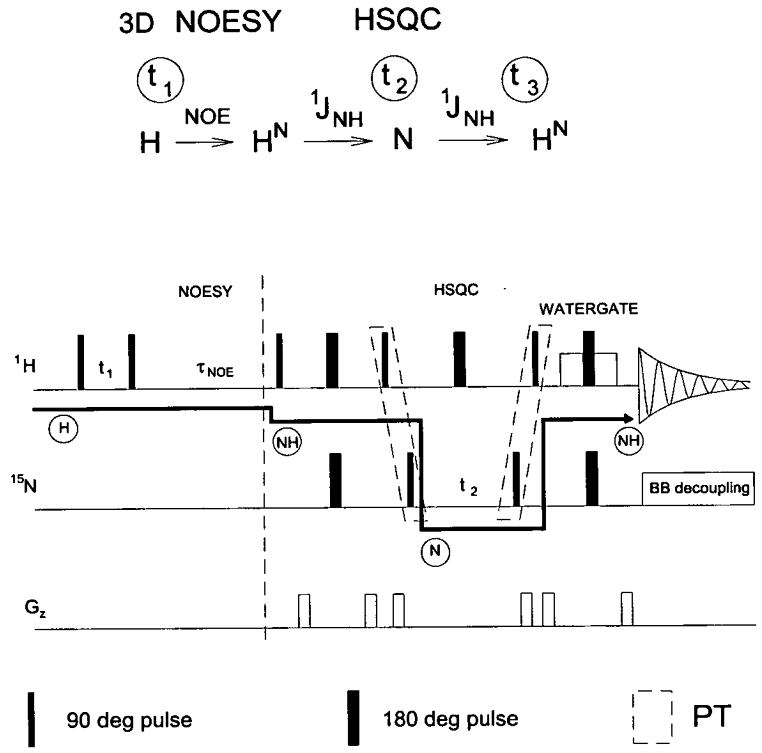

Figure 3 ix – ^{1}H-^{15}N HSQC NOESY pulse sequence. The pulse sequence for the HSQC NOESY is similar to that of the 2D NOESY initially. The magnetisation is transferred to the nitrogen nucleus and is then transferred back to the protons before acquisition. This allows connectivities involving protons bound to ^{15}N atoms to be observed.

Figure 3 x – ^{1}H-^{15}N HSQC TOCSY pulse-sequence

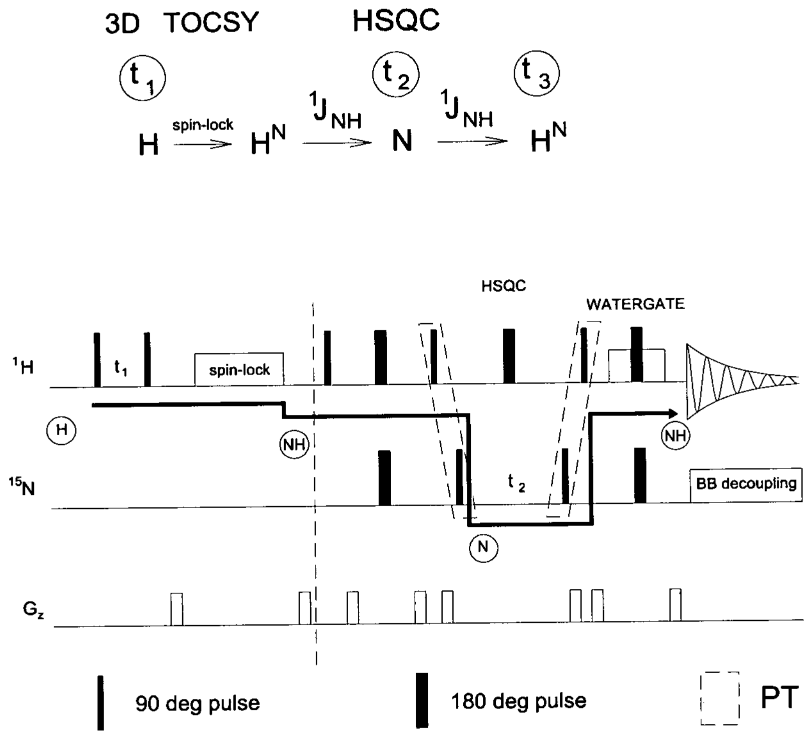

Figure 3 x – 1H-15N HSQC TOCSY pulse sequence. The pulse sequence is similar to the 2D TOCSY experiment. The magnetisation is transferred to the ^{15}N atoms and then back to protons again, as in the ^{1}H-^{15}N HSQC NOESY experiment.

the phase coherence of a group of spins is transferred to the isotopically labelled ^{15}N during t_2, allowed to evolve and then transferred back to the ^1H nuclei. This requires two different types of rf pulses, in order to manipulate the different nuclei. The result is a third frequency label, ν_3, which is dependent on the Larmor frequency of the ^{15}N atoms. Thus, the 3D matrix can be used to separate amide protons on the basis of the chemical shift of the amide proton and the chemical shift of their attached nitrogen atoms. The amide protons are labelled with the frequencies of other protons close enough in space to allow exchange to occur and the resolution is improved dramatically. The 3D cube can be split into slices in any of the two ^1H-proton dimensions or in the ^{15}N dimension. A 2D version of the spectrum can also be collected which displays each $^1H^{15}N$ chemical shift with co-ordinates F^1_{Hppm}, F^2_{Nppm}. Such 2D heteronuclear single quantum coherence spectra (HSQCs) are valuable reference spectra, as one peak should appear for each amide proton in the protein. Changes in protein structure or conformation can be monitored by acquiring a series of HSQC's under different experimental conditions and mapping the changes in position of the cross-peaks.

3.3.2 HNHA experiments

The HNHA experiment can be used to measure the $^3J_{HNH\alpha}$ scalar coupling. The pulse sequence for a HNHA experiment is shown in figure 3 xi (Kuboniwa *et al.*, 1993, Vuister & Bax, 1993). The HNHA experiment provides valuable information since the ratio of the intensity (or integral) of the diagonal (HN) peak to the Hα peak is dependant on the angle between the two bonds and can be calculated, in Hz, to give structural information. The Karplus curve (figure 3 xi) can be used to correlate the

Figure 3 xi - Karplus curve and HNHA pulse sequence

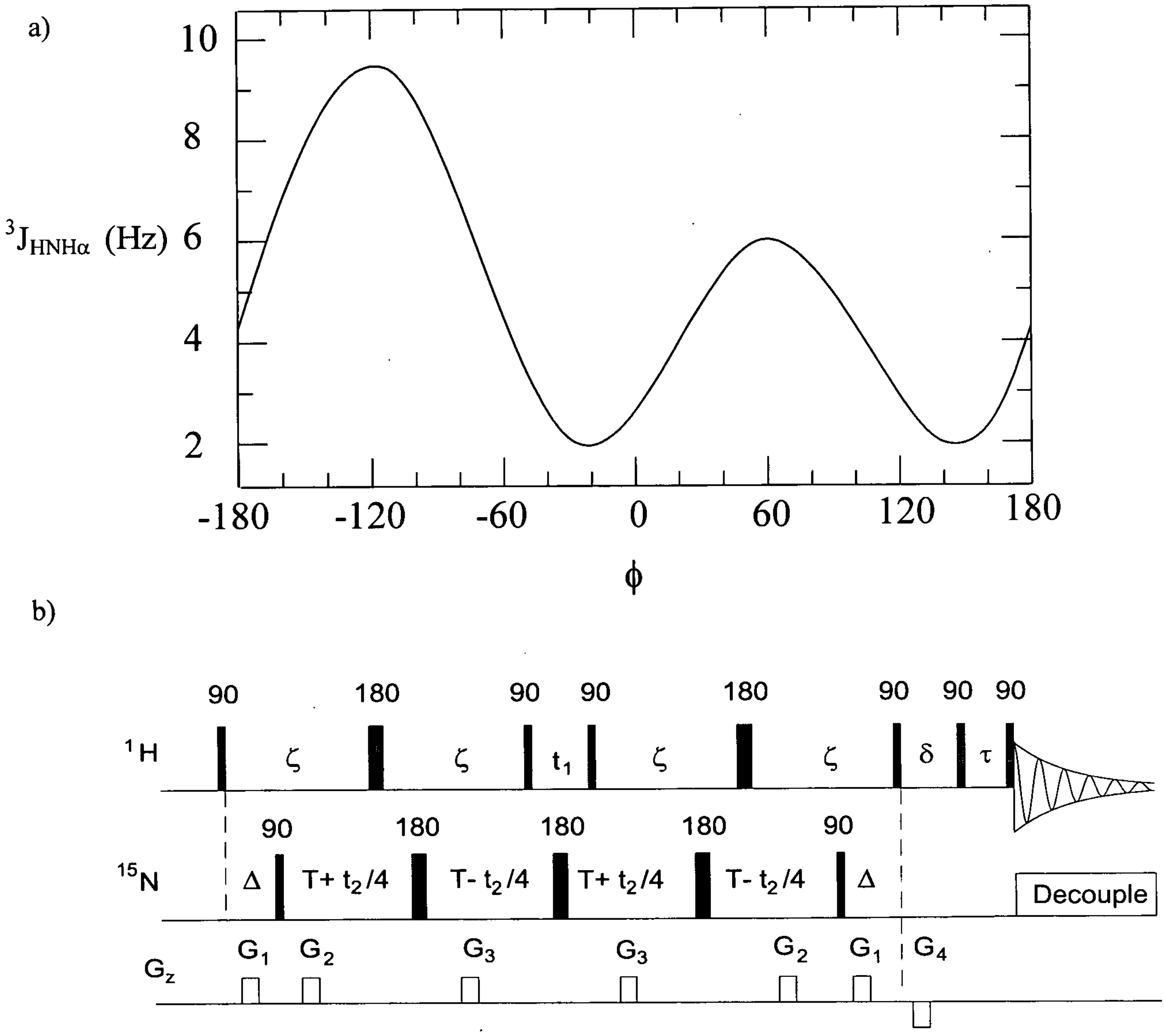

Figure 3 xi – a) The Karplus curve used to calculate ϕ from the $^3J_{HNH\alpha}$ coupling data. b) Pulse sequence for HNHA experiment.

coupling to the dihedral (ϕ) angle in the backbone of the protein. In protein structures, a ϕ angle -120 ° is inferred from a coupling constant of >7.5 Hz and an angle of -60 ° is inferred from a coupling constant of <5.5 Hz.

3.4 Water suppression

The concentration of protons in water is ~110 M, which is high compared with the ~2.5mM proton concentrations in a NMR sample. The result is a huge signal originating from the water signal, that would mask an important region (4.4-4.9 ppm) of the spectrum corresponding to peaks from Hα protons. One way of eliminating the water signal is to acquire spectra using ^2H_2O as solvent. This provides a means of assigning those resonances hidden by the water signal, but does not display resonances originating from amides that normally exchange with water but which are visible in ^1H_2O, such as backbone amides that are not participating in H-bonds and side chain asparagine HN protons.

To combat the water signal problem, a technique called presaturation can be applied, which uses a weak radio-frequency at the Larmor frequency of the water resonance. So many transitions then take place that there is no bulk magnetisation in the M_z direction corresponding to the water protons when the initial 90 ° pulse is applied and so no magnetisation in the M_{xy} plane is observable. In recent years, however, more advanced methods of suppressing the water have emerged. Double pulsed field gradient spin echo water suppression (DPFGSE) (Grzesiek & Bax, 1993) and water suppression techniques such as WATERGATE (Piotto *et al.*, 1992) can be employed in multidimensional experiments, as noted in the description of pulse sequences of previous sections. Both involve de-phasing the solvent magnetisation by applying pulse gradients to the water signal, causing the bulk magnetisation to tend towards zero.

3.5 Processing NMR spectra

Before any NMR spectrum can be inspected, the FID has to be properly processed. Each FID (intensity versus time) is Fourier transformed to give a spectrum that is a plot of intensity versus frequency. The FID can be “zero- filled” before the FT to give a better quality spectrum. This effectively adds more data points to each FID; normally the FFT (fast Fourier transformation) yields real and imaginary results for each FID, which results in two NMR spectra. They contain equivalent information, but only half the data points of the original FID are used. Zero filling, adds zeros to the data to give more data points and more efficient use of the real or imaginary data from the FID. It is also normal to apply a mathematical function to the FID, which can reduce the amount of noise and help eliminate distorted peaks that often arise at the beginning or towards the end of a FID. Examples are gaussian or sinebell functions, which change the shape of the FID to place emphasis on the 'good' parts of the FID.

After FT, each spectrum must be phased. Bloch equations tell us that the product of a FT of a FID varies between absorption and dispersion. Phase-correction is necessary to allow a phase error to be carried throughout the spectrum, resulting (ideally) in a constant baseline, instead of the fluctuating positive and negative peaks normally seen before phasing. Different spectra, however, are phased in different ways, although most spectra for VCP~2,3 were phased to produce evenly shaped, predominantly positive peaks.

References

Altieri A. S. & Byrd, R. A. (1995). *J. Magn. Reson. B.* **107** (3), 260-266

Aue, W. P., Bartholdi, E. Ernst, R.R. (1976) *J. Chem. Phys.* **64**, 2229-2246

Braunschweiler, R. R. & Ernst, R. R. (1983) *J. Magn. Reson.* **53**, 521-528

Callighan, D., West, J., Kumar, S., Schweitzer. B. I. & Logan, T. M. (1996). *J. Magn. Reson. B.* **112** (1), 82-85

Cavanagh, J., Fairbrother, W. J., Palmer III, A. G. & Skelton, N. J. (Eds) (1996). *Protein NMR spectroscopy. Principles and practice.* Academic Press Inc., San Diego, California

Ernst, R. R. (1979). *Chem.Phys.* **71**, 4546

Grzesiek, S. & Bax, A. (1993) *J. Am. Chem. Soc.* **115**, 12593-12594

Hartmann, S. R. & Hann, E. L. (1962) *Phys. Rev*, **128**, 2042-2053

Hwang, T. L. & Shaka, A. J. (1995). *J. Magn. Reson. A.* **112**, 275

Jeener, J. (1972) Lecture, Ampere Summer School, Basko Polje, Yugoslavia.

Jeener, J., Meier, B. H., Bachmann, P. & Ernst, R. R. (1979) *J. Chem. Phys.* **71**, 4546-4553

Kuboniwa , H., Grzesiek, S., Delaglip, F. & Bax, A. (1994). *J. Biomol. NMR*, **4**, 871

Mori, S., Abeyggunawardana, M.m Johnson, P. C. M. & van Zijl (1995). *J. Magn. Reson.* **108**, 94-98

Piotto, M., Saudek, V. & Sklenar, V. (1992). *J. Biomol. NMR*, **2**, 661-665

Rattle, H. (Ed) (1995). *An NMR primer for life scientists.* Partnership Press, Great Britain.

Shaka, A. J., Lee, C. J. & Pines, A. (1988). *J. Magn. Reson.* **77**, 274

Sklenar, V., Piotto, M., Leppik, R.& Saudek, V. (1993) *J. Magn. Reson.* **102**, 241-245

Vuister, G. W. & Bax, A. (1993) *J. Am. Chem. Soc.* **115**, 7772-7777

Chapter 4 - Assignment of VCP~2,3

4.1 Introduction

Despite constant technical advances, solving the 3D structure of a protein by NMR spectroscopy is still not a straightforward process. It involves the successful completion of a series of steps.

Expression and purification the protein may prove to be a time-consuming task that provides the first hurdle. Fortunately, the creation of more efficient systems to aid protein expression and purification has been the goal of many academic research groups and biotechnology companies. The result is an increase in the chances of successfully expressing and isolating suitable quantities of the protein of interest in a reasonable timescale. For VCP~2,3, the *Pichia pastoris* expression system proved to be an extremely efficient system for expressing large quantities of the protein in a convenient period of time (see Chapter 2). Purificaton also proved to be routine once suitable protocols had been established.

When studying a protein by NMR, its concentration in solution within the NMR tube is an important factor as this will affect the length of time for which a spectrum needs to be acquired in order to achieve a suitable signal-to-noise ratio. The protein must also be stable during the course of the experiment and this could be influenced by microbial contamination of the sample, air oxidation and other chemical modifications of the protein, temperature, pH and hydrolytic breakdown by proteases.

Once the protein has been obtained in milligram quantities, a 1D NMR spectrum may be acquired that provides information about the behaviour and characteristics of the protein sample in solution. A 1D NMR spectrum indicates, for example, if there are likely to be any complications such as dimerisation (or formation of higher order aggregates), unfolding or precipitation. A suitable sample will usually result in a spectrum that has a good signal-to-noise ratio and well-dispersed, sharp resonances; these indicate the protein to be in a folded state. Providing the sample is stable and does not oligomerise, further, complex spectra can be acquired that contain information necessary to assign each amino acid in the protein, a pre-requisite for any structure calculation.

4.2 The NMR spectra of VCP~2,3

4.2.1 1D NMR spectrum of VCP~2,3

A typical 1D NMR spectrum of ~2.5 mM VCP~2,3 dissolved in 550 μl of 10 mM sodium phosphate buffer, pH 6.0, with 50 μl $^{2}H_{2}O$, is shown in figure 4 i. It can be observed that the peaks are relatively sharp and appear in all regions of the spectrum in which signals would be expected from a folded protein. This indicates that the conditions used for the data collection were consistent with the structural integrity of the protein. The small peaks at the left hand-side of the spectrum represent the tryptophan side-chain NH resonances, one of which is present in each of the two modules of VCP~2,3. Those peaks that have negative chemical shifts are particularly strong indicators of a folded protein, since they are shifted away from any of the so-called 'random coil' shifts expected from an unfolded protein. The narrow line-widths of peaks in the1D spectrum of VCP~2,3, and the persistence of signals over many hours in the spectrometer indicated that the protein was not oligomerising and

Figure 4 i – 1D NMR spectrum of VCP~2,3

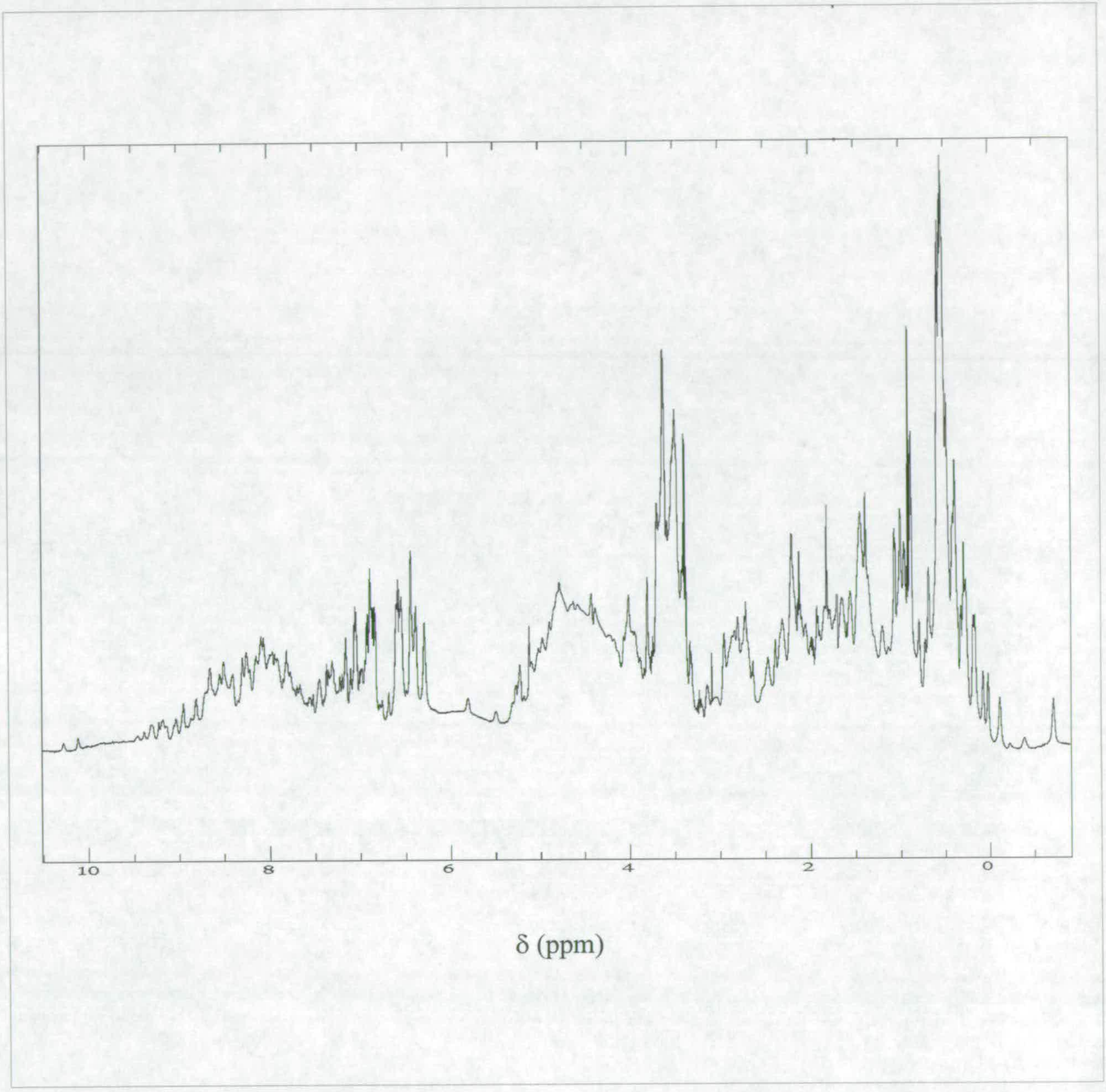

Figure 4 i - 1D NMR spectrum of VCP~2,3. The spectrum, collected over 128 scans, shows a good signal-to-noise ratio. The two peaks around 10 ppm represent tryptophan side chain resonances. The –vely shifted peaks are strongly indicative of a compactly folded protein.

remained stable over time. The signal-to-noise ratio was consistent with the estimated concentration (Chapter 2, Equation 1) of the protein.

4.2.2 2D ^{1}H NOESY and TOCSY

A set of 2D NMR experiments were acquired using ~2.5 mM VCP~2,3 and these yielded spectra that were deemed to be of sufficient quality for assignment and structural analysis, without changing any of the experimental conditions (pH 6.0, 10 mM sodium phospate, 310 K). The spectra were considered to be adequate for assignment because the cross peaks were abundant and well-dispersed in each of the spectra. NOESY and TOCSY spectra were collected with different mixing times. It was found that the longer mixing time for the NOESY of 150 ms compared with 100 ms, and shorter mixing time for the TOCSY of 39 ms compared with 63 ms, gave spectra that were more suitable for assignment purposes. These longer mixing times resulted in spectra that had stronger, and an increased number of resonances. Portions of the 2D NOESY and TOCSY spectra, along with the experimental details, are shown in figure 4 ii. A 2D COSY spectrum was also acquired.

The water signal was suppressed using a DPFGSE (see Chapter 3). Despite this, water gave such a strong signal that it was impossible to suppress completely. This is a problem when Hα resonances lie close to or under the water signal. Such proved to be the case for the 2D spectra of VCP~2,3 acquired in ^1H_2O, where many signals were lost due to such difficulties.

A 1D analysis of the VCP~2,3 sample after it had been used to collect the 2D data, revealed that no new or unusual signals were present. This showed that the protein had not undergone degradation or unfolding during the acquisition, and

Figure 4 ii – 2D NOESY and 2D TOCSY NMR spectra

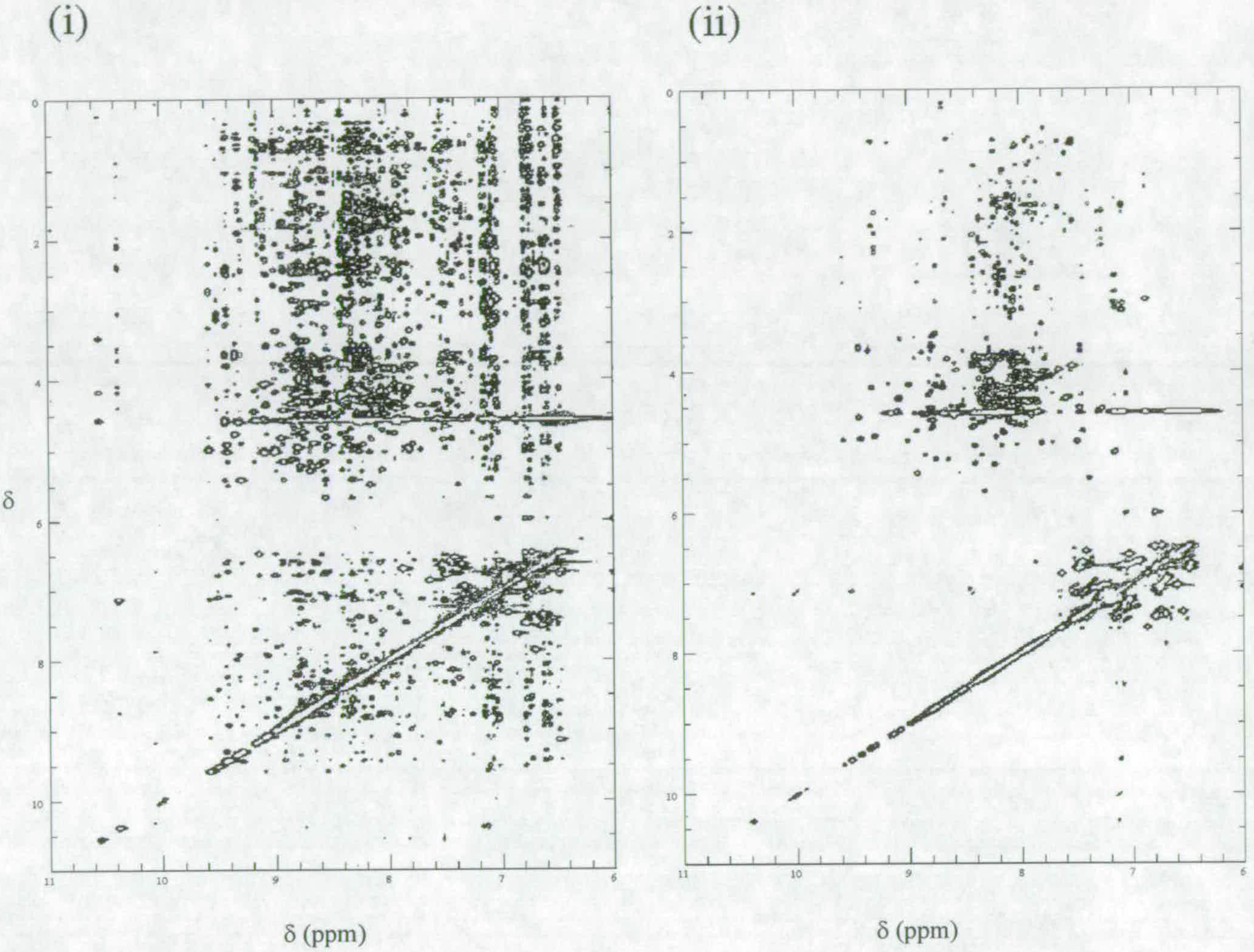

Figure 4 ii - (i) 2D ^{1}H NOESY of VCP~2,3 in 550 μl 10 mM sodium phosphate, pH 6.0 + 50 μl ^2H_2O. (ii) 2D TOCSY (conditions the same as in (i)). Spectral widths were 12001.2 Hz in F_1 and 8000 Hz in F_2. 128 transients were accumulated in each of the 512 increments. The acquired matrix was 2048 x 512 complex data points.

confirmed that there was no need to adjust any of the acquisition conditions such as buffer, pH or temperature.

4.2.3 2D NOESY, TOCSY and COSY experiments, using ^2H_2O as solvent

Acquiring spectra with ^2H_2O as the solvent is useful since ^2H_2O gives no NMR-detectable signal in a ^1H spectrum. In the case of VCP~2,3, many signals that had not been visible in the Hα-Hα region of the 2D ^1H_2O NOESY spectra (e.g those close to the water resonance), appeared in the ^2H_2O spectra. The ^2H_2O NOESY spectrum is also useful as a way of determining those cross-peaks originating from exchangable protons such as backbone or sidechain NHs. These only appear in the 2D NOESY where ^1H_2O is the solvent. The ^2H_2O NOESY, TOCSY and COSY spectra and acquisition parameters are shown in figures 4 iii. The ^2H_2O spectrum was also used to assign the cross-peaks originating from resonances from aromatic ring protons, as the aromatic region of the spectrum was less crowded (due to fewer NH signals in the 6-8 ppm region). This is evident in figures 4 iv, which displays a comparison of these regions from the ^1H_2O and ^2H_2O spectra.

4.2.4 3D spectra of VCP~2,3

The 2D NMR spectra of VCP~2,3 confirmed that the protein behaves well in solution, and would therefore be suitable for 3D NMR spectra acquisition. This was a necessity since there were many similar backbone NH chemical shifts, which could not be resolved using the 2D data alone. In larger proteins, both ^{15}N and ^{13}C data are

Figure 4.iii - NOESY, TOCSY and COSY spectra recorded with ^2H_2O as solvent

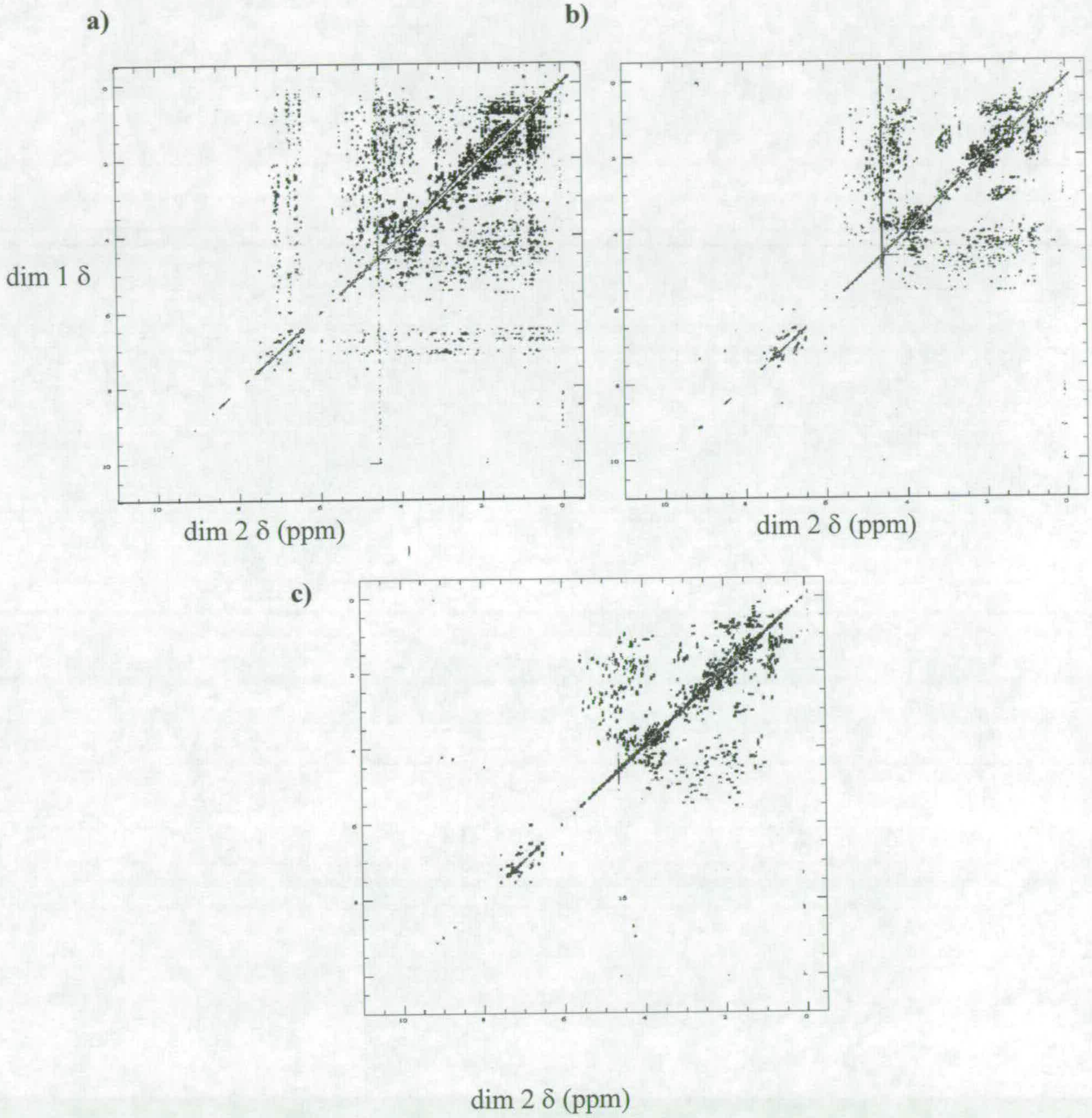

Figure 4 iii - a), b) and c) : 2D NOESY, TOCSY and COSY spectra respectively acquired with 99.99% ^2H_2O as solvent. All three spectra were recorded in 600 μl 10 mM sodium phosphate buffer, pH 6.0 in ^2H_2O. Spectral widths for NOESY, TOCSY and COSY spectra were 10000 Hz in F_1 and 6600 Hz in F_2. The spectra were processed to give matrices of 2048 x 2048 complex data points.

