UNIVERSITY
of
GLASGOW

The Biophysical Chemistry of Frog Foam Nest Proteins

A thesis presented in part fulfilment of the requirements for the Degree of Doctor of Philosophy

by

Cameron D. Mackenzie

WESTChem Department of Chemistry
April 2007

University of Glasgow

Abstract

The tropical frog, *Physalaemus pustulosus*, which is widespread throughout South and Central America, protects its offspring during development by encapsulating them in a foam nest. The nest material protects the tadpoles from environmental and predatory dangers for several days. Previous work within the Cooper group has identified the constituents of the nest to be based almost entirely on proteins and carbohydrates, and have sequenced six proteins termed *Ranaspumins*. Of these six proteins, RSN-2 (Ranaspumin 2) is of particular interest due to its lack of similarity to any other proteins published in the protein data bank. Proceeding on the hypothesis that RSN-2 has an important role in the foam structure, the recombinant protein has been cloned and expressed in bacteria and subsequently the solution structure has been derived experimentally by NMR. Using surface techniques including Langmuir-Blodgett depositions, surface pressure analysis and atomic force microscopy, the surface structure has been studied.

A proteomic and chromatographic separation and identification strategy was used to identify a large number of peptide and polypeptide fragments present in various abundances. This dramatically adds to the six proteins already discovered and the initial carbohydrate analysis performed by collaborators. The large number of peptides is typical of amphibian secretions and they are almost certainly functional to some degree.

The solution structure of RSN-2 was calculated as a five-turn helix, running perpendicular over a five-strand sheet. The elongated amino-terminus was very flexible and relatively unstructured in solution whereas the carboxy-terminus was anchored to the sheet. Analysis of nitrogen relaxation and chemical shifts suggested that there are turn regions in rapid motion: one region with slow dynamics (possibly cis-trans proline isomerisation) and a cleft undergoing slow conformational exchange. The region between the helix and first strand is flexible and is postulated as a hinge for surface unfolding.

Recombinant RSN-2 was established to be a powerful surfactant and provides surface pressures of 20-30 mNm^{-1} at approximately monolayer coverage. Hydrophobic patches were observed to partition at the air interface, and it is proposed that this is a result of reversible denaturing at the surface.

RSN-1 and other proteins were examined briefly and are good candidates for further investigation due to their interesting postulated functions.

Table of Contents

List of Tables

List of Figures

Acknowledgements

There are many people that have lent time, help and support over this time. In particular I wish to thank Alan Cooper and Malcolm Kennedy for giving me the chance to work on such a stimulating (maybe weird!) project. Special thanks are due to Alan, for providing guidance, encouragement, scientific rigour and consistently good humour. Also, to Malcolm, for his unlimited supply of tales of the unusual, and for providing expert guidance in the midnight mists of Trinidadian graveyards, not to mention deepest, darkest Tunapuna. Thanks to Margaret, for keeping the whole place from falling about our ears, and assisting with DSC.

I also owe a massive debt of gratitude to Brian Smith, for his assistance with all things NMR. The time provided on his spectrometer was critical for this project, and I know that it would not have been available at such short notice (and certainly low cost) elsewhere. Brian's expert tuition was invaluable and thanks to him, four years later, I think I finally get NMR! The time he spent helping with acquisitions, late on Friday nights, was above and beyond. I also appreciate the use of the NMR facilities at The University of Edinburgh. In particular, I am greatly indebted to Alan, Malcolm and Brian for their proof-reading of this manuscript.

To Carlos Bloch, I extend my appreciation for many things. His enthusiasm for science in general helped to ignite many new ideas, and his schooling in mass spectrometry proved to be invaluable. Although his regular browsing along maps of Recife beach was cruel, his Latin influence was always appreciated. Thanks also, to Andy Pitt and Richard Burchmore for the use of their proteomics kit and their help with experiments (I wonder if Andy would have written my Ph.D. reference, if he'd known how much I would pester him?). Nikolaj Gadegaard was extremely generous with his time and knowledge of AFM, especially given my odd requests – frog-on-a-chip? Also thanks to David Wokosin (two-photon fluorescence microscopy), Barrie Condon (for the ultimately unsuccessful MRI), Sharon Kelly and Tommy Jess (for CD spectroscopy) and Dave Knox (for enzyme inhibition assays and useful correspondence).

There are numerous characters that have passed through Alan, Malcolm and Brian's labs during my time. Special thanks to Sam (for putting me in touch with the project), Lindsay and Kevin (for office entertainment and life coaching), Rachel (for all her help and instruction), Rhona (for doing all the jobs in the lab that I tried to ignore!), Rosalind (for periodic supply of her Granny's fish crackers and exotic teas), Asia and all the others that made the lab such good fun to work in. Thanks also, to those other students that helped make Friday nights (sometimes) memorable.

Special thanks to my parents for sacrificing much to get me here, to Peter for planting the Ph.D. seed and to Kathryn for all her support and for passing on (a little) knowledge of organic chemistry. Thanks to my family, who managed to keep the ridicule over my Ph.D. to a minimum (so that it only happened when I was present), and for making sure that I remembered the absolute pointlessness of trying to get a Ph.D. if I couldn't back a trailer-load of hay into a barn. Finally, thanks to Marion, who didn't quite hold out long enough to get another photo for the shelf. She will be greatly missed.

The financial support for this work, supplied by the BBSRC, the RSC (West of Scotland section), the Elizabeth Aitken travel fund and the Keystone Conference Scholarship, is all greatly appreciated.

Author's Declaration

The work in this thesis is entirely my own, unless explicitly stated in the text.

Cameron Mackenzie. April 2007.

Definitions

2-DE	Two Dimensional Electrophoresis
AFM	Atomic Force Microscopy
ANS	Anilinonaphthalenesulfonic Acid
APS	Ammonium persulphate
APBS	Adaptive Poisson-Boltzmann Solver
ARIA	Ambiguous Restraints in Iterative Assignment
BLAST	Basic Local Alignment Search Tool
BSA	Bovine Serum Albumin
CATH	Class Architecture Topology Homologous Superfamily
CCPN	Collaborative Computing Project for NMR
CD	Circular Dichroism
CHAPS	Cholamidopropyldimethylammonio-propanesulfonic acid
CNS	Crystallography and NMR System
CPU	Central Processing Unit
CRD	Carbohydrate Recognition Domain
CSA	Chemical Shift Anisotropy
dH_2O	Deionised Water
DTT	1,4-dithiothreitol
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotide triphosphates
DSC	Differential Scanning Calorimetry
EDTA	Ethylenediaminetetraacetic Acid
ELISA	Enzyme Linked Immunosorbent Assay
FID	Free Induction Decay
FITC	Fluorescein Isothiocyanate
FPLC	Fast Protein Liquid Chromatography
FT	Fourier Transform
GFC	Gel Filtration Chromatography
GFP	Green Fluorescent Protein
GST	Glutathione S-transferase
GRAVY	Grand Average of Hydropathicity
HIC	Hydrophobic Interaction Chromatography
HPLC	High Performance Liquid Chromatography
HSQC	Heteronuclear Single Quantum Correlation
IEF	Isoelectric Focusing
IEX	Ion-Exchange Chromatography
INEPT	Insensitive Nuclei Enhanced by Polarisation Transfer
IPG	Immobilised pH Gradient
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IUPAC	International Union of Pure and Applied Chemistry
LS	Least Squares
MALDI	Matrix-Assisted Laser Desorption Ionisation
ML	Maximum Likelihood
MS	Mass Spectrometry
MWCO	Molecular Weight Cut-Off
NCBI	National Centre for Biotechnology Information
NHS	N-hydroxy Succinimide
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NTA	Nitrilotriacetic Acid
PAGE	Polyacrylamide Gel Electrophoresis

PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PES	Polyether Sulphone
PHYRE	Protein Fold Recognition Server
PMSF	Phenylmethanesulphonylfluoride
RACE	Rapid Amplification of cDNA Ends
RDC	Residual Dipolar Coupling
RF	Radio Frequency
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
RSN	Ranaspumin
RT	Reverse Transcriptase
SAIL	Stereo-Array Isotope Labelling
SCOP	Structural Classification of Proteins
SDS	Sodium Dodecylsulphate
TOF	Time of Flight
TEMED	Tetramethylethylenediamine
TFA	Trifluoroacetic Acid
TOCSY	Total Correlation Spectroscopy
TROSY	Transverse Relaxation Optimised Spectroscopy
UV	Ultraviolet
WATERGATE	Water Suppression Through Gradient Tailored Excitation

1 Introduction

1.1 Natural Products

The term "natural product" describes the range of substances derived from living organisms and, as such, encompasses a quite enormous collection of chemically diverse materials. By definition these products have been the source of the vast majority of foods, fuels, medicines and materials for much of the existence of humans and it is only relatively recently that synthetic alternatives have been used in some of these applications. The chemistries of these natural systems have been of particular interest because of their innate functions, driven to a mature stage by evolutionary pressures. Many of the planet's natural products have been discovered, studied and are well understood but there are many more that remain unknown or untested and it is these materials that represent an exciting opportunity for bridging unmet needs and advancing the understanding of the environment. In chemistry, the term "natural product" is often taken to mean a relatively small molecule, isolated from plant or animal that has been identified as potent in some form and efforts are made to understand the chemistry that produced such an entity, and how that can be transferred to modern research or manufacturing. In this work, the focus will be on a particularly curious, naturally-derived product and the inner workings and potential applications of such.

1.1.1 Pharmaceuticals

Certainly the most strenuous efforts in natural product research have been put into finding nature's medicines: that is an extract that is biologically active in the human body and can be developed for the treatment of patients. Famous examples include acetylsalicylic acid, the first known use of which was over two and a half millennia ago in ancient Greece where the scholar and physician Hippocrates described the fever and pain relief from an extract from willow bark. Later this was found to be salicylic acid, the acetyl adduct of which is common aspirin which even recently has gained popularity for further indications such as the recent data suggesting that it prevents the development of asthma in adults[1].

At the opposite end of the toxicity spectrum (although aspirin does have significant side effects), several natural products have been found to be exceptionally toxic. A pentapeptide with some unusual modified amino acids, dolastatin-10 has been isolated[2] from a marine mollusc and found to be toxic to such a degree that its antimitotic properties could not be exploited due to it being poorly tolerated[3]. New technologies allowing antibody conjugation have allowed dolastatin and the auristatin analogues to be promoted as a promising treatment for haematological cancers[4] and the same is true for the maytansinoids, extracted from tropical plants[5].

Commonly thought to be the most important pharmaceutical discovery of modern times is the isolation of penicillin from *Penicillium notatum* by Sir Alexander Fleming, in 1928[6]. Although the antibiotic properties of some moulds had been known, it was the observations of the inhibition of bacterial growth from the particular mould that allowed the particular compound to be isolated from culture broth. Large scale productions were not made possible until the flurry of work initiated by the mass casualties of World War II. Since then, penicillin and other members of the β-lactam family have allowed countless lives to be saved from otherwise fatal infections.

1.1.2 Other Materials

Traditionally, natural products such as timber were widely used in construction and the advantageous strength and weight of these materials mean that they are still used, whilst there is still availability. Other fields of use include the consumer products industry where extracts from plants are often used to add selected properties and value to shampoos, soaps etc.

The ability to tailor protein sequences for specific applications and the ability for some polypeptide chains to self-assemble in a controllable and defined mode has made them attractive in a number of other emerging fields. Silk from various silkworms is composed of a fibrous arrangement of glycine and alanine-rich protein chains arranged with a pleated sheet structure and is known for natural strength and flexibility, and similar biomimetic fibres have potential uses in a variety of fields such as textiles and tissue engineering[7]. Modern methods in protein conjugation and immobilisation also allow for the preparation of biofilms for packaging, detection and agrochemical industries.

In the field of diagnostics, proteins have already found a great number of applications. At the heart of many of these, is the extraordinary avidity of the biotin-avidin complex. First observed as a deficiency of vitamin H (biotin) in rodents being fed with high egg content (avidin being found in egg)[8], this is now known to be due the complexation of biotin by avidin. A dissociation constant of 10^{-15} mol L^{-1} is the strongest non-covalent biological affinity known[9] and the relative ease of biotinylation of molecules means that avidin (or streptavidin, the non-glycosylated extract from *Streptomyces avidinii* with similar affinity) can act as a molecular hook to capture a target from a mixture at extremely low levels. These can then be identified with antibodies and conjugated with other remarkable proteins for detection – for example, a green fluorescent protein (GFP) from *Aequorea victoria* or a bioluminescent protein such as the luciferase from *Photinus pyralis*. Such systems have been developed for use in an Enzyme-Linked Immuno-Sorbent Assay (ELISA), which has become the standard test for sensitive biochemical identity in industry and academia.

1.2 Proteins

1.2.1 Amino Acids

The word "protein" describes an organic molecule made up of a group of building blocks termed amino acids because of their amine and carboxylic acid group contents. There are a total of twenty standard amino acids, the general structure of which is shown.

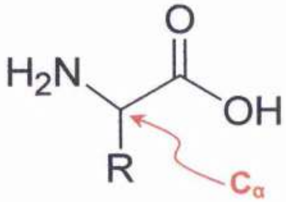


Figure 1.1. The chemical structure of a generic amino acid. "R" represents the different sidechains that make up the twenty amino acids.

Between the two functional groups is a tetrahedral carbon (the α-carbon or C_α) which is a chiral centre, the two isomers being described as L and D amino acids (normally *S* and *R* absolute configurations according to IUPAC convention). Only L isomers are usually observed in natural proteins (some enzymes can invert the configurations for specific purposes).

Variety in proteins is introduced by the twenty different sidechains possible, varying from the simple glycine (R=H) to more complicated aromatic sidechains such as tryptophan. It is the subtle diversity in sidechain structures that provide the exquisite variety in protein structure, and therefore function, allowing them to fulfil almost every conceivable biological role.

In nature, amino acids are coded for in the genetic material of the organism, being stored as DNA and transcribed into proteins via messenger RNA. In these nucleic acids, an amino acid is represented by a combination of one or more sets of three nucleotides (codons). Proteins are manufactured on the cellular ribosome where each of the constituent amino acids are fused together before being released. Each amino acid is joined to the next via the reaction of one primary amine group on one amino acid to the carboxyl group of the next through formation of an amide bond (or peptide bond). This very process is an example of the power of these natural processes in that the amide bond is created rapidly, consistently and selectively at physiological temperatures and in an aqueous environment, whereas the corresponding reaction in modern synthetic methodology would require expensive coupling components and could not be as selective or robust.

Following assembly on the ribosome, some proteins undergo modification to create the native structure. This can include enzymatic cleavage of a part of the sequence to allow membrane transport or to efficiently make fused proteins. There are numerous ways in which a protein can be added to, to insert a functional group into an otherwise inert or unsuitable biochemical environment. These modifications are termed post-translational modifications. One common modification is the addition of carbohydrate groups or chains to create glycoproteins with a range of properties. Carbohydrates (often termed sugars or glycans) can be attached to asparagine (N-linked), serine or threonine (O-linked). This addition means that proteins can be further diversified from the usual 20 amino acids and the field of glycoproteins is deeply studied in itself. Other, more simple adducts can be brought about by phosphorylation, alkylation and acylation amongst others. Further to this, natural amino acids can be chemically modified into other less common structures or transformed into other amino acids by citrullination or deamidation.

After synthesis, these linear hetero-polymers (polypeptides) fold into three-dimensional structures dictated by their amino-acid sequence (their primary structure). Amino acids can be divided into groups depending on their properties with respect to charge, hydrophobicity, size and chemical function and are shown in Appendix A. At neutral pH, the amino terminus (N-terminus) is protonated and the carboxy terminus (C-terminus) of the protein is deprotonated giving them positive and negative charges respectively. Throughout the protein, some of the other amino acid residues will possess a charge due to the weak acid/base character of their sidechains given by the individual pK_a of the group in question. The combination of all of these charges combines to give the protein a net charge given by its isoelectric point. This is the pH at which there is zero net charge on the protein i.e. the opposite charges of the groups completely balance each other out.

1.2.2 Protein Folding

The process by which the linear strip of amino acids folds into a three-dimensional structure is not fully understood but a great deal of effort has been put in to investigations and a great deal is now known. Due to the difficulties in physically imaging molecular dynamics or inferring it from other data, this field has developed with a large theoretical contribution.

Typical timescales for protein folding are in the millisecond range for a small protein[10]. It seems that the propensity for an amino acid sequence to fold is so strong that a great deal of the initial folding and sampling of conformational space occurs whilst the protein is still on the ribosome (cotranslational protein folding)[11]. Proteins are thought to follow an iterative folding path, driven by the lowering of energy, but this is not simply a case of selecting the lowest energy conformation after trying all others – this has been shown to be impossible as the random sampling of the 3^N (where N is the number of amino acids) conformations on a fast sampling regime would still take longer than the age of the universe[12]. This contradiction, known as Levinthal's paradox, has driven investigation into the true folding pathway and has resulted in the modern *Energy Funnel* model. Incidentally, computational studies have shown that relatively little extra energy would be required to reduce the sampling time covered in Levinthal's paradox to biologically significant times[13].

Simulations of protein folding are highly CPU-intensive, but dynamic simulations are being performed by utilising distributed worldwide computing. Some early results have shown it to be useful in demonstrating the formation of transient, structured species along the pathway[14,15].

Theoretical descriptions have also changed. It is no longer assumed that the polypeptide must fold in a particular way, uniform for its primary structure and rigid in progress. Instead, it is assumed that a degree of inherent randomness must occur in each fold and that some variability in the route from linear chain to folded (native) state is certain. Biophysical measurements have shown that short-range order and the generation of simple local structures (especially helices) occurs extremely rapidly, on the nano to microsecond timescale[16,17]. From this early state the protein folds in a way that is random but only within the available folds that it can make. As more native-like contacts are made, the less flexibility the protein has and the less available areas of conformational space it can sample. In this way the fold progresses down an energy funnel to an energy minima, driven by the increasing thermodynamic stability inferred by favourably electrostatic, hydrophobic and other interactions and opposed by the decreasing entropy brought upon it[18]. Simulations have suggested that initial folds are heavily influenced by a few native contacts and small proteins can adopt a largely native topology quickly and then smaller adjustments are made until the global energy minimum is reached. During these transitions, local energy minima are encountered where some stability is created from non-native contacts so the energy landscape is a rugged structure rather than a smooth funnel[19,20]. These local minima are quickly overcome by Brownian motion and more favourable native contacts elsewhere. The nature of the energy funnel and the nature by which a small protein rapidly folds into a native state is summarised in Figure 1.2, which is taken from a recent report[21] and excellent review on the subject[22].

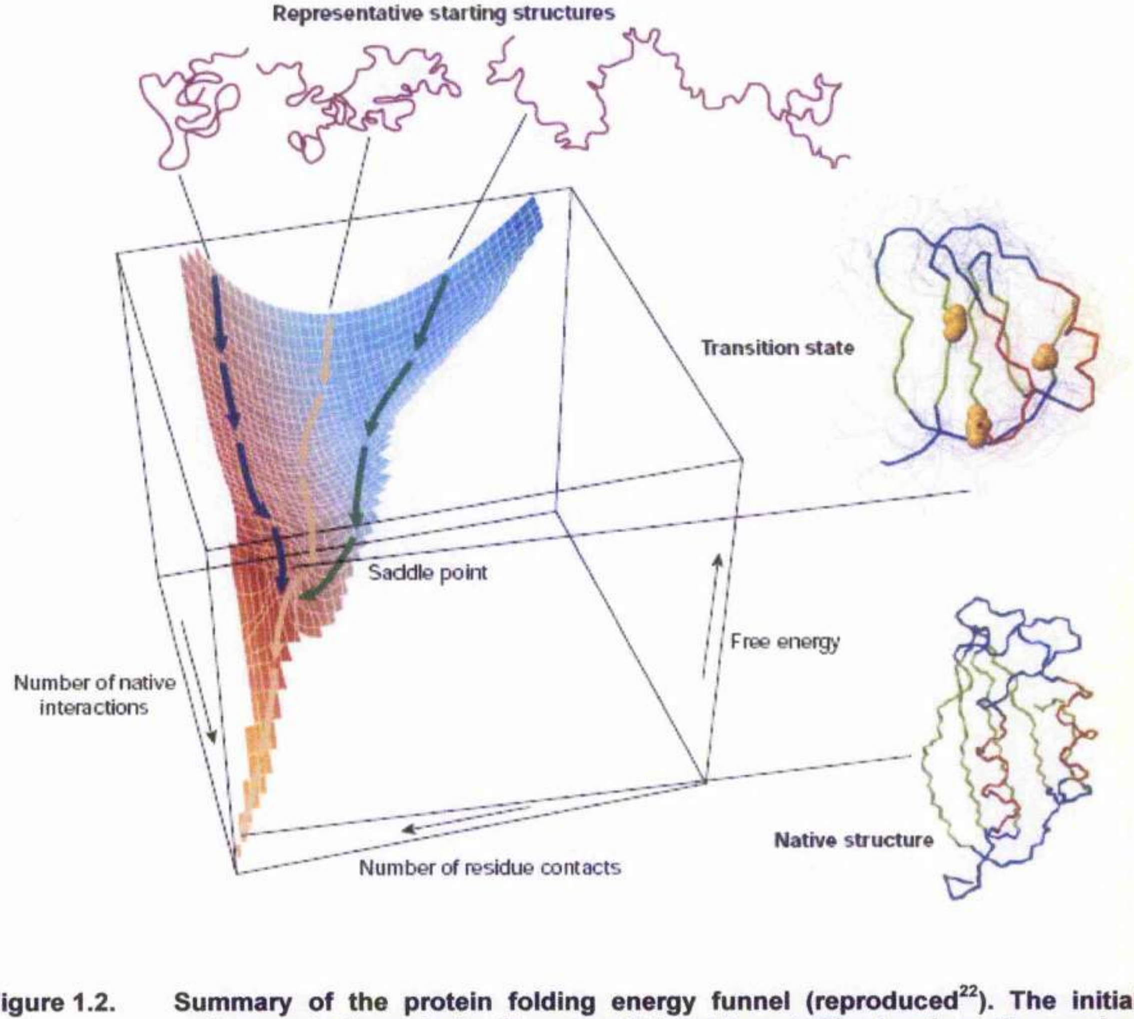

Representative starting structures
Transition state
Saddle point
Free energy
Number of native interactions
Number of residue contacts
Native structure

Figure 1.2. Summary of the protein folding energy funnel (reproduced[22]). The initial native contacts made by the essential residues (yellow) reduce the number of possible structures dramatically and allow the folding to progress rapidly.