Figure 4.iv - Aromatic regions of NOESY spectra in H_2O and ^2H_2O

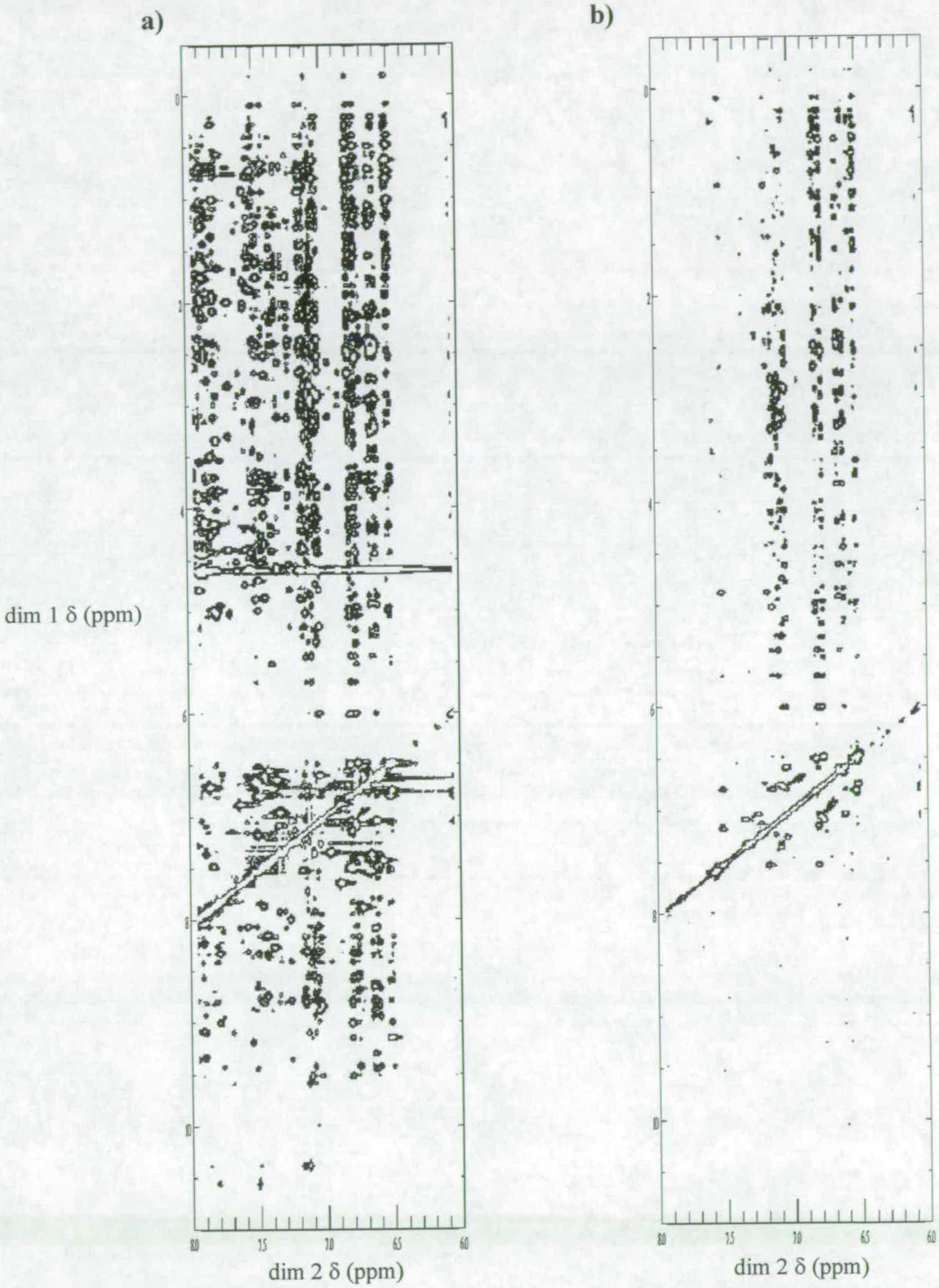

Figure 4 iv - a) and b) show the aromatic regions of ^1H_2O NOESY and ^2H_2O NOESY spectra. Because of the decrease in the number of signals in this area in the ^2H_2O NOESY, the cross-peaks arising from the aromatic resonances are easier to see and assign.

normally obtained, but this was unnecessary in the case of VCP~2,3 as ^{15}N-edited data were deemed to be sufficient to complete resonance assignment. Thus an isotopically labelled ^{15}N-^{1}H VCP~2,3 was produced after the 2D spectra had been obtained, processed and assessed for quality. This allowed an assessment of suitable conditions using the unlabelled protein that was not expensive to produce.

The 3D ^{1}H-^{15}N HSQC-NOESY and 3D ^{1}H-^{15}N HSQC-TOCSY spectra were used to initiate the backbone assignment of VCP~2,3. In these spectra, any NH proton that does not exchange with water on an immediate time-scale should give a set of cross-peaks, corresponding to the ^{15}N and ^{1}H chemical shifts of the amide. In this way, the information that is contained within the fingerprint region of the 2D ^{1}H NOESY and TOCSY spectra is resolved on the basis of the chemical shift of the attached ^{15}N (figure 4 v). The two ^{1}H, ^{1}H projections of the HSQC-NOESY and HSQC-TOCSY spectra are shown in figure 4 vi. Also shown is the 2D ^{1}H-^{15}N HSQC of VCP~2,3.

4.2.5 Other acquired spectra

In addition to the spectra used for the backbone and side-chain assignments, described above, other useful spectra were acquired. Several 2D ^{1}H-^{15}N HSQC experiments were performed, each at a different temperature, to obtain an indication of those residues whose chemical shifts are most susceptible to peturbation upon temperature-induced denaturation of the protein (figure 4 vii). This gave an insight into the relative thermal stability of the two protein modules. HSQC were also collected at increasing concentration of guanidium chloride. These data were used in combination with other biophysical studies (performed by Dr. Marina Kirkitadze in this laboratory) to characterise the intermodular interface in VCP~2,3, as discussed in Chapter 6. Finally, a series of HSQC experiments were also acquired with ^2H_2O as the solvent; the freeze-dried protein was dissolved in 99.99% ^2H_2O and HSQC spectra acquired after 5, 13, 30, 60 and 120 mins (figure 4 viii). Only those NH protons that were in the most stable chemical environments remained detectable as time progressed, since the remainder exchanged with the ^2H_2O. This provided a basis for inferring H-bonded NH protons, which is very useful information in calculating the secondary structure within the protein. The fully assigned ^{1}H-^{15}N HSQC spectrum of VCP~2,3 is shown in figure 4 ix. The last spectrum collected was an HNHA spectrum, which was used to determine values of $^3J_{hnh\alpha,}$ i.e. the three bond scalar coupling constants between the HN and the HA protons. $J_{hnh\alpha}$ is related to the ϕ angle by the Karplus equation and hence may be used to infer secondary structure. The value of $^3J_{hnh\alpha}$ (or of ϕ) may also be used directly in the structure calculation (see below).

Figure 4 v - Extracting slices from a ^{1}H-^{15}N HSQC TOCSY or NOESY NMR spectrum based on ^{15}N chemical shift

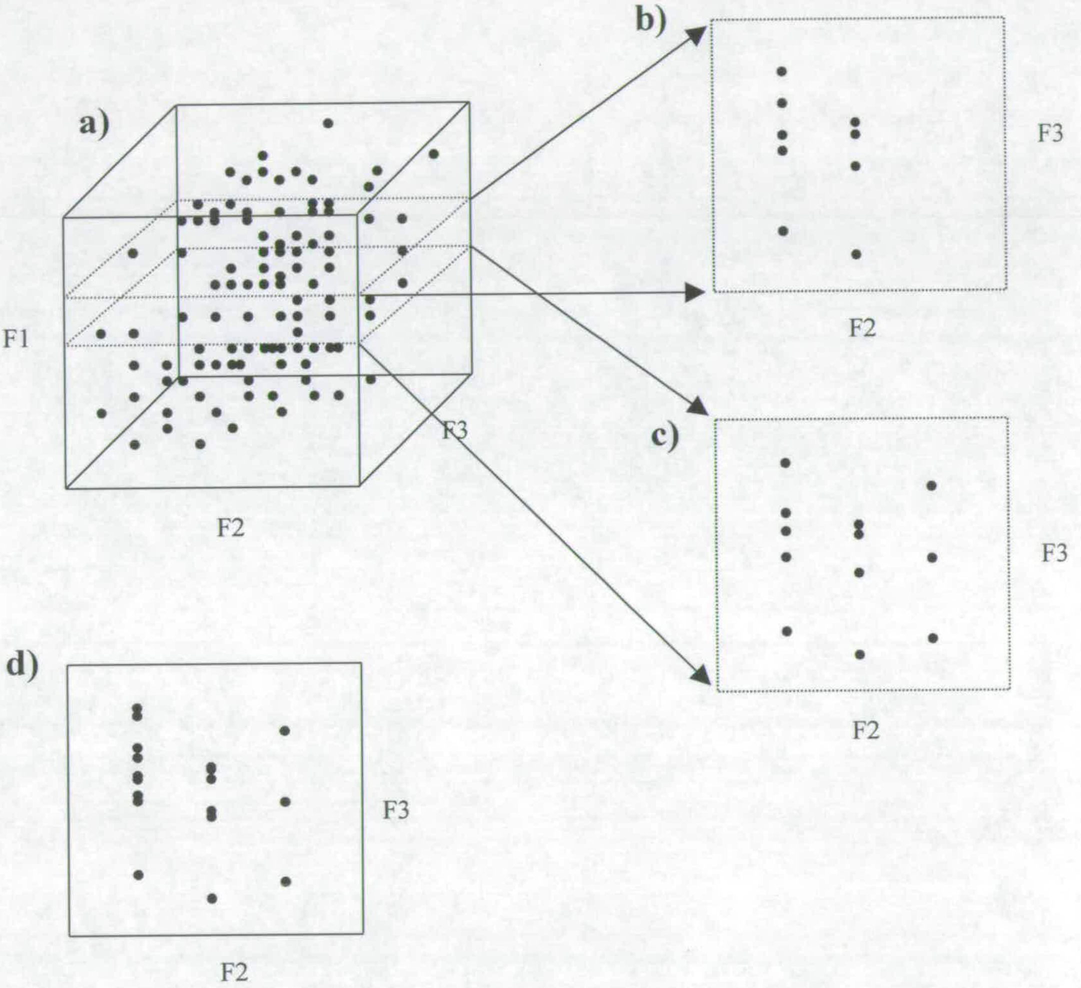

Figure 4 v - (a) The ^{15}N dimension (F1) gives each amide resonance a third co-ordinate in a 3D matrix (F1, ^{15}N; F2, ^{1}H dim2, F3 ^{1}H dim 1). (b) and (c) illustrate how slices can be extracted from this matrix based on ^{15}N chemical shift. (d) shows how the resolution is lost if the resonance patterns in (b) and (c) cannot be separated by ^{15}N shift (e.g as would occur in a ^{1}H-^{1}H 2D NOESY spectrum)

Figure 4 vi - ^{15}N-edited NMR spectra

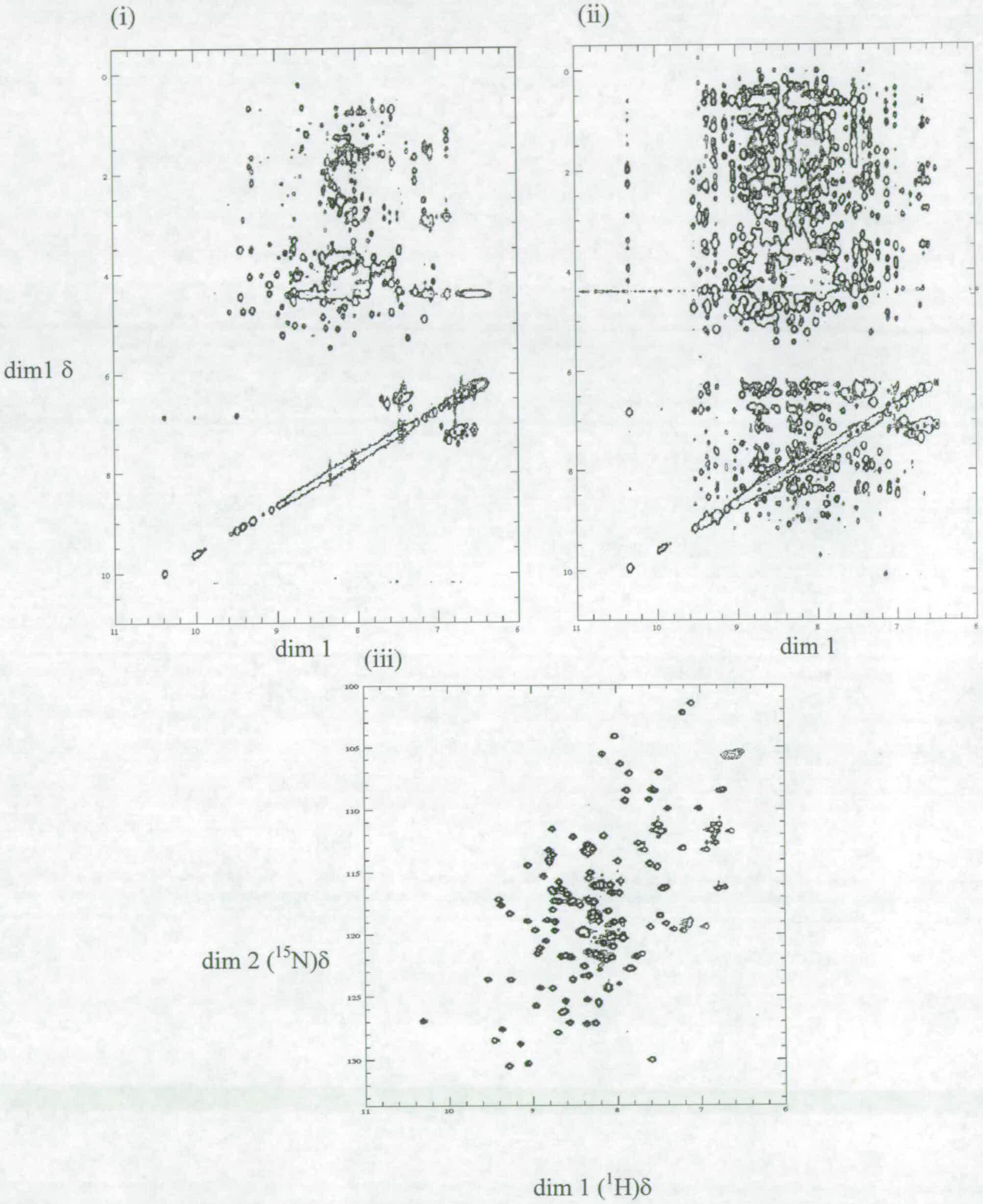

Figure 4 vi - (i) and (ii) show the ^{1}H, ^{1}H projections of the 3D ^{1}H, ^{15}N HSQC - TOCSY and NOESY spectra respectively. (iii) shows the ^{1}H, ^{15}N HSQC that is equivalent to the ^{1}H, ^{15}N projection of the 3D spectrum.

Figure 4 vii - ^{1}H-^{15}N HSQC spectra of VCP~2,3 acquired at different temperatures or in presence of GdmCl

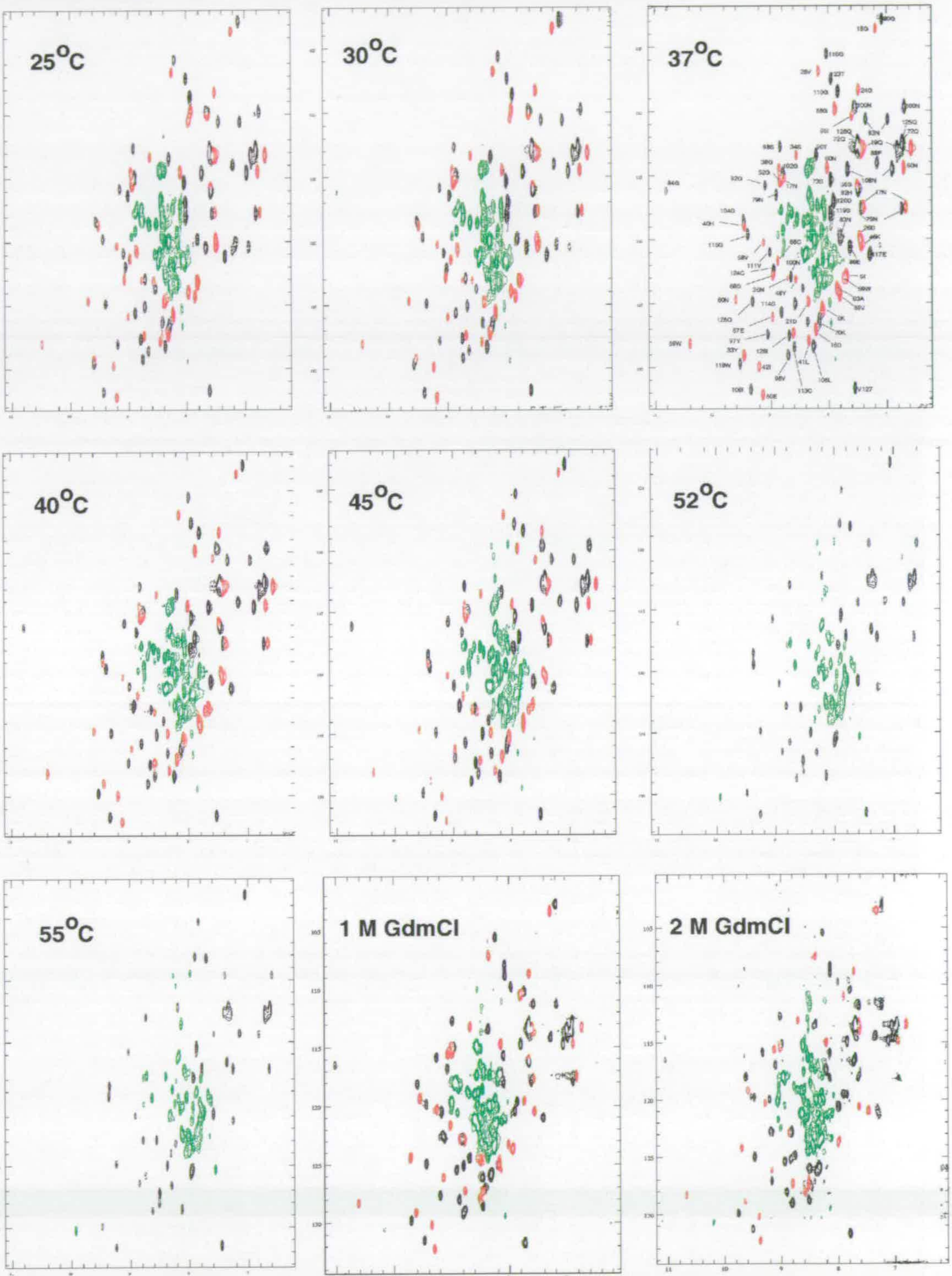

Figure 4 vii - HSQC spectra of VCP~2,3 obtained at different temperatures, or in the presence of 1 or 2 M guanidium chloride. Those HSQC peaks that could be reliably monitored at the different temperatures are shown with assignments at 37 °C (red, VCP~2, black, VCP~3, green, not assigned).

Figure 4 viii - ^2H_2O HSQC spectra of VCP~2,3

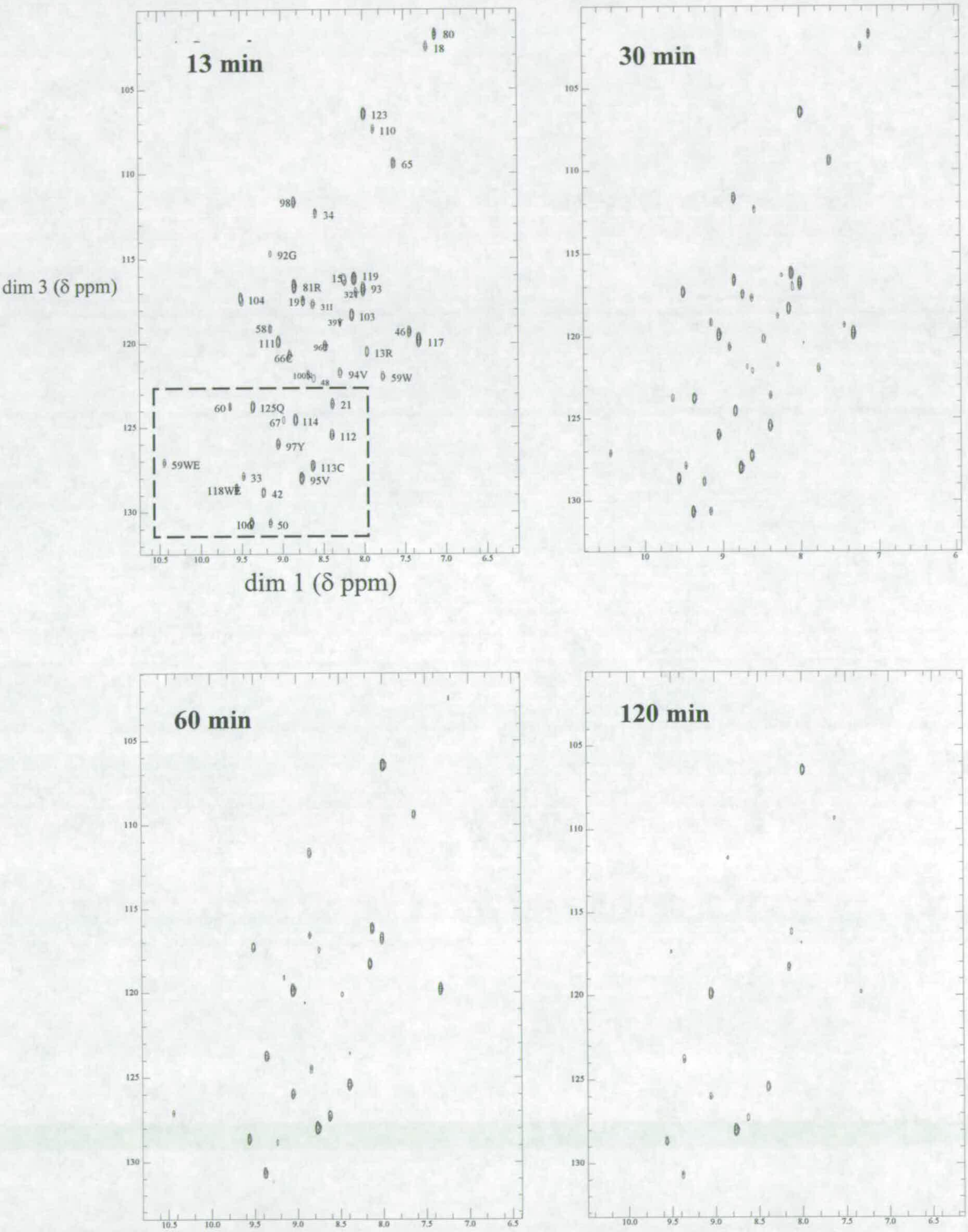

Figure 4 viii -HSQC spectra of VCP~2,3 recorded after incubating in 99.8% ^2H_2O for the time periods indicated. The amide resonances that prevail are in more stable environments since they do not exhange rapidly with the ^2H_2O. This can be used to infer locations of H-bonds.

Figure 4 ix - Assigned ^{1}H-^{15}N HSQC spectrum of VCP~2,3

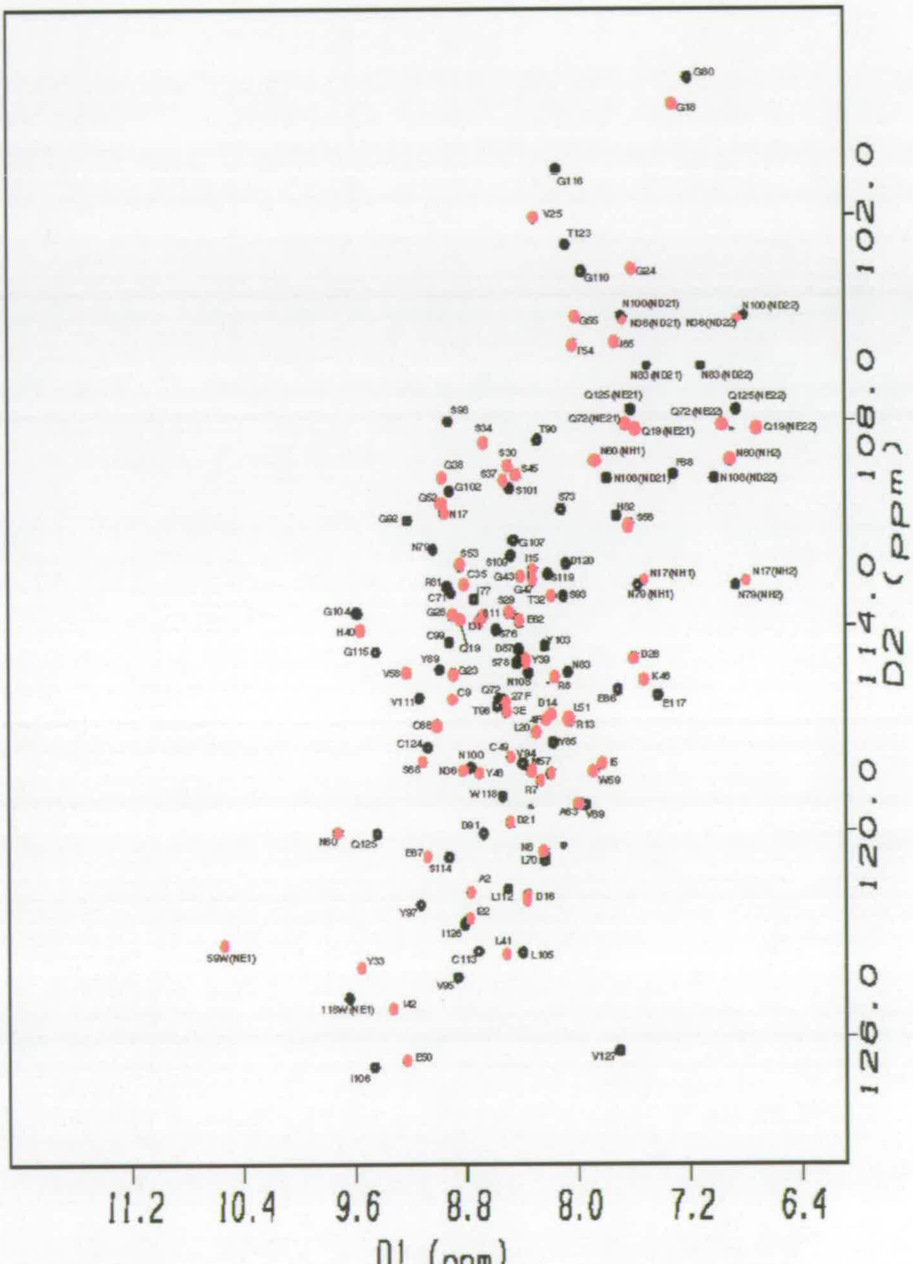

Figure 4 ix - Assignments for module 2 (pink) and module 3 (black) are shown for the ^{1}H-^{15}N HSQC spectrum of VCP~2,3.

4.3 Full proton assignment of VCP~2,3

4.3.1 Using ANSIG for backbone assignment

To facilitate resonance assignment and subsequent progress, the NMR spectra were loaded into a software package called 'ANSIG' (Kraulis, 1989, Kraulis *et al*, 1994). ANSIG incorporates a range of associated macros that were used to automate some of the more laborious peak-picking tasks associated with NMR assignment. Initially, a macro that allows automatic peak-picking was used. After defining a maximum line-width for each dimension (0.03, 0.03 and 0.3 ppm for the directly acquired ^{1}H, indirectly acquired ^{1}H and ^{15}N dimensions, respectively), a separate macro was used to integrate each NOESY or TOCSY peak, and a cross was displayed on the screen representing the centre and most intense region of the NOE or TOCSY connectivity. The program was then utilised in a user-interactive manner to allow decisions to be made about each peak, i.e. whether a peak is legitimate and should be accepted and allowed to keep its associated cross, or should be "deleted". In this way it was possible to very quickly place a cross in the centre of each legitimate peak, resulting in a series of “cross-peaks”. After each cross-peak had been picked and vetted, a series of macros were utilised to aid sequential assignment, as discussed below. These proved to be an invaluable tool.

4.3.2 Starting the assignment

The ^{1}H-^{15}N HSQC TOCSY and ^{1}H-^{15}N HSQC NOESY spectra, with their enhanced dispersion, were a good starting point for assigning residues in VCP~2,3. As originally noted by Wuthrich (1986) for a hypothetical fragment of protein, shown in figure 4 x, there are certain pairs of protons that should be near enough to each other for an NOE to occur between them. Examples of these connectivities in VCP~2,3 are shown in figure 4 xi. The Hα at residue i, should always be close enough in space to the amide proton of i+1 to produce a detectable NOE . Glycines normally have two Hα proton resonances in the range of ~3-5 ppm and (unless they are degenerate) these yield a pair of Hα-NH cross-peaks that are similar in shape and appearance in the HSQC-TOCSY spectrum, making the glycine residues a good starting point for assignment. In practice, the HSQC-NOESY peak corresponding to a Hα–NH(Gly) connectivity was selected and an ANSIG macro called 'Close in ^{1}H' macro was used to find other HSQC-NOESY cross-peaks that matched the ^{1}H chemical shift of one of the two Hα(Gly) protons. Amino acids coming after glycines in the sequence were relatively easy to identify since the second Hα cross-peak should also have a corresponding peak in the ^{1}H-^{15}N 'strip' of the following residue ('strip' refers to all the resonances from any given HN proton). The normal procedure involved establishing a few potential candidates for the sequential assignment, and then trying to verify one of these by further investigation of sequential residues. Additionally, it was possible to use other sequential NOEs, e.g. NH_i - $NH_{(i+1)}$. The main problems arose from sequential amino acids whose Hα or HN protons had the same chemical shift. When amino acids had similar HN chemical shifts, differences in ^{15}N chemical shifts were used to discriminate between resonances from the overlapping HN protons. Also, several good candidates often arose because VCP~2,3 is a β-sheet protein, and cross-sheet Hα-HN_i NOEs are common.

Figure 4 x - NOESY resonance patterns

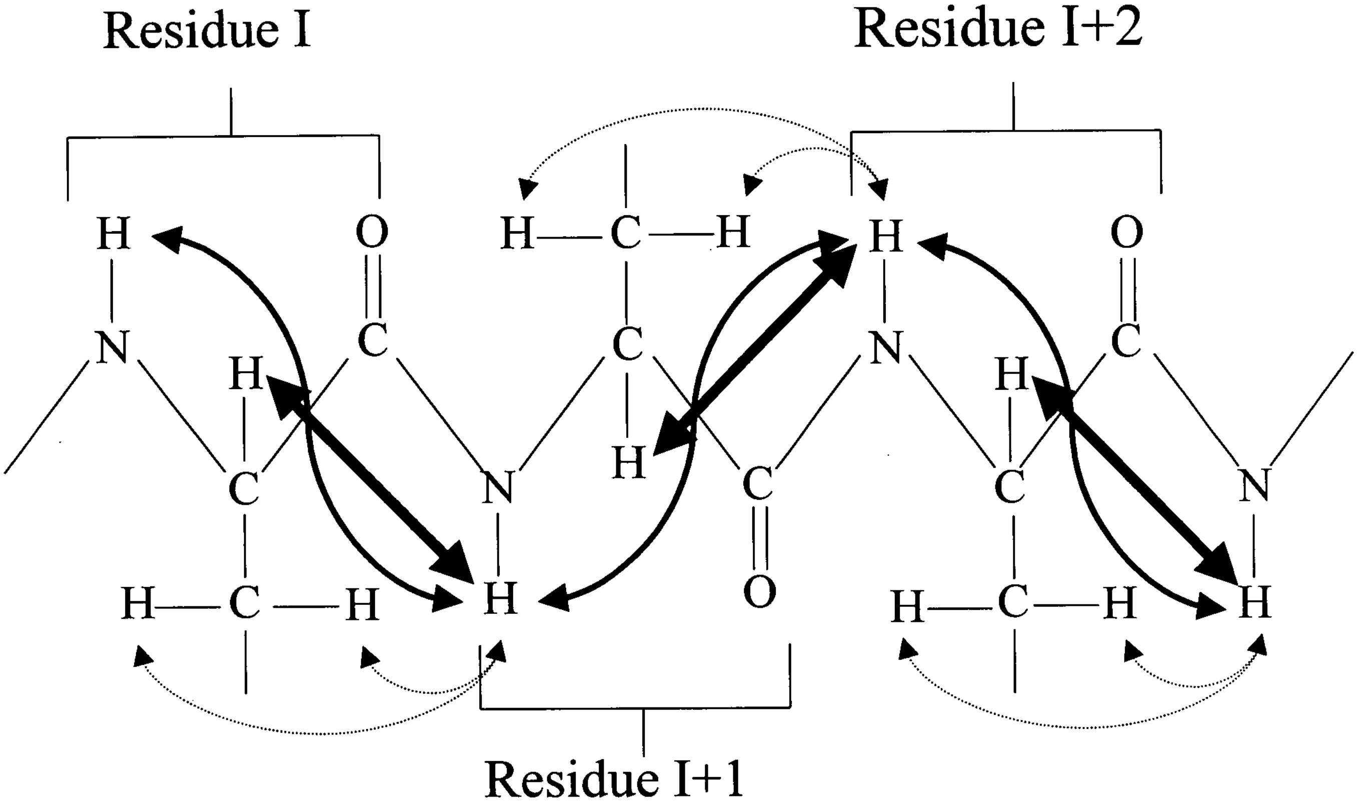

Figure 4 x - Resonance connectivities normally visible in a NOESY spectrum. Strong arrow, $H\alpha_i$-HN_{i+1}; Medium arrow, HN_i-HN_{i+1}; Dotted arrow, βH_i-HN_{i+1}.