Some proteins do not spontaneously adopt the native fold and require assistance from cellular machinery to induce the proper structure. Large (~800kDa) chaperone proteins are capable of encapsulating unfolded or misfolded proteins from the ribosome or cellular space and affect the correct fold, with the use of ATP[23,24]. Different classes of chaperone proteins are capable of transporting the correctly folded proteins across various cellular boundaries and this is an active area of research. Chaperones are often, or often associated with, heat-shock proteins because cellular stress often promotes protein misfolding which requires an increased expression of chaperone proteins.

1.2.3 Protein Structure

After the primary structure, the next step in the structural hierarchy is short-range structure driven primarily by hydrogen bonds forming defined compact shapes termed secondary structure. These structures are defined by the energy favourable contacts between the different atoms in the peptide bond rather than side-chain contacts, and so the local structures are common to proteins in general (although certain amino acids do have a bias towards certain secondary structural elements).

Most importantly in the analysis of protein structure, is the concept of electron resonance in the peptide bond and the subsequent planar structure that must be enforced. The carboxyl double bond is delocalised across the adjacent CN bond in two canonical forms in resonance.

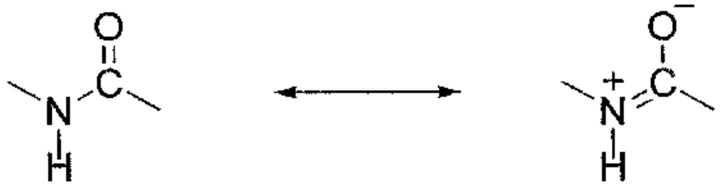


Figure 1.3. Resonance structures of amino acids through the delocalisation of the carbon-oxygen pi bond.

Electron delocalisation means that the C-N bond has approximately 40% double bond character[25] and this makes the amide bond essentially (but not always) planar. Instead, rotation in the backbone structure is brought about by the rotation of the bonds adjacent to the central carbon (the α-carbon).

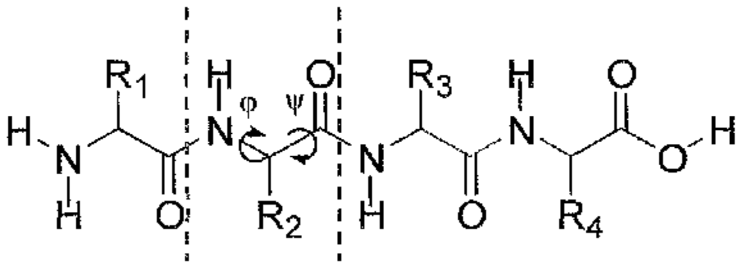


Figure 1.4. A four-residue peptide showing the backbone flexibility creating the dihedral angle (red).

The flexible bonds are denoted φ for the HN-C$_\alpha$ bond and ψ for the C$_\alpha$-CO bond. These bonds are relatively free to rotate and together form the 4-bond CONC dihedral angle. This angle is affected by the steric restraints placed on it by the particular amino acid sidechains involved but, in general, the two sidechains arrange themselves so that they are staggered in a *trans* conformation rather than the eclipsed *cis* form where unfavourable steric hindrances are more likely to occur. Exceptions to these rules are glycine, which has a single hydrogen atom as a sidechain creating minimal steric resistance, and the cyclic amino acid proline, which has much less of an energy difference between the *cis* and *trans* conformations than other amino acids, especially when the preceding residue is aromatic.

Generally, the secondary structure formed is one of three different types – helix or sheet, or the joining regions with less regular structure. These structures were elucidated by the examination of the known crystal structures of amino acids and the assumption that the peptide bond was planar[26]. Crucially, no assumption was made as to the requirement for symmetrical structures; instead structures were derived that satisfied the hydrogen bond donor and acceptor interactions most efficiently. The models created have shown to be largely correct although the original work predicted a γ-helix and more importantly the α-helix that was predicted was left-handed, with D-amino acids rather than the right-handed helix of L-amino acids now known to exist in proteins[27].

In proteins, α-helix is almost always used to describe a twisted structure where the amine proton is a hydrogen bond donor to the carbonyl acceptor of the residue four amino acids ahead, and this creates a structure where there is an average of 3.6 residues for every complete turn of the helix. A slightly tighter helix, termed the 3_{10} helix, is less commonly observed and usually not as long as the α-helix. The structure created has a packed axis and the amino acid sidechains point outwards from the helix and are surface-exposed. Because of this, the individual properties of the helices are different depending on the composition of the outer residues. Efficient packing and energetic advantages (more entropically favourable than sheet) mean that the helix is the most common secondary structure element in known proteins[28]. Because all of the charged peptide bonds are linear with the axis of the helix, a dipole moment is created along the length.

A β-sheet is the other main form of secondary structure and is different from the helix in many ways. The sheet consists of two or more constituent β-strands, which are strands of trans-conformation peptides that interact with each other to satisfy the hydrogen bond requirements of the amides. Strands can either run in the same direction (parallel) or opposite directions (anti-parallel), and many strands can gather to form considerably large sheet structures. Although usually thought of as planar they are often twisted and this may infer stability.

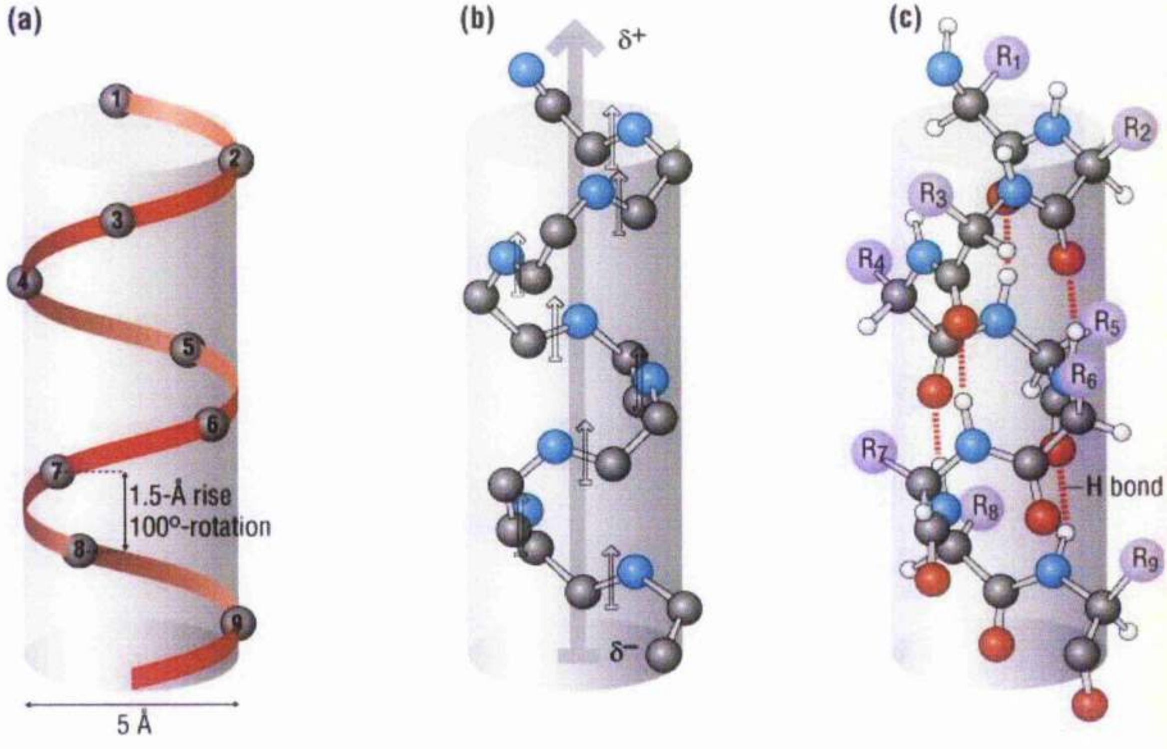


Figure 1.5. **Depiction of an α-helix showing a) the backbone trace as a cylinder, b) the helix dipole moment induced and c) the fully occupied hydrogen bond network and surface exposed sidechains (reproduced[29]).**

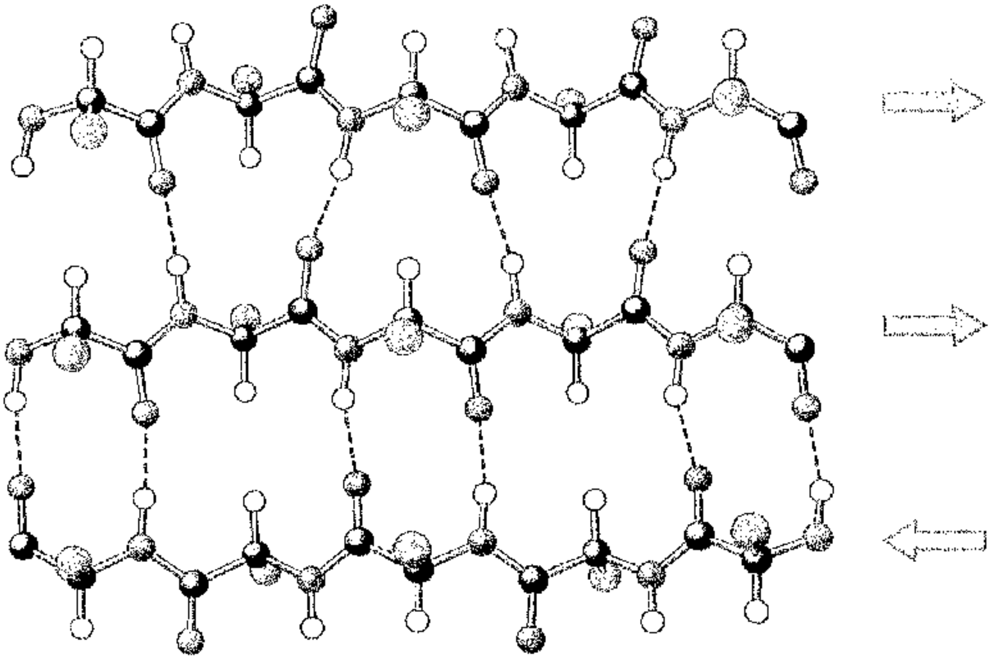


Figure 1.6. Ball and stick depiction of the parallel and anti-parallel arrangements possible in a mixed orientation β-sheet[25].

The relatively flat conformation of the sheet means that adjacent sidechains lie on opposite sides of the plane and, as such, it can often be seen that residues on one side of a plane possess similar or complementary properties rather than residues that are close to each other in primary structure. Similar structures can be formed where a strand turns back on itself to connect through hydrogen bonds and these structures are known as β-turns. The amyloid fibrils produced when proteins misfold are thought to contain fibrils of helical β-strands[30].

For a protein of *N* amino acids, there are 20^N possible combinations of possible proteins which is $1.3\text{x}10^{130}$ for a 100 amino acid protein but the human genome project has only identified 20000-25000 protein-coding genes[31]. Even allowing for the significant application of gene splicing for protein production, it would appear that not all combinations of amino acids are favoured for protein structure.

One reason for the restriction on the number of structures is that many of the possible dihedral angles created between two residues are not possible or very unfavorable due to steric clashes. These observations are depicted on a plot of the φ versus ψ torsion angles for the peptide bonds of a protein where defined "allowed" regions are identified[32]. This representation is called a Ramachandran Plot. The allowed regions can be mapped to specific secondary structure features and are a useful structure validation tool for experimentally-derived structures.

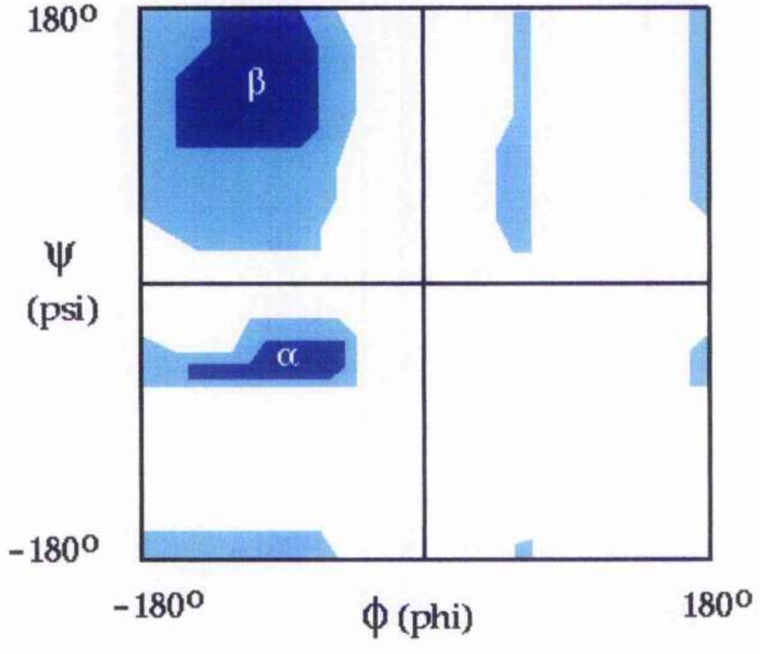


Figure 1.7. A Ramachandran plot showing the allowed region (blue) and the disallowed (white) for polypeptide torsion angles and the corresponding secondary structures associated with them (α ≡ alpha helix, β ≡ beta strand).

Further to this, certain amino acids are more prevalent on proteins than others and have a tendency to form particular structures. Most amino acids will form a stable helix with a few exceptions. Particularly large amino acids will preferentially form a strand structure where there is more space for them to extend into (Tyr, Trp), or create inter-chain links (Cys). This is also true for branched sidechains that need to occupy more space. Some other residues destabilise hydrogen bonded structures due to hydrogen bonding sidechains, rigidity, or size and these residues are more often found in structures with less order. These areas are often called random coil which falsely suggests that the structure is random – these loops and extensions are often functionally important.

1.2.4 Tertiary and Quaternary Structure

It is common for secondary structural elements to interact with each other in similar ways in numerous proteins to form small structural motifs that are often diagnostic of function. These small groups of secondary structure are known as super-secondary structure. Secondary structure elements all combine to form into a more energetically favourable environment called the tertiary structure. For a globular protein this will typically involve the grouping of non-polar (hydrophobic) residues at the core of the protein from where water is largely excluded and the promotion of polar residues to the surface where consideration of the thermodynamics of electrostatic contacts are favourable. This Hydrophobic Effect concept was developed from the simple interaction of oil and water[33].

Other interactions also stabilise the overall protein fold. The electrostatic, charge-charge, interaction between positive and negative residues (lysine and aspartic acid for example) can be relatively strong – this bond is often called a salt-bridge which is misleading due to the lack of any salt. At the hydrophobic core of the protein aromatic residues are most often observed to be stabilised by the interaction between their aromatic rings, maximising the interaction between the electron-rich and deficient regions by perpendicular edge-on packing. Alternatively the aromatic rings can be slightly offset, often termed "pi-stacking" due to the planar arrangements of the pi bonded electrons.

In oxidising environments the sulphydryl groups on cysteine residues will spontaneously dimerise to form a disulphide bond. These bonds can be formed between residues that are close in space but often not close in sequence and can help to stabilise the overall topology (fold) of the molecule. The structure of the disulphide bond normally creates a 90° dihedral CSSC bond angle as shown in Figure 1.8. The intracellular environment is mostly reducing meaning that disulphides are less likely to be stably formed until secretion although there is some evidence for the successful alteration of this[34]. Most proteins form disulphide bonds in only one possible way to give the native structure which is particularly impressive given that many proteins contain multiple free cysteine residues. The correct arrangement or rearrangement of disulphides is often mediated by the redox enzyme Disulphide Isomerase which ensures rapid native fold formation. Disulphide bonds are truly covalent and as such add strength to proteins or parts of proteins which is extremely important in many structural proteins like keratins.

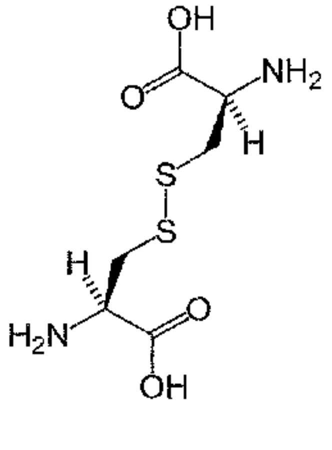


Figure 1.8. A disulphide bond between two cysteine amino acids. The four bond CSSC dihedral angle is normally 180°.

Finally, the highest level of protein structure is quaternary structure and describes the spatial arrangement of individual polypeptide structures that come together to form the native protein. These individual structures are subunits of proteins with tertiary structure as described.

A protein constructed from more than one subunit is called a multimer where two subunits make a dimer, three subunits make a trimer and a single subunit is a monomer. Often the identical subunits some together to form a homo-dimer which can be symmetrical or not and different subunits can also come together to form hetero-dimers or higher ordered multimers. Sometimes, the quaternary structure includes a component that is not an amino acid, such as a substrate or a metal ion which completes the functionally active structure such as haemoglobin, the oxygen carrier protein in red blood cells. This protein is a tetramer of two α subunits and two β subunits as well as four haem groups. Similarly, ion channels located in cellular membranes consist of a collection of subunits that alter conformation to affect the bulk structure to create an open or closed pore.

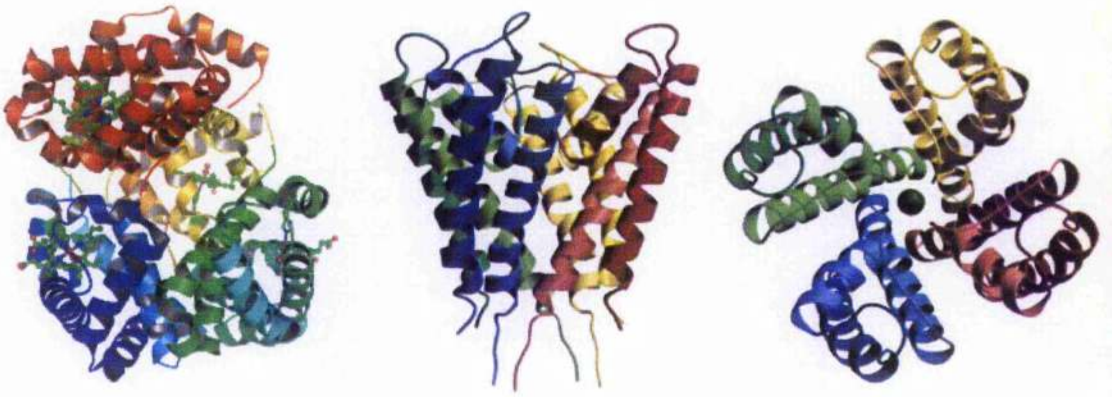


Figure 1.9. Examples of proteins with functional quaternary structure shown as cartoons. Left is haemoglobin[35], middle and right are across and above membrane views respectively, of the mammalian potassium ion channels[36].

1.2.5 Classes of Proteins

All living organisms create proteins from their genetic library and they produce a myriad of different unique structures. Several different proteins from different sources have been found to be similar in terms of structure or function and often both. Often these proteins are found to have similar sequences and this can reveal a lot about the residues that are important for either structural integrity, functional efficacy or both. These similarities can also be related to evolutionary biology and insights into species divergence can be made where other methods are ineffective.

Increasing computing power has lead to the significant increase in the use of bioinformatics as a tool for analysing and predicting protein structure and function. Derived protein structures are deposited in the Protein Data Bank (PDB)[37] and assigned a unique identifier, but smaller databases have been created to enhance the uses that can be made of the information within the PDB. Commonly, proteins are separated into families that are classified according to structural similarity. Two frequently used databases are CATH (Class, Architecture, Topology and Homologous superfamily)[38] and SCOP (Structural Classification of Proteins)[39] which both arrange protein structures into subsets but view protein fold space in different ways leading to differences in the groupings. Studies have shown that the differences are due to the definitions attributed to domain classification rather than any actual distinctions in the philosophies behind them[40].

1.2.6 Structure Elucidation

For structural studies of proteins with atomic resolution there are currently only two viable methods - X-ray crystallography and nuclear magnetic resonance (NMR). Cryo-electron microscopy has more recently been used to gain extremely useful insights into the structural mechanism of ribosomal transcription[41] but still suffers from relatively low resolution.

X-ray crystallography is the dominant method for protein structure determination and is responsible for the majority of structures in the PDB with the first crystal structure of a protein (myoglobin) being published in 1958[42]. Packed protein crystals are subjected to X-ray radiation and the regular array of electron density of the atoms diffract the radiation in patterns that can be related to structure. Difficulties in this technique can be due to the need for a high quality protein crystal which is not always possible, especially for particularly flexible or membrane bound proteins.

Nuclear magnetic resonance (NMR) can also be used to elucidate protein structures in solution. Because these are solution structures, they can be complementary to X-ray structures and are capable of providing additional data including information about protein dynamics and intra- and inter-molecular interactions. This method probes the magnetic properties of atomic nuclei with radio-frequency radiation in a powerful static magnetic field. NMR is generally less sensitive than other methods and can be labour-intensive, but promise in automation, the development of higher magnetic field instruments, and more powerful radiation pulse sequences suggests that this technique will be increasingly useful in the emergence of high throughput biology. This technique will be described in more detail in later chapters.

1.3 Amphibians, Frogs and Physalaemus Pustulosus

In this work, we are investigating proteins from an unusual source (in that the majority of known proteins are mammalian or bacterial) – the foam nests of a tropical frog. Frogs are of the *Anura* order of amphibians and have some extremely interesting features due to their ability to live in aqueous or terrestrial environments. Frogs have porous skin through which they absorb water for hydration and importantly they can absorb dissolved oxygen into the blood stream as a substitute for conventional breathing. It is for this reason that the skin is usually kept moist. The appearances, behaviour and anatomy of frogs change across the wide number of species which reflects the extremely wide range of territories over which frogs can be found.

Frogs vary greatly in size from ~1 cm (*Brachycepahlus didactylus*) to ~30 cm (*Conraua goliath*) and have different features depending on requirements. For example, tree frogs have evolved large fingers and toes containing pads with remarkable adhesive properties[43], whereas frogs that spend more of their lives in the water have developed webbed feet for more efficient movement in water. Other burrowing frogs can have a useful limb extension to assist in digging. In contrast to these mainly tropical frogs, *Rana sylvatica* ranges beyond the arctic circle and has developed strategies to survive the sub-zero temperatures - the frog's metabolism ensures a build up of glucose and urea which act as cryoprotectants for the cells and proteins and prevents dehydration despite ~70% of the body freezing[44].

In biochemical sciences there has been relatively little activity in the study of amphibians except for the in-depth genomic analysis of frogs from the *Xenopus* genus. *Xenopus laevis,* and more recently *Xenopus tropicalis,* have been used as model systems for exploring developmental biology because of the large and easy to manipulate oocyte. The most well known use is in the original pregnancy test, the popularity of which, led to the African frogs being exported in large numbers. Recently this exportation has been suggested as a principle reason for the spread of chytridiomycosis which threatens amphibian survival worldwide[45]. Many hypothetical proteins and useful genetic information has been gleaned from the ongoing project to sequence the genome of *Xenopus tropicalis*[46].

Often the frog species which are best known are those with striking colours on display such as the poison dart frogs of South America. In general these brightly coloured frogs use colour as a warning of their toxicity to deter potential predators (aposematism); *Dendrobates imitator*, whilst itself toxic, imitates the colour in other local toxic frogs – an extremely interesting case of Mullerian mimicry[47]. The more common Batesian mimicry where a non-toxic species imitates a toxic one is also found in poison dart frogs where it has been shown that a species of relatively low toxicity is copied as a more effective anti-predation strategy[48]. Toxic secretions are normally alkaloid based, and are easily biosynthesised or stored from their rich diet[49]. Most other, less brightly coloured frogs, use their darker and sometimes rougher skin as camouflage and this often affects the behaviour when confronted with a predator – camouflaged frogs may remain still whilst coloured frogs may take flight.

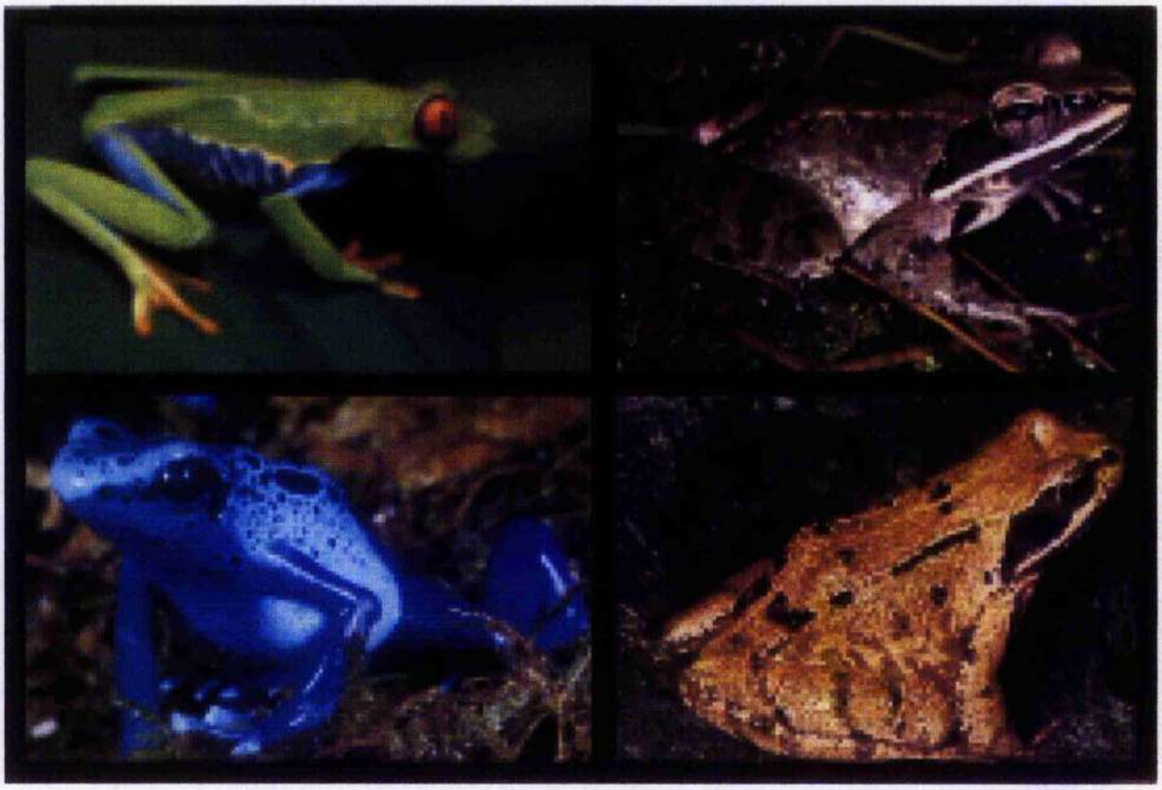


Figure 1.10. A variety of frog species with differing appearances according to their local environments. They are *Agalychnis callidryas* (top left, © Mark Kostich Photography) from Central America, *Rana sylvatica* (top right, © Walter W. Knapp) from North America, *Dendrobates azureus* (bottom left, © Ryan Photographic) from Surinam and *Rana temporaria* (© Henk Wallays) which is common to the British Isles.

A large number of other secretions from frog skin have been collected and characterised and are chemically diverse. At a fundamental level of biochemical activity, an enzyme secreted from the skin of frogs in the *Bombina* genus is capable of isomerising natural L-amino acids to D-amino acids[50]. Many antimicrobial peptides have been observed[51] and have been defined as active due to their charge properties which can disrupt bacterial membranes[52]. This is a different mechanism from most common antibacterials and may slow rate of pathogen resistance. Recently some antimicrobial peptides from frog skin have shown some promise as an inhibitor of the HIV virus[53]. Whilst the vast number of antimicrobial peptides have claimed the greatest attention, many other secretions have been identified with unusual properties such as a bradykinin-realted (a vascodilator) peptide found in the secretions of *Phyllomedusa hypochondrialis*[54].

Much less attention has been paid to other frog secretions, most notably those associated with egg and embryo protection. In most species, mating is triggered by rainfall when the frogs gather at the breeding areas which are often at or near areas of water. They tend to gather in large groups and the male frogs initiate calling in a chorus which is specific for every species and can be very diverse. Once the male locates the female it positions on top of the female and grasps her in amplexus. The female frog discharges a number of eggs and they are fertilised externally by sperm from the male and the frog spawn is complete.

After mating, different species behave very differently. In many cases the parents immediately leave the fertilised eggs although some frogs may linger for a period to protect their offspring. In general the tadpoles emerge from the eggs very quickly and are fully water-dwelling owing to their lack of limbs, their tail for movement and gills for respiration. After some time, normally a number of days, the tadpoles undergo metamorphosis into froglets in a process which is associated with some dramatic morphological and metabolic changes. Forelimbs and legs are produced, the tail is absorbed, the eyes translate around the head and the internal organs adjust to a land-based diet.

During breeding frogs create many tadpoles, often hundreds or thousands and this is necessary due to the high predation rate at this stage of the life cycle. Floating on standing water often attached to terrestrial vegetation, the tadpoles can be attacked by insects, snakes, newts, birds, lizards and even other frogs or other tadpoles (cannibalism has been observed[55] and some females lay excess, unfertilised eggs as a food source). In addition to this, tadpoles face accidental crushing from larger animals, dehydration, starvation and many other possible dangers for which many defence strategies have evolved.

Many frog spawn or nests are created in less accessible areas to protect the young - many tree frogs create nests above ground and the eggs are allowed to develop in relative safety before the tadpoles drop or are transported to the ground by a parent. However, this does not guarantee safety as predation of foam nests in elevated leaves has been observed by opportunistic monkeys[56]. Other frogs have adopted more unusual strategies for incubating the eggs or tadpoles until they are developed, keeping their offspring about their bodies. The most renowned of these is the gastric brooding frog (*Rheobatrachus* genus) which kept the froglets in the stomach of the mother where the normally acidic conditions are physiologically altered to prevent damage. It is now thought to be extinct[57]. Similarly, Darwin's frog (*Rhinoderma darwinii*) allows the tadpole to develop in the vocal sac, the male midwife frog carries the developing eggs on its back for several days and the female pouch frog carries the developing tadpoles in a marsupial-like pouch on its back. Frogs are usually very sensitive to their environments due to their porous skin and their position in the food chain. The recent decline in frog populations and the extinction of several species has been identified as a biological indicator for the changing environment.

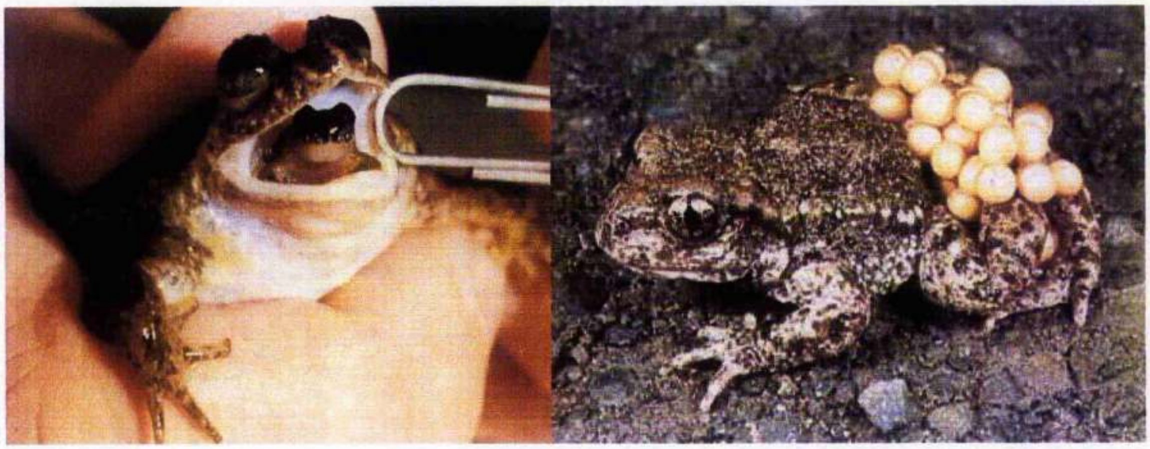


Figure 1.11. Parental behaviour of the gastric brooding frog and the midwife frog (© Michael Frede).

Foam nesting is one particularly interesting phenomenon and is the focus of the work described here, concentrating particularly on *Physalaemus pustulosus*. *P. pustulosus* is a tropical frog, prolific throughout Central and South America and the Caribbean. Also known as the Túngara frog or Mud Puddle Frog, it is part of the *Leptodactylidae* family which contains over 1,100 known species found throughout the Americas. Currently *P. pustulosus* is not an endangered species and has managed to adapt to life in areas of human disturbance, although habitat destruction may be a danger in the future. Although very similar in size, the female is marginally larger than the male at approximately 50 mm in size. The rugose skin gives it a toad-like appearance which is reflected in its name.

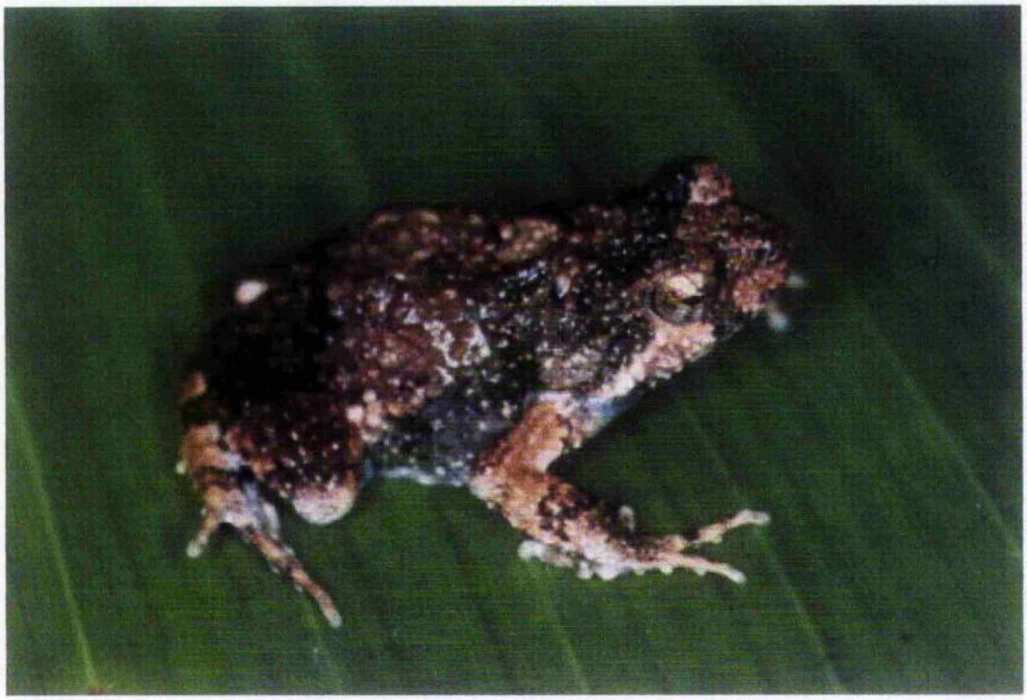


Figure 1.12. An example of an adult *Physalaemus pustulosus*.

P. pustulosus is a ground-dwelling frog, residing in leaf-litter and grassed areas or marshes. Mating is induced by heavy rains which are frequent during the Caribbean wet season commencing in approximately June. Males attract females by inflating the air sacks below their mouths and making two distinct calls which have been the subject of intense study as a model for the evolution of acoustic sexual selection in frogs. Female frogs have been found to show preference to frogs calling with a deep "chuck" sound following the standard "whine" but the deep chuck has been shown to attract predation in the males[58]. Furthermore the males have been shown to use sound to discern the size of a competitor and it uses this information to decide whether or not to compete[59].

Many tropical frogs, including *P. pustulosus*, deposit their eggs in a foam nest rather than spawn. These nests can be floating on water (as in *P. pustulosus*), in trees or elevated vegetation as in many tree frogs or even buried under the ground surface (*Leptodactylus fuscus*) and are assumed to be a defence against predation and dehydration when surrounding water evaporates. The phenomenon is observed in frogs across many different species from dispersed territories. The foam surface acts as a physical barrier against attack from insects and although attack from snakes is still observed, the foam structure is less easy to penetrate. Nests are observed to exist for several days without microbial attack and so are assumed to affect some anti-microbial property.

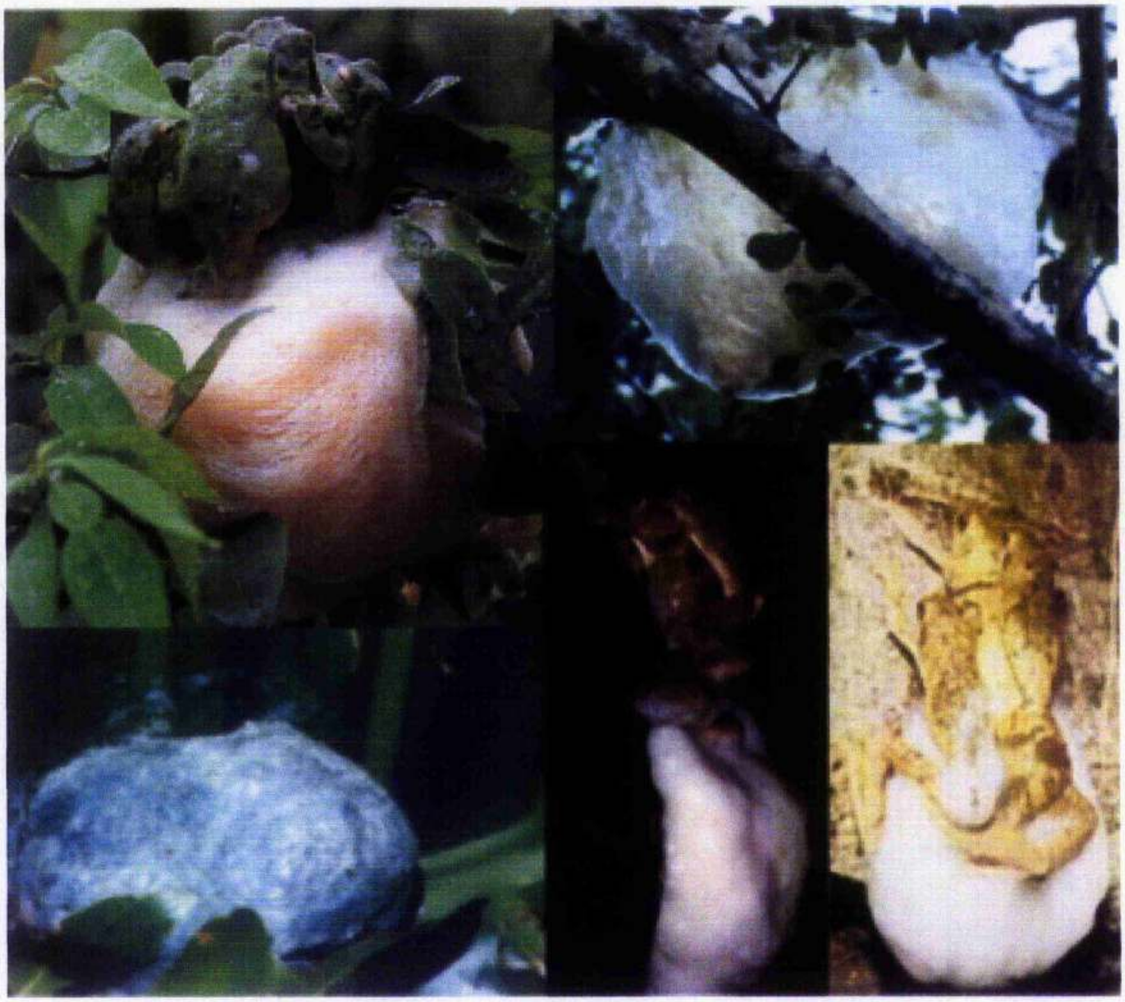


Figure 1.13. Foam nests from a number of different frog species. Clockwise starting top-left – *Rhacophorus schlegelii arborea* from Japan (© Osamu Miyake), *Chiromantis xerampelina* from South Africa, *Chiromantis xerampelina* again and in amplexus (www.thefrog.org), *Rhacophorus annamensis* from Vietnam (© Nikolai Orlov) and *Polypedates leucomystax* from Malaysia (© Monkeyfrog.com).

Further to a physical defence, the foam barrier also prevents dehydration – a crust often forms on the outside while the inside remains moist. In some cases the nests have been observed to act as incubators for the tadpoles with an increase in temperature of 8.5°C above ambient[60] and in other cases the foam acts as an insulator, regulating the tadpoles against highs and lows of temperature[61]. In *P. pustulosus* the frogs often engage in group nesting where many foam nests are produced in a small area and are attached to each other. This probably decreases the possibility of predation due to the decrease in surface area available on each nest.