Figure 4 xi - ^{15}N - ^{1}H HSQC NOESY strips

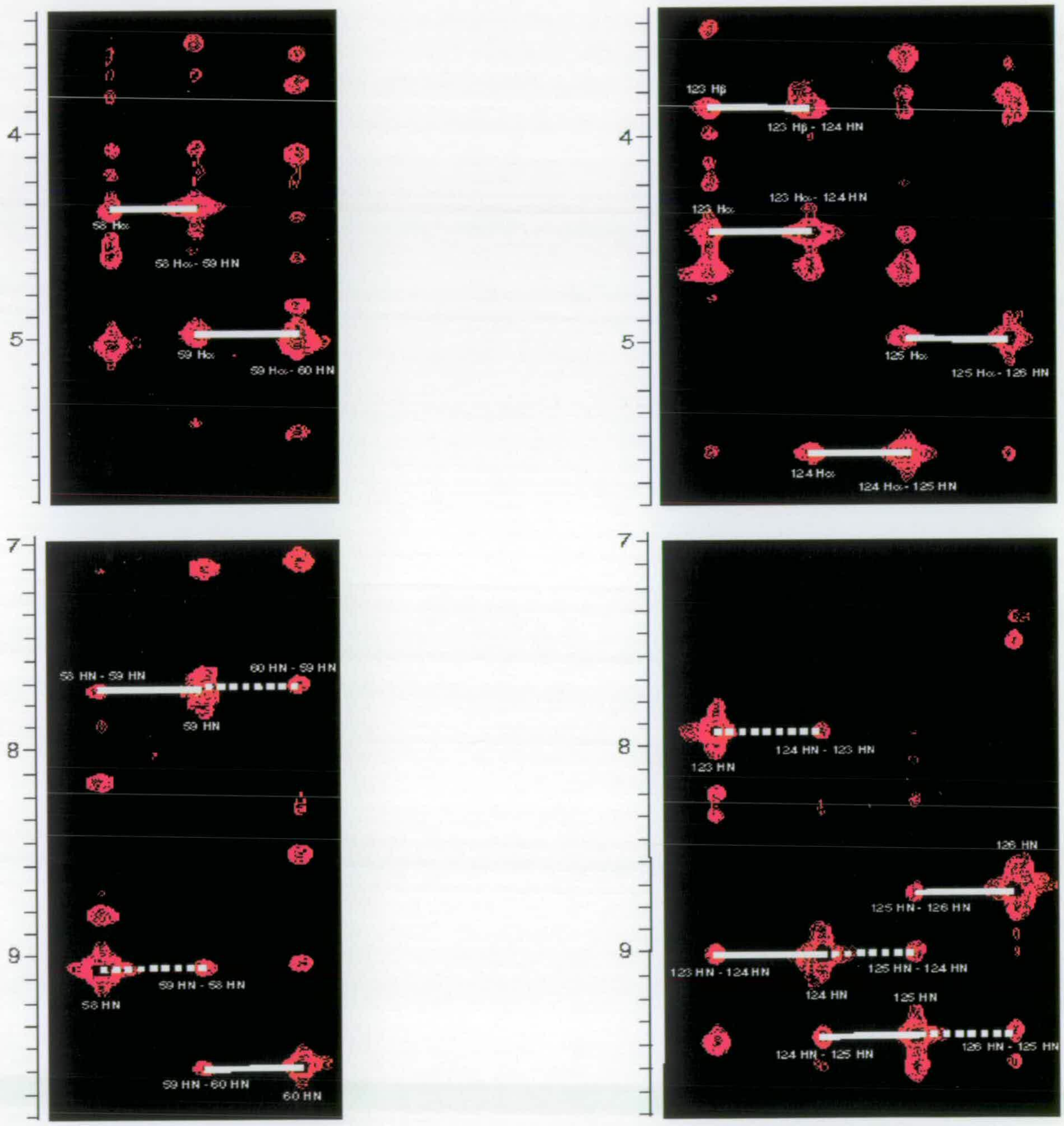

Figure 4 xi - Some of the connectivities between adjacent residues are shown. These are strips extracted from the ^{1}H-^{15}N HSQC-NOESY spectrum.

Another problem with the backbone assignment occurred when the Hα chemical shift lay directly beneath, or very close to, the water signal. This problem was overcome by reference to the $^{2}H_2O$ spectra and their use in combination with the 3D spectra. The key information that could be extracted from each spectrum (three-bond connectivity in COSY, intra-residue connectivities in TOCSY, inter- and intra-residue connectivities in NOESY) was utilised fully to collect enough evidence to deduce the assignment of backbone protons that were overlapped. Assignments were made based on two key criteria: 1) the ^{1}H-^{15}N strip of the residue must contain resonance patterns consistent with the side-chain of that particular amino acid; 2) the residue must have good sequential connections to other amino acids that are consistent with the protein sequence. The exceptions to this assignment strategy were the proline residues, which contain no HN protons and therefore do not have a ^{1}H-^{15}N strip. The assignment of the proline residues is discussed later. While this mainchain-directed approach furnished the majority of the assignments, knowledge of other CCP module structures and the developing picture of the VCP~2,3 secondary structure also proved invaluable – this is discussed below.

4.3.3 Side-chain assignment

Once nearly all of the backbone Hα and HN protons had been assigned, the goal was to assign the rest of the protons in each amino acid side-chain and hence identify a distinct chemical shift for each proton in the protein. The assignment of the β, γ and δ protons of amino acid residues required further utilisation of the 2D COSY, TOCSY and NOESY spectra. A good example of how these techniques were employed to fully assign a side-chain was the case of tryptophan assignment. The full assignment of tryptophan 118 is shown in figure 4 xii. It was found that it was important when making side-chain assignments to be intuitive, looking at peak shape/size and at other connectivities to verify previous inferences.

In proteins, strong NOEs between NH_i - NH_{i+1} or NH_{i+2} indicate potential turns. Strong $H\alpha_i$ - $H\alpha_{i+>4}$, $H\alpha_i$ - $HN_{i+>4}$ or HN_i - $HN_{i+>4}$ NOEs are indicative of residues involved in β-strands, since these NOEs originate from 'cross-sheet' connectivities. As assignment of VCP~2,3 progressed, a picture of the secondary structure begun to emerge and this provided further confidence in the procedure as described below.

4.3.4 Assignment of the proline residues

Prolines proved to be the most difficult residues to assign in VCP~2,3, as they possess no backbone amide proton. Because of this, the prolines do not have HSQC peaks or associated strips. To make matters more complicated, VCP~2,3 contains three sequential pairs of prolines and two further prolines. Assigning prolines therefore required different tactics, as the Hα-NH connections and NH-NH connectivities normally used, do not apply. Figure 4 xiii shows the different connectivities which should be detectable for a proline in *cis* or *trans* arrangement. Identification of the Hα proton of some of the prolines was not too difficult providing the proline preceeded a non-proline residue. This is because a strong $H\alpha_{(Pro)}$-$NH_{(Pro+1)}$ NOE was normally visible in the HSQC-NOESY strip of the Pro+1 residue. Identification of the proline Hδ protons was possible by looking for reasonably strong $H\alpha(_{Pro-1})$ - $H\delta(_{Pro})$ peaks in the aliphatic region of the 2D NOESY spectrum. This was

Figure 4 xii - Tryptophan assignment

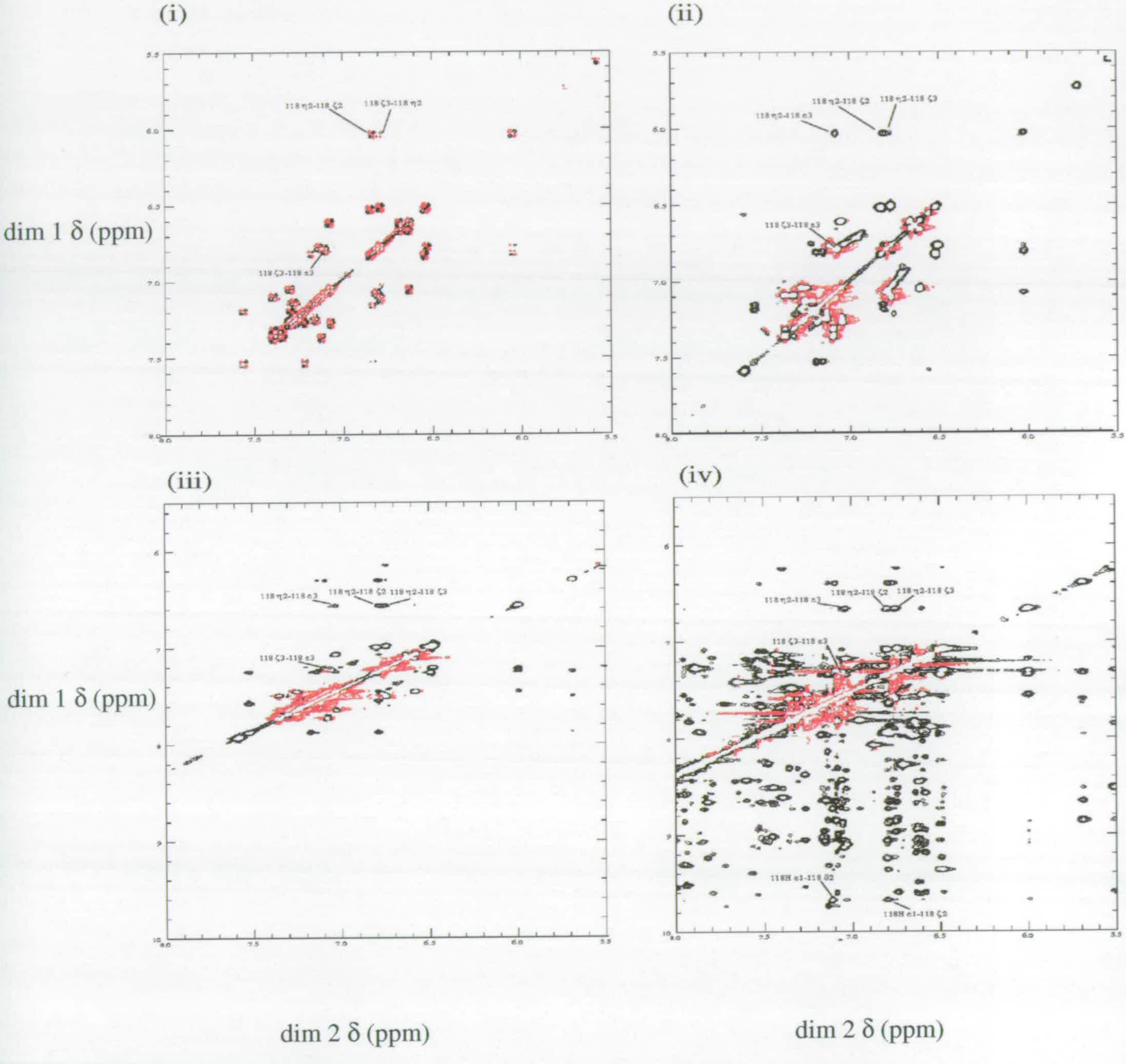

Figure 4 xii - (i) connectivities through three bonds for Trp 118 in ^2H_2O COSY. (ii) ^2H_2O TOCSY has the same connectivities as COSY plus additional peaks for connectivities > 3 bonds. (iii & iv) - ^2H_2O NOESY and ^1H_2O NOESY. Resonances from the Hε1 side chain amide are not visible in the ^2H_2O NOESY, but are present in the ^1H_2O NOESY allowing the aromatic side-chain resonances to be completed.

particularly difficult for CP modules of VCP~2,3, as the conserved prolines in the respective modules are in very similar chemical environments and overlap is present.

Assignment of the prolines was aided using an indirect approach that utilised the inferred secondary structure of the two modules. Many cross-sheet NOEs between assigned residues had been identified by this point, and NOEs that implied turns or loops between them had also been identified. In this way, an approximate picture of the secondary structure was deduced. Furthermore, examples of CP modules studied previously were examined, the consensus residues identified and their position within the crude secondary structure of VCP~2,3 was noted. In this way knowledge of residues that were likely to be in close proximity to the prolines was employed in an attempt to identify NOEs originating from the proline protons of VCP~2,3. An example is the second proline in VCP~2. In all CP module structures examined previously, the Hγ protons of this proline are in close enough proximity to the HN of the first consensus cysteine to produce an NOE. Furthermore, the Hζ protons of the conserved tryptophan are also likely to be close to the Hγ protons of the Pro. Inspection of NOEs originating from these Trp peaks revealed four cross-peaks which could correspond to the Hγ1/Hγ2(Pro) - Hζ(Trp) and Hγ1/Hγ2(Pro) - HN(Cys) connectivities. Inspection of the 2D TOCSY and ^2H_2O COSY spectra revealed that the two potential Hγ peaks were connected through-bonds and could be linked to other resonances that fell in the pattern of a proline side-chain. Using the conserved features of the CP modules as a reference, the other prolines were assigned in a similar manner.

Figure xiii - characteristic resonances indicative of *cis* or *trans* arrangement of proline residues

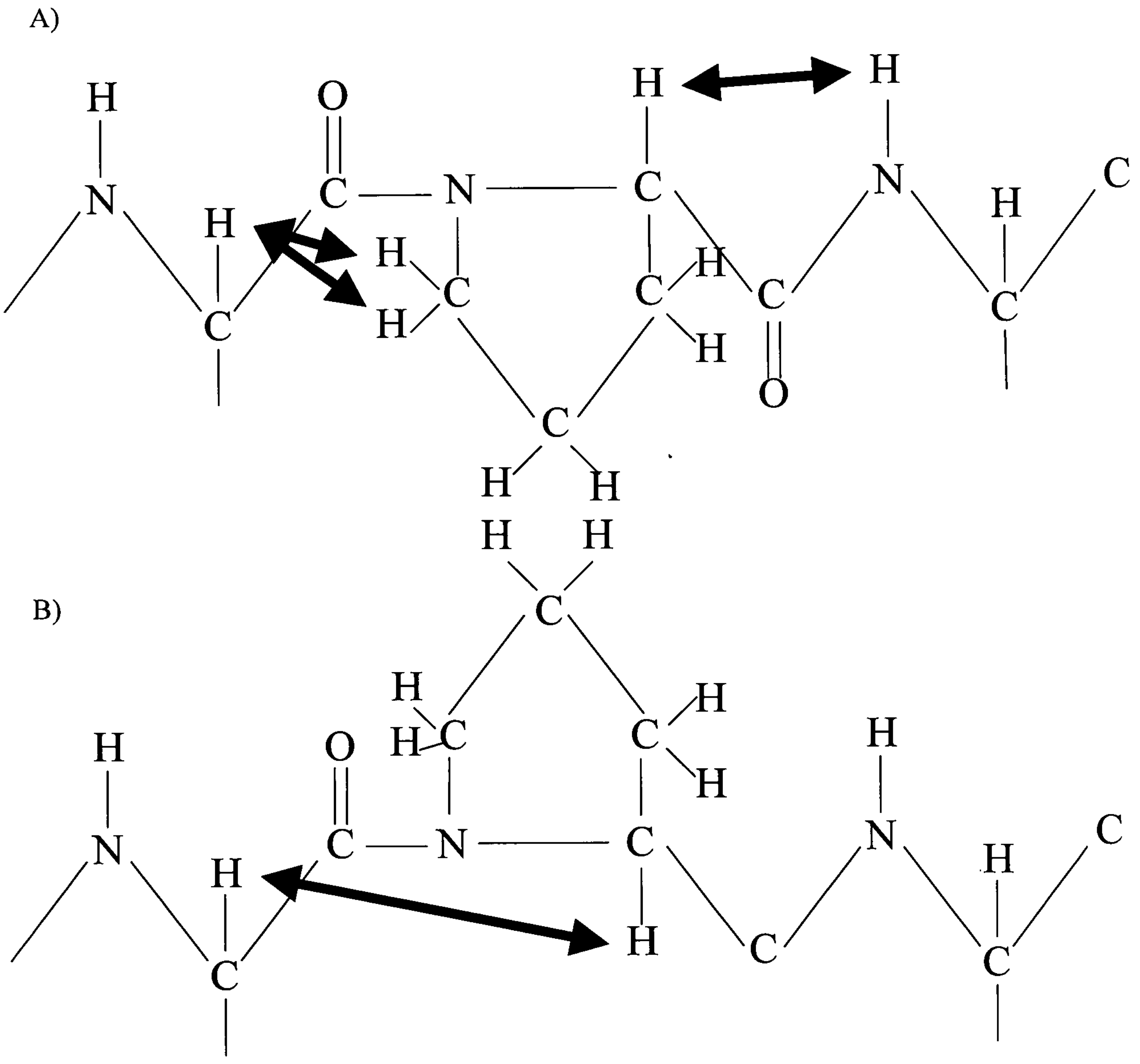

Figure 4 xiii - A) shows the resonance visible when proline is in the *trans* conformation ($H\alpha_{PRO}$-HN_{PRO+1}, $H\alpha_{PRO-1}$-$H\delta_{PRO}$). When proline is in *cis* conformation (B), a medium strength $H\alpha_{PRO-1}$-$H\alpha_{PRO}$ NOE is observed.

4.3.5 Assignment techniques particular to CP modules

Using the consensus structure of the CP modules proved to be a critical tool for assigning the prolines. The complete assignment of the remainder of the VCP~2,3 protons was also aided by the information available from other CP module structures. Because the CP module has a large number of well conserved, or invariant, residues their chemical environments are similar and correct prediction of the presence of NOEs is possible in many cases. Because of their easily recognisable side-chain NH resonance, the tryptophan residue of each module was easily assigned. On top of this, one of the invariant, disulphide bonded cysteines had an Hα proton directly across the β-sheet from the Hα proton of the conserved tryptophan in all CP modules structures examined. On that basis, using the 2D NOESY spectrum, it was possible to look for any Hα-Hα NOEs from the tryptophan 118 Hα proton. One very strong possibility was found and could be connected, using the 2D COSY to two β-protons, as would be expected for a cysteine. The amino acid was then confirmed as the cysteine (C113) by sequential assignment, or by considering other hypothetical cross-sheet NOEs, since the cysteine is always found in a β-strand. This proved to be an additional and valuable method for assigning the backbone and side-chain protons. It was very important, however, when applying this method to keep an open mind and verify the residues by the classical main-chain directed route described earlier.

4.4 Special features of the VCP~2,3 resonance assignment

Ultimately, most of the backbone and side-chain protons and ^{15}N were assigned for VCP~2,3 although many of the protons were determined to be of a degenerate nature (see Appendix 1 for the chemical shift table). Certain unusual features, besides the atypical chemical shift of the proline Hβ protons, were noted. These included the first disulphide-bonded cysteine (C9, C71) in both modules having unusual Hβ shifts, probably a result of the close proximity of the tryptophan side chain. This was also the case for one of the Hβ protons of the first proline of module 3. It was also impossible to resolve all the side-chain arginine protons using the available spectra. Tryptophan 118, was shown to have two distinct ring HN shifts in the 2D spectrum consistent with two very slowly-exchanging conformers. The cross-peaks arising from these different amides were to protons from the same residues, but showed different intensities. One of the side-chain amides, though, had a stronger set of cross-peaks, indicating that this was probably the major form.

References

Kraulis, P. J. (1989). *J. Magn. Reson. ser A.* **84**, 627-633

Kraulis, P. J. (1994).*J. Mol. Biol.* **243**, 696-718

Wuthrich, K. (1986). *NMR of Proteins and Nucleic acids*, John Wiley, New York

Chapter 5 - Structure calculation using X-PLOR and ARIA

5.1 Data required for structure calculation

A combination of NMR spectra are used to manually assign chemical shifts to each of the ^1H atoms in the protein (see Chapter 4). Each cross-peak in the NOESY spectra can then be assigned as an NOE between two protons, or two sets of protons with identical chemical shifts. The various assignment possibilities for each of the NOE cross-peaks are used as data input for the structure calculation itself. The intensity of the NOE is related to the distance between the protons involved, and NOE's are normally categorised as weak (reflecting an interproton distance of < 5 Å), medium (< 3.3 Å) or strong (< 2.7 Å). Other useful empirical information, such as the location of disulphide and hydrogen bonds or dihedral angle restraints, may also be used. In addition knowledge of the covalent chemistry of the molecule is incorporated into the calculation.

Taken collectively, the data provides a network of restraints which describe the structure of the protein.

5.2 Finalising the chemical shift

In calculating the structure of VCP~2,3, the program ARIA (Nilges *et al*, 1997) (a set of scripts written within the program X-PLOR (Brunger, 1992)) was used as there were many ambiguous cross-peaks – i.e. cross-peaks that could not be assigned to a unique pair of protons due to overlapping chemical shifts. In these cases, ARIA invokes procedures that result in some possibilities being eliminated on the basis of the emerging three-dimensional structure. In the event that two or more equally feasible NOEs exist for a given cross-peak, ARIA creates an ambiguous distance restraint.

Three NOESY-based experiments, the ^{15}N-^{1}H HSQC-NOESY, the 2D ^{1}H-^{1}H NOESY and the 2D ^{1}H-^{1}H $^{2}H_2O$ NOESY, were used as a source of NOEs in the structure calculation of VCP ~2,3, and a separate chemical shift table was generated for each of them. This was performed by exporting a selected set of assigned (see Chapter 4) cross-peaks (integrated according to peak volume) from each spectrum. Macros within a program called AZARA (Boucher, http//:www.bio.cam.ac.uk/azara) (a separate program that can also be used for viewing NMR spectra and has useful macros that are not available in ANSIG) were then used to generate the three chemical shift lists in the following way.

The first script to be used was the "make_shift_list script" (Appendix A), which reads all the cross-peaks from each spectrum and makes a table of all the chemical shifts associated with the manually assigned peaks. This table then served as input for the “make_shift_summary” macro. This reads the table, groups together all the chemical shifts for each assigned atom, and lists them in ascending order of chemical shift; although the chemical shift of a proton would be expected to be the same each time it was allocated to a particular NOE, errors in establishing the peak centres mean that there may be small differences in the chemical shifts associated with the same atom if it participates in more than one NOE within the same spectrum. A user-defined threshold (0.03 ppm for (^{1}H) dimensions 1, 2 and 0.3 ppm for (^{15}N) dimension 3) was set, and all chemical shifts associated with an atom had to be within this threshold to be included in the main output file from “make_shift_summary”. If the chemical shifts from any one atom were above the defined threshold, all assignments that included the “offending” atom were exported to a separate file for examination. Examination of the output file normally highlighted one chemical shift that was significantly different from the others, and this proved an effective way of screening for assignment mistakes that could subsequently be amended.

Ultimately, the sorted list of chemical shifts was used as input for the “make_connected_shifts” macro. This took all the chemical shifts from “make_shift_summary” and combined them to give an average chemical shift for each proton or nitrogen atom. Thus, the output from “make_connected_shifts” was taken as the final chemical shift assignment table (Appendix B). Because of the discrepancy in cross-peak positions, each chemical shift was given an associated error, which was the same error margin or tolerance as those described for “make_shift_summary”.

Once a chemical shift table had been finalised for each spectrum, restraint tables were generated for use as input into the structure calculation program. This involved a series of scripts called 'connect'. These scripts read in all the cross-peaks picked in any one spectrum and used the chemical shift table for that particular spectrum to infer restraint possibilities. Those cross-peaks that had earlier been assigned manually remained "unambiguously" assigned, whilst the remainder were connected to unambiguous or ambiguous assignments depending on the number of possibilities available to them. An example of the output from the connect scripts is shown in figure 5 i. Each of the restraint tables generated for the three NOESY spectra were treated independently by the subsequent steps of the ARIA protocol. Restraint tables that listed the inferred hydrogen bonds (chapter 4) were also created as input for the ARIA protocol (figure 5 ii). The disulphide bonds (Chapter 1) were also entered as distance restraints for the structure calculation of VCP~2,3, as shown in figure 5.iii. The dihedral angles were calculated from the coupling constants taken from the HNHα data (Chapter 4). The coupling between the two protons (in Hz) was fitted to the Karplus curve, which was used to obtain the dihedral angle restraints as additional data for the structure calculation. The coupling constants are listed in figure 5 iii.

5.3 ARIA - An automated method for assigning ambiguous restraints

5.3.1 Distance calibration - relating NOE intensity to distance

Any structure calculated by NMR spectroscopy relies on the fact that cross-peak (NOE) intensity is proportional (r^{-6} dependancy) to the distance between two protons – this is the essence of all structure calculations using NOE information. One of the most important considerations therefore is the distance calibration, which relates cross-peak intensity to inter-proton distance. ARIA uses the formulae shown in equations 1 and 2 to represent the upper and lower distance-bounds for an NOE related to the integrated volume of the cross-peak. As mentioned briefly in section

Figure 5.i - Example of output from "connect" script

```
assign ( resid 74 and name HB2
       )
       ( resid 75 and name HD2
       or resid 88 and name HB1
       or resid 89 and name HB1
       or resid 108 and name HB1
       or resid 118 and name HB2
       ) 5.0 5.0 0.0 volume=0.463 peak=3074 ppm1=-0.202 ppm2=2.888

assign ( resid 74 and name HB2
       )
       ( resid 7 and name HH1#
       or resid 7 and name HH2#
       or resid 33 and name HE1
       or resid 33 and name HE2
       or resid 97 and name HE1
       or resid 97 and name HE2
       ) 5.0 5.0 0.0 volume=0.314 peak=4225 ppm1=-0.204 ppm2=6.490

assign ( resid 74 and name HB2
       )
       ( resid 118 and name HE1
       ) 6.0 6.0 0.0 volume=0.231 peak=4204 ppm1=-0.204 ppm2=9.514

assign ( resid 31 and name HD11
       or resid 31 and name HD12
       or resid 31 and name HD13
       )
       ( resid 3 and name HA
       or resid 12 and name HA
       or resid 119 and name HB2
       ) 6.0 6.0 0.0 volume=0.191 peak=5557 ppm1=-0.517 ppm2=4.110
```

Figure 5 i - Example of output from connect script. Each cross-peak from the NMR spectra is assigned unique identifying number and chemical shifts in dim(ension)1 and dim 2. Each peak also has a volume, corresponding to intensity of the NOE between the contributing protons. The numbers on the bottom row (e.g. 5.0 5.0 0.0) refer to maximum internuclear distance for the two protons contributing to the NOE. All possible contributing protons with chemical shift, ppm1 and ppm 2, are listed.

Figure 5 ii - Some distance-restraints used as ARIA input to represent H-Bonds

```
assign (resid  19 and name O    )(resid  34 and name HN   ) 1.88 0.3 0.42
assign (resid  19 and name O    )(resid  34 and name  N   ) 1.88 0.3 1.32
assign (resid  34 and name O    )(resid  19 and name HN   ) 1.88 0.3 0.42
assign (resid  34 and name O    )(resid  19 and name  N   ) 1.88 0.3 1.32
assign (resid  48 and name O    )(resid  60 and name HN   ) 1.88 0.3 0.42
assign (resid  48 and name O    )(resid  60 and name  N   ) 1.88 0.3 1.32
assign (resid  60 and name O    )(resid  48 and name HN   ) 1.88 0.3 0.42
assign (resid  60 and name O    )(resid  48 and name  N   ) 1.88 0.3 1.32
assign (resid  21 and name O    )(resid  32 and name HN   ) 1.88 0.3 0.42
assign (resid  21 and name O    )(resid  32 and name  N   ) 1.88 0.3 1.32
assign (resid  32 and name O    )(resid  21 and name HN   ) 1.88 0.3 0.42
assign (resid  32 and name O    )(resid  21 and name  N   ) 1.88 0.3 1.32
assign (resid  65 and name O    )(resid  42 and name HN   ) 1.88 0.3 0.42
assign (resid  65 and name O    )(resid  42 and name  N   ) 1.88 0.3 1.32
assign (resid  42 and name O    )(resid  65 and name HN   ) 1.88 0.3 0.42
assign (resid  42 and name O    )(resid  65 and name  N   ) 1.88 0.3 1.32
assign (resid  125 and name O    )(resid  104 and name HN    ) 1.88 0.3 0.42
assign (resid  125 and name O    )(resid  104 and name  N    ) 1.88 0.3 1.32
assign (resid  104 and name O    )(resid  125 and name HN    ) 1.88 0.3 0.42
assign (resid  104 and name O    )(resid  125 and name  N    ) 1.88 0.3 1.32
assign (resid  119 and name O    )(resid  112 and name HN    ) 1.88 0.3 0.42
assign (resid  119 and name O    )(resid  112 and name  N    ) 1.88 0.3 1.32
assign (resid  112 and name O    )(resid 119 and name HN    ) 1.88 0.3 0.42
assign (resid  112 and name O    )(resid 119 and name  N    ) 1.88 0.3 1.32
assign (resid  106 and name O    )(resid  123 and name HN    ) 1.88 0.3 0.42
assign (resid  106 and name O    )(resid  123 and name  N    ) 1.88 0.3 1.32
assign (resid  123 and name O    )(resid  106 and name HN    ) 1.88 0.3 0.42
assign (resid  123 and name O    )(resid  106 and name  N    ) 1.88 0.3 1.32
assign (resid  113 and name O    )(resid  93 and name HN    ) 1.88 0.3 0.42
assign (resid  113 and name O    )(resid  93 and name  N    ) 1.88 0.3 1.32
assign (resid  93 and name O    )(resid  113 and name HN    ) 1.88 0.3 0.42
assign (resid  93 and name O    )(resid  113 and name  N    ) 1.88 0.3 1.32
assign (resid  111 and name O    )(resid  95 and name HN    ) 1.88 0.3 0.42
assign (resid  111 and name O    )(resid  95 and name  N    ) 1.88 0.3 1.32
assign (resid  95 and name O    )(resid  111 and name HN    ) 1.88 0.3 0.42
assign (resid  95 and name O    )(resid  111 and name  N    ) 1.88 0.3 1.32
assign (resid  81 and name O    )(resid  98 and name HN   ) 1.88 0.3 0.42
assign (resid  81 and name O    )(resid  98 and name  N   ) 1.88 0.3 1.32
assign (resid  98 and name O    )(resid  81 and name HN   ) 1.88 0.3 0.42
assign (resid  98 and name O    )(resid  81 and name  N   ) 1.88 0.3 1.32
```

Figure 5 ii - Data from slowly exchanging amide experiments were used to create restraints for inferred hydrogen bonds between the amide and carbonyl groups of two different residues. The three columns to the right are distances in Å, and these three distances, x, y, z, are interpreted as representing a distance > (x-y) and < (x-z) by X-PLOR.

Equation 1

$$U = \left(\frac{d_{ref}^{-6}}{V_{ref}} V\right)^{-\frac{1}{6}} + \Delta^{+}$$

Equation 2

$$L = \left(\frac{d_{ref}^{-6}}{V_{ref}} V\right)^{-\frac{1}{6}} - \Delta^{-}$$

U = Upper bound for interproton distance
L = Lower bound for interproton distance
d_{ref} = reference distance between contributing protons (in this case, an average of all unambiguous distance restraints)
V_{ref} = reference peak volume (in this case, an average of all peak volumes from unambiguous restraints)
Δ^{+} = error for upper bound (in this case $0.125D^2$ where D is the calibrated distance)
Δ^{-} = error for lower bound (in this case $0.125D^2$ where D is the calibrated distance)

5.1, a typical approach is to classify an NOE as being medium, weak, or strong, and the distance between two protons is correlated to these distance boundaries. The ARIA method, instead, uses the distances and cross-peak volumes of a series of "known" distance restraints (from unambiguous assignment data) and calibrates every other cross-peak accordingly. These reference restraints that are used to relate cross-peak volume to distance, are recalculated on an iterative basis as the automated assignment proceeds. The distances estimated from the NOEs have set errors (Δ^{+} and Δ^{-} (see equations 1 and 2)), that were defined in this case as $0.125D^2$ where D is the estimated distance according to the peak volume. ARIA proceeds in an iterative manner, using information from each previous iteration to generate more accurate and lower energy structures each time; the closest structure to the actual structure will have the lowest energy, so ARIA uses the lowest energy structures from one iteration to help guide the subsequent calculations.