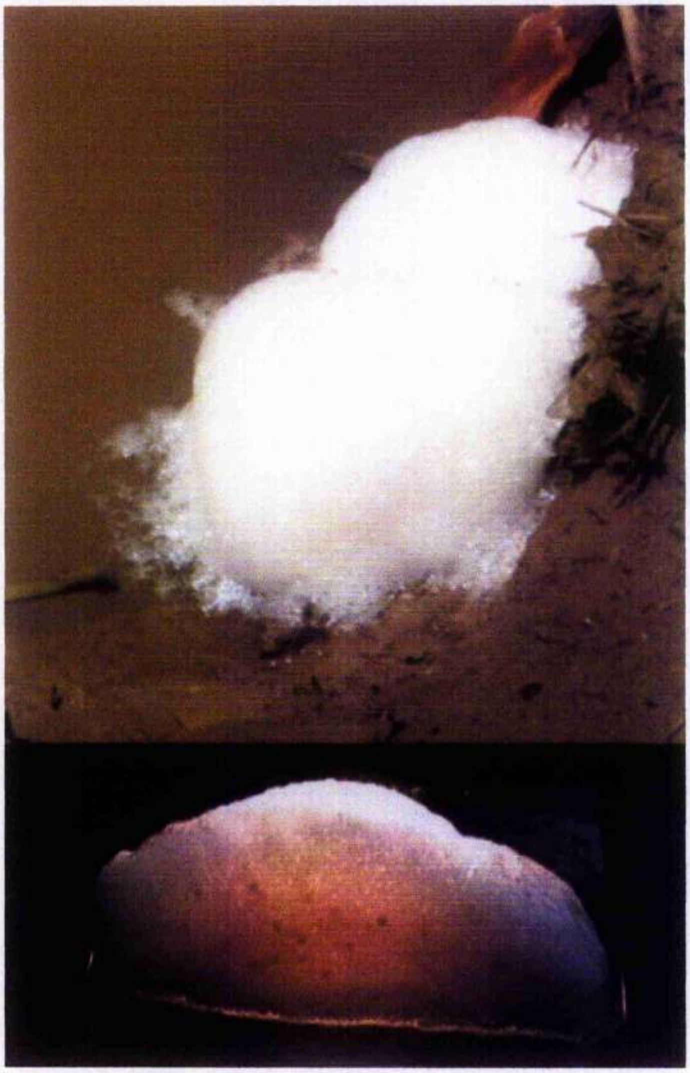


Figure 1.14. Two foam nests from *P. pustulosus* attached to the bank of an area of standing water (top) and an illuminated cross section of a foam nest (bottom, courtesy of A. Cooper).

P. pustulosus eggs are consistently arranged so that they are gathered at the core of the nest and we assume that this is a deliberate effort to prevent surface exposure or damage. Nests are generally 10-20cm in a diameter with a vaguely hemi-spherical shape, they weigh approximately 20-40 g depending on associated moisture and each nest contains ~400-500 eggs which is lower than in some larger species. Eggs are off-white in colour and approximately 2 mm in diameter.

1.4 Foam Constituents and Ranaspumins

1.4.1 Previous Work

The unusual properties of the foam material have been the subject of investigation in this group for some time and much is already known. Biophysical analysis of the whole foam has shown that it is a natural surfactant which preferentially self-assembles at the air-water interface[62]. Chemical analysis has shown that the composition of the foam is an aqueous solution of protein and carbohydrate with proteins being present at ~1-2 mg/mL concentration and carbohydrate at a similar level[63].

Recently the foam has been shown to have potential in clinical use as an anti-adhesive[64] and even more recently the group has crystallised[65] and solved the structure of a remarkable blue protein named *Ranasmurfin*, from the *Polypedates leucomystax* foam nest.

Field trips to Trinidad have allowed collection of *P. pustulosus* foam samples and frog specimens for captive breeding programs. Collaborative studies of the carbohydrate fractions have indicated diverse and high levels of different carbohydrate moieties including fucose, mannose, galactose and glucose. Digestion and mass spectrometry revealed O- and N- glycans constructed of numerous different complex chains[66]. Some of these were assigned but no complete analysis has been performed yet.

Denaturing electrophoresis of the raw foam showed a variety of discrete proteinaceous bands, six of which were excised and sequenced. These six proteins were termed *Ranaspumins* collectively and named RSN-1 to RSN-6 (technically Pp-RSN-1 to Pp-RSN-6 as these are inferred from the genetic information rather than protein sequence) based loosely on molecular masses. A great deal of work has gone into the recombinant expression of these six proteins with varying success but the eventual aim of the project is to investigate the structures and functions of all of the constituents and understand the interactions between them. By relating these properties to biological function is it hoped that useful systems might be created with applications in biomedical, environmental and chemical applications.

Each of the six sequenced proteins are individually interesting and warrant further investigation. The amino acid sequences are given in Appendix B. Using bioinformatics, each of the proteins can be compared to databases for similarities, and functionality can be inferred. Prior to starting the work detailed in this thesis, the six proteins had been identified and sequenced. Although the sequencing and initial sequence analysis was performed by others, it is included here as background information. The sequence analysis described here is an extension on what had gone previously (BLAST searches and SWISS-MODEL structural modelling).

1.4.2 RSN-1

RSN-1 is a small protein at 11.4 kDa with a slightly acidic estimated pI at 5.1. It has a glycine rich N-terminus which may indicate flexibility and a high number of proline residues, particularly in two discrete regions. There are two cysteine residues, presumably forming a cystine linkage in the native fold, and no tryptophan residues.

A similarity search using BLAST[67] indicates a similarity to mammalian cysteine proteinase inhibitors, although the similarities are not high enough to have confidence in any translation to structural or functional similarity. However, use of the Phyre (Protein HomologY Recognition Engine) to investigate similarity gives a much more confident prediction of a similarity to the same proteinase inhibitors. This method models the sequence against known structures and generates structural motifs based on similar stretches of primary sequence. Phyre is the successor to the 3D-PSSM software[68].

We hypothesise that RSN-1 may contribute towards anti-microbial behaviour in the nest by inhibition of foreign proteinases. Proteolytic enzymes are found in many microbial organisms including viruses, protozoa, bacteria, yeast and fungi and the inhibition of these can prevent pathogenic advance.

1.4.3 RSN-2

Although named RSN-2, this protein has a lower molecular mass than RSN-1 but they are very difficult to distinguish on SDS-PAGE gels. It has a relative molecular mass of 10.8 kDa and a theoretical pI of 5.2 which is also similar to the acidic RSN-1. There are a number of striking features in the primary structure of RSN-2, most obviously at the carboxy terminus where there is a sequence of six consecutive aspartic acid residues (poly-Asp_6) which is a unique motif, as far as can be determined from database searches.

Six aspartic acid residues will create a region of intense negative charge in neutral solution (the aspartic acid sidechain pK_a is 3.9) and it is unlikely that this sequence would be entirely random, and so some structural or functional properties are likely to be induced on the protein that are necessary for its role in the natural foam. In many cases a grouping of charged residues creates a metal binding domain where a central positively charged metal ion is at the centre of a metal-ligand complex. In this case, this may take the form of an octahedral complex with the charged aspartic acid residues creating six dative bonds with a suitable metal ion. Most proteins that contain a metal complex have the requirement of a redox centre for electron transport or for the electronic absorption of visible light for colouration although sometimes they are present to stabilise the overall fold. In the foam nest there is no obvious requirement for such a function although the charge interactions may be related to foaming.

Analysis of the RSN-2 sequence also indicates the highly charged nature of the constituent amino acids. The negative charge is almost entirely brought about by the extremely large number of aspartic acid residues which are distributed throughout the protein to a total of 21% compared to the normal average of 5%[69]. The negative charge is largely countered by a 19% lysine content (average is 6%). The highly charged, polar residues are distributed quite evenly throughout the sequence but there are some regions with a more hydrophobic character (mostly at the amino terminus) and the sequence has a high leucine, isoleucine and valine content totalling 23.2% which is only slightly above the 21% average but is more meaningful when considered in the otherwise extremely polar environment. The highly charged nature of the protein is demonstrated by the GRAVY (Grand Average of Hydropathicity) value of -0.9[70] suggesting a highly soluble protein and the distribution of this charge as well as the hydrophobic regions are shown in Figure 1.15 as a hydrophobicity plot.

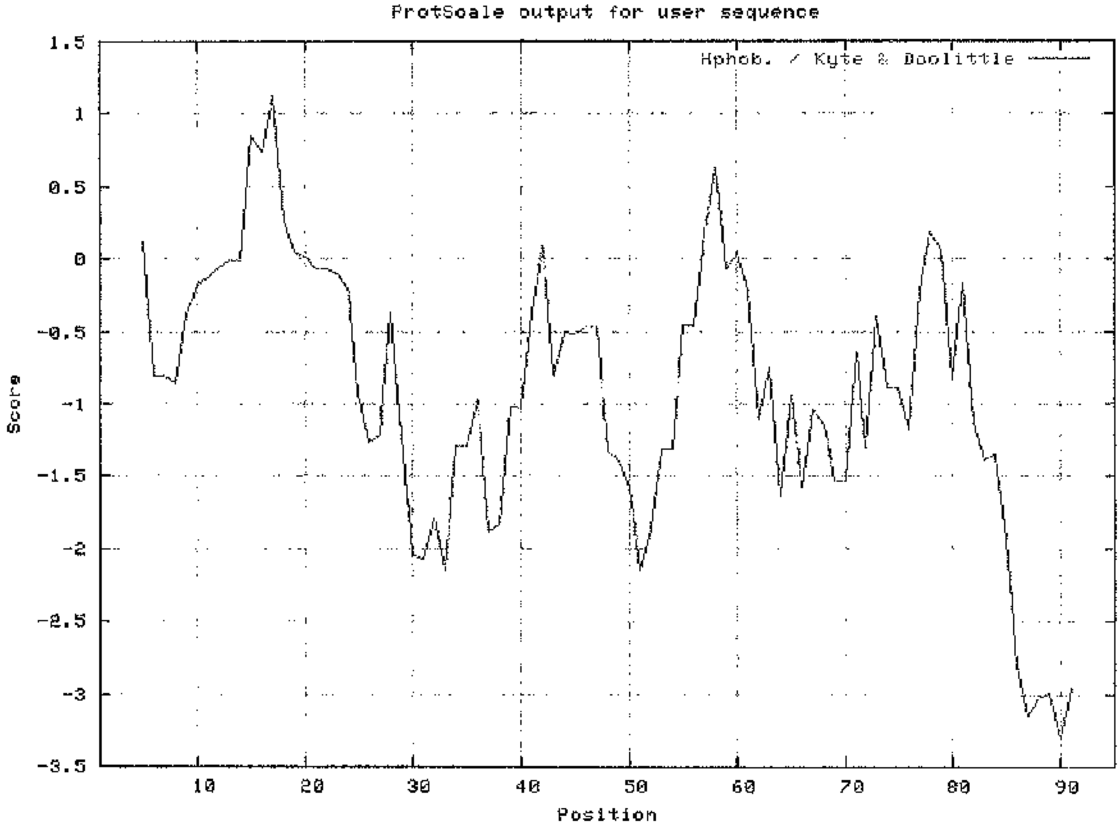


Figure 1.15. Hydrophobicity plot for the amino acid sequence of RSN-2. The residues are numbered from the N- to the C-terminus.

Database searches have confirmed the highly unusual nature of the RSN-2 sequence. A simple BLAST search revealed no similarity to any known protein and only one similarity to a hypothetical protein – a 34 kDa thioredoxin reductase from *Thermosipho melanesiensis* (a bacterium) with a very low confidence level and this is unlikely to have any significance. Using PHYRE, some predicted folds could be modelled but not with any degree of confidence. Interestingly, the PHYRE predecessor 3D-PSSM identified a similarity to the same cystatins identified for RSN-1 but with much less confidence. Close examination of the results of these searches suggested that short peptide sequences could have some similarity to known proteins but when taken as a whole the protein could not be confidently mapped to any known protein family.

Although similarity searches did not suggest any biological function for RSN-2 it was hypothesised that the charge distribution in the sequence may be simplistically indicative of a source for the foaming ability of the foam fluid. This idea is based largely on the hydrophobicity plot showing a relatively hydrophobic N-terminus and an extremely charged C-terminus and this division of sequence into an amphiphilic whole may mimic classic surfactants such as SDS (sodium dodecylsulphate). It has also been postulated that a highly charged molecule is more likely to be surface active if the charges are nearly balancing as they are in RSN-2[71]. SDS contains a charged sulphate "head group" and a long, oily, hydrophobic tail which is driven to the air-water interface (due to water repulsion) whilst maintaining solubility. Whilst these detergent molecules are highly efficient, they disrupt biological membranes and as such are not compatible with natural systems.

The concept of protein surfactants is not new and the foaming of any concentrated protein solution is well known. This foaming is thought to be brought about by irreversible denaturation from shear forces although some recent evidence appears to suggest that the shear forces involved in normal laboratory procedures would not be enough to denature most proteins[72]. Irrespective of the mechanism of typical protein foaming, true protein surfactants operate in a different manner and are characterised by surface activity at a relatively low concentration. Surface activity is often accompanied by denaturation and often these structure alterations are irreversible. For example, shear denaturing of lysozyme in egg whites.

There are several examples of protein surfactants found in nature, with varying levels of activity and a range of mechanisms. The term "protein surfactant" is often used to describe the group of proteins found in mammalian lungs where they allow full inflation. In some cases, such as prematurely born infants, the lung surfactants are not present and must be replaced with fully or partially synthetic alternatives. The structural basis for the activity in these proteins is based on cooperative association in multimers and with other lipid and metal ion components[73].

Caseins are responsible for the foaming behaviour found in milk. β-casein has a native unfolded structure (very little secondary structure in the native fold), and associates with charged calcium and phosphate groups to form an emulsion of micelles in solution, although the exact nature of this is still not exactly known[74]. Although not as strong as in some other surfactants, bovine serum albumin is a well-known protein surfactant and the ability to bind to surfaces is used widely as a blocking agent in immunochemistry.

Easily the most powerful natural protein surfactants known are the hydrophobins, secreted from fungi to modify aqueous surfaces and promote growth. At the air water interface the hydrophobins dramatically reduce surface tension[75], allowing the fungus to escape the aqueous environment and then form a protective, hydrophobic film on the surface to protect the fungus in the air[76]. Analyses of these solutions have shown that the protein exist as oligomers in solution but contain a hydrophobic surface patch that drives migration to the air-water interface. The structure is stabilised by a network of four disulphide bridges (a large number for these ~100 amino acid proteins) which prevents internalisation of the hydrophobic patches and preserves the rigid globular structure[77] and a recent surface study has indicated that the film created is an extremely elastic monolayer, capable of supporting a variety of different shapes[78].

It is proposed here that RSN-2 is a protein surfactant, either in its own right or by interaction with other components and its unusual sequence indicates a possible new mechanism. As such, RSN-2 may represent an extremely important part of the foam matrix and could present an avenue for the development of potential applications.

1.4.4 RSN-3 to RSN-6

The remaining four *Ranaspumins* span a molecular mass range of 18 – 26 kDa and estimated pI range of 5.8 – 7.7. All of the proteins are related because database searches show that they are all similar to lectins. These similarities are much more confident than the matches found for RSN-1 and 2, with strong similarity between each of the other four proteins and a large number of known lectins

Lectins are proteins that specifically bind carbohydrates via carbohydrate recognition domains (CRDs) and are ubiquitous throughout living organisms. Family divisions are usually made based upon the particular sugar moieties recognised and a very large number have been described – a search for "lectin" on the NCBI protein database returns 4095 reference sequence entries. Although the biological functions of lectins are unknown in many cases, they undoubtedly hold a role in cell adhesion and it is this function that has led to their widespread use in human blood-group typing. This ability is due to the different specificity of lectins for surface carbohydrates on different red blood cell types; different blood groups are agglutinated by different lectins.

Each of the four *Ranaspumins* (3, 4, 5 and 6) were compared by a multiple sequence alignment shown in Appendix C and it was observed that RSN- 3, 4 and 5 have a similar sequence at their core and it would be reasonable to assume that they have similar structures. RSN-6 appears to align poorly with the other three and it is unlikely that they share structure.

A BLAST search indicates that RSN-3, 4 and 5 are similar to the fucolectins, that is lectins that bind fucose or the terminal fucose on a glycan chain, and a PHYRE search predicted similarity to a galactose binding domain with more certainty. Further analysis of the BLAST results gave some confidence that these three proteins were all members of the F-type lectin family represented by similarity to the most well characterised member of the family of agglutinins, from *Anguilla anguilla*, the European fresh water eel. This protein binds fucose-termini on pathogens in serum and the crystal structure has been solved[79] to show the novel fold binding fucose. The crystal structure also shows how small sequence deviations at the binding site can significantly alter sugar recognition.

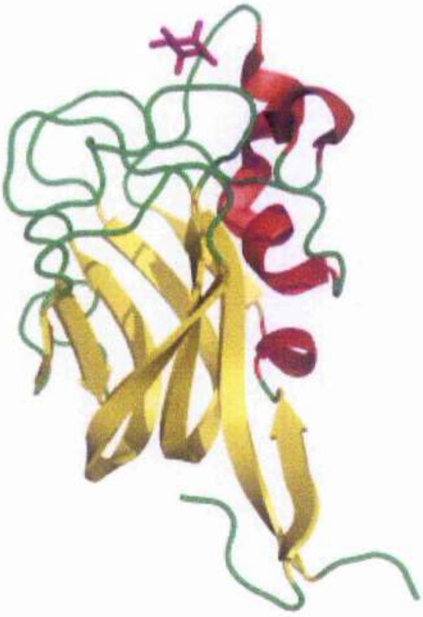


Figure 1.16. The crystal structure of an F-type Lectin binding fucose (PDB 1K12).

The structure is a β-barrel, held in place by two disulphide bridges allowing presentation of a collection of five loops for fucose binding. In the crystal it was found to form a non-covalent homotrimer which may enhance affinity. The affinity may also be enhanced by calcium binding in a sub-domain. By making the reasonable assumption of structural conservation in this family, the structures of RSN-3-5 have been modelled[63].

F-type lectins are most commonly found in fish and many other animals and bacteria but it has not been observed in plants, fungi, mammals, birds or reptiles. They are often observed as tandem domains thought to be due to gene duplication and sometimes as fusions with other functionalities[80]. The number of observed family members is growing, most recently with the discovery of an F-type lectin from gilt head bream (*Sparus aurata*) serum[81]. Using multiple structure alignments (Clustal_X[82]) RSN-3-5 have been shown to belong firmly to the F-type lectin family due to generally good alignment and particular conservation of specific residues for structure and sugar binding. This is shown in Appendix C with the following abbreviations for different organisms, following the alignment in reference 80: Msa, *Morone saxatilis* (striped bass); Mch, *Morone chrysops*(white bass); Gac, *Gasterosteus aculeatus* (stickleback); Fru, *Fugu rubripes* (tiger pufferfish); Aja, *Anguilla japonica* (Japanese eel); Omy, *Oncorhyncus mykiss* (rainbow trout); Dre, *Daniorerio* (zebrafish); Cca, *Cyprinus carpio* (carp); Ler, *Leucoraja erinacea* (skate); Xla, *Xenopus laevis* (tetraploid clawed frog); Xtr, *Xenopus tropicalis* (diploid clawed frog); Ame, *Ambystoma mexicanum* (axolotl); Spu, *Strongylocentrotus purpuratus* (sea urchin); Dme, *Drosophila melanogaster* (fly); Aga, *Anopheles gambiae* (mosquito); Ttr, *Tachypleus tridentatus* (horseshoe crab); Cvi, *Crassostrea virginica* (oyster); Dja, *Dugesia japonica* (flatworm); Spn, *Streptococcus pneumoniae*; Mde, *Microbulbifer degradans*; SS, Sargasso sea. The number and diversity of these family members suggests an important functional role.

It is common for members of this family to function as part of the organism's innate immune system, by recognising bacterial lipopolysaccharides and other surface structures. The diversity of the F-type lectins found in different organisms, and in many cases within the same organism, are thought to be present for defence against a wide variety of pathogens and this may be the case in RSN-3, 4 and 5[83]. The alignments suggest that RSN-3 and 5 are very similar, RSN-4 is more divergent within the family and together they possibly present protection against nest invasion from a large number of bacterial and other pathogens.

RSN-6 belongs to another lectin family and probably has a similar biological function with a different target. BLAST searches confidently predict that RSN-6 is a member of the C-type lectin (calcium dependant) family and it shows especially close similarity (41% amino acid identity) to the fish egg lectin (FEL) protein derived from the eggs of *Cyprinus carpio*[84]. Although not as numerous as members of the F-type lectin family, there are several known members of the C-type lectin family which are subdivided into other sub-groups based on function[85] and it may be that RSN-6 may incorporate a galactose or mannose binding domain to support the fucosyl binding of other nest proteins. This mechanism is more likely than the suggestion that these are membrane-bound proteins, as suggested by a PHYRE search and elsewhere[86], as the protein freely exists in the nest solution (although it may bind foreign membranes).

Members of the family that are similar to RSN-6 are from fish or crabs but there are also members (called Tectonins) from a slime mould. Recently these have been added to by a lectin from *Ephydatia fluviatilis*, a fresh water sponge[87]. An alignment of the family members confirms that RSN-6 has the tachylectin repeat sequence highlighted in Appendix C. Interestingly, all members of this family are aquatic or have a high exposure to water which corresponds well to foam nest proteins.

It is likely that RSN-6 is another piece of armour in the nest for the defence against pathogens in the environment. The combination of the RSN-6 (C-type lectin), RSN-3, 4 and 5 (F-type lectins) and RSN-1 (cystatin) is likely to infer strong protection for the developing tadpoles against microbial invaders.

These proteins may have host defence functions in the nest, but it has also been suggested[62] that they may form part the structural matrix of the nest. The related theory is that the lectins bind the incorporated carbohydrates, which are known to be present in high concentration in the foam fluid, and multiple bindings can create a three dimensional structure. In this way, the identified proteins have all been postulated to have useful nest functions with RSN-1 and the lectins being defence proteins, RSN-2 creating foaming activity and the lectins and carbohydrates inferring stability on the macrostructure.

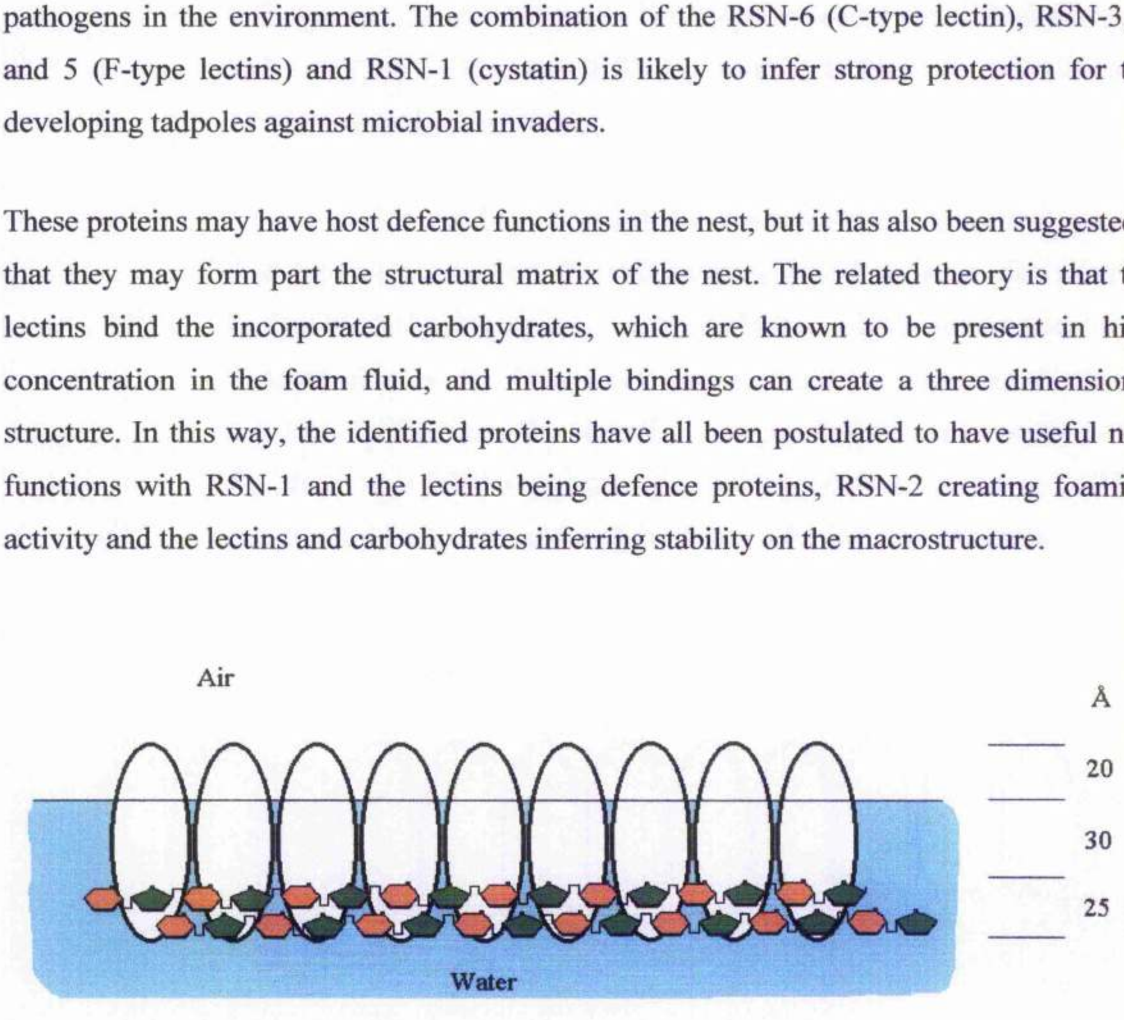


Figure 1.17. A model of possible assembly of foam nest proteins and carbohydrates at the air interface (courtesy of A. Cooper).

1.5 Objectives

In this thesis, the hypotheses for the protein functions will be tested. This will focus on the particularly unusual protein RSN-2 primarily, but will undoubtedly encroach on the other functions and structures of the other proteins in such a complex system. The content of the nest will be investigated using proteomic techniques and compared with the known constituents.

The investigation of RSN-2 will use recombinant protein technology and this will allow the structure to be probed at the atomic level using high resolution NMR. In addition, the structure and function of RSN-2 will be examined at the air-water interface to examine the presence of any of the proposed surface activity and the results will be examined with respect to the foam properties.

2 Protein Production

2.1 Introduction

From an initial observation of the unusual foam material, deconvoluting the functions of individual components or combinations of these components was the next phase in the discovery process. *P. pustulosus* frogs are not an endangered species and are relatively common in their natural habitat and, as such, the foam material could be collected in large enough quantities for investigation.

An alternative, parallel approach was to use recombinant protein technologies to produce synthetic versions of the natural proteins. This approach is essential in structural studies where molecular substitutions or isotopic enrichments are required. Recombinant protein expression is routine practice in modern biochemical laboratories, with numerous different expression and purification systems now being available meaning that the recombinant route is often preferential to purification of native proteins; this is especially true where samples are limited.

2.2 Collection of Foam

Pairs of *P. pustulosus* were collected, with permission from the Trinidad and Tobago Wildlife Authority, from the Northern Range of Trinidad during the wet season (July – August) 2004. During the hours of approximately 2200 – 0200, known breeding areas were examined for mating pairs. Once in amplexus, the pairs were carefully placed in a container for transport to laboratories at the University of the West Indies, St. Augustine and transferred to a glass tank with ~3cm water. Usually, the pairs produced foam nests during the night. After several hours, the foam nest was collected and manually separated from the eggs with spatulas before being frozen at -20°C and eventual shipment to the UK on dry ice. It was important to remove the eggs as the freezing process bursts the outer membrane and would contaminate the nest with albumins and other egg proteins.

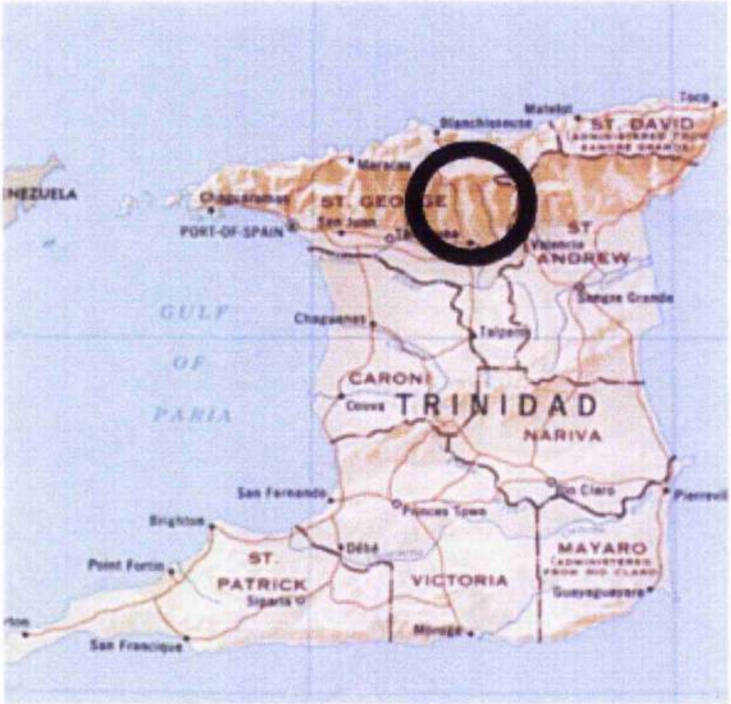


Figure 2.1. Map of Trinidad with the Northern Range highlighted.

Approximately 5 litres of foam were collected and transported in this way. Once in the UK foam samples were stored in various sized aliquots and in various freezers at -20°C or -80°C along with eggs and frog tissue samples. This foam supply was used for at least 18 months with no obvious change in appearance or composition.

For general experiments, the foam was sampled whilst still frozen, as it was more manageable in this form, and then allowed to thaw at room temperature. Various strategies were employed to convert the foam into a fluid including dissolution and sonication but the most convenient way was found to be centrifugation. This resulted in a foam fluid with a protein content >1mg/mL (by the Bradford assay[88]) and a small amount of undissolved material on top. The solid portion could be dissolved by mixing, followed by further centrifugation, sonication, mixing or NaOH wash and was seen to be of identical composition to the soluble fraction by electrophoresis.

Some foam samples from a laboratory breeding colony of *P. pustulosus*, for comparison, were also obtained as a gift from Dr. Mike Ryan, University of Texas, Austin.

2.2.1 SDS-PAGE for Foam Analysis

Foam samples were analysed and compared by SDS-PAGE (Sodium DodecylSulphate PolyAcrylamide Gel Electrophoresis). This method uses an electric potential gradient to move charged proteins through a porous matrix which behaves like a sieve. Proteins migrate according to their charge (roughly correlated with size) and their mobility is retarded by physical contacts with the gel pores. In this work the gels were prepared and visualised using the general method of *Laemmli*[89].

Proteinaceous solutions were treated with a mixture of reagents to ensure that they were in the correct state for analysis. SDS was added to 0.5-1% by weight, which is thought to denature proteins by fully encapsulating the surface via hydrophobic interactions between the detergent tail and the peptide sidechains. Each protein is evenly coated in SDS molecules which are highly negatively charged and mask any protein charges. The result of this is that the proteins are all unfolded into different lengths of linear chain dependant on size, and all of the proteins have the same charge per surface area without any contribution from intrinsic protein charge. When applying an electric charge across these proteins, they all move at the same rate due to their identical electrophoretic mobilities but by blocking their path with a porous gel the proteins move at different rates due to their varying degree of retardation.

A reducing agent, dithiothreitol (DTT) at 50-100 mM final concentration, was sometimes added to break disulphide bonds in the proteins and ensure complete unfolding. This is a redox agent that is capable of reducing the protein disulphides in favour of oxidation of free thiols and the formation of an internal disulphide bond in itself – thiol-disulphide exchange. The disulphide is preferentially formed in the DTT molecule rather than the protein due to the energetic favourability of forming a six-membered ring in the DTT molecule[90].

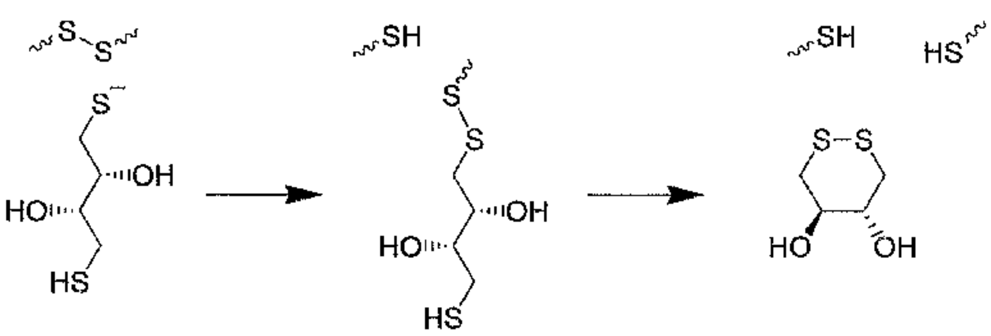


Figure 2.2. The reduction of protein disulphide bonds by DTT.

Also added to the solutions was 20% by volume glycerol to make the solution dense and easy to apply to the gels, 0.05% bromophenol blue as a small molecule marker dye to track the progress of the electrophoresis and a concentrated buffer to maintain pH control (1.5M

Tris.HCl pH 8.5). Typically, the solution was mixed and then boiled in a block heater for approximately ten minutes before applying to the gels.

Gels were usually cast by polymerising an acrylamide solution between glass plates but on some occasions pre-cast gels were purchased (Bio-Rad Ready Gels or Invitrogen NuPage) to take advantage of the increased resolution of gradient gels. Gradient gels have a higher acrylamide concentration at the bottom of the gel relative to the top and give better resolution of proteins over a wide molecular weight range compared to the gels prepared in-house which were of a single acrylamide concentration – typically 15% compared to 8-16% in pre-cast. Homemade gels were prepared in two parts: the resolving gel and the stacking gel. The resolving gel was the bulk of the slab and is the area where separation takes place and the stacking gel was poured on top to create an area of large pore size where diffuse samples were concentrated to a tight band prior to separation. Acrylamide monomer (30% solution acrylamide:bis-acrylamide 37.5:1, Bio-Rad) was mixed with SDS and buffer (1.5M Tris.HCl pH 8.5 for the resolving gel, 0.5M Tris.HCl pH 6.8 for stacking) to the appropriate concentration and the polymerisation was initiated with 0.5% APS (ammonium persulphate) and TEMED (tetramethylenediamine) was included in catalytic quantities for rapid polymerisation. Polymerisation was allowed to proceed in a gel mould until use.

To use the gels the electrophoresis tank was filled with running buffer (25mM Tris.HCl, 192mM Glycine 0.1% SDS, pH 8.3) and the gels were immersed. After sample loading the electrophoresis was allowed to proceed under an applied voltage of 150V until the dye front was nearly at the bottom of the gel, at which time the gels were removed. Proteins were fixed and stained by shaking in a Coomassie blue solution (0.125% Coomassie brilliant blue R-250, 50% methanol, 10% acetic acid) and then destained in methanol/acetic acid solution. For recording, the gel images they were photographed using a Kodak Image Station 4000 system and analysed with Kodak 1D 3.5 software. An example of an SDS-PAGE gel of foam nest fluid is shown in Figure 2.3.

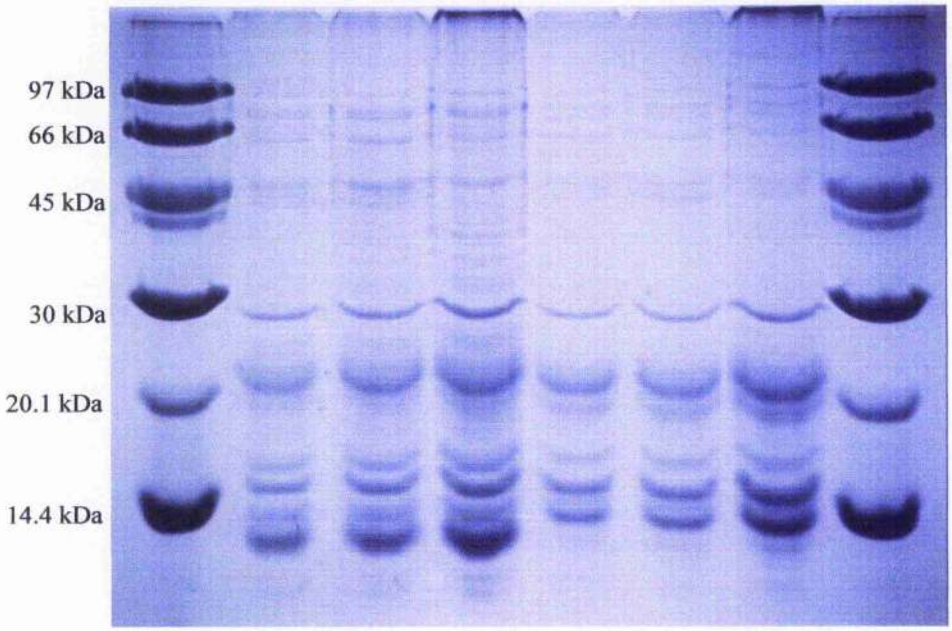


Figure 2.3. **SDS-PAGE of *P. pustulosus* foam nests from Trinidad and Texas (gift from Mike Ryan, University of Texas at Austin). Lanes 1 and 8 are molecular weight markers. Lanes 2-4 are 5µL insoluble Trinidadian foam, 5 and 10 µL soluble Trinidadian foam respectively. Lanes 5-7 are duplicated with Texas foam.**

2.3 Molecular Biology and Recombinant Protein Expression

2.3.1 Recombinant Protein Technologies

The advances in solid phase peptide synthesis and in couplings means that small proteins can now be synthesised from freely available starting materials but the relatively low yields, high expense and time taken for such reactions means that a biochemical route is normally preferred. These techniques use genetic engineering to take advantage of the highly efficient and robust protein production facilities already existing within cellular bodies. There are a number of steps which are followed to produce a protein of interest from a foreign expression system and the following steps which were used in this project are typical. It should be noted that these procedures were the work of Dr. R. Fleming (although many of the steps were replicated later by the author) and are only included for background information and as means of an introduction.

From the SDS-PAGE gels, the principle bands were excised and subjected to N-terminal sequencing (University of Leeds). Usually this is done by Edman degradation[91] which involves cycles of derivitisation and hydrolysis of the terminal N-terminal peptide bond and sensitive detection. This technique doesn't provide full length sequence information because the hydrolysis reaction is not complete in every cycle, and gets progressively less accurate as sequencing proceeds, so molecular biology techniques are used to yield the full sequence instead.

One method for extending the known sequence information is Rapid Amplification of cDNA Ends (RACE). This technique produces complementary DNA (cDNA) from the transcribed messenger RNA (mRNA) for the target gene. This avoids problems associated with using genomic DNA as a template for gene duplication because mRNA doesn't contain any of the redundant information contained in the introns that can be present in genomic DNA of higher organisms. In this case Trizol (Invitrogen) reagent was used to extract total RNA (tRNA) from female *P. pustulosus* oviducts and were assumed to contain mRNA as a genetic template. RACE is performed in two directions of reverse transcription – 3' and 5'. The easiest to perform is 3'-RACE because it can take advantage of the natural poly(A) tail of mRNA as a generic site for amplification.

Nucleic acid amplification is performed by variations of the Polymerase Chain Reaction (PCR) which has allowed recombinant technologies to be performed as part of typical laboratory procedures. The history of the invention of this technique is somewhat controversial and it can be dated back to the early nineteen seventies[92] but the most important part of the development was the adaptation of the polymerase enzyme (the catalyst for nucleic acid polymerisation) from *Thermophilus aquaticus* (Taq) for use in *in vitro* DNA replication by variable temperature cycles. This enzyme remains stable at the elevated temperatures required for the essential DNA denaturation and annealing that allow the technique to proceed. Before proceeding with PCR experiments, something must be known about the gene sequence of interest so that short sequences of complementary DNA can be designed called primers. These primers should bracket the sequence of interest so that it can be multiplied.

The reagents required for PCR experiments are a DNA source, primers, polymerase and a source of nucleic acid building blocks (dNTPs). In the first step, DNA is heated to ~95°C to separate the double strands. The temperature is then lowered to an annealing temperature which is dependant on the primers used but is between 50 and 75°C and the primers in solution hybridise to the target DNA. After annealing, the complementary DNA is manufactured during an extension step at 72°C during which time the polymerase builds the nucleic acid chain along the strands from the bases provided in the initial solution mixture. The number of base pairs produced in this process is dependant on the time allowed for extension and the efficiency of the polymerase but it is only required to extend past the binding site of the other primer. These steps are repeated for many cycles to produce many millions or billions of copies of the target sequence because the target sequence is amplified exponentially (due to the use of the bracketing primers) whereas the other lengths of DNA (from only one primer) are only amplified linearly. A diagram of this process is given in Figure 2.4.