Figure 5 iii - SS bond restraints and dihedral angles

a)

```
assign (resid  9 and name sg) (not resid  9 and name sg) 2.08 0.1 0.1
assign (resid 49 and name sg) (not resid 49 and name sg) 2.08 0.1 0.1
assign (resid 35 and name sg) (not resid 35 and name sg) 2.08 0.1 0.1
assign (resid 66 and name sg) (not resid 66 and name sg) 2.08 0.1 0.1
assign (resid  71 and name sg) (not resid  71 and name sg) 2.08 0.1 0.1
assign (resid 113 and name sg) (not resid 113 and name sg) 2.08 0.1 0.1
assign (resid  99 and name sg) (not resid  99 and name sg) 2.08 0.1 0.1
assign (resid 124 and name sg) (not resid 124 and name sg) 2.08 0.1 0.1
```

b)

```
ASSIgn (residue   8 and name C ) (residue   9 and name N)
       (residue   9 and name CA) (residue   9 and name C) 1 -120 50 2 !
ASSIgn (residue  18 and name C ) (residue  19 and name N)
       (residue  19 and name CA) (residue  19 and name C) 1 -120 50 2 !
ASSIgn (residue  20 and name C ) (residue  21 and name N)
       (residue  21 and name CA) (residue  21 and name C) 1 -120 50 2 !
ASSIgn (residue  21 and name C ) (residue  22 and name N)
       (residue  22 and name CA) (residue  22 and name C) 1 -120 50 2 !
ASSIgn (residue  24 and name C ) (residue  25 and name N)
       (residue  25 and name CA) (residue  25 and name C) 1 -120 50 2 !
ASSIgn (residue  25 and name C ) (residue  26 and name N)
       (residue  26 and name CA) (residue  26 and name C) 1 -120 50 2 !
ASSIgn (residue  29 and name C ) (residue  30 and name N)
       (residue  30 and name CA) (residue  30 and name C) 1 -120 50 2 !
ASSIgn (residue  30 and name C ) (residue  31 and name N)
       (residue  31 and name CA) (residue  31 and name C) 1 -120 50 2 !
ASSIgn (residue  32 and name C ) (residue  33 and name N)
       (residue  33 and name CA) (residue  33 and name C) 1 -120 50 2 !
ASSIgn (residue  33 and name C ) (residue  34 and name N)
       (residue  34 and name CA) (residue  34 and name C) 1 -120 50 2 !
ASSIgn (residue  34 and name C ) (residue  35 and name N)
       (residue  35 and name CA) (residue  35 and name C) 1 -120 50 2 !
ASSIgn (residue  38 and name C ) (residue  39 and name N)
       (residue  39 and name CA) (residue  39 and name C) 1 -120 50 2 !
ASSIgn (residue  40 and name C ) (residue  41 and name N)
       (residue  41 and name CA) (residue  41 and name C) 1 -60 50 2 !
ASSIgn (residue  41 and name C ) (residue  42 and name N)
       (residue  42 and name CA) (residue  42 and name C) 1 -120 50 2 !
ASSIgn (residue  45 and name C ) (residue  46 and name N)
       (residue  46 and name CA) (residue  46 and name C) 1 -120 50 2 !
ASSIgn (residue  46 and name C ) (residue  47 and name N)
       (residue  47 and name CA) (residue  47 and name C) 1 -120 50 2 !
ASSIgn (residue  47 and name C ) (residue  48 and name N)
       (residue  48 and name CA) (residue  48 and name C) 1 -120 50 2 !
ASSIgn (residue  57 and name C ) (residue  58 and name N)
       (residue  58 and name CA) (residue  58 and name C) 1 -120 50 2 !
ASSIgn (residue  58 and name C ) (residue  59 and name N)
       (residue  59 and name CA) (residue  59 and name C) 1 -120 50 2 !
ASSIgn (residue  59 and name C ) (residue  60 and name N)
       (residue  60 and name CA) (residue  60 and name C) 1 -120 50 2 !
ASSIgn (residue  61 and name C ) (residue  62 and name N)
       (residue  62 and name CA) (residue  62 and name C) 1 -60 50 2 !
ASSIgn (residue  62 and name C ) (residue  63 and name N)
       (residue  63 and name CA) (residue  63 and name C) 1 -120 50 2 !
ASSIgn (residue  64 and name C ) (residue  65 and name N)
       (residue  65 and name CA) (residue  65 and name C) 1 -120 50 2 !
ASSIgn (residue  65 and name C ) (residue  66 and name N)
       (residue  66 and name CA) (residue  66 and name C) 1 -120 50 2 !
ASSIgn (residue  66 and name C ) (residue  67 and name N)
       (residue  67 and name CA) (residue  67 and name C) 1 -120 50 2 !
ASSIgn (residue  71 and name C ) (residue  72 and name N)
       (residue  72 and name CA) (residue  72 and name C) 1 -120 50 2 !
ASSIgn (residue  72 and name C ) (residue  73 and name N)
       (residue  73 and name CA) (residue  73 and name C) 1 -60 50 2 !
ASSIgn (residue  76 and name C ) (residue  77 and name N)
       (residue  77 and name CA) (residue  77 and name C) 1 -120 50 2 !
ASSIgn (residue  80 and name C ) (residue  81 and name N)
       (residue  81 and name CA) (residue  81 and name C) 1 -120 50 2 !
```

Figure 5.iii cont'd

```
ASSIgn (residue  85 and name C ) (residue  86 and name N)
       (residue  86 and name CA) (residue  86 and name C) 1 -120 50 2 !
ASSIgn (residue  86 and name C ) (residue  87 and name N)
       (residue  87 and name CA) (residue  87 and name C) 1 -120 50 2 !
ASSIgn (residue  87 and name C ) (residue  88 and name N)
       (residue  88 and name CA) (residue  88 and name C) 1 -120 50 2 !
ASSIgn (residue  88 and name C ) (residue  89 and name N)
       (residue  89 and name CA) (residue  89 and name C) 1 -120 50 2 !
ASSIgn (residue  89 and name C ) (residue  90 and name N)
       (residue  90 and name CA) (residue  90 and name C) 1 -120 50 2 !
ASSIgn (residue  92 and name C ) (residue  93 and name N)
       (residue  93 and name CA) (residue  93 and name C) 1 -60 50 2 !
ASSIgn (residue  93 and name C ) (residue  94 and name N)
       (residue  94 and name CA) (residue  94 and name C) 1 -120 50 2 !
ASSIgn (residue  94 and name C ) (residue  95 and name N)
       (residue  95 and name CA) (residue  95 and name C) 1 -120 50 2 !
ASSIgn (residue  96 and name C ) (residue  97 and name N)
       (residue  97 and name CA) (residue  97 and name C) 1 -120 50 2 !
ASSIgn (residue  97 and name C ) (residue  98 and name N)
       (residue  98 and name CA) (residue  98 and name C) 1 -120 50 2 !
ASSIgn (residue  98 and name C ) (residue  99 and name N)
       (residue  99 and name CA) (residue  99 and name C) 1 -120 50 2 !
ASSIgn (residue  99 and name C ) (residue 100 and name N)
       (residue 100 and name CA) (residue 100 and name C) 1 -60 50 2 !
ASSIgn (residue 102 and name C ) (residue 103 and name N)
       (residue 103 and name CA) (residue 103 and name C) 1 -120 50 2 !
ASSIgn (residue 103 and name C ) (residue 104 and name N)
       (residue 104 and name CA) (residue 104 and name C) 1 -120 50 2 !
ASSIgn (residue 105 and name C ) (residue 106 and name N)
       (residue 106 and name CA) (residue 106 and name C) 1 -120 50 2 !
ASSIgn (residue 108 and name C ) (residue 109 and name N)
       (residue 109 and name CA) (residue 109 and name C) 1 -60 50 2 !
ASSIgn (residue 110 and name C ) (residue 111 and name N)
       (residue 111 and name CA) (residue 111 and name C) 1 -120 50 2 !
ASSIgn (residue 111 and name C ) (residue 112 and name N)
       (residue 112 and name CA) (residue 112 and name C) 1 -120 50 2 !
ASSIgn (residue 112 and name C ) (residue 113 and name N)
       (residue 113 and name CA) (residue 113 and name C) 1 -120 50 2 !
ASSIgn (residue 113 and name C ) (residue 114 and name N)
       (residue 114 and name CA) (residue 114 and name C) 1 -120 50 2 !
ASSIgn (residue 116 and name C ) (residue 117 and name N)
       (residue 117 and name CA) (residue 117 and name C) 1 -120 50 2 !
ASSIgn (residue 117 and name C ) (residue 118 and name N)
       (residue 118 and name CA) (residue 118 and name C) 1 -120 50 2 !
ASSIgn (residue 122 and name C ) (residue 123 and name N)
       (residue 123 and name CA) (residue 123 and name C) 1 -120 50 2 !
ASSIgn (residue 123 and name C ) (residue 124 and name N)
       (residue 124 and name CA) (residue 124 and name C) 1 -120 50 2 !
ASSIgn (residue 124 and name C ) (residue 125 and name N)
       (residue 125 and name CA) (residue 125 and name C) 1 -120 50 2 !
```

Figure 5 iii - a) Restraint input table for residues involved in disulphide bonds. b) Restraint input table for dihedral angle restraints. The two columns on the right (between 1 and 2) e.g. -120 50 are interpreted as e.g. -120 ± 50 ° by X-PLOR.

5.3.2 Automatic assignment of ambiguous NOE's

At the start of the structure calculation in ARIA, the NOEs that have been unambiguously assigned during an initial manual stage, define the initial fold. As the calculation proceeds, ARIA creates additional assignments on the basis of the chemical shift tables and by reference to inter-proton distances in the emerging structure. ARIA works on an iterative basis and generates a new assignment table in each iteration, based on the calculated structures from the previous round. ARIA also eliminates duplicate restraints that are present in the same, or in different spectra, pooling the data from the different spectra into one final restraints table. Each spectrum is considered individually during the structure calculation, but the resulting NOE tables are a combination of them all.

In ARIA, a user-specified number of the structures from the previous iteration are used to generate the assignment tables for subsequent iterations; the lowest energy structures will have the most correct restraints, so only the lowest energy structures are used to create the subsequent assignment tables (the actual number used can be changed from iteration to iteration). ARIA compares all the possible assignments for an NOE with the measured distances in the structures available and considers the contribution that each restraint is making to the cross-peak. In the restraints tables generated from the connect scripts, many of the possible restraints will be unfeasible, as two protons may lie at opposite ends of the structure, yet would still be listed as a possible restraint. ARIA filters out these assignment possibilities since their contribution to the overall cross-peak intensity will be insignificant. Those restraints that appear likely to contribute to cross-peak intensity are assigned an appropriate contribution value (C). A cut-off value (p) can be applied to specify the percentage a

NOE must contribute to the cross-peak intensity to be included as apossibility within the distance restraint.

In initial iterations, all possible restraints are considered. In later iterations, the value of p is lowered - typically, from 1 (100%) to 0.9 (90%). All assignment possibilities are ordered such that the highest contributing restraint possibility is given priority, followed by all other contributing restraints in order. The contributions are defined as outlined in equation 3. The fewest number of contributions are chosen such that the condition shown in equation 4 is satisfied. The effect is to keep the shortest restraints that cumulatively acount for 90% (0.9) of total cross-peak intensity and discard the remaining restraints. If the automatically assigned restraints are accurate, the energy of the system is lowered and the structures will improve – the ambiguous restraints are re-considered in each iteration. Occasionally, assignment possibilities are chosen which do not tie in with other restraints. In these circumstances, the energy

Equation 3

$$C_n = \frac{\hat{d}_n^{-6}}{\sum_{a=1}^{N_\delta} \hat{d}_a^{-6}}$$

C_n - Contribution to cross-peak from assignment n

$\hat{d}_n^{-6}$ - minimum or average distance for the restraint(s)

$\sum_{a=1}^{N_\delta} \hat{d}_a^{-6}$ - NOE intensity (a runs through N_δ contributions; d_a is distance between two protons corresponding to ath contribution)

Equation 4

$$\sum_{a=1}^{N_p} C_a > p$$

of the system will go up and the structures will not be used for subsequent iterations. These errors are not carried forward to the next iteration, as only those structures which have the lowest E_{total} are used for the calibration of restraints in the next round.

In the structure calculation of VCP~2,3, four different parameters in ARIA were altered at each iteration. These parameters were:

- The number of structures generated in each iteration (N)
- The number of lowest energy structures from the previous iteration kept as starting structures (S)
- The number of lowest energy structures from iteration (i) used to create the restraint tables for i+1 (Equation 4) (**A**)
- The value of p (Equation 4)

Eleven iterations were performed in total with the respective values of N, S, **A** and p set according to table 1 (these values were arrived at after significant amounts of trial and error). In the first iteration, p is set at 0.999 so that almost all contributing assignments are used to generate the assignment table for iteration 2. This value is progressively lowered so that those assignments contributing very little to the overall NOE intensity are filtered out. **A**, the number of lowest energy structures used for creating the restraint tables for iteration (i+1) is also systematically altered. In the final iteration, no starting structures are used. Thus, all 100 final structures are generated from random starting structures using restraint tables generated from the 25 lowest energy structures from iteration 10.

Table i - Values of N, S, A and p are altered at each iteration to assist the structure calculation and reduce computational cost.

ITERATION	N	S	A	p
1	20	0	0	0.999
2	20	15	10	0.999
3	20	15	10	0.99
4	20	15	5	0.98
5	20	15	5	0.97
6	20	15	10	0.96
7	20	15	10	0.95
8	30	15	5	0.94
9	50	20	5	0.93
10	100	25	5	0.90
11	100	0	25	0.90

5.3.3 Simulated annealing

The basis for solving a structure by simulated annealing involves optimising a target function E_{total} that represents the overall energy of a protein structure. This total energy will take into account the covalent bond energy, energy associated with bond angles, planarity, and van der Waals' interactions in addition to energies associated with the experimentally derived restraints used as input such as the NOEs, H-bonds, dihedral angles and the disulphide bonds. The structure calculation aims to find a global minimum for all atoms, representing the lowest energy conformation of the protein. It uses molecular dynamics to solve Newton's equations of motion for each atom in the protein to get E_{total}. One problem is that atoms can find themselves in local minima and require kinetic energy to escape. To avoid this, the simulated annealing protocol incorporates provision of a constant temperature water bath that provides kinetic energy to the system, allowing atoms to escape from the local minima. In order to do this effectively, certain energy parameters such as the van der Waals' repulsion terms are initially switched off, allowing the atoms to pass through each other in the early stages. This decreases the computational time required for the initial minimisation. When the global minimum has been attained, the system is cooled and the van der Waals' potential is re-introduced. X-PLOR is a widely utilised software tool that runs the complex algorithms to calculate the energy of the system and to perform the simulated annealing process.

5.4 Structure calculation and troubleshooting

5.4.1 Distance re-calibration

During a structure calculation, many restraints are violated, i.e the calibrated distance is larger than the measured inter-proton distance. Within ARIA, when this occurs in more than a user-defined proportion of the ensemble of calculated structures, the restraint is subjected to a re-calibration procedure, whereby the distance is set to a pre-defined value plus a defined error margin. The re-calibrated distance represents the upper bound for the longest distance for which an NOE should be visible. This allows effects such as spin-diffusion, whereby an NOE may appear stronger than it would normally be due to effects of other atomic spins in the vicinity, to be taken into account. These re-calibrated peaks are incorporated into the structure, if possible. If the algorithm cannot restrain it to the appropriate distance, then the cross-peak is considered to be a genuine violation and will be excluded from the final set of structures and the output restraints file.

In the case of VCP~2,3, many NOE-derived distance restraints that were consistently violated in different structure calculations were thoroughly examined and, in some cases, excluded from the input tables in subsequent calculations on the grounds of being artefactual peaks or too overlapped to allow reliable assignment or integration. Dealing with violated NOEs turned out to be one of the most important and time-consuming parts of the structure calculation. Extra peaks may arise due to artefacts of the spectral processing, bad water suppression, sample impurities, peak splitting due to strong coupling, or from the presence of different conformations of.

During successive iterations of the structure calculation of VCP~2,3, and as the structures improved (in terms of total energy and the number of violations), the ARIA parameter for violation tolerance was modified. In early iterations and structure calculations, the violation tolerance was defined such that NOE's causing violations in 50% of the calculated structures and violated by more than 0.5 Å were listed in a violation output file for subsequent analysis. In the final stages of structure calculation, stricter violation tolerances were applied so that restraints violated by more than 0.1 Å in any structure were listed. These output files highlighted some of the problems with the different NMR spectra, peak-picking or assignment, as discussed below.

With VCP~2,3, peak splitting (due to very strong scalar coupling between the geminal α protons) was a problem in the case of glycine residues, resulting in two separate cross-peaks emanating from one resonance. Once identified, this problem was overcome by manually determining the centre of the two peaks. Sequential assignments were used as an extra checking mechanism to ensure the cross-peak centre was correct.

The analysis of violated peaks often led to identification of previously unassigned atoms, which were thought to be degenerate. In several cases, the unassigned atom was missed in the initial assignment stages because the chemical shift was unusual; an approximate reference of chemical shift exists for each atom in an amino acid (Wuthrich, 1986) and those atoms with chemical shifts outside of the "expected" range were much less easy to assign.

When dealing with violations, it was important to consider how bad the violation was in terms of distance, how realistic the different restraint possibilities were, and whether the violated peak could in fact be coming from an unassigned proton. It was possible to make decisions about violated restraints by looking at the spectrum and finding all the cross-peaks that could emanate from a resonance at that particular chemical shift. In some cases, several different violated cross-peaks would have the same chemical shift in one dimension. Examination of the spectrum then revealed that these peaks had strong connections to one particular amino acid so the full assignment of that residue was then checked to see if all atoms had been assigned or if there were degenerate atoms. In other cases, incorrect unambiguous manual assignments were removed to resolve the problem.

Cross-peaks with very strong spin-systems in the COSY and TOCSY spectra, but very little NOESY information were signs of an artefact. Peak shape and size was also a good indication of an artefact, which underlined the importance of effective peak-picking at the start of the spectral analysis. The "normal" cross-peaks were almost circular in shape and showed a degree of consistancy throughout the spectrum.

NOEs arising from artefacts had a tendancy to be noticeably different from the others in shape (often elongated and pointed) or intensity.

5.4.2 Dealing with pro-chiral groups

Another obstacle that must be overcome during the structure calculation is the assignment of the pro-chiral atoms. If a pair of pro-chiral atoms can be resolved, then to use a pseudo atom (an imaginary atom placed midway between the pro-chiral protons) results in a loss of information. The system that is used in the ARIA protocol is to 'float' the pro-chiral centres. Using this method, the atoms can be given an arbitrary assignment initially and the algorithm subsequently allows them to be swapped over during the structure calculation so that both possibilities are considered. This means that unambiguous manual assignment of protons with regard to chirality was not necessary in the structure calculation of VCP~2,3.

5.4.3 Tolerance levels

An additonal problem that had to be overcome in the structure calculation of VCP~2,3 arose from the tolerance level of 0.03 ppm (i.e. the tolerance allowed during automated assignment on the basis of the chemical shift table) in the proton dimensions. This caused several NOE violations as follows. Occasionally, overlap would result in a cross-peak being picked in the centre of what was in reality a group of two or more cross-peaks that were not completely resolved. Because of the tolerance level, such a cross-peak could be attributed by ARIA to an incorrect, non-existent restraint. When the tolerances in chemical shifts were subsequently lowered in an attempt to circumvent the problem, some assignment possibilities were excluded by ARIA that were in fact the correct restraints. If on the other hand the tolerances in chemical shifts were raised sufficiently for inclusion of the unresolved peak, the number of possibilities for ambiguous restraints increased, raising computational cost and providing a greater potential for computational error in the automated assignment procedure. For this reason, manual peak picking is probably more preferable to automated peak picking, as the human eye can often make better judgements that the algorithm. Moreover, by manually picking the cross-peaks, many peaks that were in fact actually attributable to noise were also eliminated.

The problems of chemical shift tolerances also highlighted the necessity of having separate chemical shift tables for each spectrum - differences in cross-peak positions as little as 0.015 ppm from one spectrum to the next could potentially cause many violation problems.

References

Brunger, A. (1992). *X-PLOR Version 3.1: A system for X-ray crystallography and NMR,* Yale University Press, New Haven and London

Nilges, M. & O'Donoghue, S. I. (1997). *Ambiguous NOE's and automated NOE assignment*. ARIA manual, Structural Biology Programme, European Molecular Biology Laboratory, Heidelberg.

Nilges, M., Macias, M. J., O'Donaghue, S. I. & Oschkinat, H (1997) *J. Mol. Biol.* **269**, 408-422

Wuthrich, K. (1986). *NMR of Proteins and Nucleic acids*, John Wiley, New York

Chapter 6 - The solution structure of VCP~2,3

6.1 Secondary structure

The secondary structure of a protein may be inferred from chemical shifts, slowly-exchanging amides, coupling constants and the pattern of backbone-backbone NOEs. This is a useful complement and precursor to a full determination of 3D structure based on all of the experimental data.

A chemical shift index (CSI) (Wishart *et al.*, 1992) was calculated for VCP~2,3 on the basis of Hα shifts (Figure 6 i). It is consistent with the prevalence of short β-strands, and a lack of helices, now recognised as being typical of CP modules (Barlow & Campbell, 1994) and many other small extracellular module-types (Bork *et al.*, 1996). The incidence of residues with large (> 8 Hz) $^3J_{H\alpha N}$ and strong $H\alpha_iHN_{(i+1)}$ NOEs that are associated with extended regions of polypeptide (Wüthrich, 1986), largely support this pattern of strands (Figure 6 ii). Smaller (< 5 Hz) $^3J_{H\alpha N}$ and strong $HN_iHN_{(i+1)}$ NOEs that are generally seen in turns (Wüthrich, 1986), are observed for some residues that lie between the inferred strands, or at their extremities – again this is mostly consistent with the CSI. Amide protons that exchange with solvent relatively slowly and might participate in H-bonds are located mostly within the inferred β-strands, as expected (Figure 6 iii). Taken together, the sequential NOEs, coupling constants and slowly exchanging amides imply that there are three strands in module 3 not suggested by the CSI. This is not altogether

Figure 6 i - Sequence of VCP~2,3 and CSI plot

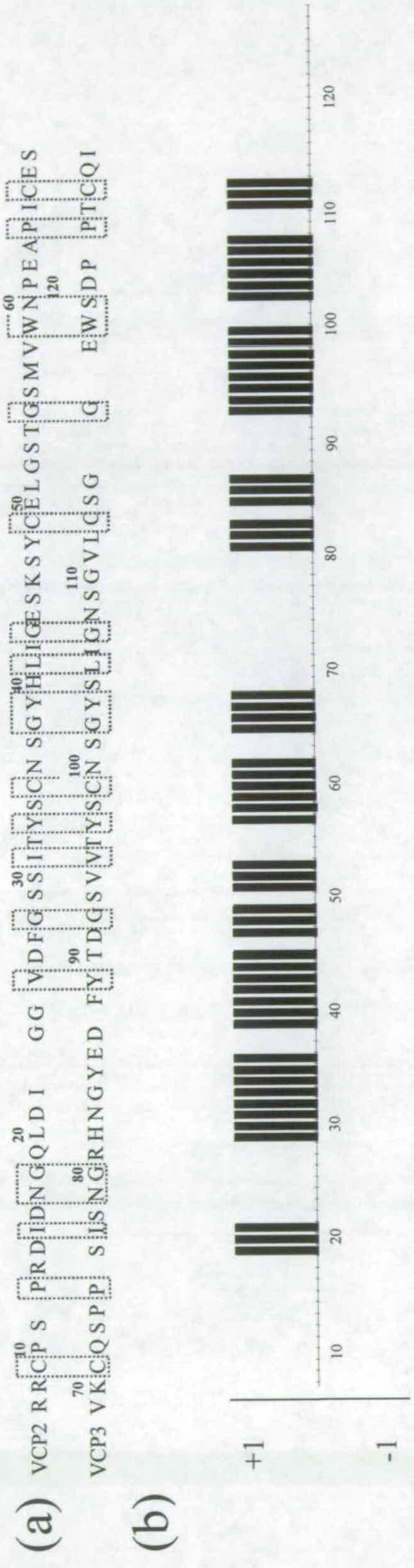

Figure 6 i - a) Sequence of VCP~2,3 (residues conserved in RCA proteins are boxed). b) CSI plot of VCP~2,3 showing areas of predicted β-strands (+1). The CSI plot also indicated that there were no predicted helices (-1).

Figure 6 ii - NOE patterns and coupling constants in VCP~2,3

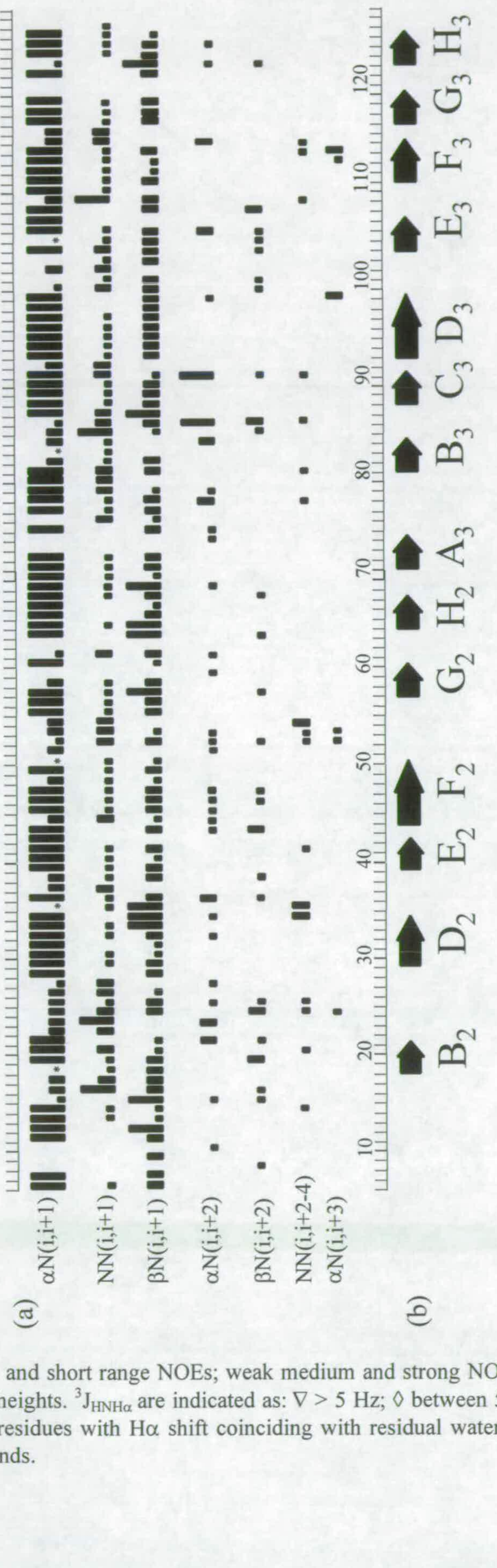

Figure 6 ii - a) Sequential and short range NOEs; weak medium and strong NOEs are represented by black boxes of increasing heights. $^3J_{HNH\alpha}$ are indicated as: ∇ > 5 Hz; ◊ between 5 and 8 Hz; Δ > 8 Hz. Presence of a * indicates residues with Hα shift coinciding with residual water signal b) Position in sequence of inferred β-strands.

surprising since the CSI is not completely reliable when ^{13}C shifts are not included in the calculation. Moreover, CP modules have in general rather short strands and the CSI is best suited to the identification of longer regions of regular secondary structure. From a consideration of all this data, together with the networks of long-range NOEs represented in Figure 6 iii, the locations of β-strands may be inferred as: strand B_2, 19-21; strand D_2, 30-34; strand E_2, 40-42; strand F_2, 45-50; strand G_2, 58-60; strand H_2, 65-67; strand A_3, 71-73; strand B_3, 81-83; strand C_3, 88-90; strand D_3, 93-98; strand E_3, 104-106; strand F_3, 111-114; strand G_3, 117-119; and strand H_3, 123-125. Module 2 differs from module 3 in that it lacks strands A and C, which form small anti-parallel β-sheets in some but not all CP module structures solved previously (Barlow & Campbell, 1994; Wiles *et al.*, 1997; Casasnovas *et al.*, 1999). In both of the modules β-strand H extends one residue beyond the fourth cysteine, thereby (in the case of module 2) effectively reducing the degrees of freedom available to residues of the linker (residues 67-70). This is a critical consideration when considering module-module flexibility and orientations (see below). Figure 6 iii shows the arrangement of strands inferred from manually assigned backbone-backbone NOEs and amide-exchange data.

In the case of VCP~2,3 it was surprising to find that very few amide hydrogens persisted for the long time-periods typical of other globular proteins – this is discussed further below. Module 2 contains only one HN that persists for > 30 minutes after dissolving VCP~2,3 in D_2O (and this does not survive for 2 h). On the other hand module 3 has seven HNs that may be observed in the spectrum after 2 h in D_2O and eight further examples that persist for > 30 minutes. An additional 18 HNs, from both modules, are visible after 15 minutes. The difference in the behaviour of modules 2 and 3 in this respect has been noted previously (Kirkitadze *et al.*, 1999b).

Figure 6 iii - NOEs observed in NMR spectra of VCP~2,3

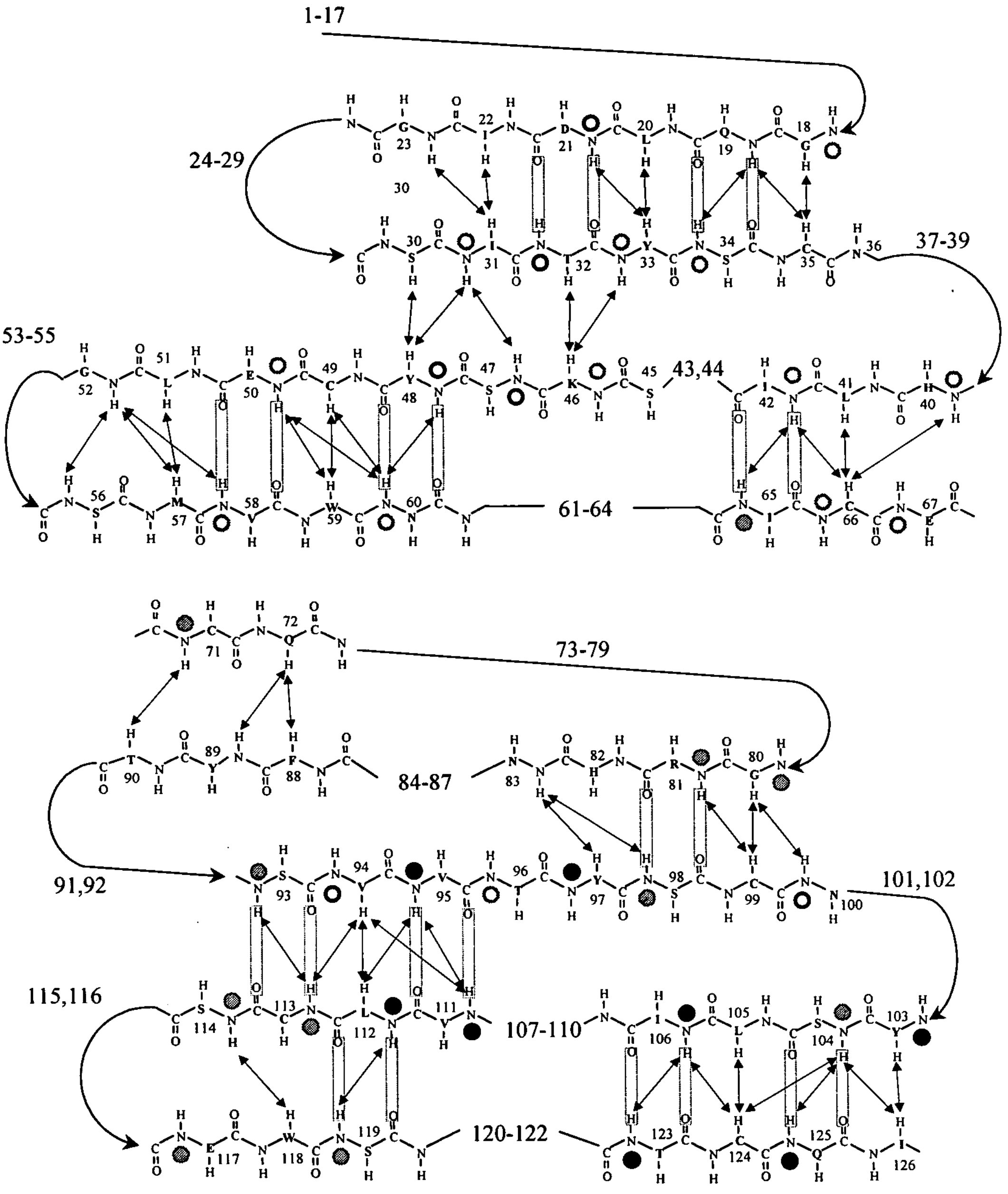

Figure 6 iii - NOEs observed in NMR spectra of VCP~2,3. The residues are labelled by single letter code and sequence number at the Cα's. The persistence of backbone amide protons after dissolving in ^2H_2O is indicated: present after (open grey circles) 13, (filled grey circles) 30, (white filled circles) 60, (black filled circles) 120 minutes. A double headed arrow indicates an NOE. Residues involved in inferred H-bonds are boxed.

Nonetheless, the available information was used to infer H-bonds with the help of the supporting networks of NOEs, as shown in Figure 2a. The Asn60-Pro61 bond was found to be in the *cis* configuration based on the presence of a characteristic Hα(Asn)$_{60}$-Hα(Pro)$_{61}$ NOE and absence of Hα(Asn)$_{60}$-Hδ(Pro)$_{61}$ NOEs. Other X-Pro bonds were in the *trans* conformation.

6.2 Experimentally derived data used for the calculation

A total of 1096 manually assigned NOEs, and the NOEs assigned automatically by ARIA on the basis of their chemical shift (see Chapter 5), served as input for structure calculation together with 54 ϕ dihedral restraints (based on coupling constants), 42 distance restraints representing the inferred H-bonds, and four restraints representing the disulphide bonds. Following ten cycles of simulated annealing performed within X-PLOR (Brünger, 1992) using the ARIA scripts (Nilges *et al.*, 1997) (Chapter 5), a total of 2187 unambiguous NOEs and 896 ambiguous NOEs were used in the final (eleventh) round of structure calculations. As expected for a compactly folded protein, amino acids contributing to inter-residue NOEs are distributed throughout the peptide sequence with an average of 20 inter-residue NOEs per residue (Figure 6 iv)). This is a healthy number of distance restraints and would be expected to lead to a high-resolution solution structure. Some residues, however, have only one non-sequential inter-residue NOE (Arg7, Asp16, Ser37, Thr54, Gly55, Asp87, Ser101, Gly102) while Ser76 has none. The residues with the most of this category of NOEs are Trp59 with 66 and Trp118 with 53. Although there was a disparity in the number of slowly-exchanging amides, approximately equivalent

Figure 6 iv - NOEs at the interface of VCP~2,3

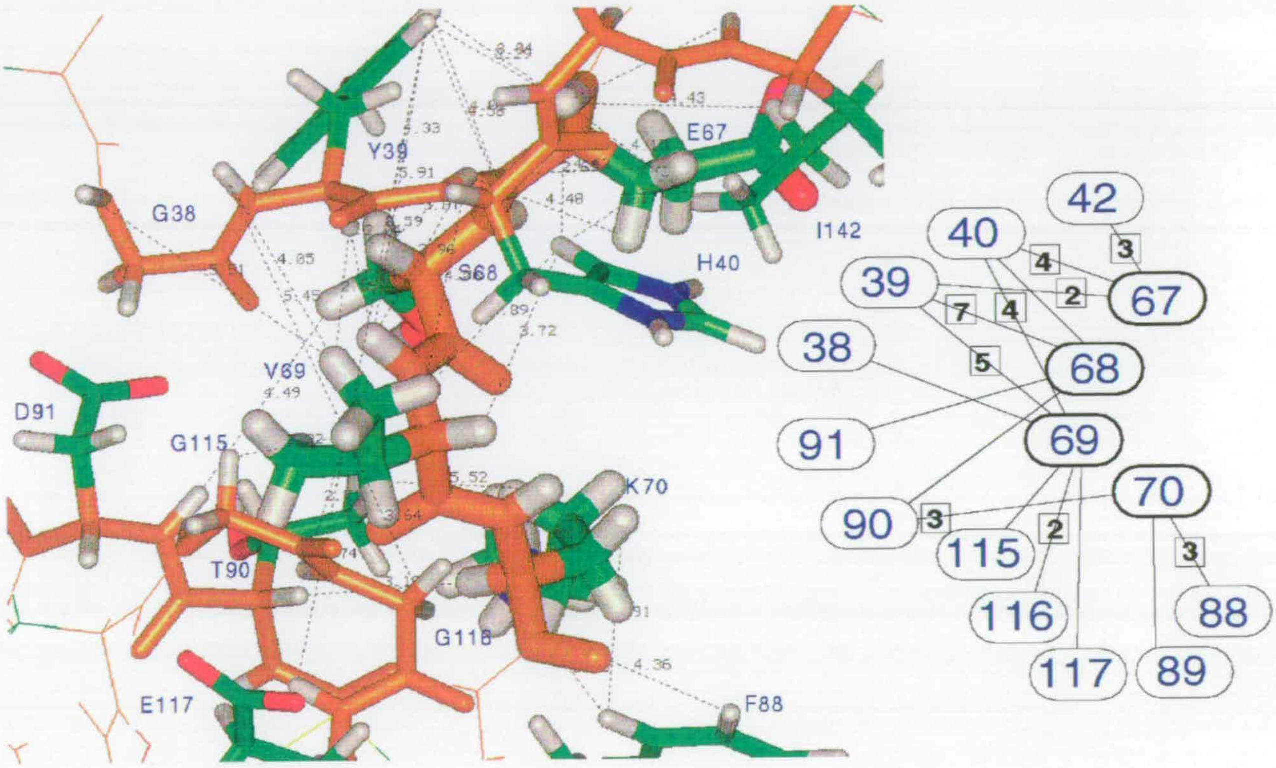

Figure 6 iv - Residues involved in the interface between VCP~2 and 3. The backbone of the protein is shown in orange, and the sidechains are green. Dotted lines and distances Å are shown for observed NOEs between the linker and modules 2 and 3. The number of connectivities observed between the residues is also illustrated (diagram generated using the program INSIGHT) numbers of inter-residue NOEs were identified in each of the two modules. Because of the interest in intermodular orientation it is worth noting that Val69 in the linker has 20 non-sequential inter-residue NOEs. In total, 27 NOEs were identified between the body of module 2 (*i.e.* up to and including Cys66) and the linker (residues 67-70), and 13 NOEs connected the linker and the body of module 3 (*i.e.* Cys71 and subsequent residues). Of these 40 NOEs (represented in Figure 6 iv), 20 had been unambiguously assigned and 20 were ambiguous but contributed > 50% to the intensity of the cross-peak (calculated within ARIA). Only two NOEs were potentially inter-modular, both of them ambiguous. One (Gly38Hα - Asp91Hβ) contributed less than 5% of the intensity of the ambiguous restraint, and the other (Tyr39Hε- Thr90Hβ) less than 25%.

The structure calculation of VCP~2,3 was seeded with a total of 1096 manually assigned NOE's that were used as initial input for the ARIA protocol. A total of 54 dihedral angle restraints and 42 distance restraints representing the inferred hydrogen bonds were also used along with four restraints representing the disulphide bonds. An ensemble of the final 50 structures calculated for VCP~2,3 (following ten cycles of simulated annealing performed within X-PLOR using the ARIA scripts, see Chapter 5), selected, on the bases of total potential energies, out of 100 structures calculated in iteration 11, are shown in figure 6 v. The final list of restraints used to calculate the structure consisted of 2187 unambiguous restraints and 887 ambiguous restraints (Table i). The distribution of the NOE's in the unambiguous restraints are shown in fig 6 vi and resulted in an average of 20 inter-residue restraints per amino acid. Amino acids that had few unambiguous NOE's showed the highest deviation when compared in different structures from the ensemble. The least structured residues, which had

Figure 6 v - Overlay of the 50 lowest energy structures calculated for VCP~2,3

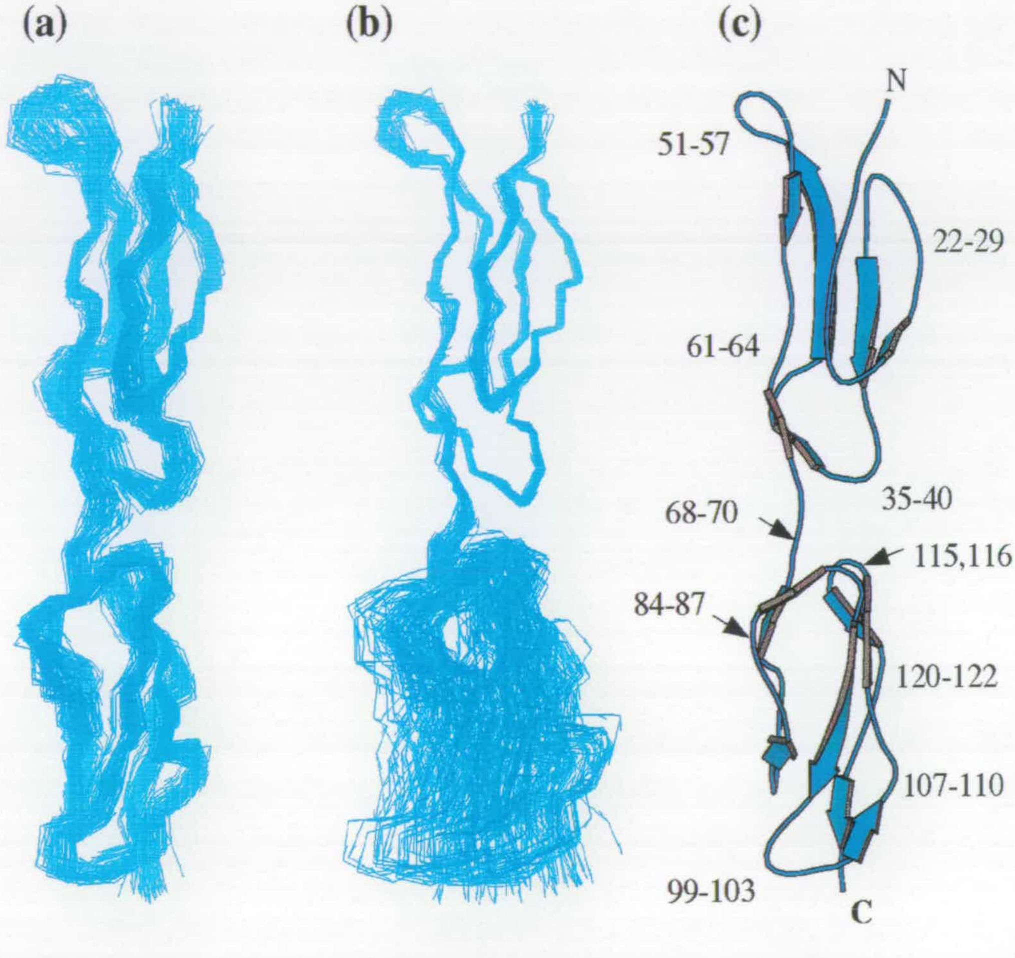

Figure 6 v - Overlay of final 50 structures shown as backbone traces, superimposed on Cα's on a) both modules; b) module 2. c) Molscript (Kraulis, 1991) representation of VCP~2,3 in the same orientation as b). Co-ordinates have been deposited at EBI and assigned the PDB ID code 1e5g.

Figure 6 vi - Unambiguous NOE distribution and regions of highest deviation in the final 50 VCP~2,3 structures

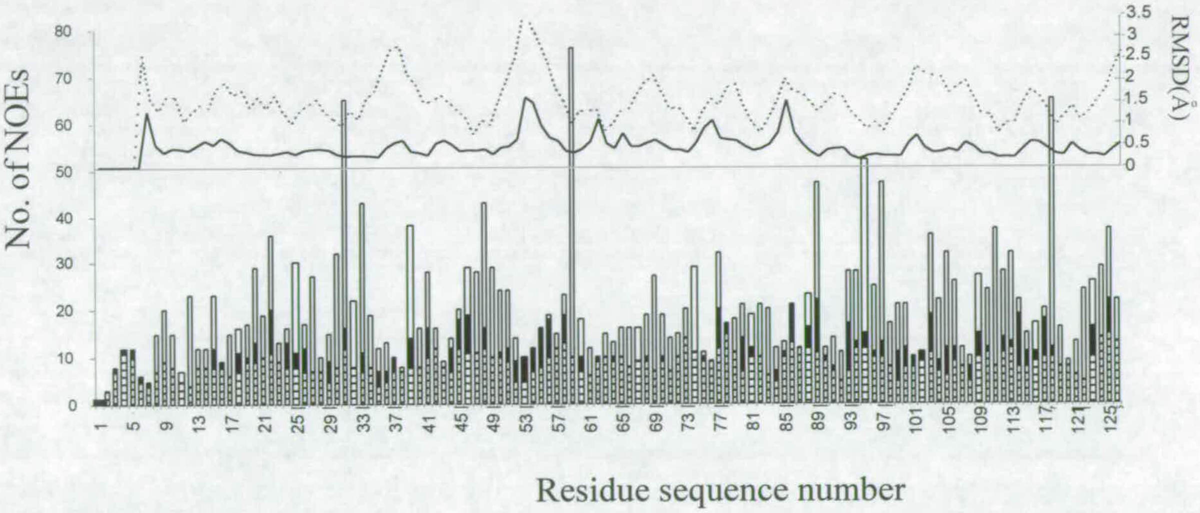

Figure 6 vi - The hatched bars indicate numbers of sequential NOEs, black bars medium range (i-i + (2-4)) NOEs, and white bars long range NOEs. Only unambiguous NOEs are shown. The dotted line represents the rmsd (backbone) from the average structure for the ensemble of structures; continuous line shows the rmsd (backbone) when each module is considered separately.

only one non-sequential inter-residue NOE were Arg7, Asp16, Ser37, Thr54, Gly55, Asp87, Ser101, and Gly102. Ser76 had no non-sequential NOE's. The residues with the most inter-residue NOE's were Trp59, which had 66, and Trp 118, which had 53. Roughly equal numbers of inter-residue NOE's occur in both modules. A total of 27 NOE's were identified between the body of module 2 (ie, up to and including Cys 66) and the linker (residues 67 to 70), while 13 NOEs connected the linker and the body of module 3 (i.e Cys 71 and subsequent residues). Of these 40 NOEs, 20 were unambiguous and 20 were ambiguous but contributed >50% to the intensity of the cross-peak (calculated within ARIA). Only two NOEs were potentially inter-modular, both of them ambiguous. One (Gly38Hα-Asp91Hβ) contributed less than 5% of the intensity of the ambiguous restraint, and the other (Tyr39Hε-Thr90Hβ) less than 25%. The favoured orientation of the modules (see Figure 6 v) is presumably defined by the 40 NOE's between the bodies of the modules and the linker (figure 6 iv) and connectivities within the linker. There was no significant difference between the structures calculated with, or without, the Tyr 39-Thr90 ambiguous NOE (data not shown).

6.3 The structures of the two CP modules in VCP~2,3

A set of 50 structures that were consistent with the experimentally derived distance and angle restraints (Table 1) were chosen from a total of 100 structures calculated in the eleventh and final iteration of the ARIA protocol. Selection was on the grounds of lowest total energy. Amongst the 50 selected structures, the structures of individual CP modules converged well. The mean root mean square deviation (rmsd) from the average for Cα atoms was 0.56 Å in module 2 and 0.49 Å in module

Table 1 - Structural statistics for the final 50 structures

Total number of NOE restraints	3082
Unambiguous	2195
Ambiguous	887
For Unambiguous NOE's:	
Intra-residue	879
Sequential	548
Short range (i-j<4)	346
Long range (i-j>4)	422
Root Mean Square deviations:	
Module 2 (backbone atoms C^{α}, N, CO)	0.552
Module 3	0.468
Modules 2,3	1.407
Module 2 (C^{α} atoms only)	0.562
Module 3	0.490
Modules 2,3	1.422
Angles (°)	0.422 ± 0.07
Bonds (Å)	0.0028 ± 0.0006
NOE's (Å)	0.023 ± 0.0056
Dihedrals (°)	0.238 ± 0.148
Total Energy (kJmol^{-1})	240.6 ± 99.98

Average intermodular orientations

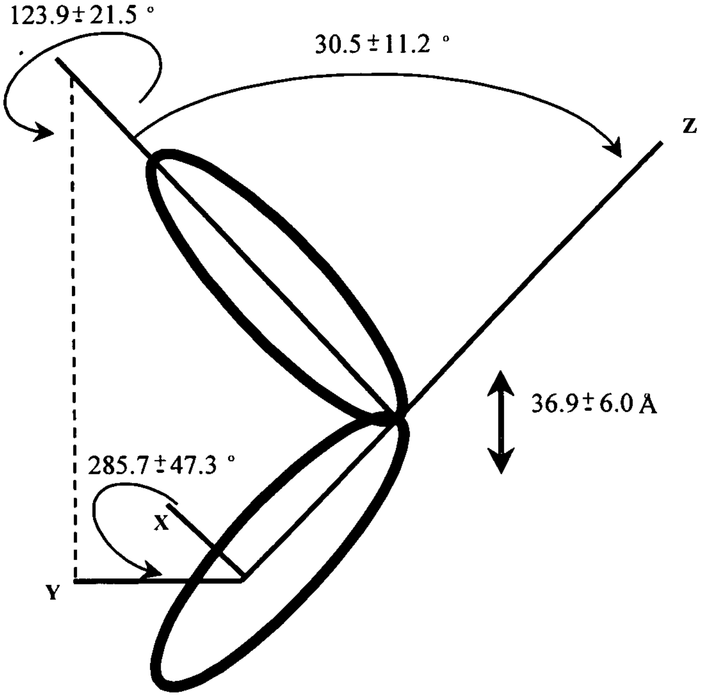

Table 1 - Structural statistics for 50 final structures. Also shown are the intermodular angles with respect to twist, tilt and skew.

3. The region of least convergence was the loop between strands F_2 and G_2 (the F_2G_2 loop), - this is to be expected since these residues have few detectable long-range NOEs (Figure 6 vi) and this loop is the site of a five-residue insertion (Gly52 - Ser56). A second region that exhibits poor convergence is found between strands B_3 and C_3 (Gly84 - Asp87). This corresponds to a region of CP-modules that in general is seen to be poorly conserved, is in many instances a site of insertions and deletions, is variable in terms of its structure, and has previously been termed the "hypervariable loop" (Barlow *et al.*, 1993).

Each of the two modules has a similar overall structure that is becoming recognised as typical of CP modules. Each module has an elongated shape with one long axis and two short ones. Both modules are composed of short β-strands surrounding a hydrophobic core, with the amino- and carboxy-termini lying at opposite ends of the long axis of the module. The β-strands are, in general, aligned approximately with the long axis and form short anti-parallel sheets. The Gly52 - Ser56 insertion extends the F_2G_2 loop at the N-terminal end of the long axis of module 2 by ~8 Å. The hypervariable loop of module 2 is atypically short; in module 3 its length is more typical and it projects an additional 3 Å in a direction perpendicular to the long axis. Consequently module 2 appears more elongated than module 3.

In addition to the four invariant cysteines, the core of each module contains the following buried (or only slightly surface exposed) hydrophobic side chains (conserved or partially conserved residues are underlined) - module 2; Ile15, Leu20, Ile31, Tyr33, Leu41, Trp59, Pro64; module 3; Ile77, Tyr89, Val95, Tyr97, Leu105, Val111, Trp118, Pro122. Tyr39 is partly exposed (Figure 6 vii(a)). Of the non-conserved core residues, Leu20 lies towards the edge of module 2's core and is obscured from solvent by Asp14, while Val111 is buried in the core of module 3 and is replaced by Ser47 (with a potential H-bond between its γ-hydroxyl and the C=O of Phe60) in module 2. In each module the Trp residues, that is an almost invariant component of the consensus sequence of CP-modules, occupies a buried position at one end-of the core directly adjacent to the CysI-III disulphide. Other hydrophobic side chains are mainly exposed at the module surface, these include - module 2; Ile22, Val25, Phe27, Ile42, Tyr48, Leu51, Met57, Val58; module 3; Tyr85, Phe88, Val94, Tyr103, Ile106, Leu112 (Figure 6 vii(b)). Val25 is a partially conserved residue of module 2 that is positioned such that it might form an interface with module 1 within the intact VCP. The equivalent residue in module 3 is Tyr89, which is buried within the lateral bulge of the hypervariable loop. Tyr103 is known to be buried in the 3-4 interface in VCP~3,4 (Wiles *et al.*, 1997).

Of those residues that form the intermodular linker, the side-chain of Val69 is mainly buried within a small hydrophobic pocket that lies between the modules. There appear to be no other hydrophobic side-chains in van der Waals' contact but the CH_2 groups of Gly38, His40, Asp91, Gly115 and Gly116 are near-by.

6.4 Intermodular orientation

The orientation of module 3 relative to module 2 varies amongst the ensemble of 50 calculated structures of VCP~2,3. One way to quantitate this is to calculate an overall mean rmsd for the whole structure instead of calculating values for the individual modules (see previous section). The mean rmsd is 1.42 Å from the co-ordinate averaged structure for backbone atoms in the β-strands of both modules. This relatively low value confirms what is apparent from inspection of Figure 6 v - that there is a distinct, favoured set of

Figure 6 vii - Location of buried and surface exposed residues in VCP~2,3

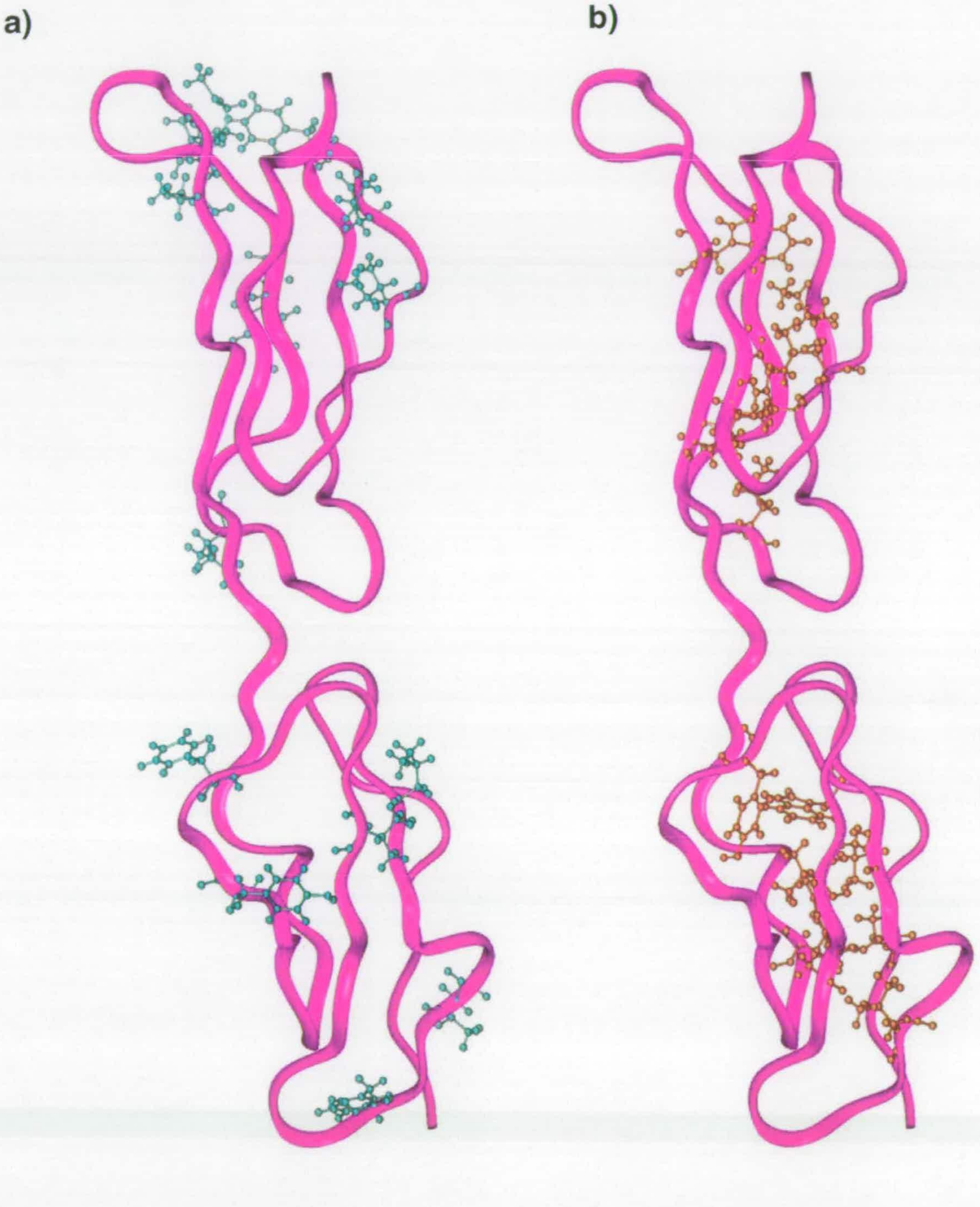

Figure 6 vii - a) Hydrophobic residues that are substantially surface exposed in VCP~2,3 (cyan) b) Hydrophobic residues that are buried (or only slightly surface exposed) in VCP~2,3 (orange).

intermodular orientation amongst the members of the ensemble. A convenient way to summarise the orientation of a module with respect to its neighbour, where the two modules are broadly similar in structure, is to calculate the values of tilt, twist and skew for an ensemble of structures as described previously (Barlow *et al.*, 1993). The results of this procedure for VCP~2,3 are incorporated in Table 1. Each structure within the ensemble was calculated separately, starting from a random structure. There was no significant difference between the structures calculated with, or without, the Tyr39-Thr90 ambiguous NOE. Hence, the favoured orientation is presumably defined by the 40 NOEs between the bodies of the modules and the linker (Figure 6 iii) and connectivities within the linker.

The sample of VCP~2,3 prepared as part of the current study was also used in a set of biophysical experiments carried out by Dr Marina Kirkitadze in this laboratory. These were considered together with the NMR experiments that were carried out over a range of temperatures in order to gain insight into the nature of modular stability and intermodular contacts. In further NMR relaxation studies carried out on this sample of VCP~2,3 by Krystyna Bromek in this laboratory, considerable insight was gained into the flexibility of the 2-3 intermodular junction of VCP. Finally, a sample of VCP~1-4 produced in Kotwal's laboratory from the clone provided by the Barlow laboratory was crystallised by Dr Krishna Murthy of the University of Alabama and the structure was solved (after completion of the NMR studies). All of this work, and its impact on the solution structure of VCP~2,3 described in this Chapter, is discussed in Chapter 7.

References

Barlow, P. N. & Campbell, I. D. (1994) *Method Enzymol.* **239,** 464-485

Barlow, P. N., Steinkasserer, A., Norman, D. G., Kieffer, B., Wiles, A. P., Sim R. B. & Campbell, I. D (1993). *J. Mol. Biol* **232** 268-284.

Bork, P., Downing, A. K., Keiffer, B. & Campbell, I. D. (1996). *Quart. Rev. Biophys.* **29**, 119-167

Brunger, A. (1992). *X-PLOR Version 3.1: A system for X-ray crystallography and NMR,* Yale University Press, New Haven and London

Casasnovas, J. M., Larvie, M. & Stehle, T. (1999). *EMBO J*, **18,** 2911-2922.

Kirkitadze, M. D., Henderson, C., Price, N. C., Kelly, S. M., Mullin, N. P., Parkinson, J., Dryden, D. T. F. & Barlow, P. N. (1999b)*Biochem. J.* **343**, 167-175

Kraulis, P. J. (1991). *J. Appl. Crystallog.* **24**, 946-950

Nilges, M., Macias, M. J., O'Donaghue, S. I. & Oschkinat, H (1997) *J. Mol. Biol.* **269**, 408-422

Wiles, A. P., Shaw, G., Bright, J., Perczel, A., Campbell, I. D. and Barlow, P. N. (1997). *J. Mol. Biol.* **272**, 253-265.

Wishart, D. S., Sykes, B. D. & Richards, F. M. (1992) *Biochemistry* **31**, 1647-1651

Wuthrich, K. (1986). *NMR of Proteins and Nucleic acids*, John Wiley, New York

Chapter 7 - Discussion

7.1 Protein expression

Pichia pastoris was successfully engineered to produce a KM71 cell line with a mut^s/his$^+$ phenotype that expressed recombinant VCP~2,3 in milligram quantities. Two plasmids, pPICZαA and pPIC9K, were incorporated into the yeast genome with the gene encoding VCP~2,3 cloned into the pPICZαA plasmid. The plasmid pPIC9K allowed the yeast to metabolise histidine intracellularly, since the KM71 strain has a deficient histidinol dehydrogenase gene *(his4)*, which prevents the cells from synthesising histidine. Expression of VCP~2,3 was induced by providing methanol as the sole carbon source.

Initially, two different cell lines (KM71 and GS115) had been transformed with the pPICZα/VCP~2,3 construct. The GS115 strain had previously been used in studies of VCP modules 3 and 4 (VCP~3,4) and successfully expressed the two CP modules at yields of 2 mg/l, using a pPIC9/VCP~3,4 construct (Wiles *et al.*, 1997). The KM71 mut^s strain expressed VCP~2,3 in higher yields throughout the test inductions and emerged as the clear favourite.

One disadvantage of using the plasmid pPICZαA as the vector for transformation into *P.pastoris* was the need for further transformation of pPIC9K to confer the his$^+$ phenotype. With this in mind, use of a plasmid that confers the his$^+$ phenotype in the initial transformations would be a logical choice for producing protein with isotopic labels. Positively transformed cells could have been selected on their ability to grow in a histidine-deficient medium.

VCP~2,3 was expressed in yields of ~25 mg/l thanks to isolation of a 'jackpot' clone of KM71. The increased level of expression for VCP~2,3 compared with those reported in the literature for VCP~3,4 is attributable to multiple integration events in VCP~2,3 production. On the other hand, KM71 has been shown in previous studies to be better at producing foreign proteins than wild-type strains that lack the mutation in the *AOX1* gene (Tschopp *et al.*, 1987, Cregg *et al.*, 1987, Chiruvolu *et al.*, 1997). Use of the KM71 strain requires less growth media during the induction stages, which is more cost-effective if producing isotopically labelled protein for NMR studies. Indeed, levels of protein production of 400 mg/l for hepatitis surface antigen produced in *P. pastoris,* (Chiou *et al.*, 1997) and 930 mg/l for bovine pancreatic trypsin inhibitor (Vedvick *et al.*, 1991), are examples of the very high levels of expression in *P. pastoris* that have been reported in the literature. This expression system has now been used to produce recombinantly, proteins from bacteria, fungi, protists, plants, invertebrates and humans.

Production of CP modules was originally accomplished using a *Saccharomyces cerevisiae* system. The published yields of protein using this yeast could be very low e.g. only around 0.1 mg/l growth in the case of factor H modules 15 and 16 (Barlow et al., 1993). *P. pastoris* appears to represent a significant improvement although expression yields of the VCP~2,3 CP module pair are relatively high compared with other examples of CP modules expressed in this system. In this laboratory, MCP~1 was also expressed in *P. pastoris* using a KM71/pPIC9 expression system that yielded 5-7 mg/l of the recombinant protein (O'Leary, 2000). However attempts to produce useful levels of MCP~1,2 failed. Similarly C4bBP~1,2 did not express well in *P. pastoris*, although yields were more respectable when a fermentor rather than shake-flasks was used. Expression of

VCP~1-4 gave modest yields – this protein was subsequently purified and crystallised (Murthy et al, 2001). The CP module-pair from the GABA$_b$ receptor was expressed at levels of 12 mg/l using the same expression system as used for VCP~2,3. Module 2 from the GABA$_b$ receptor pair was expressed individually in yields of 20 mg/l. In other laboratories,. *P. pastoris* has been used successfully to express CP modules 15-17 of CR1 (glycosylation sites deleted), with yields of greater than 30 mg/l obtained using the pPIC9 plasmid (Kirkitadze *et al.*, 1999a). On the other hand the same group has experienced difficulty with expression modules 1-3 of CR1 in the same system (Malgorzata Krych, personal communication). The four CP-modules of DAF and first two CP modules of CR2 have also been expressed at useful levels in *P. pastoris* (Lin et al., 2001, Szakonyi *et al.*, 2001). The *P. pastoris* expression system, therefore, seems well suited to the recombinant production of CP module-containing protein fragments.

E. coli has also been used for CP module over-expression. For example, the N-terminal CP module of membrane co-factor protein, MCP~1, with a GST fusion protein attached to aid purification was expressed in this laboratory at levels of only ~1 mg/l. Much greater success was achieved with modules 1-3 of CR1 and (recently) modules 1-4 of DAF (now crystallised, Susan Lea, Oxford, personal communication). In these cases special re-folding protocols have been devised since in general *E. coli* is not nowadays the vector of first choice for the expression of proteins containing disulphides (VCP~2,3 has four). Finally CHO cells have proved effective for production of CP module-containing proteins that are intended for crystallography, examples being MCP~1,2 (Casasnovas et al., 1999) & β2GPI (Bouma et al., 1999).

P. pastoris has a tendency to hyperglycosylate and sugars must be removed enzymatically, or N-glycosylation sites engineered out. Unlike the mammalian CCP-containing proteins, however, VCP has no N-glycosylation sites and glycosylation was therefore not a problem in this particular case. Another potential problem of using the *Pichia* system for CP module expression is the tendency for non-covalently bound carbohydrate to co-elute with the protein during purification. These carbohydrates are secreted from *P. pastoris* and proved to be a problem for most of the recombinantly produced batches of CP modules produced in this laboratory. Subsequent purification protocols using a heparin column as the main purification step helped to solve the problem (Smith *et al.*, 2000).

In summary, a good choice for future production of CP module-containing proteins would be *P. pastoris* KM71 cells transformed with a pPIC9 plasmid containing the protein of interest and, if possible, purified using a heparin column or other affinity method. In general, experience has shown that N-glycosylation sites are best removed by mutaganesis provided they can be shown to be physiologically irrelevant.

7.2 NMR data and assignment of VCP~2,3

A sample of VCP~2,3 in a buffer of 10 mM sodium phosphate, pH 6.0 was used for all NMR experiments and resulted in good quality NMR spectra, with many dispersed resonances. The concentration of the protein was ~2.5 mM for all 2D and 3D spectra and mixing times of 150 ms for NOESY spectra and 63 ms for TOCSY spectra proved to be more effective than shorter mixing times of 100 ms for NOESY and 39 ms for TOCSY. Longer mixing times were also used for acquisition of spectra in the VCP~3,4 module pair (Wiles et al., 1997), although in that case a pH of 4.0 was used and some spectra were acquired at 750 MHz rather than the 600 MHz used for

VCP~2,3. In the structure determination of Factor H modules 15 and 16 (fH~15,16) somewhat longer mixing times of 75 ms for the TOCSY and 250 ms for the NOESY spectra were used – spin-diffusion is likely to occur under these circumstance. fH15-16, which was not isotopically labelled due to the very low expression yields. Instead, due to overlap in the 2D spectra, structures of fH15 and fH16 were solved individually before the module pair was fully assigned. Although this would have been a time consuming step, it might have proved a useful strategy in the case of VCP~2,3, if the individual modules had been available - there was significant overlap in most regions of the 2D NOESY spectra of VCP~2,3, and even some overlap in the 3D NOESY spectrum, which required painstaking analysis. Assignment of the proline residues in particular proved more difficult and would have been quicker had the spectra of individual modules been available.

Despite the overlap mentioned in the previous paragraph, it proved not necessary to double-label VCP~2,3 with ^{15}N and ^{13}C as the NMR spectra contained many dispersed peaks that were sufficient for assignment and structure calculation. On the other hand, double labelling could have speeded up the whole process and improved the number and quality of experimental restraints.

The ^{1}H-^{15}N 3D spectra of VCP~2,3 were particularly useful in the initial stages of assignment, as was the program ANSIG, which allowed connectivities between sequential residues to be assigned relatively quickly. The NMR spectra acquired with ^{15}N labelled VCP~2,3 provided a means of assigning cross-peaks that were difficult to resolve using the ^{1}H-^{1}H spectra alone. The $^{2}H_2O$ spectra were particularly useful for aromatic assignment, as there was less overlap in the aromatic region due to the absence of signals originating from side chain NH protons. The $^{2}H_2O$ and $J_{hnh\alpha}$ spectra proved useful for confirmation of Hα assignments that were very close to the water signal and were not readily assignable in the 2D and 3D spectra. The $J_{hnh\alpha}$ data also highlighted a mistake in assignment of one of the serine residues, for which an Hβ proton had been mis-assigned as an Hα proton due to a low chemical shift for the Hβ. A series of scripts was used to convert the exported cross-peak lists form ANSIG into chemical shift lists for each NOESY spectrum, followed by generation of restraint tables. Analysis of the output of these scripts, however, highlighted the necessity of editing the automatically picked cross-peaks in ANSIG, many of which had errors in the peak position - these were corrected manually.