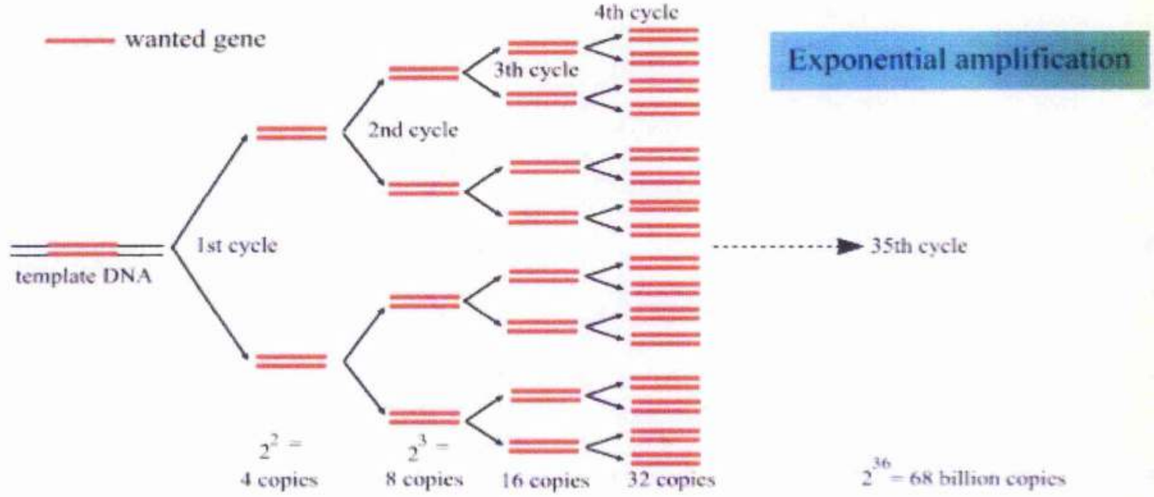


Figure 2.4. Progression of the polymerase chain reaction to create billions of copies of a single gene of interest.

Taking advantage of PCR techniques the 3'-RACE experiment uses a poly(T) primer to bind the poly(A) motif on the mRNA and a gene specific primer (GSP) to bind to the known N-terminal sequence for amplification of a stretch of nucleic acid containing the gene of interest. cDNA is created from mRNA using the reverse transcriptase enzyme rather than the polymerase enzyme, accordingly the technique is termed reverse transcriptase PCR (RT-PCR). The previous work in this group has used a RACE kit (Clontech) that also transfers a poly(C) motif to the end 3' end of the mRNA, allowing the use of a generic poly(G) primer for 5'-RACE. Using the GSP and the generic primers the complete cDNA for the gene of interest could be amplified. The full protein sequences were then identified by cloning the cDNA into plasmid DNA and obtaining nucleic acid sequences through which protein sequence could be inferred.

Plasmids are DNA molecules that can be present in organisms but are not part of the chromosomal DNA. They can be present in high numbers and take advantage of the replication machinery of the host. In molecular biology, the plasmids used for genetic engineering are called vectors and in this work, and all previous work on this project, all vectors have been specific for *Escherichia coli* bacteria. These vectors are typically double stranded and circular in shape and contain restriction sites which can be cut by restriction enzymes. Following the cut (linearisation), DNA of interest can be inserted into the vector using the DNA ligase enzyme which joins the ends of the DNA strands if the restriction sites and enzymes are designed in a specific way. In this project these time-consuming steps were avoided by use of the TOPO® cloning vectors (Invitrogen) which are supplied as linearised DNA with overhanging thymidine (T) bases at the ends which efficiently ligate to the adenosine (A) overhang created from the use of Taq polymerase in PCR amplification.

Successfully cloned PCR products can then be introduced into an expression system for cellular production of the protein of interest by transfection. This was accomplished practically by altering the cellular membrane of competent cells by heat shock to make them permeable to the vector. Different bacterial strains suitable for different purposes such as storage or over-expression can be used as the host organism and an isolated plasmid can be retained and stored for long periods and transfected to many different strains. Alternatively, a plasmid can be recovered from cellular stocks by DNA extraction (miniprep) and purification (agarose gels which are similar to SDS-PAGE but with a smaller pore size and no denaturing agent).

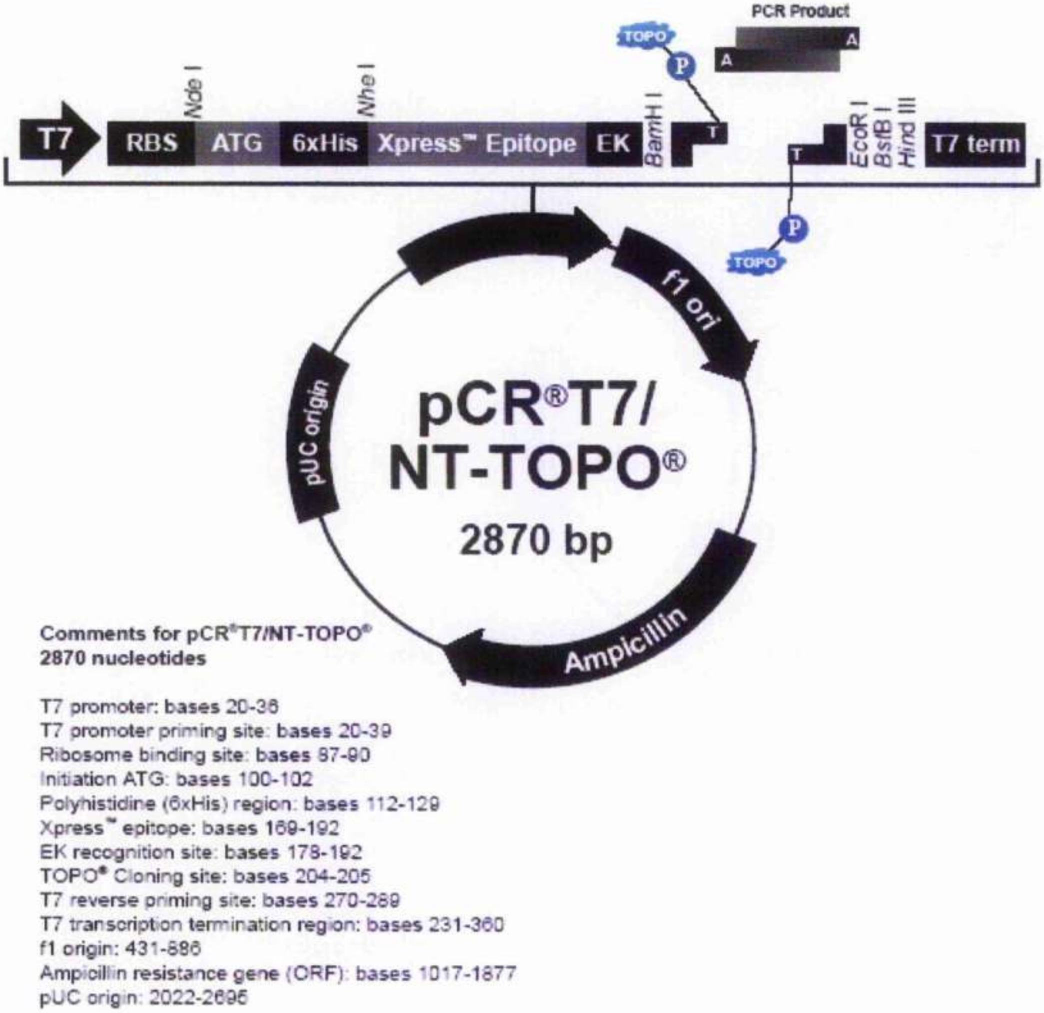
2.3.2 Existing Genetic Constructs

Using the steps described, the six *Ranaspumins* were isolated, sequenced and cloned into expression vectors by Dr. R. Fleming who made kind gifts of them for this work. The RSN-2 gene, like all of the other *Ranaspumin* gene sequences, had been cloned into an expression vector (pCR® T7/NT-TOPO®, Invitrogen) using RSN-2-specific primers and mRNA from a female *P. pustulosus* oviduct. In addition to the RSN-2 gene, elongated primers had been used to add a stop codon to the C-terminus and bases coding for an extra isoleucine amino acid adduct on the N-terminus. In addition to the PCR product, this vector contains a number of other useful features:

- An ampicillin resistant genetic sequence coding for the β-lactamase enzyme which inhibits the antibacterial action of ampicillin, an analogue of penicillin.
- A pUC origin sequence which ensures high copy numbers of the gene of interest (provided the correct bacterial strain is chosen).
- A 6xHis sequence which codes for six histidine residues together on the amino terminus of the expressed protein. The 6xHis (poly-His$_6$) residues strongly bind divalent nickel ions and so behave as a useful tag for protein purification.
- An Xpress™ epitope which is another affinity tag.

- An enterokinase (Ek) cleavage recognition site for removal of the fusion tag post purification.

The combination of these genes in a commercial vector is designed for efficient and rapid over-expression of the target protein fused with an extra polypeptide tail containing sequences for easy purification and subsequent cleavage to yield the target protein in a format that is very close to the native structure. The vector map is shown in Figure 2.5 with the point of insertion of the PCR product clearly identified with thymidine overhangs.



Comments for pCR®T7/NT-TOPO®
2870 nucleotides

T7 promoter: bases 20-36
T7 promoter priming site: bases 20-39
Ribosome binding site: bases 87-90
Initiation ATG: bases 100-102
Polyhistidine (6xHis) region: bases 112-129
Xpress™ epitope: bases 169-192
EK recognition site: bases 178-192
TOPO® Cloning site: bases 204-205
T7 reverse priming site: bases 270-289
T7 transcription termination region: bases 231-360
f1 origin: 431-886
Ampicillin resistance gene (ORF): bases 1017-1877
pUC origin: 2022-2696

Figure 2.5. The expression vector used to clone the *Ranaspumin* genes (from Invitrogen).

2.3.3 His$_6$-Ek-RSN-2 Overexpression

RSN-2 was overexpressed as the fusion-tagged form (His$_6$-Ek-RSN-2) in this vector and *E. coli* as the host cell. The particular strain chosen was BL21(DE3) (Novagen) which was chosen for high expression levels due to chromosomal expression of the T7 polymerase enzyme which is, in turn, under control of the *lacUV5* promoter. Other bacterial strains possess more complicated genetic features for tighter control of expression and inhibition of the basal expression of toxic or proteolytic products but it was decided to use the simplest host in the first instance. Because the T7 expression is controlled by a *lac* promoter, the expression can effectively switched on by addition of the lactose analogue IPTG (Isopropyl β-D-1-thiogalactopyranoside). IPTG imitates lactose for induction but is not a substrate for β-galactosidase which would normally "switch off" the expression. Instead of this, expression continues throughout bacterial life.

All materials were sterilised either by autoclaving or 0.22μm filtering solutions, and manual handling was performed in an aseptic manner. Frozen glycerol stock of RSN-2 in this vector and BL21(DE3) *E. coli* strain were sampled and grown overnight in ~20mL LB (*Lysogeny Broth*) with 100μg/mL ampicillin at 37°C. LB broth was prepared by dissolution of 10g/L Tryptone (Duchefa), 5g/L (Difco) and 10g/L sodium chloride (Fisher) in deionised water and provides the required nutrients for a rich growth medium for bacteria. The following day the cell suspension was evenly distributed amongst two 2 L conical flasks, each containing 500 mL of LB-amp selective growth medium and the flasks were shaken at 37°C and open to air (cotton wool plug to prevent contamination) to promote cell growth. An overnight culture was used as a starting point because it contains a high number of cells to initiate further growth. The cell multiplication was measured by optical density of a solution at visible wavelengths using a spectrophotometer. The measured absorbance is actually scatter from the cells and so is related to the number of cells in solution. Cells were observed to grow slowly at first as they adapted to their environment and began to divide until they reached an optimum concentration for growth when they grew much more rapidly, known as the log (logarithmic) growth phase. This continued until the cells excretions became toxic and cells numbers gradually depleted.

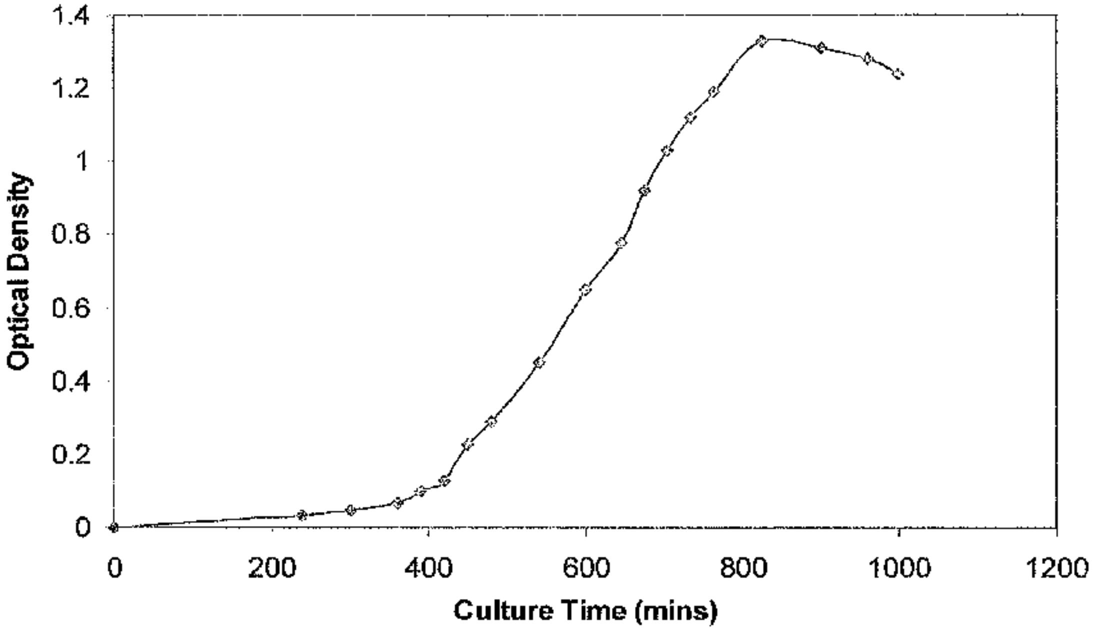


Figure 2.6. Cell growth curve for *E. coli* in LB broth during RSN-2 expression.

This growth curve is typical for cells and the procedure for inoculation and harvesting is designed to take advantage of the observed stages. During the initial growth period the T7 polymerase was switched off so as not to stress the cells any more than necessary. During the mid-log phase ($OD_{550} \approx 0.5$) the polymerase was switched on by addition of a 1 M sterile solution of IPTG (Melford Labs.) to a final concentration of 1 mM and this induced RSN-2 expression during a high division stage for maximum protein yield. The media was incubated and shaken for a further four hours to allow cell numbers to grow to a maximum but minimise cell death. Cells were harvested by centrifugation on a swing-bucket centrifuge (Sartorius) at 4000 x g (gravity) for 20 minutes and the cell pellet was collected.

2.3.4 Purification

Although these expression systems were designed to express the target protein in high concentration, the cells will also express other proteins native to their organism and purification is required. Because *E. coli* expresses the proteins in the intracellular space the first stage of the purification is to break open (lyse) the cells to access the collection of expressed proteins. The harvested cells were dispersed in a small volume (~30 mL) of NTA binding buffer (20 mM Tris.HCl, 500 mM NaCl, 5 mM imidazole, pH 7.9) and cell lysis was produced by high energy ultrasound using a sonic probe. The sonication process was exothermic and a large amount of heat was produced due to the high cell count and the heat could cause protein denaturation. To avoid this, the cell suspension was immersed in a double ice bath for cooling and the ultrasonic energy was only used in short bursts – thirty cycles of 20 second bursts with 30 second rest periods. The resulting protein solution was separated from cell debris by centrifugation at 10000 rpm in a fixed rotor centrifuge for 20 minutes and the supernatant was retained for purification by nickel affinity chromatography.

Recombinant RSN-2 was separated from other unwanted protein by binding the poly-His$_6$ region of the N-terminal fusion through nickel chelation. Initially the resin was activated by flushing the column with 50 mM nickel sulphate solution and then equilibrated with binding buffer before passing the crude RSN-2 solution through. Other weakly binding proteins were removed by step increases in imidazole concentration (competitive nickel binder) in the wash buffer (20 mM Tris.HCl, 500 mM NaCl, 120 mM imidazole, pH 7.9) and then elution buffer (20 mM Tris.HCl, 500 mM NaCl, 1 M imidazole, pH 7.9) which was used to elute the recombinant protein. Following purification the column was stripped for re-use by rinsing with a 100 mM EDTA solution.

The poly-His$_6$ containing fusion tag was useful for initial bulk purification but before studying the protein it would be useful to remove the tag as it represents a significant amount of the total protein (28% of residues) and may interfere with the native fold or function. Fusion tag cleavage was provided for by the inclusion of a proteolytic enzyme recognition sequence near the N-terminus of the native protein (i.e. the C-terminus of the fusion tag). In particular, the Asp-Asp-Asp-Asp-Lys↓ sequence is recognised by enteropeptidase (often called enterokinase but it does not phosphorylate) which is a naturally occurring serine protease found in the mammalian digestive system[93].

Crude RSN-2 solution was exchanged into an optimised buffer (according to the manufacturer's instructions) by dialysis. Cleavage buffer was 10 mM Tris.HCl 10 mM CaCl$_2$ pH 8.0 and dialysis was achieved using a 2 kDa MWCO dialysis cassette (Slide-A-Lyzer, Pierce). 50 μL of this solution was treated with 1 unit of enteropeptidase (purified from bovine intestine, Sigma) and cleavage was allowed to proceed at room temperature. The resulting mixture showed a size shift on SDS-PAGE gel relative to the fusion tagged RSN-2 but the protein band was slightly blurred and inconclusive as to the efficiency of the protein cleavage.

The enteropeptidase reaction was further investigated by mass spectrometry (Sir Henry Welcome Functional Genomics Facility, University of Glasgow). In the particular system used (Applied Biosystems Q-Star pulsar) the solution was introduced to the spectrometer as a spray and then ionised with a strong electric field. Detection was by a Time of Flight (TOF) module, which is described in more detail elsewhere in this thesis. The spectrum showed no presence of the correct cleavage product and suggested that the protein had been cleaved predominantly at the incorrect site as well as numerous secondary cleavage sites. From the mass/charge value obtained, it appeared that cleavage took place near the C-terminus of the native sequence where a natural enterokinase recognition site exists. Several attempts were made to vary pH, temperature and buffer composition but none of these parameters affected the cleavage product.

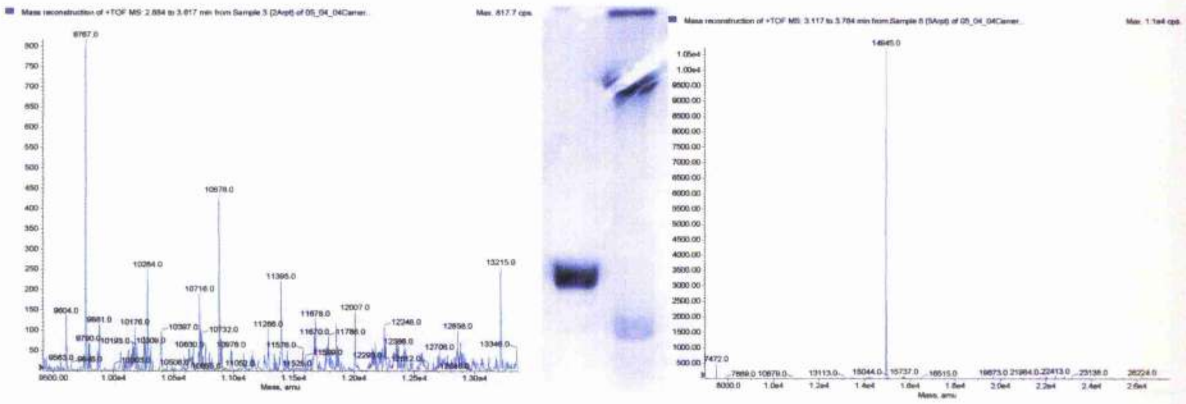


Figure 2.7. Mass spectrum of RSN-2-tag (right) and RSN-2-tag with enterokinase cleavage (left). Complementary SDS-PAGE is also shown (lane 1 – uncleaved, lane 2 – cleaved).

2.3.5 Engineering of Thrombin Cleavage Site

In order to create recombinant material as similar as possible to the native material, an alternative strategy was devised whereby a cleavage recognition site would be engineered into the protein between the existing enterokinase recognition site and the native N-terminus. It was decided to introduce a thrombin recognition site because there is no natural thrombin recognition site in the RSN-2 sequence or the existing fusion tag and because high quality thrombin is freely available.

To test the stability of the recombinant protein with respect to thrombin cleavage, a ~5mg His_6-Ek-RSN-2 sample was incubated for 48 hours with 20 units of thrombin (Sigma, purified from human plasma). Size analysis of the thrombin-RSN-2 sample and the control (no thrombin) by SDS-PAGE suggested that there was no thrombin-induced hydrolysis in the existing recombinant protein. Accordingly, thrombin was deemed to be a good candidate for accurate fusion tag hydrolysis and this approach was continued.

Like enteropeptidase, thrombin is a serine proteinase and is part of the coagulation cascade that results in blood clotting. In particular thrombin is responsible for the transformation of soluble fibrinogen molecules to insoluble fibrin which cross-links to strengthen the clotted platelets. The recognition site for thrombin mediated hydrolysis can be diverse but an arginine residue followed by a hydrophobic residue is the generalised substrate and the Pro-Arg-↓-Gly-X sequence is often used as a functional insert where X is a hydrophobic residue and cleavage occurs between the arginine and glycine residues[94]. Thrombin recognition sites are commonly engineered into genetic constructs and are often combined with Glutathione S-Transferase (GST) tags.