7.3 Brief review of relaxation studies of VCP~2,3 and intermodular flexibility

A major series of relaxation studies were preformed on the 2.5 mM ^{15}N-labelled VCP ~2,3 sample that had been used for the structure determination. These studies were performed by Krystyna Bromek and are described in detail in a recent paper (Henderson *et al.*, 2001). It is instructive at this point to review the relaxation studies as a preamble to discussion of the solution structure in the next section.

It has been established that multiple modules are required for functional activity in VCP and in other RCA proteins - thus the dynamical and orientational relationship between modules is of great interest. To properly characterise the solution structure of VCP~2,3 it is therefore important to determine whether the range of conformations represented amongst the ensemble of 50 calculated structures genuinely reflects the situation in solution. While poor convergence amongst the calculated structures might result from flexibility, it could instead be a consequence of a lack of experimentally-derived restraints. On the other hand, the spectrum of intermodular orientations inferred from the structure calculation might comprise just a subset of a much more diverse collection of conformations that are exchanging rapidly compared to the build-up time (50-150 ms) of the ^{1}H-^{1}H NOE.

The ^{15}N T$_1$ relaxation rate and heteronuclear NOE of the backbone ^{1}H-^{15}N pair are sensitive to motion on the rapid, 10^{-12} s to 10^{-9} s, time-scale; in addition, chemical exchange can be sampled on the 10^{-6} s - 10^{-3} s time-scale as an $R_{ex} > 0$ addition to T_2 [$(T_2 = (R_2 + R_{ex})^{-1}$] (Lipari & Szabo, 1982). The exchange of amide protons with solvent water is a slower process that reflects 10^{1} - 10^{-3} s^{-1} motion. Most importantly for the purposes of this study, it is possible to characterise the anisotropy of rotational diffusion from ^{15}N relaxation studies. This is a function of the time-averaged shape of the protein, which in turn reflects the mobility between modules on a time-scale slower than the overall correlation time.

Based on the assignment of backbone ^{15}N and ^{1}H nuclei described in Chapter 4, the ^{15}N relaxation rates for a large number of residues in VCP~2,3 are given in Henderson *et al.*, 2001. The relaxation properties of some individual residues are mentioned below in the discussion of the structure of VCP~2,3. The most important result from the point of view of characterising intermodular motion was obtained by best-fitting a rotational diffusion tensor to the relaxation properties of each individual module treated as a rigid rotor (Dosset *et al.,* 2000). Module 2 was best-fitted using an axially symmetric model (i.e. z > x, x = y) for VCP~2,3 with overall correlation time = 7.3 ns and $D_{\|}/D_{\perp} = 1.6$ – this is the ratio of the long (z) axis of the molecule to its two shorter axes (x,y). This model was shown to be a significant improvement on an isotropic (x = y = z) fit (passed F-test at 5% probability) (Bevington & Robinson, 1992), while no further improvement was observed when fitting a fully anisotropic rotor (x < y < z). Data from module 3, on the other hand, were best fitted to an isotropic diffusion model with a correlation time = 6.6 ns; the best χ^2 fit of an axial rotor yielded $D_{\parallel}/D_{\perp} = 1.1$ and the F-test for this axial ratio failed. It is important to appreciate that although this calculation is based on individual modules, in each case the values obtained relate to the whole molecule. The lower experimental values of $D_{\parallel}/D_{\perp}$ (between 1.1 and 1.6) would therefore suggest strongly that the modules are not arranged in an extended manner (with small tilt angle) as indicated by the structure calculations based on the NMR data. This inconsistency is discussed in the next section.

7.4 The structure and dynamics of VCP~2,3; comparison with other structures

The solution structure of VCP~2,3 reveals two elongated domains fused head-to-tail at a small interface. The consensus cysteines, glycines, and prolines together with 12 out of the 14 consensus hydrophobic residues play important structural roles within the two domains, each of which has a typical CP module-like structure. It appears from an inspection of the exchangeability of backbone amide protons that both modules of VCP~2,3 undergo "breathing" motions on the 10^{-3} - 10^{1} s time-scale. Module 2 in particular is notably lacking in slowly exchanging amides and this is consistent with its melting temperature being lower than that of module 3 (Chapter 5, and Kirkitadze *et al.*, 1999b). An analysis of relaxation data presented in Henderson *et al.*, 2001 revealed more 10^{9}-10^{12} s^{-1} motion in module 3 than in module 2. Despite rather low levels of sequence identity, both modules superimpose well with each of the other 12 CP-module structures that had been published by the beginning of 2001. (see Table 1; <rmsd> for pair-wise superpositions = 0.9 Å ± 0.4 Å (s.d.) for the Cαs and Cβs of the invariant Cys and Trp residues). VCP~2, like fH~16, MCP~1 and the modules of β_2GPI has only 6 β-stands and lacks strands A_2 and C_2. These strands normally run anti parallel to each other and form the first β-sheet near the N-terminus. All other CP modules studied to date have 7 or 8 β-strands. Module 2 is most similar to MCP~1 in these comparisons (0.49 A) and least similar to fH~5 (2.02 A). Module 3 deviates the least from B_2GPI~3 and is most dissimilar to MCP~2.

The structure of module 3 in the context of VCP~2,3 has small but significant differences in structure compared to its structure in the context of VCP~3,4 (Figure 7 i). The pair-wise rmsd for all backbone atoms from Cys71 (Cys3 in VCP~3,4) to Cys124 (Cys56 in VCP~2,3), inclusive, is 1.45 Å (based on a comparison of lowest energy structures from the two ensembles). The D_3E_3 turn (Asn100-Tyr103) appears to have experienced a hinge-like movement resulting in displacement by up to 3 Å of the Gly102 backbone. At the other end of strand E_3 there are differences in the position of the backbone of residues Ile106-Asn108 by up to 3.5 Å. All of these residues are proximal to the interface with module 4. For example, Ile106 (Leu38 in VCP~3,4) is a key interface residue, as is Tyr103 (Tyr35). Several of them (Tyr103, Ile106, Gly107 and Asn108) exhibit chemical shift differences between VCP~2,3 and VCP~3,4 ($\Delta\delta^{H\alpha}$ > 0.1 ppm) (figure 7 ii). Residues 101 and 107 also have significantly higher ^{15}N T_1s when in VCP~2,3 compared to VCP~3,4 - indeed Gly107 has the highest T_1 amongst residues of module 3 in VCP~3,4. At the opposite end of the module, near the amino terminus, Asp91 in the mobile C_3D_3 turn and Gly115 in the F_3G_3 turn differ in

Table I - RMSD comparisons of CP modules

Module	VCP module 2 (rmsd)*	VCP module 3 (rmsd)*
fH 5	2.02	1.51
fH 15	1.57	1.18
fH 16	1.41	1.35
VCP2	-	0.93
VCP3#	1.32	0.90
VCP3	0.93	-
VCP4	1.84	1.48
MCP1	0.49	1.02
MCP2	1.59	1.44
B_2GPI 1	0.96	1.18
B_2GPI 2	1.61	1.30
B_2GPI 3	0.94	0.93
B_2GPI 4	1.33	1.16
C1s	1.16	1.11

*rmsd based on a pair wise comparison using Cα, Cβ and HN atoms of conserved Cys and Trp residues

#Module 3 from VCP~3,4 as described by Wiles *et al.*

Table I - RMSD comparisons for VCP modules 2 and 3 with other CP modules.

Figure 7 i - Differences between VCP~3 from VCP~2,3 and VCP~3,4

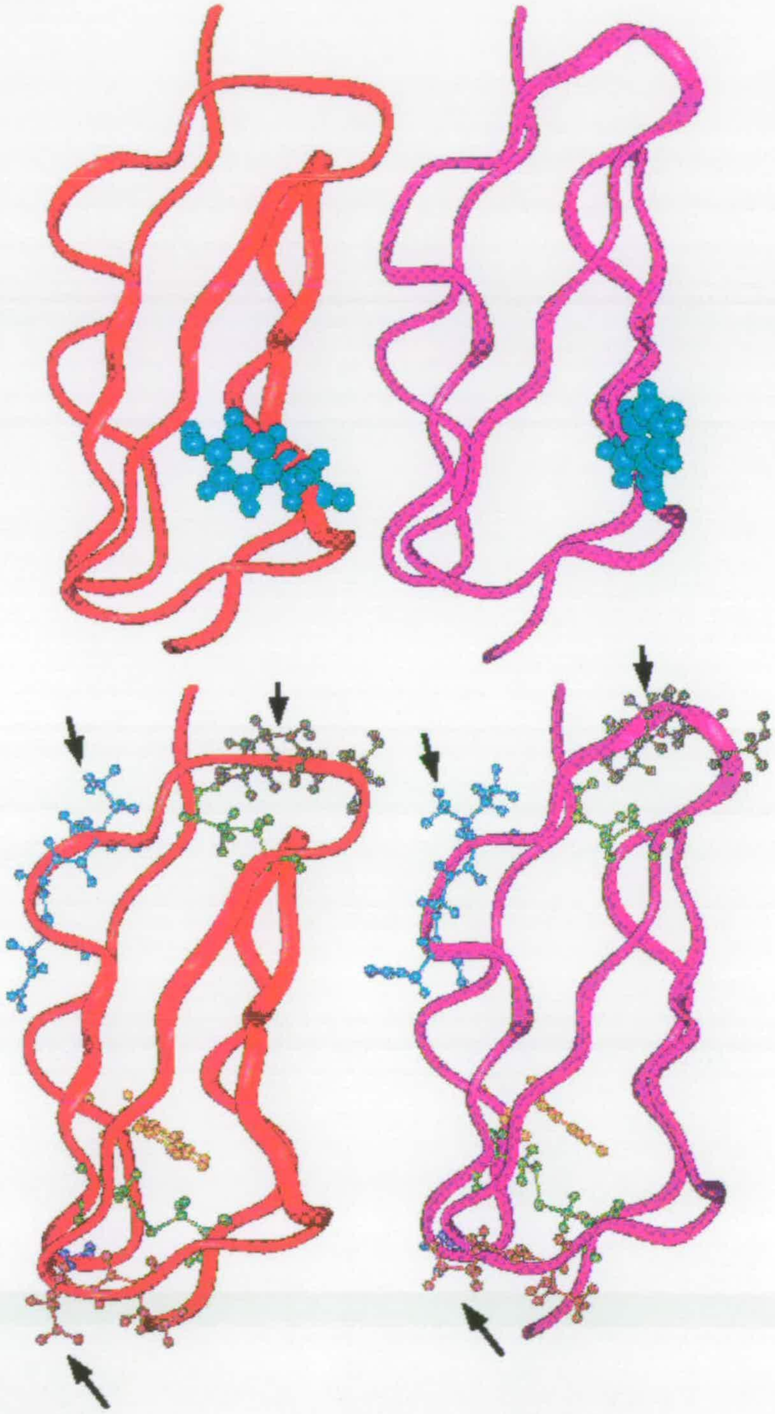

Figure 7 i - ribbon representation of VCP~3 from VCP~2,3 (red) and VCP~3,4 (purple) showing the different conformations of Tyr 85 in the hypervariable loop. Arrows indicate those residues that are involved in 2,3 or 3,4 interfaces and have different conformations. Cys residues, green;Trp, orange;residues 101-103, black,; residues 106-108, light blue; residues 90, 91, brown; Gly115, dark blue.

Figure 7 ii - Chemical shift differences between module 3 from VCP~2,3 and VCP~3,4

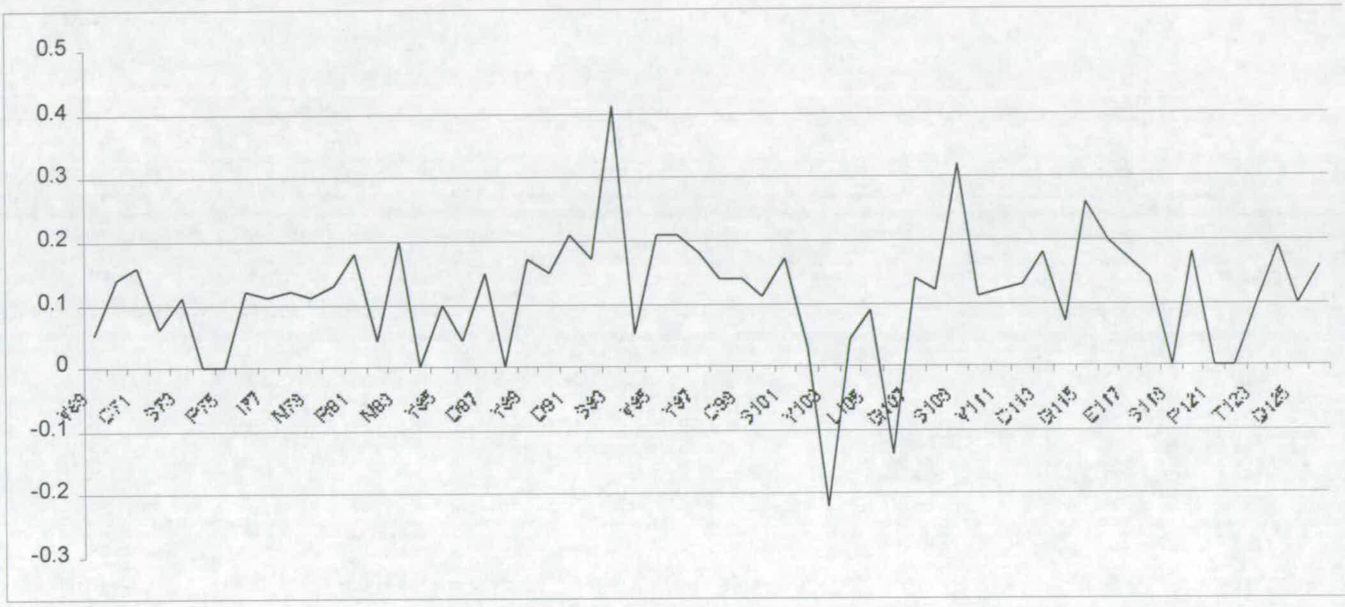

Figure 7 ii - Chemical shift differences between module 3 assignments from VCP~2,3 and VCP~3,4. X-axis values were calculated as module 3 (VCP~3,4) Hα shift - module 3 (VCP~2,3) Hα shift.

position by 1.5-2.0 Å. These residues are close to the small interface with module 2; the $\delta^{H\alpha}$ of Gly115 has changed (~0.1 ppm) compared to its chemical shift in the context of VCP~3,4. These structural differences are not just an artefact of the structure calculation protocols - numerous NOEs found in the spectra of VCP~2,3 and assigned unambiguously to inter-residue connectivities within module 3 were absent in the VCP~3,4 spectra; the relevant distance restraints were also violated by the structure of VCP~3,4. The converse was also observed - *i.e.* module 3 NOEs occurred in the spectra of VCP~3,4 that were absent in the spectra of VCP~2,3.

As might be expected the major differences in structure between module 3 from VCP~2,3 and VCP~3,4 occur close to the intermodular interfaces. When the residues concerned (91-92,101-103,106-108,115) are excluded from the comparison, the pair-wise rmsd (backbone atoms) decreases to 1.16 Å (*i.e.* based on a comparison of the lowest energy structures from each of the VCP~2,3 and VCP~3,4 ensembles). In the case of Tyr 85 (figure 7 i) the difference is probably due to additional assignments in VCP~2,3 of Gly 84. Elsewhere, however, the differences are probably too great to be attributable to experimental uncertainty, meaning that changes of structure near the interface must somehow be propagated into the body of the module. That the presence or absence of neighbouring modules has a small but significant affect upon a module's 3D structure is in agreement with previously published biophysical studies of CP modules (Kirkitadze *et al.*, 1999a,b,c,d) that demonstrated a context-dependency of melting temperatures, and enthalpies of unfolding, of individual modules. These studies also showed co-operativity of melting amongst linked modules. On the other hand, an earlier study of the 15-16 CP-module pair from factor H (Barlow *et al.*, 1993) revealed no such structural changes - this study however was based on a smaller data set since the protein was not ^{15}N-labelled, and in any case these modules are not tightly interfaced as are modules 3 and 4 of VCP.

Module 2 is more elongated than module 3 due to the five-residue insertion in the F_2G_2 loop. This lies close to the amino terminus of module 2 and it was speculated that it could conceivably interact with module 1 in VCP (subsequently confirmed by the crystal structure (Murthy *et al.* 2001)). These five residues have few long-range NOEs, and are very mobile on the 10^{-12}-10^{-9} s time-scale. The loop has the sequence LGSTGSM, and its equivalents are GG in module 3 of VCP, PGN in VCP module 4 and GT in VCP module 1. Both Leu (51) and Met (57) side-chains are solvent-exposed in VCP~2,3. This loop is conserved amongst the second CP modules of other poxvirus complement control proteins but insertions of similar length are found in few human CP-modules. Examples of note are the second (and the third) modules of C4bBP, chain α, and the second module of MCP. In the crystal structure of MCP modules 1 and 2 (MCP~1,2) this loop (sequence LKGSVA) has a similar conformation to that of VCP module 2, but has swung in towards the body of the module. The Leu (109) and Val (114) of MCP are solvent-exposed and there is no contact with MCP module 1, which is tilted away from the loop. Module 2 of VCP also appears elongated due to a deletion, with respect to nearly all other CP-modules, in the hypervariable region. As a result, unlike other CP-modules whose structures have been solved previously, module 2 has no lateral bulge here. Consequently, the semi-conserved Val25 is solvent exposed rather than buried in the bulge like, for example, Tyr85 of VCP module 3 and Tyr83 of MCP module 2. Gly24 (in the hypervariable region), Val25 and Asp27 each undergo motion on the fast time-scale and Val25 experiences additional 10^{6}-10^{3} s^{-1} motion. Residues 25 and 27 of VCP module 2 correspond to residues that form β-strand C in module 3 but they do not form a β-strand in module 2. The equivalent region of module 3 (residues 85-90) is also highly mobile on the 10^{-12} - 10^{-9} s time-scale.

The "top" of module 2 – *i.e.* the region that comes closest to module 3 - is similar in structure and sequence to the "top" of module 3. Residues 35-42, which have the sequence CNSGYHLI and correspond to the D_2E_2 loop and strand E_2, and residues 66-67 (the fourth Cys and the first residue of the linker, Glu) of VCP~2,3 overlay on the equivalent ten residues of VCP~3,4 (CNSGY*S*LI + CQ) with an rmsd of 0.92 Å (Cα and Cβ atoms). The backbone at the "bottom" of module 3 follows a similar course to its equivalent in module 4 - thus the Cαs of 70,71 (last residue of linker and first Cys), 89-92 (last residue of β-strand C_3 and the C_3D_3 loop), and 113-118 (from strand F_3 into strand G_3) overlay on the equivalent residues of VCP~3,4 with an rmsd of 0.88 Å. The sequences, however, are different; in particular the key interface residues of module 4 (Pro106 and Tyr81, VCP~3,4 numbering) are replaced by a Gly and an Asp in module 3. Another key interface hydrophobic side chain of the 3,4 junction that is lost in VCP~2,3 is that of the linker residue Ile58 (VCP~3,4 numbering) that is replaced by Ser. The principal hydrophobic content of the small 2,3 interface is provided by the buried side chain Val69 (equivalent to Val59 of VCP~3,4). It is therefore the replacement of three hydrophobic side chains (Pro, Tyr and Ile) by smaller or polar ones (Gly, Asp, Ser) that appears to be responsible for the dearth of intermodular contacts between modules 2 and 3 compared to the 3,4 situation. A lack of well-defined structure in the 2,3 junction is also indicated by the high mobility of β-strand C_3 (88-90), and the C_3D_3 loop (91-93) on both the fast (88, 90, 91 and 93) and chemical exchange (91 and 93) time scales, and the high backbone ^{15}N T_1 of linker residue Lys70.

A comparable dearth of contacts were detectable between the 15th and 16th modules of factor H (Barlow *et al*, 1993), and there are few contacts between the first and second modules of MCP according to the crystal structure (Casasnovas *et al.*, 1999). In the crystal structure of β_2-glycoprotein I (β_2GPI) (Bouma *et al.*, 1999) there are only limited hydrophobic contacts and a single hydrogen-bond within each of the three junctions between the four adjacent classical CP-modules - to take the most extreme example, only 242 $Å^2$ of surface area is buried between β_2GPI modules 2 and 3. Despite the scarcity of forces stabilising intermodular junctions, both MCP~1,2 and β_2GPI have been crystallised and display quite well defined intermodular orientations. Structures of VCP~2,3 calculated on the basis of NOEs also converge to a limited range of intermodular conformations despite the fact that only 195 $Å^2$ of surface area is buried in the lowest energy structure. In all of these cases it is possible that the actual situation, in solution, differs from the outcome of the structure determination - just a few out of many conformations of MCP~1,2 may have been selectively crystallised for example, or intermodular movement on a time-scale rapid with respect to the build up of the ^{1}H-^{1}H NOE might not be apparent in the NMR-derived structures.

It was therefore considered important to exploit the ability of NMR to provide insights into intermodular motion. This was done, according to the method recently reported by Copie *et al.* (1998), which involved comparing the anisotropy of rotational diffusion derived from an analysis of relaxation data for VCP~2,3 on the one hand, and from the calculated structure of VCP~2,3 on the other. Despite the fact that analysis of ^{15}N relaxation times was made more complicated by the disparities between the internal fast motions of the two modules, it was obvious that the value of $D_{\|}/D_{\perp}$ (~1.6) obtained (based on module 2) was too small to be consistent with the extended head-to-tail conformation indicated by the structure calculation. To explain this, it is necessary to invoke mobility between the modules and therefore attention focuses on the intermodular linking sequences. Since linker residues show no evidence of mobility on the chemical exchange time-scale, and little evidence of 10^9-10^{12} s^{-1} motion (in Val69 and Lys70), any intermodular motion is likely to be on a time-scale that is faster than chemical exchange (10^6 s^{-1}) but slower than the overall correlation time (~10^{-8} s). According to this model, the NOEs that promote convergence of the structures during simulated annealing – and it is important to remember that all of these are between the bodies of the modules and the linker or within the linker, rather than between modules – represent, in fact, average distances between two mobile segments of the molecule that share only a small interface. It is impossible to distinguish, on the basis of an NOE, between the case of two protons that spend only a fraction of time relatively close together in space (e.g. 3.3 Å apart), and a pair of protons that are at a fixed but longer distance (e.g. 5 Å). Hence, to explain the $D_{\|}/D_{\perp}$ ratio obtained from the relaxation studies it is necessary to invoke a large number of interconverting orientations, many of which are kinked (large tilt angles), whose time-averaged structure is more compact than the calculated structures (figure 7 iii). It is interesting to discuss how these findings fit with other biophysical studies of VCP.

Figure 7 iii - Differences in time-averaged structures

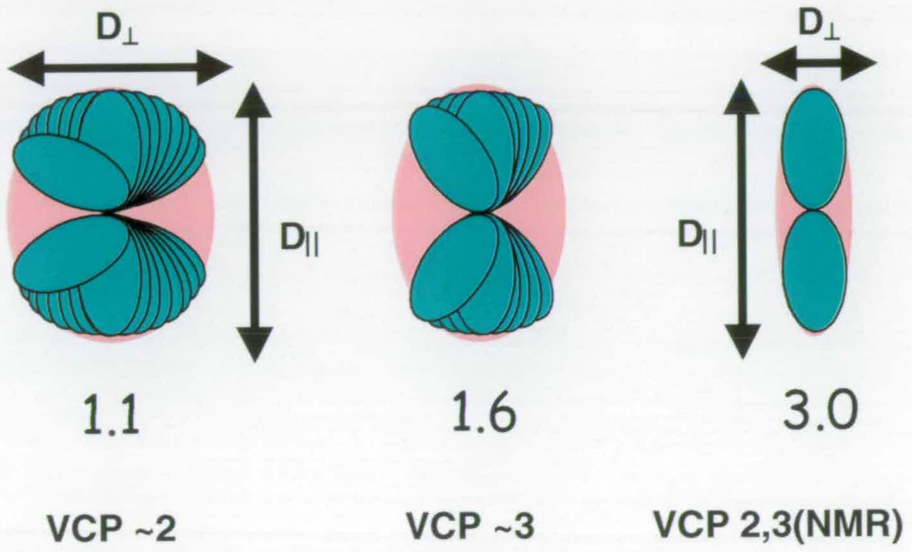

Figure 7 iii - A flexible junction between modules 2 and 3 results in a more compact time averaged structure. The number below each illustration is the axial ratio ($D_{||}/D_{\perp}$).

7.5 Biophysical studies of VCP~2,3

Using the sample prepared for the NMR studies, a series of biophysical measurements were made on VCP~2,3, and these are discussed in detail in a publication by Kirkitadze *et al.* (1999b). The calorimetric profile of VCP~2,3 showed a melting transition at ~ 57 °C and a calorimetric enthalpy (86 kcal/mol) consistent with each of the two modules having a compact fold. The ratio of calorimetric and van't Hoff enthalpies was greater than one, together with a distinct shoulder on the main peak within the profile, indicated that melting of CCP modules 2 and 3 occur at similar, but not identical temperatures. In addition to the major 57 °C transition, a minor transition in the 30-40 °C range was visible in the calorimetric profile. In repeated scans, this appeared either as a discrete peak or as a shoulder on the edge of the main feature. As described in Chapter 4, heteronuclear NMR revealed the melting pathway of VCP~2,3. Between 25 and 40 °C there was very little change in the chemical shifts of amide protons and ^{15}N nuclei implying that negligible unfolding of the two modules occurs over this temperature range. Between 40 and 45 °C, resonances that had been assigned to module 2 lost intensity while new cross-peaks consistent with the presence of random-coil structure emerged. As the temperature increased above 45 °C, signals arising from module 2 faded rapidly, while module 3 resonances remained largely unperturbed even at 52 °C. These data indicate that module 2 is less stable to temperature than module 3 within the context of the 2,3 fragment and are in agreement with experiments using chemical denaturants, and with the greater abundance of slowly (> 1 hour) exchanging amide protons in module 3 compared to module 2 described above.

Since neither module unfolds below 40 °C, the minor calorimetric feature in the 30-40 °C range could be assigned to the small interface between modules 2 and 3 of VCP, as also seen from the NMR studies. This means that unless the presence of modules 1 and/or 4 stabilise the 2-3 junction within the context of the intact VCP molecule, then it will be partially melted at physiological temperature and will act as a flexible hinge at the centre of the molecule. Ultracentrifugation studies conducted at 4 °C revealed the 2,3 module pair to be highly asymmetric with an axial ratio of approximately 5.3:1. Thus below the putative melting point of the junction, modules 2 and 3 seem to adopt a very extended head-to-tail arrangement with a very limited intermodular junction. According to this interpretation, melting of the junction between modules 2 and 3 of VCP takes place at a lower temperature, and with an apparently smaller change in enthalpy, than the junction between the homologous modules 16 and 17 of CR1 for which a similar set of biophysical measurements have been made (Kirkitadze *et al.*, 1999c).

Hence the biophysical data strongly support the idea of a flexible hinge between modules 2 and 3 of VCP at 37 °C as suggested by the NMR studies. As the temperature is lowered, Krystyna Bromek showed that VCP~2,3 becomes progressively more elongated (unpublished data) presumably due to the stabilisation of the 2-3 intermodular junction below its calorimetrically determined melting point (30-40 °C); and this agrees with the ultracentrifugation data at 4 °C.

7.6 Comparison of NMR and crystal structures

It proved possible (after the work described in this thesis was completed but before it was published) to produce crystals of intact VCP (recombinant, fully functional material produced from a *P. pastoris* construct prepared in this laboratory by Dr. Nick Mullin) of sufficient quality to solve its structure at 2.2 Å resolution (Murthy *et al.*, 2001, Figure 7 iv). Crystallisation took place at 20 °C. This is the first 3D structure of any intact RCA protein and it revealed for the first time the structures of module 1 of VCP and the 1-2 intermodular junction. Under the conditions used for crystallisation, all the intermodular junctions of VCP appear to be "frozen" since the four CCP modules are arranged in a nearly identical manner within the unit cells of two different crystal forms and in a total of five crystallographically independent examples of the molecule. The structure (Figure 7 iv) may be regarded as consisting of two pairs of closely interacting modules (1,2 and 3,4) with very few contacts between the bodies of modules 2 and 3 - in excellent agreement with previous results. The 2-3 junction of intact VCP, however, is different from the junction in VCP~2,3 calculated from NOEs; on the other hand the 3-4 junction of intact VCP is very similar to that of VCP~3,4 but somewhat more tilted (Figure 7 iv). It is interesting to note that the structure of module 2 in the crystal has some significant differences from its structure in VCP~2,3. Of special note, the Leu109-Ser114 insertion in module 2 (see Figures 6 i and 7 iv) is a mobile and unstructured loop in VCP~2,3 but has moved by some 15 Å in the crystal structure to contact the Gln42, Lys43 insertion of module 1, and stabilise the 1-2 intermodular interface (Figure 7 iv). Figure 7 v shows a representation of the crystal structure of VCP with regard to twist, tilt and skew angles (see Chapter 1 for definitions). It is not known, but is a matter of interesting speculation, whether module 2 of VCP~2,3 is in fact destabilised by the lack of module 1. If so, then it is conceivable that the 2-3 junction is also destabilised by the lack of module 1 which in turn implies that the flexibility between these modules may not be as great as indicated in solution studies.

Figure 7 iv - Crystal structure and NMR fragments of VCP

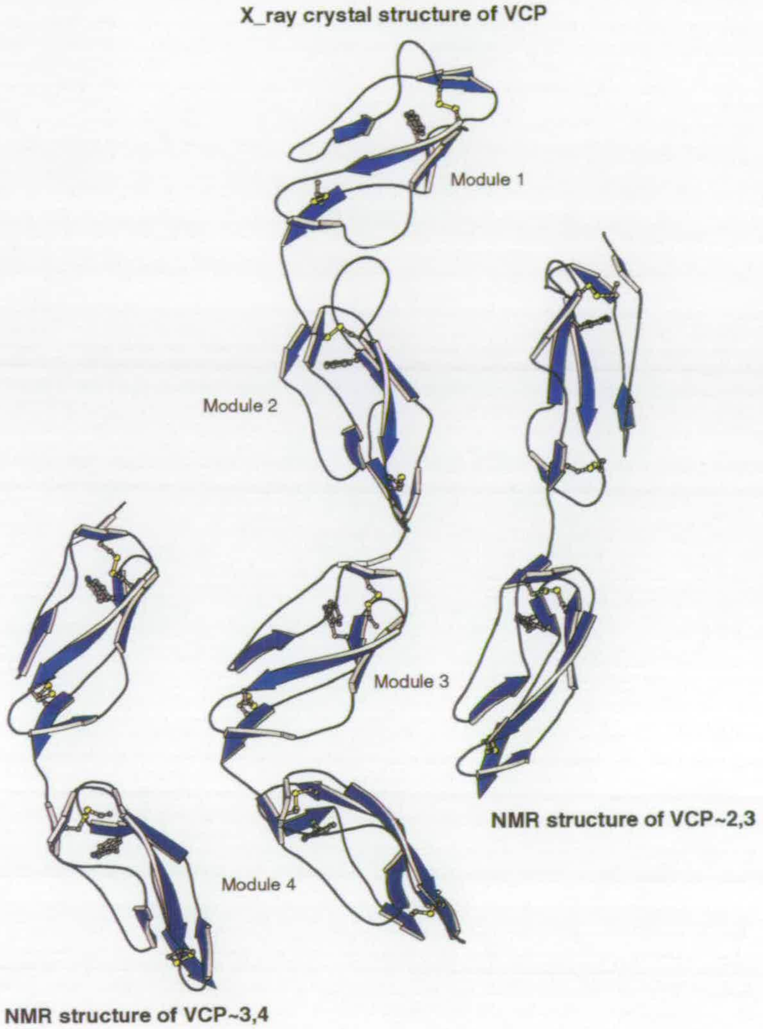

Figure 7 iv - Molscript representations of VCP~3,4 (NMR), intact VCP (X-ray crystallography) and VCP~2,3 (NMR). The alignments represent the best fit orientations adopted for VCP~2,3 and VCP~3,4 when compared with the crystal structure.

Figure 7 v - Intermodular geometry in crystal structure of VCP

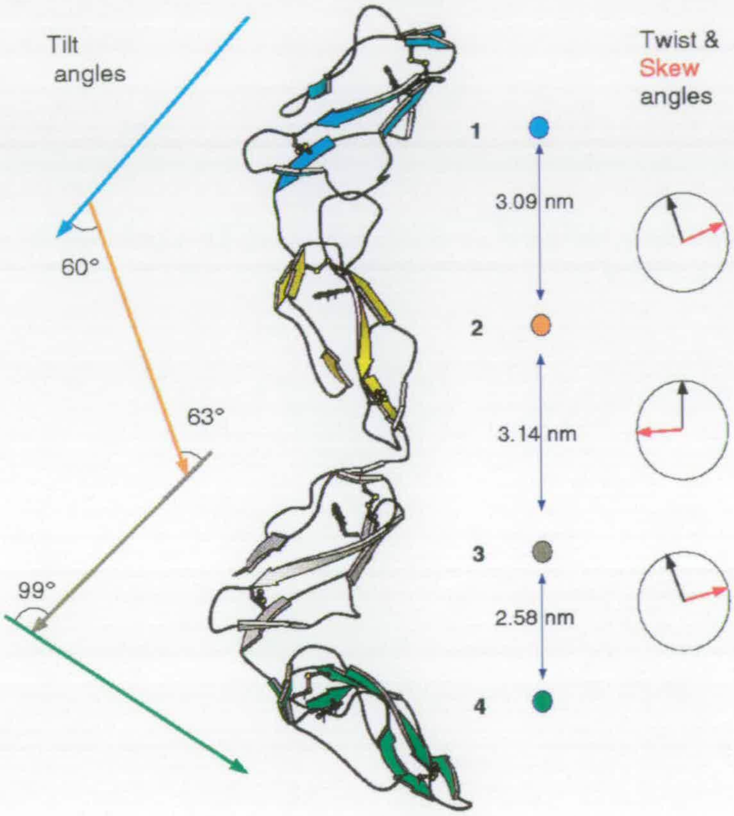

Figure 7 v - Intermodular angles in crystal structure of VCP defined by twist, tilt and skew.

On balance, however, the likelihood is that VCP does indeed have hinge- or pivot-like flexibility as a result of the dearth of contacts between its central modules.

In all cases where examination of module-module interfaces is possible, recurring themes are present. The linker residues almost always have side chains that are entirely or partly hydrophobic, and two or more aromatic or histidine side chains lie in close proximity; one at the second turn towards the C-terminus of the preceding module and one near the first turn towards the N-terminus of the succeeding one. As mentioned previously, the relative orientations of one module with respect to the next is varied, even though the sequences close to the interfaces show a degree of similarity. This indicates that modelling other CP module pairs on the basis of homology could be difficult.

7.7 Recreation of SPICE using VCP~2,3 and VCP~3,4

The sequence of complement control proteins is highly conserved amongst the poxviruses. Comparison of VCP with SPICE (Rosengard & Ahearn, 1998), the complement control protein of Variola virus, indicates 11 substitutions (Figure 7 v) in modules 2-4, while module 1 is identical in the two proteins. SPICE has been reported to have four times the potency of VCP in complement inhibition, although these are preliminary findings and have not yet been published (Rosengard & Ahearn, 1998) and this might assist in the greater virulence of the Variola (small pox) virus. It is believed that loss of certain immunomodulatory components from Variola compared with other poxviruses, however, may be the dominant factor in increased virulence of smallpox. It is therefore of

Figure 7 v - Recreation of SPICE using NMR derived structures of VCP~2,3 and VCP~3,4

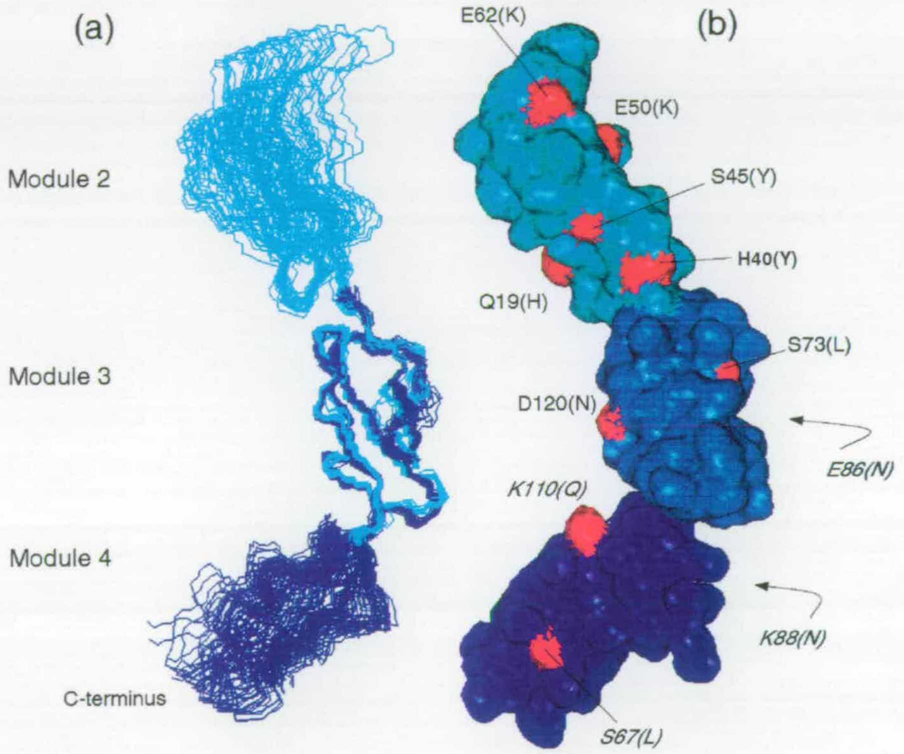

Figure 7 vi - a) shows backbone trace of VCP~2,3 and VCP~3,4 solved by NMR. b) Recreation of possible structure of SPICE. Module 1 of SPICE is unchanged from VCP sequence - residues in SPICE that differ from VCP are annotated and shown in red.

interest to review the locations of the SPICE substitutions in the light of the 3D structures of VCP modules 2-4. Figure 7 vi indicates the positions of the substituted residues on a surface representation of the three modules. This reconstruction was created by overlaying structures of VCP~2,3 and 3,4 using module 3 for superposition. All 11 of the substituted residues are surface exposed - none appear to be of structural importance or (with the possible exception of His40) involved in the intermodular junctions. Nine of the residues concerned are visible in the view of the molecule presented - Glu86 of VCP~2,3 and Lys88 of VCP~3,4 are the exceptions. None of the substitutions are conservative but H40 -> Y, S45 -> Y, S73 -> L and S67 (of VCP~3,4) -> L would add significant hydrophobicity to a surface that already exposes a high proportion of non-polar side chains. Other replacements in SPICE would facilitate a difference in overall surface charge in module 2 due to the addition of two lysine and one histidine residues, replacing three acidic residues (fig 7v). Interestingly, replacing His->Tyr or Tyr->His at an analogous position in two separate CR1 binding regions has been shown to affect the C4b binding (Krych *et al*, 1998).

7.8 - Other CP module studies and concluding remarks

As discussed earlier, comparisons of VCP~2,3 structures with the intact VCP structure (Murthy *et al.*, 2001) and VCP~3,4 (NMR) (Wiles *et al.*, 1997) highlights some of the structural changes that may occur due to the presence of other modules. If change in one module can affect the structure of others, this has potential functional implications. The crystal structure of VCP has two regions that are likely candidates for specific interactions with C3b and C4b. These two sites are large positive charged patches on modules 1 and 4 respectively that could be involved in binding to negatively charged regions on C3b/C4b and/or heparin. Regions of positive charge in CP modules have also been shown to be important in C3b and C4b binding in other mutagenesis studies. In the case of active sites 1 and 2 of CR1 (CP modules 1-3 and 8-10 respectively) for example, mutagenesis of residues with no charge, to basic residues, was shown to increase affinity for both C3b and C4b (Krych *et al.*, 1994, 1998). As mentioned previously, it is tempting to speculate that intact VCP may show functional diversity by virtue of having some flexibility around the 2,3 interface. This would allow the putative ligand-binding areas of VCP to change position relative to each other.

We have seen that VCP~2,3 appears relatively mobile in solution with a possible hinge formed by the lack of a well defined 2,3 interface. Studies of the C4bBP α-chain showed that a region important for binding is a cluster of predominantly basic side-chains at the CP 1, 2 interface (Blom *et al.*, 1999) with a number of R, K and H residues important for binding activity. VCP 2,3 has different, uncharged residues at some of the equivalent interface positions and retains only two of the five charged residues implied in C4bBP binding to its ligand. This indicates that the 2,3 interface is probably not directly involved in specific protein-protein interactions that have a strong dependency on the presence of basic residues.

However, a cautious approach is necessary when considering structure/function relationships - results of recent studies with CP modules 1 and 2 from CR2 bound to C3d have shown the module/ligand interaction is not dependant on clusters of basic residues. The two functional CP modules from CR2 were crystallised bound to the ligand, C3d (Szakonyi *et al.*, 2001). This crystal structure revealed that the modules were arranged side by side, rather than in the head to tail arrangement observed in all other structural studies of CP modules. This was possible due to a long (eight residue) linker, allowing module 2 to bend back towards module 1 in a "V" shape - a non-consensus tryptophan residue was key to holding the two modules together. This was interesting in its own right, but perhaps even more interesting was the interface between CR2 and C3d. Residues in the B/C loop of CP 2 were involved in formation of H-bonds with main chain carbonyl atoms of C3d - only one amino acid from C3d had any direct interaction with CR2. Several carbonyl groups formed an anion pocket, which housed R83 from CR2, a charged residue at the inter-protein interface. Only modules 2 had direct interaction with C3d, whilst module 1 had an apparent scaffolding role stabilising module 2 and masking a non-binding surface. Perhaps this could help to explain why interfaces in module pairs from VCP show marked differences - VCP~1,2 and VCP~3,4 orientations are well defined and can be thought of as the scaffold-type arrangement seen in CR2 binding, stabilising the modules and for orientating them appropriately to allow interactions to occur, whilst the centre of the protein is less well defined and has limited movement to provide functional diversity.

In mutagenesis studies of CR1, mentioned briefly earlier, some of the residues mutated and shown to be key in binding were those that would be expected, by sequence homology, to form the intermodular interface (Krych *et al.*, 1994, 1998). Loss of binding may be because alteration in these key residues disrupts the interface One could infer then that altering one amino acid at the interface may disrupt the orientation of the two modules, moving the charged areas out of alignment required for binding. The change in interface residues could also alter the structure of the modules themselves, which could affect the loops and turns. This could, in turn, change the position of individual residues that may be important.

Mutagenesis studies by Liszewski *et al* (2000) on MCP have shown that mutation of residues 94-103 abolished co-factor activity, whilst C4b activity is retained. Mutation of E102A affected C3b binding. These residues have very strong sequence homology when compared to VCP~2,3, and lie at the VCP~2,3 interface. Changes in these residues are likely to disrupt the structure and intermodular orientation of MCP. Perhaps the interaction between VCP and C4b is similar and is dependent on charge rather than orientation of modules, as mentioned earlier. Thus by changing residues near the interface, thereby altering the flexibility or interdomain movement at the proposed hinge in VCP~2,3, it might be possible to abolish co-factor activity either by decreasing the amount of intermodular flexibility, which could limit involvement of residues from the C-terminal half of VCP, or allowing too much freedom of movement of the modules that may twist key residues from normal site of interaction. This would therefore be a useful future mutagenesis experiment.

Until more structural studies and rational structure-based mutagenesis have been carried out on VCP, the importance of intermodular interface residues, or intermodular flexibility remains unknown. The structure of CR2 bound to C3d added a surprising "twist" to the vast range of data available - it re-enforces the need for NMR and crystallography based structures. This work has highlighted the benefit of using different techniques in conjunction with one another to gain the best possible insight into the way that proteins such as VCP may look, or act, in solution. The plot thickens.

References

Barlow, P. N., Steinkasserer, A., Norman, D. G., Kieffer, B., Wiles, A. P., Sim R. B. & Campbell, I. D (1993). *J. Mol. Biol* **232** 268-284.

Bevington, P. R. & Robinson, D. A. (1992). *Quart. Rev. Biophys.* **29**, 119-167

Bouma B., De Groot P. G.,.Van Den Elsen J. M. H, Ravelli R. B. G., Schouten A., Brady, R. L., Dodson, E.J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). *Science*, **260**, 979-983

Casasnovas, J. M., Larvie, M. & Stehle, T. (1999). *EMBO J*, **18**, 2911-2922.

Chiou, H. L., Lee, T. S., Kuo, J., Mau, Y. C. & Ho, M. S. (1997) *J. Gen. Virol.* **78**, 2639-2645

Chiruvolu, V., Cregg, J. M. & Meagher, M. M. (1997) *Enzyme Microb Technol* **21**, 277-283

Copi, V., Tomita, Y., Akiyama, S. K., Aota, S., Yamada, K. M., Venable, R. M., Pastor, R. W., Krueger, S. & Torchia, D. A. (1998) *J. Mol. Biol.* **277**, 663-682

Cregg, J. M & Madden, K.R (1987) *In: Biological Research on industrial yeasts (Stewart et al.)* CRC Press, BOCG Raton, Fl **Vol 2,** 1-18

Dosset, P., Hus, J. C., Blackledge, M. & Marion, D. (2000) *J. Biomol. NMR.* **16**, 23-28

Henderson, C. E., Bromek, K., Mullin, N. P., Smith, B. O., Uhrin, D. & Barlow, P. N. (2001) *J. Mol. Biol* **307**, 323-339

Kirkitadze, M. D., Dryden, D. T. F., Kelly, S. M., Price, N. C., Wang, X., Krych, M., Atkinson, J. P. & Barlow, P. N. (1999d) *FEBS Letters*, **459**, 133-138

Kirkitadze, M. D., Henderson, C., Price, N. C., Kelly, S. M., Mullin, N. P., Parkinson, J. P., Dryden, D. T. F. & Barlow, P. N. (1999b) *Biochem. J*, **343,** 167-175

Kirkitadze, M. D., Jumel, K., Harding, S., Dryden, D., Krych, M., Atkinson, J. P. & Barlow, P. N. (1999c) *Prog. Polymer Colloid. Sci.* **113**, 164-167

Kirkitadze, M. Krych, M., Uhrin, D., Dryden, D., Cooper, A., Wang, X., Hauhart, R., Atkinson, J. P. & Barlow, P. N. (1999a) *Biochemistry*, **38**, 7019, 7031

Krych, M., Hauhart, R., & Atkinson, J. P. (1998). *J. Biol. Chem.* **273**, 8623-8629

Lin, F., Immormino, R. M., Shoham, M., Medof, M. E. (2001) *Arch Biochem Biophys* **393,** 67-72

Lipari, G. & Szabo, A. (1982) *J. Am. Chem. Soc.* **104**, 4546-4559;4559-4570.

Liszewski, M. K., Leung, M. K. & Atkinson, J. P. (2000) *J.Biol. Chem.* **275**, 37692-37701

Murthy, K. H., Smith, S. A., Ganesh, V. K, Judge, K. W., Mullin, N., Barlow, P. N, Ogata, C. M and Kotwal, G. J (2001). *Cell* **104**, 301-311

O'Leary, J. PhD Thesis, 2000

Rosengard, A. M. & Ahearn, J. M. (1998). *Mol. Immun.*, **35**, 397

Smith, S. A, Mullin, N. P., Parkinson, J., Shchelkunov, S. N., Totmenin, A. V., Loparev, V. N., Srisatjaluk, R., Reynolds, D. N., Keeling, K. L., Justus, D. E., Barlow, P. N. & Kotwal, G. J. (2000). *J. Virol.* **74**, 5659-5666

Szakonyi, G., Guthridge, J. M., Li, D., Young, K. Holers, V. M. & Chen, X. S. (2001) *Science*, **292**, 1725-1728.

Tschopp, J. F., Brust, P. F., Cregg, J. M., Stillman, C. A & Gingeras, T. R. (1987) *Nucleic Acids Res* **15**, 3859-3876

Vedvick, T., Buckholtz, R. G., Engel, M., Urcan, M., Kinney, J., Provow, S., Siegel, R. S. & Thill, G. P. (1991) *J. Ind. Microbiol.*, **7**, 197-201

Wiles, A. P., Shaw, G., Bright, J., Perczel, A., Campbell, I. D. and Barlow, P. N. (1997). *J. Mol. Biol.* **272**, 253-265.

Appendix A

make_shift_list

The make shift list macro reads in all chemical shifts from spectra exported from ANSIG.

```
#!/bin/csh -f
# makeShiftlist
# Makes assignment tables of shifts based on the individual spectra
# which have crosspeaks - they don't need to have assignments
# The outputs are 1) A table which summarizes unassigned atoms:
$$_shifts_unassigned.rdb
# 2) A table containing all the shifts entires which is sorted on the
basis of resid atnam
# priority and shift: $$_shifts_all_sorted.rdb and 3) A unique table
sorted on resid
# and atnam (sorted on the previous sort) which contains only one
entry for the chemical shift:
# $$_shifts_sorted.rdb.  The all_sorted table will be used for
subsequent operations.
#
# We have mades 2 versions of this to account for N15-based spectra
requiring
# a chemical shift database from H2O spectra and C13-NOESY's having
come
# from D2O spectra. The priorities for establishing this database
will thus
# be different. For N15-based spectra the order is
#
#     (1)15N HSQC    (2)13C TOCSY in H2O (3)4D_HCCONH (4)4D 15N/13C
NOESY (5)aromatics
#       using n15-hsqccn, hcchtoc,           4d_hcconh,   4d_cn_noe,
arom_cthsqc, cbhd
#
#     C-C NOESY   !          13C TOCSY/D2O   !         2d long-range
HMQC (For His etc.)
#     4d_c13noe, c13cosy, M31Cs10toc       c13noesy, hmqcH
#

#
# Priorities for selecting shift values:
# Use NHSQC, CBCACONNH and CBCANNH = 99 for c-noesy
# Use HNSQC2 and CNOESY = 99 for rest
#
#


#set CBHDSP_P = 1
set DCOSY_P = 19
set DNOESY_P = 19
set DTOCSY_P = 19
set NOESY_P = 19
set TOCSY_P = 19
```

```
#set CNOESY_P = 19
#set CBCACONNH_M_P = 4
#set CBCANNH_M_P = 4
#set HCCHCOSY_P = 1
#set HCCH_M_P = 1
set NHSQC_P = 19
#set NHSQC2_P = 99
set NTOCSY_P = 19
set NNOESY_P = 19
set NNOESYMEM_P = 1



# n-hsqc200499

column -c resname1 resname -c resid1 resid -c atname1 atname \
       -c atomType1 atomType -c shift1 shift spectrum xpk -a priority
N \
    < n-hsqc.rdb | compute priority = $NHSQC_P \; >
VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    < n-hsqc.rdb | compute priority = $NHSQC_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

# deucosy200499

column -c resname1 resname -c resid1 resid -c atname1 atname \
       -c atomType1 atomType -c shift1 shift spectrum xpk -a priority
N \
    < deucosy.rdb | compute priority = $DCOSY_P \; | headchg -del >>
VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    < deucosy.rdb | compute priority = $DCOSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb



# deunoesy200499

column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
    < deunoesy.rdb | compute priority = $DNOESY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    <  deunoesy.rdb | compute priority = $DNOESY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb


# deutocsy200499

column -c resname1 resname -c resid1 resid -c atname1 atname \
       -c atomType1 atomType -c shift1 shift spectrum xpk -a priority
N \
    < deutocsy.rdb | compute priority = $DTOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    < deutocsy.rdb | compute priority = $DTOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

# noesy200499


column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
    < noesy.rdb | compute priority = $NOESY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    < noesy.rdb | compute priority = $NOESY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb


# c-noesy

#column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
 #   < c-noesy.rdb | compute priority = $CNOESY_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
  #  < c-noesy.rdb | compute priority = $CNOESY_P \; | \
   # headchg -del >> M31Cs_2shifts.rdb

#column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
 #   < c-noesy.rdb | compute priority = $CNOESY_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb


# cbcaconnh_m

#column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
 #   < cbcaconnh_m.rdb | compute priority = $CBCACONNH_M_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
 #   < cbcaconnh_m.rdb | compute priority = $CBCACONNH_M_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
 #   < cbcaconnh_m.rdb | compute priority = $CBCACONNH_M_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

# cbcannh_m

#column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
 #   < cbcannh_m.rdb | compute priority = $CBCANNH_M_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
 #   < cbcannh_m.rdb | compute priority = $CBCANNH_M_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
 #   < cbcannh_m.rdb | compute priority = $CBCANNH_M_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb


# hcch-cosy

#column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
 #   < hcch-cosy.rdb | compute priority = $HCCHCOSY_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
 #   < hcch-cosy.rdb | compute priority = $HCCHCOSY_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
 #   < hcch-cosy.rdb | compute priority = $HCCHCOSY_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

# n-tocsy200499

column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
    < n-tocsy.rdb | compute priority = $NTOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    < n-tocsy.rdb | compute priority = $NTOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
    < n-tocsy.rdb | compute priority = $NTOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

# tocsy160499

column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
    < tocsy.rdb | compute priority = $TOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
    < tocsy.rdb | compute priority = $TOCSY_P \; | \
    headchg -del >> VCP_23N220200shifts.rdb


# n-hsqc2
```

```
#column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
 #    < n-hsqc2.rdb | compute priority = $NHSQC2_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb

#column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
 #    < n-hsqc2.rdb | compute priority = $NHSQC2_P \; | \
  #  headchg -del >> M31Cs_2shifts.rdb


# n-noesy200499

column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
     < n-noesy.rdb | compute priority = $NNOESY_P \; | \
     headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
     < n-noesy.rdb | compute priority = $NNOESY_P \; | \
     headchg -del >> VCP_23N220200shifts.rdb

column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
     < n-noesy.rdb | compute priority = $NNOESY_P \; | \
     headchg -del >> VCP_23N220200shifts.rdb


# n-noesymem200499

column resname1 resid1 atname1 atomType1 shift1 spectrum xpk -a
priority N \
     < n-noesymem.rdb | compute priority = $NNOESYMEM_P \; | \
     headchg -del >> VCP_23N220200shifts.rdb

column resname2 resid2 atname2 atomType2 shift2 spectrum xpk -a
priority N \
     < n-noesymem.rdb | compute priority = $NNOESYMEM_P \; | \
     headchg -del >> VCP_23N220200shifts.rdb

column resname3 resid3 atname3 atomType3 shift3 spectrum xpk -a
priority N \
     < n-noesymem.rdb | compute priority = $NNOESYMEM_P \; | \
     headchg -del >> VCP_23N220200shifts.rdb


# Save the unassigned shifts
# Either residues with no assignements or crosspeaks with no
# assignments

row resid eq null or atname eq null < VCP_23N220200shifts.rdb \
     ➢ VCP_23N220200shifts_unassigned.rdb


# now sort the table
# Note the tee means that the uniqtable will be based on the sort
# used to make the shifts_all table so that the entires in the unique
table
```

```
#(# M31Cs_shifts_sorted.rdb) will be the higest priority entry for a
given
# resid and atnam.

row resid ne null and atname ne null < VCP_23N220200shifts.rdb |\
    sorttbl resname | jointbl resname=resname translate.rdb |\
    row atname eq new.atname |\
    column resname resid -c old.atname atname atomType shift spectrum
\
           xpk priority | sorttbl resid atname priority shift | \
    tee VCP_23N220200shifts_all_sorted.rdb |\
    uniqtbl resid atname > VCP_23N220200shifts_sorted.rdb

# That's it!
```

make_shift_summary

This macro takes the output of make_shift_list and sorts the chemical shifts by atom, grouping together all chemical shifts assigned to an atom.

```
#!/bin/tcsh -f
# uses tsch rather than csh in cases where the buffer size is too
large
#
# Takes output from makeShiftList which contains all assigned
crosspeaks
# Gives a summary table which has the mean chemical shift
# and the spread of the chemical shifts.  This table is then split
into
# two tables, one containing H shifts and their spreads, and one
containing
# N and C chemical shifts and their spreads: this two tables
# are the input to the next script "makeConnectedShifts".
# Also puts out a list of "offender" or
# assignments with an unacceptable spread, in order of how bad they
are.
# This is extremely useful for checking assignments. Note that the
priority
# of the experiment is carried along. Priorities can also be given to
dimensions
# and used to weight various parameters such as the tolerance given
to a chemical
# shift in the ambiguous treatment.

######## RDB and CSH
operations##################################################
#
#             RDB operator: sorttbl
#
#Usage:  sorttbl  [options]  [-r]  column  [[-r]  column]  ...
#
#Options:
#    -c       Check that the rdbtable is sorted on the selected
columns.
#    -d       Dictionary order. Only letters, digits, and SPACE are
significant
#             in comparisons.
#    -f       Fold in lower case. Treat upper- and lower-case letters
equally.
#    -help    Print this help info.
#    -i       Ignore characters outside the ASCII range 040-0176 in
non-numeric
#             comparisons.
#    -r       Reverse order. Applys to the following column only.
#    -T dir   Place temporary files in directory 'dir'.
#    -u       Make rows unique on selected columns.
#
#Sorts an rdbtable on one or more columns. Each column may be sorted
in normal
#(ascending) or reverse (descending) order. Note that the '-r' on the
command
#line applies only to the next column; the absense of a '-r' means
the next
#column will sort in normal order.
#
#Also a column of monthnames (Jan, Apr, ...) in any case letters, may
be sorted.
#
#This operator reads an rdbtable via STDIN and writes an rdbtable via
STDOUT.
#Options may be abbreviated.

######################################################################
#######
#                   RDB operator: searchtbl
#
#Usage:  searchtbl  [options]  rdbtbl  <  keytbl
#        searchtbl  [options]  -ind  index_file  [rdbtbl]  <  keytbl
#
#Options:
#    -help    Print this help info.
#    -ind     Index file search.
#    -part    Partial (initial) match. Applies to string type data
only.
#    -sgl     Only a single row match is needed.
#    -rev     Reverse sort option. File 'rdbtbl' is sorted in reverse
order.
#    -vom     Verify only mode. If an item of info from keytbl is
valid prints
#             "ok", else an error message, on STDERR. NO new rdbtable
is
#             produced. Used by another process for verification.
#
#This operator does a fast search of 'rdbtbl' (or index_file) using a
binary
#search on a key of of one or more columns. The 'rdbtbl' (or
index_file) must
#be sorted on the key columns.  Each column in the key may be of type
string
#or type numeric (but be carefull with numeric data and exact
matches). In the
#second form of usage for this operator if 'rdbtbl' is not given its
name will
#be inferred from the name of index_file. For example if index_file
is
#'skb.x.typ' then the rdbtbl name inferred will be 'skb.rdb'.
#
```

```
#The column(s) in the file 'keytbl' specify both the key column
name(s) and the
#argument values to search for. File 'Keytbl' is in rdbtable format.
#
#Normally an argument value and a data field must compare exactally
for a match
#to occur (exact match). If the paritial match otpion (-part) is
selected, and
#if the argument value compares with the initial part of the data
field it is
#considered a match. This applies to string type data only. Note that
for
#numeric type data an exact match is always necessary.
#
#Normally all rows that match will be written to the new rdbtable, in
the
#same order as in the old rdbtable. If only a single row key match is
#appropriate some execution time can be saved by specifing the '-sgl'
option.
#
#This operator writes an rdbtable via STDOUT.  Options may be
abbreviated.
#Returns the number of non-finds at exit.

######################################################################
######
#            RDB operator: summm
#
#Usage:  summ  [options]  [column ...]
#
#Options:
#    -cu       A Count of the unique values for each column given.
#    -cun      Like option '-cu' but also shows counts of null (empty)
and
#              blank values (have only space chars), if either exist.
#    -cuu      A Count of each unique value for each column given.
#    -cu2      Like option '-cuu' but shows only counts greater than
one.
#    -help     Print this help info.
#    -m        The min, average, max, total for each column given.
#    -v        Inverse option. Selects all columns except those named.
#
#Produces "summary" information about the rdbtable. If no columns are
given
#then information about all columns is produced.  A Count of the data
rows
#is always shown.
#
#This operator reads an rdbtable via STDIN and writes a summary
report via
#STDOUT.  Options may be abbreviated.
#
######################################################################
#######
## CSH Commands
#
#
#     foreach var (wordlist)
#         ...
#     end       The variable var is successively set to each member
of
```

```
#               wordlist.  The sequence of commands between this
command and
#               the matching end is executed for each new value of
var.  (Both
#               foreach and end must appear alone on separate lines.)
#
#               The built-in command continue may be used to continue
the loop
#               prematurely and the built-in command break to
terminate it
#               prematurely.  When this command is read from the
terminal, the
#               loop is read up once prompting with ?  before any
statements in
#               the loop are executed.
#
#      set [var [ = value ] ]
#      set var[n] = word
#               With no arguments, set displays the values of all
shell
#               variables.  Multiword values are displayed as a
parenthesized
#               list.  With the var argument alone, set assigns an
empty (null)
#               value to the variable var.  With arguments of the
form var =
#               value set assigns value to var, where value is one
of:
#
#                    word        A single word (or quoted string).
#                    (wordlist)  A space-separated list of words
enclosed in
#                                parentheses.
#
#               Values are command and filename expanded before being
assigned.
#               The form set var[n] = word replaces the n'th word in
a
#               multiword value with word.
#
#               Multiple assignments can be performed with a single
set
#               command:
#
#                    set notify mail=(30 /usr/mail/nemo)
#
#########Begin the
script##################################################

#
# Do All shifts
#

set SHIFT_TABLE = VCP_23N220200shifts_all_sorted.rdb    #Generated
from makeShiftList
set SHIFT_SUMMARY_TABLE = VCP_23N220200_summary.rdb
set SHIFT_SUMMARY_TABLE_H = VCP_23N220200_summary_h.rdb    #H atoms
set SHIFT_SUMMARY_TABLE_NC = VCP_23N220200_summary_nc.rdb   #N and C
atoms
set OFFENDERS_H = offenders_1_N.rdb
set OFFENDERS_NC = offenders_1_N2.rdb
set MAX_PRIORITY = 20    # Exclude shifts from indirect proton
                         # dimensions in 4D spectra
goto start

start:

#
# Get header. Adds a column "delta" for the spread of chemical
shifts.
# Takes the top 2 lines from $SHIFT_TABLE to make header for
$SHIFT_SUMMARY_TABLE
#

column resid resname atname shift -a delta N < $SHIFT_TABLE |\
    head -2 > $SHIFT_SUMMARY_TABLE


#
# Resids
# "row resid lt 200" -->  lt the number of residues in the protein.
#

sorttbl -u resid < $SHIFT_TABLE | row resid ge 1 and resid le 127 | \
    column resid resname > resids_2.rdb


#
# For each Resid:
# For each Resid, find a unique list of the atnames. Then search the
$SHIFT_TABLE
# for each atname and sum the shifts (getting Min Avg Max and Total).
Dump the Avg
# shift and the difference between the Min and Max Value (delta) into
the $SHIFT_SUMMARY
# _TABLE.
#

foreach R (`column resid < resids_2.rdb | headchg -del`)
    set RN = `row resid eq $R < resids_2.rdb | column resname |
headchg -del`

    row resid eq $R < resids_2.rdb | /home2/brian/IRIX/bin/searchtbl
$SHIFT_TABLE |\
        sorttbl -u atname | column resid atname > atnames_2.rdb


    #
    # For each atomname
    #

    foreach A (`column atname < atnames_2.rdb | headchg -del`)
        printf "%s\t%s\t%s\t" $R $RN $A >> $SHIFT_SUMMARY_TABLE

      row atname eq $A < atnames_2.rdb |
/home2/brian/IRIX/bin/searchtbl $SHIFT_TABLE |\
          row priority lt $MAX_PRIORITY > temp_table.rdb
            set PRIO = `uniqtbl atname < temp_table.rdb | column
priority | headchg -del`

            row priority eq $PRIO < temp_table.rdb|\
            summ -m shift | \
            nawk -F'[ ,][ ,]*' '/Min, Avg, Max, Total/ {printf
"%.4f\t%.4f\n", $8, $9-$7}' \
            >> $SHIFT_SUMMARY_TABLE
    end
end

# Er, that's it!

#
# Split out a  summary table for H
# and a separate one for N and C
#

row atname mat '"^H"' < $SHIFT_SUMMARY_TABLE | \
    sorttbl -r delta > $SHIFT_SUMMARY_TABLE_H

row atname nmat '"^H"' < $SHIFT_SUMMARY_TABLE | \
    sorttbl -r delta > $SHIFT_SUMMARY_TABLE_NC

#
# Now find out what is causing the worst offenders
#

row delta gt 0.030 < $SHIFT_SUMMARY_TABLE_H | sorttbl resid |\
    column -c resid s.resid -c atname s.atname \
         -c shift s.shift -c delta s.delta |\
    jointbl s.resid=resid $SHIFT_TABLE | row s.atname eq atname | \
    column resname s.resid s.atname s.shift s.delta shift spectrum
xpk |\
    sorttbl -r s.delta shift > $OFFENDERS_H

row delta gt 0.30 < $SHIFT_SUMMARY_TABLE_NC | sorttbl resid |\
    column -c resid s.resid -c atname s.atname \
         -c shift s.shift -c delta s.delta |\
    jointbl s.resid=resid $SHIFT_TABLE | row s.atname eq atname |\
    column resname s.resid s.atname s.shift s.delta shift spectrum
xpk |\
    sorttbl -r s.delta shift > $OFFENDERS_NC
```

make_connected_shifts

This script uses the output from make_shift_summary to give an average chemical shift to each proton or nitrogen to be used as the final chemical shift value for each assigned atom.

```
#!/bin/csh -f

# makeConnectedShifts
#
# This script makes a master table of proton shifts and the
# shifts of their attached heavy atoms, using the individual
# _h and _nc tables output from the makeShiftSummary script.
#
# Requires the table connected.rdb which is a list of what
# H's are attached to what C's and N's. Also requires a
# pdb file with no coordinates ( M31Cs_dummy.pdb which was
# generated in xplor with the inp_file generate.inp)
# Also requires the fortran program pdb2rdb.


#
# make dummy shift table with ALL nuclei.
#
# ################################################################
# NOTE we should use the BMRB database values, with large delta's in
the
# dummy shift table.
# See /zeus/arcrl/nmrprogs/assign4d/newshifts.rdb as a first pass
# at a list of these values.
# ################################################################
#

# pdb2rdb takes an input pdb file with no coodinates and converts to
rdb
# format. The point is to get all the atnames/renames/resids for the
protein
# of interest.  Entries with C(CO), O and S are removed since they
are not used
# for the NMR constraints. A column shift is added and set to -99.0.
A column
# delta is also added --this can be made large where there are no
assignments!!!!
# The statement "\;resname = ucfirst \( lc \( resname \) \) \;" is a
perl
# operation which takes the first letter of resname and makes it
uppercase (ucfirst)
# and the rest lower case (lc).

pdb2rdb < VCP_23200499_dummy.pdb | row atname ne C and atname nmat
'"^O"' and atname nmat '"^S"' | \
    column resname resid atname -a shift -N -a delta -N |\
    compute shift = -99.0 \; resname = ucfirst \( lc \( resname \) \)
\;  > VCP_2300499_dummy_shifts.rdb

# Now make H shift table starting with the summary table made from
# the rdb script makeShiftSummary and with the dummy table

# The summary table containing experimental shifts and deltas
# set minimum delta

sorttbl resname atname < VCP_23N220200_summary.rdb |\
    column resname resid atname shift delta |\
    compute if \( delta lt 0.030 \) { delta = 0.030 } \; |\
    row atname mat '"^H"' > t.rdb

# Append the dummy table with shift=-99.0

sorttbl resname atname < VCP_23200499_dummy_shifts.rdb |\
    column resname resid atname shift delta |\
    row atname mat '"^H"' | headchg -del >> t.rdb

# Sort the combined table recursively based on resid and atname
#  and then recursively based on shift.  Those with no assignments
#  will be given Shift=-99.0
# Note this table is sorted on the basis of resname which is
# necessary to the last step where the H's and C's are connected.

sorttbl resid atname -r shift < t.rdb | uniqtbl resid atname |\
    sorttbl resname > VCP_23N220200_shifts_h.rdb

# Now make NC shift table

#Get all thr rows whose atname doesn't have H as the first letter
"^H"
# set minimum heavy.delta

sorttbl resname atname < VCP_23N220200_summary.rdb |\
    column resname resid -c atname heavy.atname -c shift heavy.shift
-c delta heavy.delta |\
    compute if \( heavy.delta lt 0.30 \) { heavy.delta = 0.30 } \; |\
    row heavy.atname nmat '"^H"' > t2.rdb

#Append heavy atnam rows to the t2.rdb file.

sorttbl resname atname < VCP_23200499_dummy_shifts.rdb |\
    column resname resid -c atname heavy.atname -c shift heavy.shift
-c delta heavy.delta |\
    row heavy.atname nmat '"^H"' | headchg -del >> t2.rdb

# Sort the combined table recursively based on resid and atname
#  and then recursively based on shift.  Those with no assignments
#  will be given Shift=-99.0.  Note this table is sorted on the basis
of
# resid which is necessary to the last step where the H's and C's are
connected.


sorttbl resid heavy.atname -r heavy.shift < t2.rdb | uniqtbl resid
heavy.atname | \
    sorttbl resid > VCP_23N220200_shifts_nc.rdb

# Join appropriate Hs to Ns or Cs
# This is where the file connected.rdb is required which shows what
C/N atoms are
# connected to what H's. Note the connected and M31Cs_shifts_h have
already been
# sorted based on resname, and M31Cs_shifts_nc.rdb has already been
sorted
# based on resid.

column resname -c atname c.atname -c heavy.atname c.heavy.atname <
connected.rdb |\
    jointbl resname=resname VCP_23N220200_shifts_h.rdb | row c.atname
eq atname |\
    column resname resid c.heavy.atname c.atname shift delta | \
    sorttbl resid | jointbl resid=resid VCP_23N220200_shifts_nc.rdb
|\
    row heavy.atname eq c.heavy.atname |\
    column resname resid -c c.atname atname shift heavy.atname
heavy.shift \
    delta heavy.delta > VCP_23N220200_HX_Shifts2.rdb
```