A sample of cDNA derived via mRNA from oviducts of a female *P. pustulosus* (provided by Dr. R. Fleming) was used for RT-PCR as per the vector manufacturers' instructions with modified primers. The C-terminus primer was left unchanged (coding for a STOP) but the N-terminus primer was designed to code an additional thrombin recognition site at the N-terminus:

Forward -	CTG	GTT	CCG	CGT	GGA	TCC	TTA	ATA	TTA	GAT	GGG	GAC
5'3' Translation -	L	V	P	R	G	S	L	I	L	D	G	D
Reverse -	CTA	ATA	TCC	ATC	ATC	ATC	ATC					
3'5' Translation -	D	D	D	D	G	Y	**Stop**					

The PCR product was separated by agarose gel electrophoresis and visualised by soaking with ethidium bromide. Ethidium bromide is a planar fluorescent molecule that binds in between the bases of nucleic acids (intercalation). Treated gels were exposed to Ultra Violet (limited to short times as DNA can be damaged by UV exposure) and the fluorescence observed in the visible range of light. The major band was excised from the gel with a scalpel and the DNA was extracted using DNA-affinity spin columns in a DNA gel extraction kit (Qiagen). This DNA sequence was then transformed into competent *E. coli* cells as described previously and spread on an LB-Amp infused micro agar plate and allowed to grow overnight.

Eight single colonies were picked from the plate and grown separately in a small volume of LB media for analysis. Plasmid DNA was extracted from the cells by a miniprep kit (Qiagen) which employed alkaline cell lysis and column DNA purification. A portion of the plasmid DNA was linearised by treatment with a restriction enzyme EcoR1 (Roche) at 37°C for 2 hours and subjected to agarose gel electrophoresis. The observed size correlated well with the expected 3200 base pairs for the cloned RSN-2 gene in the expression vector. PCR was used as a diagnostic for the presence of the RSN-2 sequence by analysing the product when performing PCR of the plasmid DNA with the RSN-2 forward primer and a universal T7 reverse primer. In this case, the PCR product appeared to be the correct size (291 bp). Confirmation of the success of the cloning was made by DNA sequencing (DBS Genomics) with the fluorescent dye terminator method.

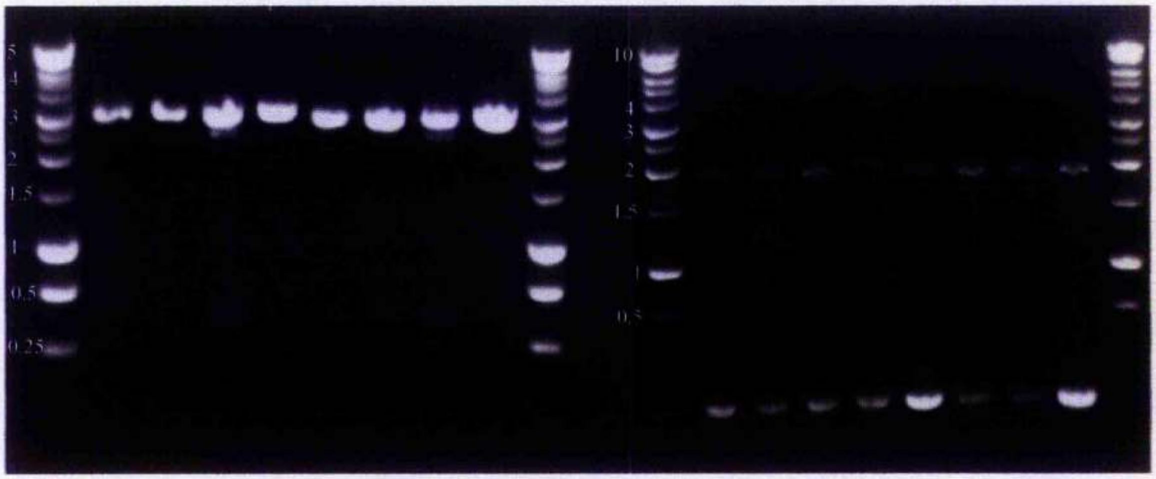


Figure 2.8. Agarose gels showing DNA size of full plasmid (left) and amplified insert (right). The outside lanes are DNA mass ladders (Promega) with the sizes in kilobases annotated.

For RSN-2 over-expression cells were grown on 2 L media scale and were observed to behave in a very similar manner to the previous construct. Cells were harvested and the protein purification was largely based on the previous methods used for the RSN-2 recombinant protein with soluble cell extracts being purified by Nickel affinity chromatography to give reasonably pure His$_6$-Throm-RSN-2 at 10-20 mg per litre of media yield.

2.3.6 Thrombin Cleavage and Purification

Small aliquots of nickel-affinity purified RSN-2 were incubated with thrombin to test the fusion tag cleavage procedure. The reaction produced three discrete bands on an SDS-PAGE gel corresponding to uncleaved protein and the two cleavage products but MALDI-TOF mass spectrometry analysis showed a mixture of products. By analysing the mass calibrated spectrum, peaks corresponding to various N-terminus cleavage products were observed.

Observed Peak	% Intensity	Proposed Fragment	Calculated Remaining Protein Mass
11264.36	97	Native	11264.87
10780.23	21	GSLIL	10781.26
10666.11	100	GSLILD	10666.18
10023.98	6	GSLILDGDLLKD	10024.45

Table 2.1. Fragmentation observed in mass spectrum of thrombin treated recombinant RSN-2. The proposed fragments are also included.

The mass spectrum suggested that the N-terminus of the protein was being non-specifically cleaved with some fragments having greater stability than others and so were present in greater intensity. Although the native protein was present at high levels some effort was made to reduce the level of fragmentation so that yields were higher and subsequent purification more simple. Variations in pH, temperature and buffer composition were investigated for affects on fragmentation, and only temperature was seen to have a significant influence. It was found that the extent of secondary cleavage at the N-terminus could be controlled by performing cleavage at low-temperature (4°C) but it was found (not surprisingly) that this also reduced reaction rate, by a factor of approximately four.

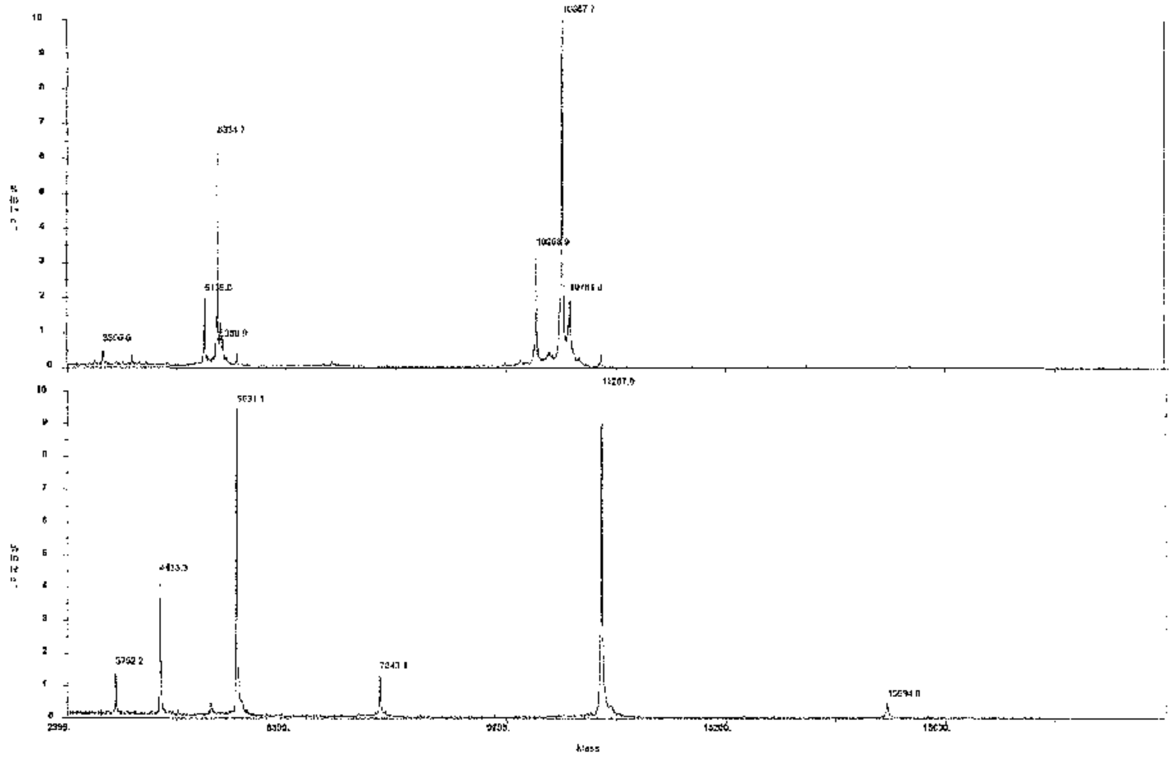


Figure 2.9. MALDI-TOF MS of thrombin cleavage of His$_6$-Throm-RSN-2 at room temperature (upper) and 4°C (lower). In the lower spectrum, some of the uncleaved product can still be observed.

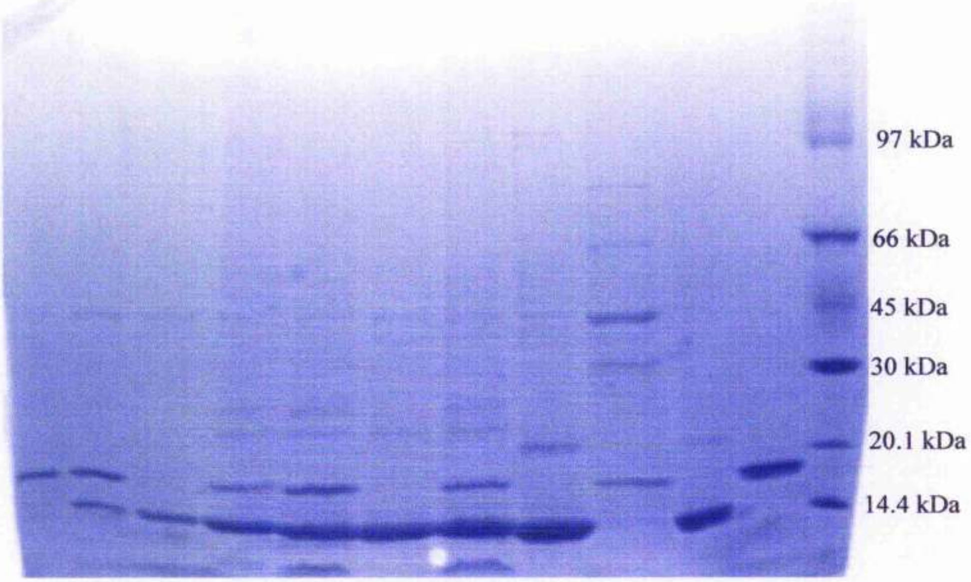


Figure 2.10. SDS-PAGE with Coomassie staining. Lanes 1-8 are RSN-2-tag with increasing thrombin loading, lane 9 is a negative control, lanes 10 and 11 are positive controls and lane 12 is a Low Molecular weight marker kit (GE Healthcare).

Using the Amersham Low Molecular Weight markers as references, densitometry tracings (using the Kodak Imager and associated software) of the contents of the SDS-PAGE gels suggested a molecular mass of cleaved RSN-2 ~15 kDa (~18 kDa before cleavage of the 4.4 kDa fusion tag) compared with the true molecular mass of 11268 Da. The difference between reducing and non-reducing (using DTT) SDS-PAGE was a shift to slightly higher molecular weight in the reduced sample suggesting that the two cysteine residues are involved in a disulphide bridge which, when broken, allows the protein to unfold in such a way as to give it a larger apparent molecular radius.

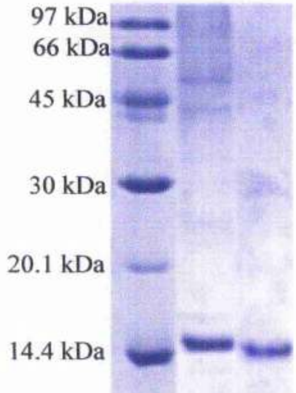


Figure 2.11. SDS-PAGE of recombinant RSN-2 reduced (lane 2) and not reduced (lane 3) with molecular weight markers (lane 1).

To streamline the purification process, it was found that the RSN-2 fusion tag could be cleaved whilst still bound to the nickel affinity column but the time taken for cleavage was found to increase by a factor of approximately 3 compared with the solution reaction. It was also noted that the efficiency and purity of the enzyme varied between different supplier batches: in one case (Sigma) cleavage was observed exclusively at D/K (elucidated from mass spectrometry) indicating Factor Xa contamination. The use of restriction grade recombinant thrombin (Novagen) eliminated this problem. It was also found that cleavage proceeded readily in nickel affinity elution buffer or PBS as well as the recommended optimum buffer (30 mM Tris, 150 mM NaCl, 2.5 mM $CaCl_2$ pH 8.2).

After complete cleavage was confirmed by mass spectrometry, the residual thrombin was removed by passing the solution through an immobilised benzamidine resin (GE Healthcare). Benzamidine reversibly binds thrombin at neutral pH and the column can be regenerated by washing with low and high pH cycles. The flow-through from the column containing RSN-2 was collected and augmented with PMSF (phenylmethylsulfonyl fluoride) to a final concentration of 0.5 mM. Like benzamidine, PMSF binds serine proteinases such as thrombin and in this case it acts as an inhibitor against any further proteolytic activity. Cleaved fusion tag was removed by passing the solution through the nickel affinity column as described previously. In doing this, the fusion tag was retained on the column and the target protein was collected in the flow through and wash fractions.

Gel filtration chromatography was used as a final polishing step to remove residual small molecules and aggregate. In this method the chromatographic resin is composed of porous beads of defined sizes and porosities. When mixtures flow through the resin the molecules that are small enough to access the pores enter them and flow through a longer path than the large molecules that do not enter the pores. Media can be prepared in two different ways – a defined single particle and pore size where molecules are separated by an absolute molecular weight cut-off (analogous to dialysis) or alternatively particles and pores are created in a distribution of sizes resulting in a resolving column where all different sizes within a range are separated relative to each other. In a system such as this, small molecule contaminants are removed as well as any large contaminants and multimers/aggregate to yield a highly pure final product.

Typically, 1mL of cleaved recombinant RSN-2 (rRSN-2) solution was introduced onto a Superdex™ 75 240 x 10mm (GE Healthcare) column using a BioRad Biologic F FPLC or an ÄKTA Purifier system. Flow-rate was 0.5 mL/min resulting in a pressure of approximately 50 psi and online monitoring of A_{280} was used to collect the RSN-2 fraction of ~1.5 mL (3 x 0.5 mL fractions). An example chromatogram is shown in Figure 2.12.

Using a calibration solution (Amersham Low Molecular Weight Gel Filtration Calibration Kit) containing proteins of known and well defined molecular masses, the apparent molecular mass could be estimated from the time taken to elute from the column. In these native conditions RSN-2 appeared as a protein with an apparent molecular mass of ~18 kDa. This way of estimating mass is inherently inaccurate as it is actually measuring the ability or inability of a protein to enter the pores in the media relative to proteins that behave as ideal, globular spherical proteins rather than making any actual measurement of mass. For this reason, this estimation is as much of shape as it is of mass. However, the fact that RSN-2 appears to be considerably larger than its true mass (native in gel filtration and denatured in SDS-PAGE) suggests that it is either i) an elongated shape, ii) associated as a dimer or iii) interacting with the media in the column and the gel in such a way that entry to the pores is retarded.

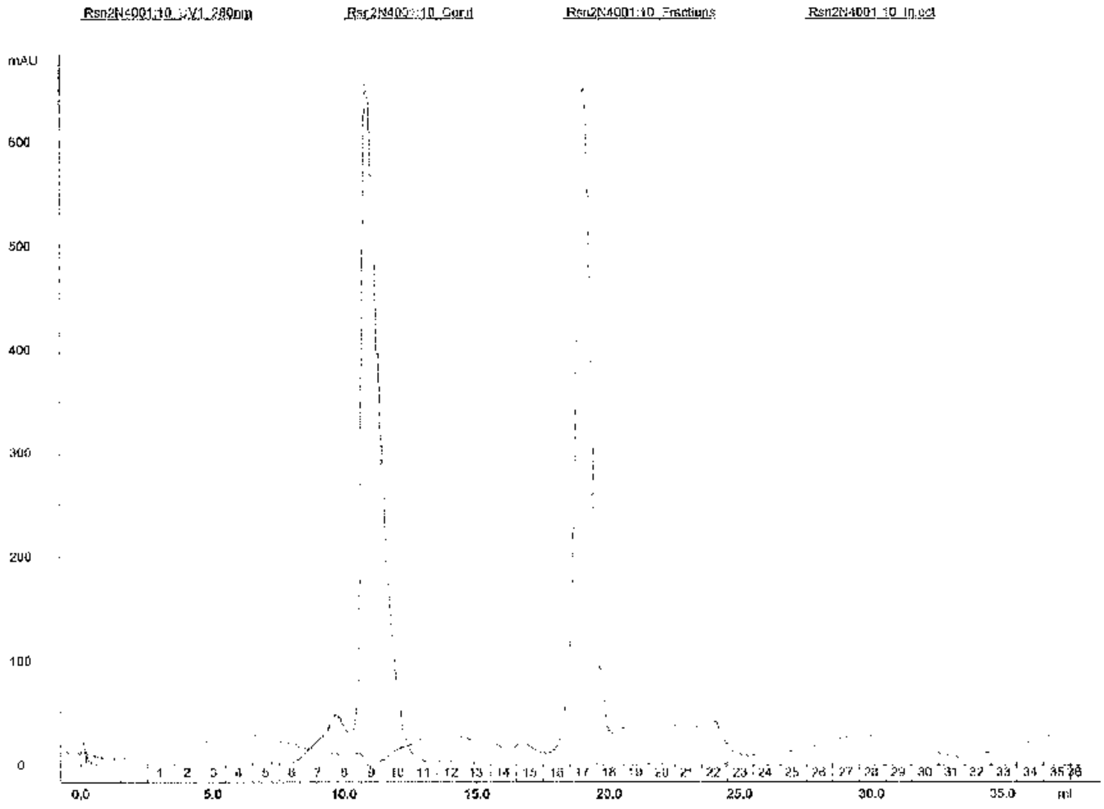


Figure 2.12. Gel filtration chromatography of rRSN-2. The aggregate peak can be seen eluting before the main RSN-2 peak by A_{280} (blue) and the small molecule elution is indicated by an increase in conductivity (brown).

After purification, the RSN-2 fractions were pooled and 20mL Vivaspin centrifugal concentrators (Sartorius) were used for ultrafiltration and/or diafiltration to give a protein solution at the appropriate concentration and buffer composition. Samples were normally used immediately but were found to be stable to storage at -20°C or -80°C for prolonged periods.

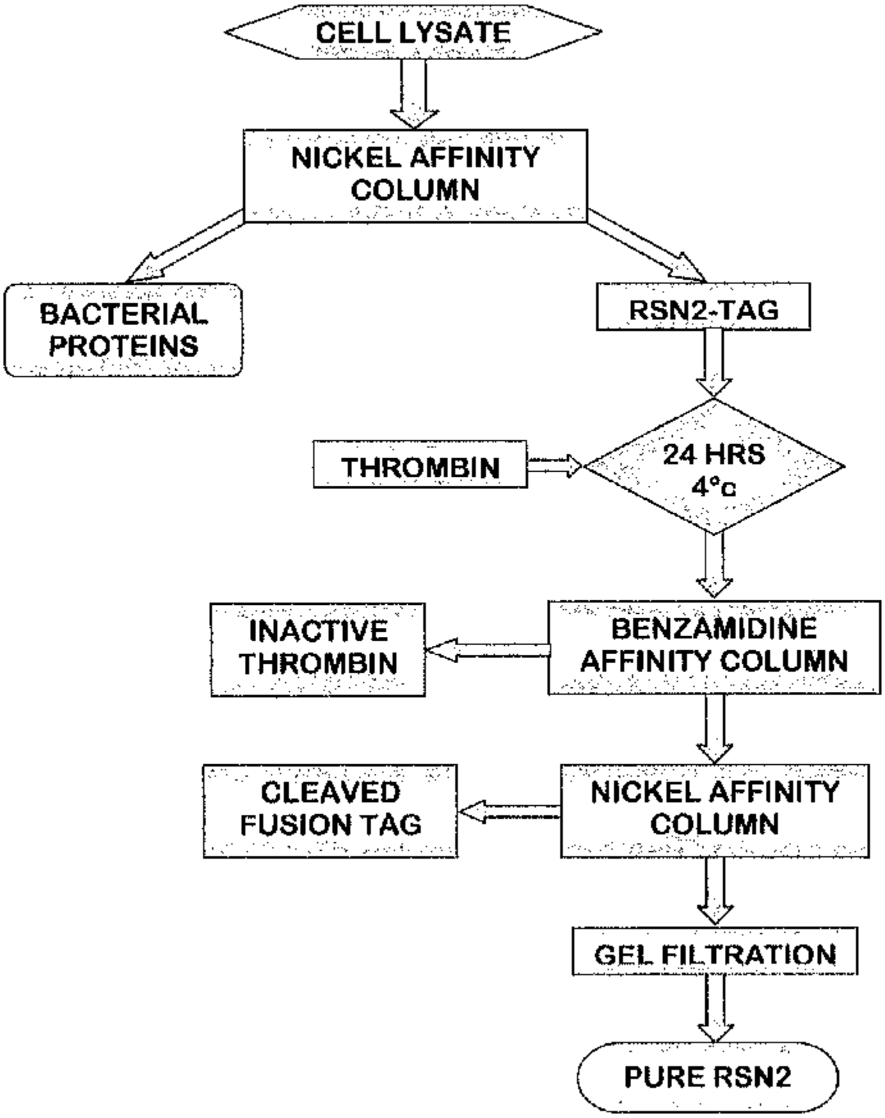


Figure 2.13. Scheme summarising the production and purification of rRSN-2.