Appendix B - Chemical shift table

Res Type	Res No.	Atom type	Shift (ppm)	Atom Type	Shift (ppm)
Arg	7	HA	4.33		
Arg	7	HB1	1.8412		
Arg	7	HB2	1.6938		
Arg	7	HD1	3.0553		
Arg	7	HD2	3.0553		
Arg	7	HG1	1.636		
Arg	7	HG2	1.537		
Arg	7	HN	8.1172	N	122.2407
Arg	7	HE	7.0775	NE	119.1195
Arg	7	HH1#	6.509	NH1	
Arg	7	HH11		NH1	
Arg	7	HH12		NH1	
Arg	7	HH2#	6.509	NH2	
Arg	7	HH21		NH2	
Arg	7	HH22		NH2	
Arg	8	HA	4.773		
Arg	8	HB1	1.856		
Arg	8	HB2	1.592		
Arg	8	HD1	3.036		
Arg	8	HD2	3.036		
Arg	8	HG1	1.451		
Arg	8	HG2	1.451		
Arg	8	HN	8.0026	N	119.2322
Arg	8	HE	7.155	NE	119.889
Arg	8	HH11		NH1	
Arg	8	HH12		NH1	
Arg	8	HH21		NH2	
Arg	8	HH22		NH2	
Cys	9	HA	4.2027		
Cys	9	HB1	2.42		
Cys	9	HB2	2.131		
Cys	9	HN	8.7546	N	119.8453
Pro	10	HA	4.49		
Pro	10	HB1	2.36		
Pro	10	HB2	1.9835		
Pro	10	HD1	3.5707		
Pro	10	HD2	3.079		
Pro	10	HG1	2.099		
Pro	10	HG2	2.099		
Ser	11	HA	4.4137		
Ser	11	HB1	3.838		
Ser	11	HB2	3.7203		
Ser	11	HN	8.5175	N	117.4073
Pro	12	HA	4.1117		
Pro	12	HB1	1.415		
Pro	12	HB2	1.1085		
Pro	12	HD1	3.732		
Pro	12	HD2	3.623		

Pro	12	HG1	1.554		
Pro	12	HG2	1.2645		
Arg	13	HA	4.161		
Arg	13	HB1	1.761		
Arg	13	HB2	1.761		
Arg	13	HD1	3.152		
Arg	13	HD2	3.152		
Arg	13	HG1	1.654		
Arg	13	HG2	1.654		
Arg	13	HN	7.908	N	120.4387
Arg	13	HE	7.088	NE	119.3385
Arg	13	HH1#	6.519	NH1	
Arg	13	HH11		NH1	
Arg	13	HH12		NH1	
Arg	13	HH2#	6.519	NH2	
Arg	13	HH21		NH2	
Arg	13	HH22		NH2	
Asp	14	HA	4.413		
Asp	14	HB1	2.612		
Asp	14	HB2	2.373		
Asp	14	HN	7.9828	N	120.3667
Ile	15	HA	4.434		
Ile	15	HB	1.551		
Ile	15	HD11	0.484		
Ile	15	HD12	0.484		
Ile	15	HD13	0.484		
Ile	15	HG11	1.167		
Ile	15	HG12	1.0802		
Ile	15	HG21	0.7212		
Ile	15	HG22	0.7212		
Ile	15	HG23	0.7212		
Ile	15	HN	8.2420	N	116.0894
Asp	16	HA	4.2087		
Asp	16	HB1	2.384		
Asp	16	HB2	1.739		
Asp	16	HN	8.2010	N	125.6158
Asn	17	HA	3.6623		
Asn	17	HB1	2.0464		
Asn	17	HB2	1.136		
Asn	17	HN	8.8003	N	114.4106
Asn	17	HD21	7.3660	ND2	116.4135
Asn	17	HD22	6.6420	ND2	116.4135
Gly	18	HA1	4.3575		
Gly	18	HA2	4.004		
Gly	18	HN	7.1580	N	102.4957
Gln	19	HA	4.6017		
Gln	19	HB1	1.632		
Gln	19	HB2	1.632		
Gln	19	HG1	2.1215		
Gln	19	HG2	1.914		
Gln	19	HN	8.6770	N	117.5724
Gln	19	HE21	7.4310	NE2	111.95
Gln	19	HE22	6.5500	NE2	111.95

Leu	20	HA	4.5429		
Leu	20	HB1	1.1417		
Leu	20	HB2	1.1417		
Leu	20	HD11	0.466		
Leu	20	HD12	0.466		
Leu	20	HD13	0.466		
Leu	20	HD21	0.374		
Leu	20	HD22	0.374		
Leu	20	HD23	0.374		
Leu	20	HG	1.143		
Leu	20	HN	8.1625	N	120.8599
Asp	21	HA	4.7135		
Asp	21	HB1	2.592		
Asp	21	HB2	2.304		
Asp	21	HN	8.3350	N	123.5129
Ile	22	HA	3.894		
Ile	22	HB	1.674		
Ile	22	HD11	0.4665		
Ile	22	HD12	0.4665		
Ile	22	HD13	0.4665		
Ile	22	HG11	1.36		
Ile	22	HG12	0.676		
Ile	22	HG21	0.64		
Ile	22	HG22	0.64		
Ile	22	HG23	0.64		
Ile	22	HN	8.5924	N	126.3286
Gly	23	HA1	4.0582		
Gly	23	HA2	4.0565		
Gly	23	HN	8.7473	N	119.1449
Gly	24	HA1	4.334		
Gly	24	HA2	3.916		
Gly	24	HN	7.4530	N	107.3457
Val	25	HA	4.868		
Val	25	HB	2.6737		
Val	25	HG11	0.923		
Val	25	HG12	0.923		
Val	25	HG13	0.923		
Val	25	HG21	0.72		
Val	25	HG22	0.72		
Val	25	HG23	0.72		
Val	25	HN	8.1514	N	105.7912
Asp	26	HA	4.8535		
Asp	26	HB1	2.687		
Asp	26	HB2	2.508		
Asp	26	HN	7.4457	N	118.6895
Phe	27	HA	3.786		
Phe	27	HB1	2.975		
Phe	27	HB2	2.918		
Phe	27	HD1	7.0818		
Phe	27	HD2	7.0818		
Phe	27	HE1	7.3531		
Phe	27	HE2	7.3531		
Phe	27	HZ	7.3095		

Phe	27	HN	8.3767	N	119.9556
Gly	28	HA1	4.083		
Gly	28	HA2	3.474		
Gly	28	HN	8.749	N	117.4296
Ser	29	HA	4.4612		
Ser	29	HB1	3.716		
Ser	29	HB2	3.716		
Ser	29	HN	8.331	N	117.3816
Ser	30	HA	5.6892		
Ser	30	HB1	3.3777		
Ser	30	HB2	3.291		
Ser	30	HN	8.3277	N	113.0245
Ile	31	HA	4.509		
Ile	31	HB	1.016		
Ile	31	HD11	-0.5192		
Ile	31	HD12	-0.5192		
Ile	31	HD13	-0.5192		
Ile	31	HG11	0.803		
Ile	31	HG12	0.3717		
Ile	31	HG21	0.0728		
Ile	31	HG22	0.0728		
Ile	31	HG23	0.0728		
Ile	31	HN	8.566	N	117.6027
Thr	32	HA	5.0385		
Thr	32	HB	3.863		
Thr	32	HG21	1.072		
Thr	32	HG22	1.072		
Thr	32	HG23	1.072		
Thr	32	HN	8.038	N	116.8395
Tyr	33	HA	4.955		
Tyr	33	HB1	2.5272		
Tyr	33	HB2	2.5272		
Tyr	33	HD1	6.7409		
Tyr	33	HD2	6.7409		
Tyr	33	HE1	6.4789		
Tyr	33	HE2	6.4789		
Tyr	33	HN	9.4244	N	127.7223
Ser	34	HA	4.49		
Ser	34	HB1	3.844		
Ser	34	HB2	3.721		
Ser	34	HN	8.5192	N	112.3459
Cys	35	HA	5.241		
Cys	35	HB1	2.941		
Cys	35	HB2	2.394		
Cys	35	HN	8.6727	N	116.5423
Asn	36	HA	4.5287		
Asn	36	HB1	2.9415		
Asn	36	HB2	2.397		
Asn	36	HN	8.6741	N	121.9529
Asn	36	HD21	7.504	ND2	108.717
Asn	36	HD22	6.713	ND2	108.717
Ser	37	HA	4.232		
Ser	37	HB1	3.808		

Ser	37	HB2	3.808		
Ser	37	HN	8.3705	N	113.5187
Gly	38	HA1	4.15		
Gly	38	HA2	3.421		
Gly	38	HN	8.8254	N	113.3988
Tyr	39	HA	4.9315		
Tyr	39	HB1	3.221		
Tyr	39	HB2	2.424		
Tyr	39	HD1	6.5846		
Tyr	39	HD2	6.5846		
Tyr	39	HE1	6.6407		
Tyr	39	HE2	6.6407		
Tyr	39	HN	8.2265	N	118.7353
His	40	HA	4.937		
His	40	HB1	3.104		
His	40	HB2	3.014		
His	40	HD2	7.089		
His	40	HE1			
His	40	HN	9.4255	N	117.8479
His	40	HD1		ND1	
His	40	HE2		NE2	
Leu	41	HA	4.4712		
Leu	41	HB1	1.4897		
Leu	41	HB2	1.4033		
Leu	41	HD11	0.5777		
Leu	41	HD12	0.5777		
Leu	41	HD13	0.5777		
Leu	41	HD21	0.3835		
Leu	41	HD22	0.3835		
Leu	41	HD23	0.3835		
Leu	41	HG	1.17		
Leu	41	HN	8.3520	N	127.3552
Ile	42	HA	4.176		
Ile	42	HB	1.837		
Ile	42	HD11	0.602		
Ile	42	HD12	0.602		
Ile	42	HD13	0.602		
Ile	42	HG11	1.234		
Ile	42	HG12	1.1		
Ile	42	HG21	0.752		
Ile	42	HG22	0.752		
Ile	42	HG23	0.752		
Ile	42	HN	9.1839	N	128.9366
Gly	43	HA1	4.514		
Gly	43	HA2	3.61		
Gly	43	HN	8.2490	N	116.2869
Glu	44	HA	4.404		
Glu	44	HB1	2.18		
Glu	44	HB2	1.8205		
Glu	44	HG1	2.391		
Glu	44	HG2	2.297		
Glu	44	HN	8.3668	N	119.923
Ser	45	HA	4.1715		

Ser	45	HB1	4.252			
Ser	45	HB2	3.795			
Ser	45	HN	8.2730	N	113.3324	
Lys	46	HA	5.5085			
Lys	46	HB1	1.6325			
Lys	46	HB2	1.4585			
Lys	46	HD1	1.555			
Lys	46	HD2	1.555			
Lys	46	HE1	2.835			
Lys	46	HE2	2.835			
Lys	46	HG1	1.335			
Lys	46	HG2	1.2223			
Lys	46	HN	7.3919	N	119.2821	
Lys	46	HZ1			NZ	
Lys	46	HZ2			NZ	
Lys	46	HZ3			NZ	
Ser	47	HA	4.8267			
Ser	47	HB1	3.7416			
Ser	47	HB2	3.5995			
Ser	47	HN	8.1780	N	116.2535	
Tyr	48	HA	5.4387			
Tyr	48	HB1	2.95			
Tyr	48	HB2	2.793			
Tyr	48	HD1	7.0926			
Tyr	48	HD2	7.0926			
Tyr	48	HE1	6.7543			
Tyr	48	HE2	6.7543			
Tyr	48	HN	8.5544	N	122.0209	
Cys	49	HA	4.084			
Cys	49	HB1	2.495			
Cys	49	HB2	1.4195			
Cys	49	HN	8.3290	N	121.5754	
Glu	50	HA	4.5595			
Glu	50	HB1	2.131			
Glu	50	HB2	1.73			
Glu	50	HG1	1.976			
Glu	50	HG2	1.587			
Glu	50	HN	9.103		N	130.4167
Leu	51	HA	4.601			
Leu	51	HB1	1.644			
Leu	51	HB2	1.531			
Leu	51	HD11	0.687			
Leu	51	HD12	0.687			
Leu	51	HD13	0.687			
Leu	51	HD21	0.658			
Leu	51	HD22	0.658			
Leu	51	HD23	0.658			
Leu	51	HG	1.246			
Leu	51	HN	7.9214		N	120.4138
Gly	52	HA1	4.3272			
Gly	52	HA2	3.833			
Gly	52	HN	8.8175		N	114.2375
Ser	53	HA	4.1605			

Ser	53	HB1	3.95		
Ser	53	HB2	3.921		
Ser	53	HN	8.6921	N	115.9721
Thr	54	HA	4.408		
Thr	54	HB	4.499		
Thr	54	HG21	1.154		
Thr	54	HG22	1.154		
Thr	54	HG23	1.154		
Thr	54	HN	7.867	N	109.5237
Gly	55	HA1	4.257		
Gly	55	HA2	3.4895		
Gly	55	HN	7.8613	N	108.72
Ser	56	HA	4.519		
Ser	56	HB1	3.725		
Ser	56	HB2	3.648		
Ser	56	HN	7.4777	N	114.7306
Met	57	HA	5.046		
Met	57	HB1	1.672		
Met	57	HB2	1.586		
Met	57	HE1			
Met	57	HE2			
Met	57	HE3			
Met	57	HG1	2.294		
Met	57	HG2	2.0315		
Met	57	HN	8.173	N	121.9561
Val	58	HA	4.393		
Val	58	HB	1.979		
Val	58	HG11	0.747		
Val	58	HG12	0.747		
Val	58	HG13	0.747		
Val	58	HG21	0.6775		
Val	58	HG22	0.6775		
Val	58	HG23	0.6775		
Val	58	HN	9.0955	N	119.0844
Trp	59	HA	5.007		
Trp	59	HB1	3.292		
Trp	59	HB2	2.798		
Trp	59	HD1	7.141		
Trp	59	HE3	7.1377		
Trp	59	HH2	7.5182		
Trp	59	HZ2	7.1746		
Trp	59	HZ3	6.7868		
Trp	59	HN	7.7207	N	121.9619
Trp	59	HE1	10.3884	NE1	127.0873
Asn	60	HA	4.97		
Asn	60	HB1	2.7414		
Asn	60	HB2	2.371		
Asn	60	HN	9.5845	N	123.7945
Asn	60	HD21	7.882	ND2	112.8656
Asn	60	HD22	6.71	ND2	112.8656
Pro	61	HA	5.033		
Pro	61	HB1	2.4977		
Pro	61	HB2	2.2895		

Pro	61	HD1	4.2423		
Pro	61	HD2	3.7383		
Pro	61	HG1	1.892		
Pro	61	HG2	1.892		
Glu	62	HA	4.136		
Glu	62	HB1	2.069		
Glu	62	HB2	1.958		
Glu	62	HG1	2.378		
Glu	62	HG2	2.378		
Glu	62	HN	8.2665	N	117.5942
Ala	63	HA	3.69		
Ala	63	HB1	0.993		
Ala	63	HB2	0.993		
Ala	63	HB3	0.993		
Ala	63	HN	7.8315	N	122.9688
Pro	64	HA	4.5732		
Pro	64	HB1	1.852		
Pro	64	HB2	1.54		
Pro	64	HD1	2.977		
Pro	64	HD2	2.977		
Pro	64	HG1	1.487		
Pro	64	HG2	0.3845		
Ile	65	HA	4.3665		
Ile	65	HB	1.646		
Ile	65	HD11	0.488		
Ile	65	HD12	0.488		
Ile	65	HD13	0.488		
Ile	65	HG11	1.267		
Ile	65	HG12	0.987		
Ile	65	HG21	0.769		
Ile	65	HG22	0.769		
Ile	65	HG23	0.769		
Ile	65	HN	7.5682	N	109.4444
Cys	66	HA	5.1736		
Cys	66	HB1	2.737		
Cys	66	HB2	2.475		
Cys	66	HN	8.8600	N	120.6586
Glu	67	HA	4.8387		
Glu	67	HB1	1.827		
Glu	67	HB2	1.771		
Glu	67	HG1	2.264		
Glu	67	HG2	2.219		
Glu	67	HN	8.9290	N	124.4372
Ser	68	HA	3.816		
Ser	68	HB1	3.527		
Ser	68	HB2	3.527		
Ser	68	HN	8.9520	N	121.6756
Val	69	HA	3.829		
Val	69	HB	1.649		
Val	69	HG11	0.5625		
Val	69	HG12	0.5625		
Val	69	HG13	0.5625		
Val	69	HG21	0.5485		

Val	69	HG22	0.5485		
Val	69	HG23	0.5485		
Val	69	HN	7.7907 N	122.9853	
Lys	70	HA	4.6513		
Lys	70	HB1	1.821		
Lys	70	HB2	1.49		
Lys	70	HD1	1.277		
Lys	70	HD2	1.203		
Lys	70	HE1	2.5823		
Lys	70	HE2	2.5823		
Lys	70	HG1	1.11		
Lys	70	HG2	1.11		
Lys	70	HN	8.1170 N	124.5894	
Lys	70	HZ1		NZ	
Lys	70	HZ2		NZ	
Lys	70	HZ3		NZ	
Cys	71	HA	4.277		
Cys	71	HB1	3.0855		
Cys	71	HB2	2.15		
Cys	71	HN	8.7970 N	116.7848	
Gln	72	HA	4.7812		
Gln	72	HB1	2.054		
Gln	72	HB2	2.054		
Gln	72	HG1	2.51		
Gln	72	HG2	2.4125		
Gln	72	HN	8.4213	N	119.8458
Gln	72	HE21	7.49	NE2	111.959
Gln	72	HE22	6.798	NE2	111.959
Ser	73	HA	4.4355		
Ser	73	HB1	3.904		
Ser	73	HB2	3.859		
Ser	73	HN	7.9505	N	114.3212
Pro	74	HA	3.1667		
Pro	74	HB1	0.87		
Pro	74	HB2	-0.206		
Pro	74	HD1	3.199		
Pro	74	HD2	3.158		
Pro	74	HG1	0.618		
Pro	74	HG2	0.3017		
Pro	75	HA	4.2535		
Pro	75	HB1	2.2425		
Pro	75	HB2	1.65		
Pro	75	HD1	3.0743		
Pro	75	HD2	2.886		
Pro	75	HG1	1.9335		
Pro	75	HG2	1.76		
Ser	76	HA	4.5874		
Ser	76	HB1	3.822		
Ser	76	HB2	3.775		
Ser	76	HN	8.411	N	117.8555
Ile	77	HA	4.616		
Ile	77	HB	1.7466		
Ile	77	HD11	0.452		

Ile	77	HD12	0.452		
Ile	77	HD13	0.452		
Ile	77	HG11	1.011		
Ile	77	HG12	0.7925		
Ile	77	HG21	0.618		
Ile	77	HG22	0.618		
Ile	77	HG23	0.618		
Ile	77	HN	8.5985	N	116.9143
Ser	78	HA	4.142		
Ser	78	HB1	3.7317		
Ser	78	HB2	3.666		
Ser	78	HN	8.283	N	118.7716
Asn	79	HA	3.738		
Asn	79	HB1	2.122		
Asn	79	HB2	1.243		
Asn	79	HN	8.8833	N	115.5845
Asn	79	HD21	7.415	ND2	116.5215
Asn	79	HD22	6.721	ND2	116.5215
Gly	80	HA1	4.328		
Gly	80	HA2	4.0315		
Gly	80	HN	7.0543	N	101.719
Arg	81	HA	4.689		
Arg	81	HB1	1.7397		
Arg	81	HB2	1.5925		
Arg	81	HD1	2.9915		
Arg	81	HD2	2.9915		
Arg	81	HG1	1.3953		
Arg	81	HG2	1.2147		
Arg	81	HN	8.788	N	116.5706
Arg	81	HE	6.8805	NE	119.581
Arg	81	HH11		NH1	
Arg	81	HH12		NH1	
Arg	81	HH21		NH2	
Arg	81	HH22		NH2	
His	82	HA	4.988		
His	82	HB1	1.9153		
His	82	HB2	1.3434		
His	82	HD2	7.59		
His	82	HE1	6.5388		
His	82	HN	7.4984	N	114.484
His	82	HD1		ND1	
His	82	HE2		NE2	
Asn	83	HA	4.5875		
Asn	83	HB1	3.103		
Asn	83	HB2	2.572		
Asn	83	HN	7.8376 N	119.0528	
Asn	83	HD21	7.3400 ND2	110.0915	
Asn	83	HD22	6.9595 ND2	110.0915	
Gly	84	HA1	4.188		
Gly	84	HA2	3.428		
Gly	84	HN	10.5547 N		
Tyr	85	HA	4.7287		
Tyr	85	HB1	3.145		

Tyr	85	HB2	2.852			
Tyr	85	HD1	7.06			
Tyr	85	HD2	7.06			
Tyr	85	HE1	6.77			
Tyr	85	HE2	6.77			
Tyr	85	HN	8.0632	N	121.1487	
Glu	86	HA	4.393			
Glu	86	HB1	1.777			
Glu	86	HB2	0.909			
Glu	86	HG1	2.033			
Glu	86	HG2	1.996			
Glu	86	HN	7.6000	N	119.5635	
Asp	87	HA	4.37			
Asp	87	HB1	2.346			
Asp	87	HB2	2.346			
Asp	87	HN	8.3120	N	118.4497	
Phe	88	HA	5.1385			
Phe	88	HB1	2.879			
Phe	88	HB2	2.651			
Phe	88	HD1	7.034			
Phe	88	HD2	7.034			
Phe	88	HE1	7.257			
Phe	88	HE2	7.257			
Phe	88	HZ	7.23			
Phe	88	HN	7.1480	N	113.2987	
Tyr	89	HA	4.6143			
Tyr	89	HB1	2.9025			
Tyr	89	HB2	2.512			
Tyr	89	HD1	7.0342			
Tyr	89	HD2	7.0342			
Tyr	89	HE1	6.5989			
Tyr	89	HE2	6.5989			
Tyr	89	HN	8.8432	N	119.0051	
Thr	90	HA	4.318			
Thr	90	HB	4.209			
Thr	90	HG21	1.1987			
Thr	90	HG22	1.1987			
Thr	90	HG23	1.1987			0.020 1.0240
Thr	90	HN	8.1710	N	112.2744	
Asp	91	HA	4.075			
Asp	91	HB1	2.528			
Asp	91	HB2	2.528			
Asp	91	HN	8.5430	N	123.8181	
Gly	92	HA1	4.378			
Gly	92	HA2	3.545			
Gly	92	HN	9.0763	N	114.6371	
Ser	93	HA	4.414			
Ser	93	HB1	4.081			
Ser	93	HB2	3.8798			
Ser	93	HN	7.9480	N	116.8436	
Val	94	HA	5.3107			
Val	94	HB	1.818			

Val	94	HG11	0.8813		
Val	94	HG12	0.8813		
Val	94	HG13	0.8813		
Val	94	HG21	0.7203		
Val	94	HG22	0.7203		
Val	94	HG23	0.7203		
Val	94	HN	8.2377	N	121.7191
Val	95	HA	4.218		
Val	95	HB	1.1547		
Val	95	HG11	0.2646		
Val	95	HG12	0.2646		
Val	95	HG13	0.2646		
Val	95	HG21	0.1985		
Val	95	HG22	0.1985		
Val	95	HG23	0.1985		
Val	95	HN	8.6941	N	128.0055
Thr	96	HA	5.0457		
Thr	96	HB	3.794		
Thr	96	HG21	1.0777		
Thr	96	HG22	1.0777		
Thr	96	HG23	1.0777		
Thr	96	HN	8.411	N	120.0093
Tyr	97	HA	4.973		
Tyr	97	HB1	2.6808		
Tyr	97	HB2	2.462		
Tyr	97	HD1	6.7991		
Tyr	97	HD2	6.7991		
Tyr	97	HE1	6.4914		
Tyr	97	HE2	6.4914		
Tyr	97	HN	9.0102	N	125.8777
Ser	98	HA	4.514		
Ser	98	HB1	3.7195		
Ser	98	HB2	3.617		
Ser	98	HN	8.7987	N	111.7201
Cys	99	HA	5.232		
Cys	99	HB1	3.014		
Cys	99	HB2	2.496		
Cys	99	HN	8.761	N	118.1932
Asn	100	HA	4.491		
Asn	100	HB1	2.9377		
Asn	100	HB2	2.3575		
Asn	100	HN	8.6242	N	121.8873
Asn	100	HD21	7.5313	ND2	108.613
Asn	100	HD22	6.624	ND2	108.613
Ser	101	HA	4.197		
Ser	101	HB1	3.819		
Ser	101	HB2	3.819		
Ser	101	HN	8.341	N	113.6614
Gly	102	HA1	4.205		
Gly	102	HA2	3.478		
Gly	102	HN	8.7645	N	113.7785
Tyr	103	HA	4.6		
Tyr	103	HB1	3.255		

Tyr	103	HB2	2.457		
Tyr	103	HD1	6.6262		
Tyr	103	HD2	6.6262		
Tyr	103	HE1	6.629		
Tyr	103	HE2	6.629		
Tyr	103	HN	8.0898	N	118.2788
Ser	104	HA	4.643		
Ser	104	HB1	3.652		
Ser	104	HB2	3.652		
Ser	104	HN	9.4416	N	117.3629
Leu	105	HA	4.6115		
Leu	105	HB1	1.624		
Leu	105	HB2	1.505		
Leu	105	HD11	0.672		
Leu	105	HD12	0.672		
Leu	105	HD13	0.672		
Leu	105	HD21	0.515		
Leu	105	HD22	0.515		
Leu	105	HD23	0.515		
Leu	105	HG	1.261		
Leu	105	HN	8.248	N	127.2661
Ile	106	HA	4.227		
Ile	106	HB	1.9267		
Ile	106	HD11	0.5857		
Ile	106	HD12	0.5857		
Ile	106	HD13	0.5857		
Ile	106	HG11	1.271		
Ile	106	HG12	1.179		
Ile	106	HG21	0.736		
Ile	106	HG22	0.736		
Ile	106	HG23	0.736		
Ile	106	HN	9.3240	N	130.6267
Gly	107	HA1	4.4177		
Gly	107	HA2	3.544		
Gly	107	HN	8.3030	N	115.2172
Asn	108	HA	4.576		
Asn	108	HB1	2.892		
Asn	108	HB2	2.674		
Asn	108	HN	8.2127	N	119.0638
Asn	108	HD21	7.6460	ND2	113.3435
Asn	108	HD22	6.8750	ND2	113.3435
Ser	109	HA	4.0045		
Ser	109	HB1	4.2		
Ser	109	HB2	3.863		
Ser	109	HN	8.3275	N	115.6792
Gly	110	HA1	5.025		
Gly	110	HA2	3.333		
Gly	110	HN	7.8182	N	107.3744
Val	111	HA	4.193		
Val	111	HB	2.0947		
Val	111	HG11	0.7873		
Val	111	HG12	0.7873		
Val	111	HG13	0.7873		

Val	111	HG21	0.618		
Val	111	HG22	0.618		
Val	111	HG23	0.618		
Val	111	HN	8.9959	N	119.8112
Leu	112	HA	5.1773		
Leu	112	HB1	1.5755		
Leu	112	HB2	1.3725		
Leu	112	HD11	0.769		
Leu	112	HD12	0.769		
Leu	112	HD13	0.769		
Leu	112	HD21	0.7155		
Leu	112	HD22	0.7155		
Leu	112	HD23	0.7155		
Leu	112	HG	1.183		
Leu	112	HN	8.3530	N	125.4265
Cys	113	HA	3.9842		
Cys	113	HB1	2.332		
Cys	113	HB2	1.0547		
Cys	113	HN	8.5714	N	127.2747
Ser	114	HA	4.7835		
Ser	114	HB1	3.648		
Ser	114	HB2	3.504		
Ser	114	HN	8.7860	N	124.4573
Gly	115	HA1	3.682		
Gly	115	HA2	3.638		
Gly	115	HN	9.3558	N	118.4684
Gly	116	HA1	3.871		
Gly	116	HA2	3.119		
Gly	116	HN	7.9982	N	104.3903
Glu	117	HA	4.54		
Glu	117	HB1	2.006		
Glu	117	HB2	1.731		
Glu	117	HG1	2.221		
Glu	117	HG2	2.133		
Glu	117	HN	7.2672	N	119.7915
Trp	118	HA	5.2927		
Trp	118	HB1	3.251		
Trp	118	HB2	2.881		
Trp	118	HD1	7.1155		
Trp	118	HE3	7.058		
Trp	118	HH2	5.9959		
Trp	118	HZ2	6.798		
Trp	118	HZ3	6.752		
Trp	118	HN	8.4093	N	122.7505
Trp	118	HE1	9.5048	NE1	128.6522
Ser	119	HA	4.438		
Ser	119	HB1	4.376		
Ser	119	HB2	4.11		
Ser	119	HN	8.067	N	116.2191
Asp	120	HA	4.629		
Asp	120	HB1	2.534		
Asp	120	HB2	2.474		
Asp	120	HN	7.9055	N	115.8752

Pro	121	HA	4.147		
Pro	121	HB1	1.4877		
Pro	121	HB2	1.4877		
Pro	121	HD1	3.893		
Pro	121	HD2	3.6865		
Pro	121	HG1	1.825		
Pro	121	HG2	1.679		
Pro	122	HA	4.6247		
Pro	122	HB1	1.933		
Pro	122	HB2	1.596		
Pro	122	HD1	3.08		
Pro	122	HD2	2.2297		
Pro	122	HG1	0.997		
Pro	122	HG2	0.672		
Thr	123	HA	4.448		
Thr	123	HB	3.898		
Thr	123	HG21	1.077		
Thr	123	HG22	1.077		
Thr	123	HG23	1.077		
Thr	123	HN	7.9347	N	106.5773
Cys	124	HA	5.4316		
Cys	124	HB1	2.7345		
Cys	124	HB2	2.4733		
Cys	124	HN	8.9275	N	121.2655
Gln	125	HA	4.915		
Gln	125	HB1	2.04		
Gln	125	HB2	1.746		
Gln	125	HG1	2.306		
Gln	125	HG2	2.238		
Gln	125	HN	9.3012	N	123.8339
Gln	125	HE21	7.466	NE2	111.39
Gln	125	HE22	6.709	NE2	111.39
Ile	126	HA	3.8295		
Ile	126	HB	1.5403		
Ile	126	HD11	0.4855		
Ile	126	HD12	0.4855		
Ile	126	HD13	0.4855		
Ile	126	HG11	1.06		
Ile	126	HG12	0.707		
Ile	126	HG21	0.7085		
Ile	126	HG22	0.7085		
Ile	126	HG23	0.7085		
Ile	126	HN	8.6608	N	126.4859