2.4 Separation of Native Foam

2.4.1 Separation by Membrane Filtration

Having established that the foam material contains a variety of proteins with a range of molecular masses of 11kDa to 66kDa, attempts were made to separate the material into fractions of varying masses. Centrifugal concentrators were used first as a crude cut-off technique. These concentrators contain a membrane of defined pore size, allowing molecules with a Stokes radius smaller than the pore size to pass through the membrane and larger molecules are retained. Trans-membrane pressure is applied by centrifugal force perpendicular to the membrane.

Using concentrators of varying pore size or Molecular Weight Cut Off (MWCO), the foam mixture was dissolved and passed across different membranes to create solutions containing molecules of defined hydrodynamic size. It was intended that the larger proteins might be separated from the smaller ones, and at the least, proteins would be separated from peptides and small carbohydrates. Filtrate and retentate samples from each of the concentrators were analysed by SDS-PAGE and/or MALDI-TOF mass spectrometry to detect biomolecules.

A major difficulty in this technique was encountered due to the cellulose-based membranes interacting with some of the proteins. This was assumed to be a sugar-lectin binding due to RSN-3, 4, 5 and 6. The effect of this was to form a gel-layer across the membrane which blocked the pores and prevented any protein flux. To prevent this binding, a PES (Polyethersulfone) membrane was used in the concentrators. Despite this, it was found that no significant proteins or peptides were detected in the filtrate in all but the largest pore size (100 kDa), in which case only one peptide was found to have eluted. No glycans were detected, although most effort was concentrated on the search for proteins and peptides.

These results suggest that the proteins do not behave as ideal single molecules expected from their molecular mass. Instead they may interact with each other or with surfaces to create a system of molecules which behave as a collective matrix rather than individual molecules.

2.4.2 Separation by Gel Filtration Chromatography

Gel filtration chromatography (GFC, sometimes called gel permeation or size exclusion chromatography) was used to separate the foam constituents into molecular size ranges. Foam fluid was passed through a Superdex™ 75 (GE Healthcare) column in 1 mL injections at 0.5mL/min using PBS as the eluant. Protein elution was monitored by A_{280}. Some separation was observed with two large peaks and several smaller peaks being resolved. The largest of the protein peaks (earliest eluting) had an apparent molecular mass of approximately 65kDa which is in good agreement with the largest of the proteins found on SDS-PAGE. The next detected peak had a M_r of approximately 15kDa and was most likely to contain the RSN-2 fraction. Those peaks eluting after this were probably smaller than could be detected on the SDS-PAGE gels run (15% Tris-Glycine) and are more likely to be peptides of varying sizes. The relative size of the peak and the wide distribution of sizes suggest that these peptides make up a large amount of the foam fluid.

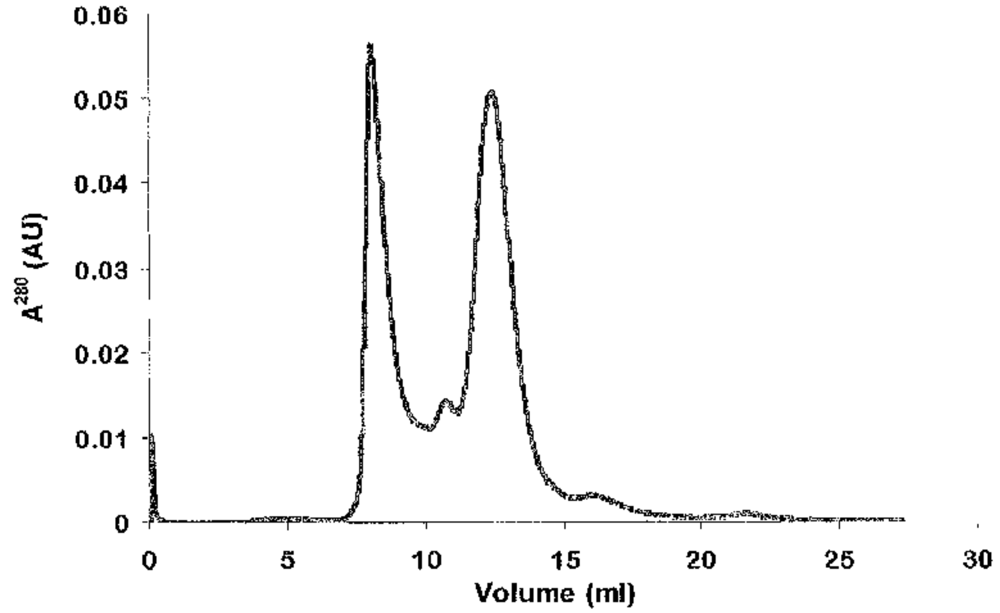


Figure 2.14. Gel filtration chromatography of the foam fluid.

2.4.3 Other Chromatographic Methods

Three fractions were collected from gel filtration chromatography but SDS-PAGE revealed that these were only crudely separated by size and were mixtures of a number of proteins/peptides. As the next step in the purification, other chromatography methods were investigated.

Trial experiments were conducted on ion-exchange chromatography to separate each fraction by relative charge. Ion exchange chromatography uses charge-charge attractions or repulsions to separate molecules with different dielectric properties. In the case of proteins these properties extend to pI, charge distribution and hydrophobicity/hydrophobicity. For ion exchange to be successful it is normally initiated by absorbing the target molecule on to the charged media by an electrostatic interaction (purification can be obtained if the target flows through and all contaminants adsorb to the media but this is unusual). This is achieved by selecting the appropriate media (positive media for anion exchange and negative media for cation exchange) and the appropriate buffer system for the sample in terms of pH and ionic strength so that the target molecule is primed for resin adsorption. After selective binding the resin is washed and then the adsorbed molecules are eluted by changing either the pH or the ionic strength. By altering the pH the protein charge is altered so that it no longer binds to the column. Increasing ionic strength causes desorption because the high concentration of charged ions interact with the column and displace the proteins which have weaker interactions.

Foam proteins were bound to Mono-Q or Mono-S (GE Healthcare) columns depending on their predicted pI and the buffer pH. Bound proteins were eluted by a gradient change of pH or increasing ionic strength. A number of different systems were used with altering combinations of media, buffer pH and elution conditions but none of these were found to produce pure proteins at the first pass.

Predicted pI's of the known *Ranaspumins* are all in the range of 5-7 so the primary strategy used was anion-exchange chromatography of proteins buffered at a mildly basic pH and this was the most successful strategy. In the most promising of the runs, proteins were bound to the Mono-Q (quaternary ammonium group) column by equilibrating the system in 50mM Tris.HCl pH 8.0 and eluted with a 0→1 M NaCl gradient over 30 minutes. Whilst this system showed promise, it was found to be very difficult to separate some of the proteins. The eventual aim of this work was to purify native RSN-2 from the foam but ion-exchange was found to be unsuccessful for this, probably because the predicted[95] pI's of RSN-1 and RSN-2 are very similar (5.1 and 5.5 respectively). It is likely that the correct conditions and a specifically tailored gradient elution could allow individual components to be separated by IEX but time and resource constraints did not allow this. Further development of this system would be a source of potentially useful future work.

On analysis of the primary sequences of RSN-1 and RSN-2, it was noted that the main differentiating factor of the intrinsic physical properties was the highly charged nature of RSN-2 with respect to RSN-1. It was therefore proposed to first fractionate the foam by means of size by gel-filtration chromatography and then apply the RSN-2-containg fraction to hydrophobic interaction chromatography (HIC). In this form of chromatography, dissolved proteins adsorb onto the media through hydrophobic interactions, driven by the high salt concentration of the initial binding conditions. The salt is thought[96] to interact with the water to decrease order at the surfaces of the solutes and media and the decrease in entropy from this promotes protein binding. Although not exactly the same, it can be considered to be a similar mechanism to the salting-out of proteins in solution.

Different proteins bind to HIC resins with different affinites and a range of substituted media is available to give optimum separation. Proteins are eluted by decreasing ionic strength in the mobile phase, therefore increasing the water interaction with the media ligand and the solutes and the proteins desorb. HIC is similar to reverse phase chromatography in that it binds proteins through hydrophobic interactions but is (usually) a much less harsh technique in terms of protein denaturing.

Because HIC functions by binding hydrophobic patches, it was assumed that the more hydrophobic proteins in the foam would bind more strongly or at lower salt concentration than more hydrophilic proteins (especially RSN-2). Therefore, to effectively separate RSN-2 from the other proteins, all proteins would need to be bound to the column and then lowering the salt concentration in the buffer would elute proteins. To find the correct conditions for high protein binding capacity and selective desorption, the HiTrap™ HIC test kit (GE Healthcare) was used. Of all the different ligands used to substitute the polymeric resin, only the phenyl sepharose showed encouraging signs of protein binding at a starting ammonium sulphate concentration of 1.5 M.

Phenyl sepharose Fast Flow™ (GE healthcare) was obtained as a pre-swollen slurry and was packed in a 10mm internal diameter, 20cm bed height column. Foam fractions from the gel filtration chromatography elutions were diluted in binding buffer (starting high salt concentration) and injected onto the column. Various conditions were used to try to bind RSN-2 to the column but none were found to be successful. This was probably due to the extremely charged nature of RSN-2, making it more stable in high salt concentrations and less inclined to bind to the media. At a starting ammonium sulphate concentration of 2M a large proportion of the proteins bound to the column but the eluted proteins (and large peptides) were still impure.

Due to time constraints and equipment accessibility, it was not possible to further increase the starting salt concentration or adjust the other method parameters. RSN-2 was not fully purified from the foam material but this is an area that could be investigated further. Because of the narrow size range of proteins and the large amount of peptides, it is almost certain that complementary IEX and/or HIC methods would need to be used, especially to fully separate RSN-1 and RSN-2 from each other. The most efficient method would probably be using immuno-chromatography developed via recombinant proteins but this would require an initial outlay in time and finance.

2.5 Summary

RSN-2 was cloned into an altered bacterial expression system including a thrombin recognition site at the amino terminus. A series of chromatographic steps were used to purify the recombinant protein in yields that were high enough for further analysis. Native RSN-2 could not be successfully purified from the foam nest fluid but some promising inroads were made into separation of the foam components.

3 Proteomics

3.1 Rationale

Having discovered that a large portion of the foam material was composed of low molecular weight proteins or peptides, the logical next step was to investigate the composition of these and also the less abundant higher molecular weight protein components. It was hypothesised that, because the major bands had been excised from an SDS-PAGE and sequenced, the six known *Ranaspumins* were the main proteinaceous species in the foam. A proteomic approach was chosen to test this hypothesis.

Modern proteomics has been largely driven by the advances in molecular biology, mass spectrometry and genomic studies, but originates from the early reports of two-dimensional electrophoresis[97] and particularly the realisation that it could be used to separate thousands of proteins from complex biological sources[98]. The advances in robotics, mass spectrometry and the increased use of bioinformatics has meant that it is possible to analyse and interpret the spots on the gels on a reasonable time-scale and it has been accepted as the starting point for a systematic analysis of protein-containing systems.

Two-dimensional electrophoresis (2-DE) is a technique that separates proteins based on two different physico-chemical properties - size (as in SDS-PAGE) and also by the pH-dependent net charge of the individual protein using isoelectric focusing (IEF). Similarly to SDS-PAGE, IEF separates proteins by applying a high voltage to the solution. There is a pH gradient between the electrodes where the local environment has a specific pH. The pH gradient is usually created by a range of ampholytes which are small molecules containing both positive and negatively charged groups. These have traditionally been in the format of a solution but modern systems use immobilised ampholytes on a permeable solid (gel) strip that is more easy to use and more reproducible due to controlled commercial manufacturing.

Protein samples are introduced to the pH gradient in solution where they adopt a charge which is dependant on the pI of the specific protein. When a high voltage is applied to the solution the charged protein moves towards the electrode with opposite charge and passes through the pH gradient. This continues until the protein reaches the point in the gradient where the pH is equal to the pI of the protein. At this point the net charge on the protein is zero and so there is no longer any attraction or repulsion to either of the electrodes and the protein is essentially immobilised. In a solid (gel) strip, this is especially useful as the separated proteins can be easily transferred for further manipulation. As each protein in a solution undergoes this effect (as long as the pI is within the pH gradient selected), a mixture of many proteins can be separated with extremely high resolution.

IEF is the first step, or dimension, in 2-DE where the second step is an orthogonal separation by SDS-PAGE. Individual proteins retain the separation obtained in IEF and are further resolved by size to produce a slab of individual spots rather than the bands normally seen in single dimension separations.

3.2 2D Electrophoresis

3.2.1 Proof of Concept

Some samples prove extremely difficult to separate on 2-DE due to protein-gel interactions or ionic strength of the solution so, as a proof of concept exercise and as a guide for future experiments, the initial experiment was chosen to be on a small scale with a broad pH and molecular weight range. For efficient IEF it is important to prepare the solution in a relatively low-conductivity solution (i.e. low ionic strength) because the high voltages employed would otherwise create excess current and therefore heat in the system. As the ionic strength of the foam fluid was unknown, it was decided to use undiluted foam as a starting point and this would also give an indication of protein concentration and act as a guide for future dilutions.

For the same reason, ionic materials such as SDS could not be used so 8 M urea and 4% w/v CHAPS (3-[(3-Cholamidopropyl dimethylammonio]-1-propanesulfonate), a zwitterionic detergent, were used to ensure complete dissolution and denaturation of all of the proteins. Also included in the sample solution were 60 mM DTT for disulphide reduction and 2% of the proprietary carrier ampholyte solution Pharmalyte™ of the appropriate pH range which also prevents protein precipitation and ensures a smooth pH gradient. 0.002% w/v bromophenol blue was added as a tracker dye.

All of the sample components were added to a 1.5 mL microcentrifuge tube and then made up to a nominal 1 mL with foam fluid so that there was minimal dilution. The solution was thoroughly vortexed until no particulate matter was visible and was allowed to equilibrate for thirty minutes. The solution was centrifuged briefly and 125μL of the supernatant was evenly loaded across an IEF tray. All of the IEF equipment and associated proprietary solutions were from the Ettan™ IPGphor™ IEF system (GE Healthcare). A 7cm pH 3-10 (non-linear) immobilised IEF strip was placed on top for hydration by the sample solution and then covered with a light silicon oil to prevent dehydration. During rehydration a relatively low voltage (30V) was applied across the strip for thirteen hours before a step gradient increased the voltage to 5000V and the IEF was allowed to continue for a total of 10.5 kVh. The maximum current recorded during the IEF was 35μA which suggests that the foam solution had a low ionic strength.

After the completion of the IEF step the voltage was switched off and at this point the proteins immediately start to diffuse, so the next steps were performed quickly. In a 15mL falcon tube, the strip was equilibrated with SDS-PAGE (50 mM Tris.HCl, 6 M urea, 30% v/v glycerol, 2% w/v SDS, 0.002% w/v bromophenol blue pH 8.8) equilibration buffer and 50 mM DTT for fifteen minutes and then in SDS-PAGE equilibration buffer with 100mM iodoacetamide. The iodoacetamide alkylates any free thiol giving a predictable mass for subsequent mass spectrometry rather than a mixture of thiol oxidation states or adducts. After a further fifteen minutes the strip was transferred to a single well on a freshly cast 10 x 10 cm SDS-PAGE 15% polyacrylamide gel. To maximise the efficiency of protein transport from the first to the second dimension the strip was placed against the top of the gel with any visible air bubbles being eliminated. A 0.5% agarose solution with 0.002% bromophenol blue was melted and applied over the strip to seal the strip against the gel.

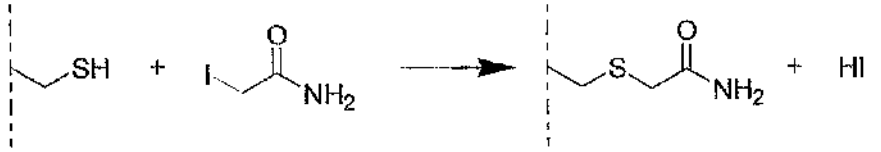


Figure 3.1. Alkylation of a cysteine free thiol by iodoacetamide.

Once transfer was complete and the sealing solution had set (less than five minutes), the tris-glycine SDS-PAGE mini-gel was run under normal conditions i.e. 50 V for 20 minutes then 150V until completion. The gel was immediately removed, fixed with 40% ethanol 10% acetic acid solution for 30 minutes and the stained overnight with colloidal Coomassie stain (1.6% v/v phosphoric acid, 8% w/v ammonium sulphate, 0.08% w/v Coomassie Brilliant Blue G-250) for increased sensitivity relative to standard Coomassie staining[99]. The gel was rinsed with several changes of dH_2O before being visualised.

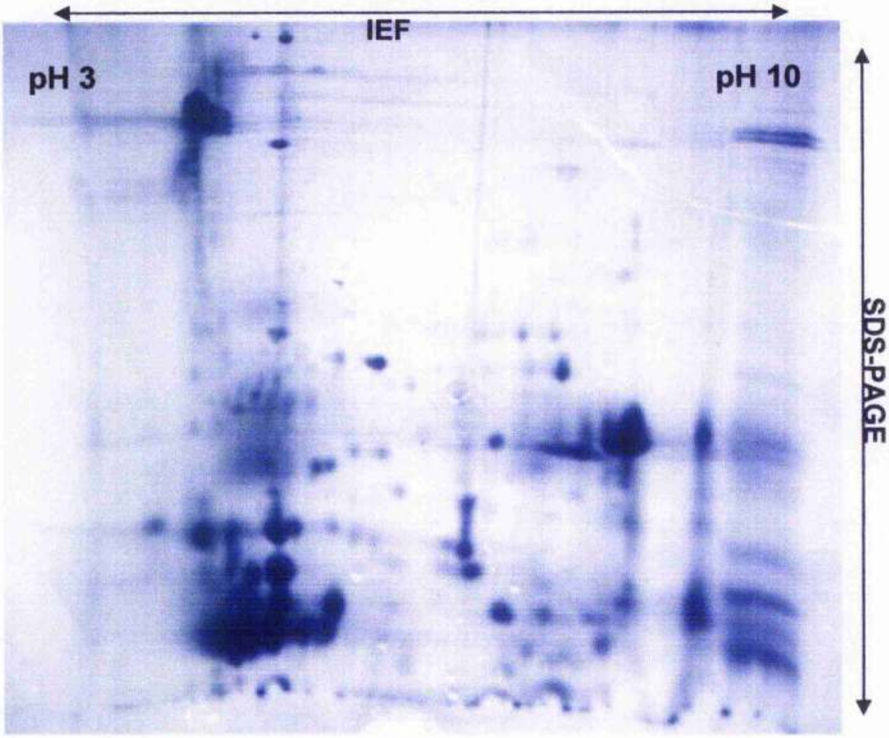


Figure 3.2. Picture of 2-DE of *P. pustulosus* foam nest fluid. The horizontal separation is a pH range from pH 3 to pH 10 (left to right). The vertical separation is based on relative molecular mass with higher mass at the top of the gel. No mass ladder was applied to this gel.

A number of observations were apparent from the gel:

- Many more spots were apparent than had been suggested from the 1D gels.
- Some proteins were observed in a ladder pattern suggesting either localised precipitation during IEF or that some proteins exist in different isoforms
- The majority of the proteins appeared to have an acidic pI (as predicted from the *Ranaspumin* sequences) but there was a banding pattern at the basic end of the strip. This suggests that some of the proteins were precipitated as they had a pI higher than pH 10.

Based on the encouraging results it was decided to progress to a larger gel system to resolve more of the spots observed for mass spectrometric analysis.

3.2.2 Full Size 2D Electrophoresis

Sample preparation was performed in exactly the same way as in the smaller scale experiment. For the IEF dimension, a 24cm IPG strip was used with a pH 3-10 non-linear immobilised pH gradient and the sample solution was scaled up to 450 μL. The increased size meant that a total of 73 kVh was used for separation which is consistent with the manufacturers' instructions. Strip equilibration and transfer to the second dimension was completed in the same way as previously described except that a shallow tray was used to soak the strips that were too long for a falcon tube.

SDS-PAGE was performed on an Ettan™ DALT 12.5% pre-cast polyacrylamide gel (GE Healthcare) with dimensions of 26 x 20 cm. The strip was carefully placed against the gel that is specifically manufactured with a flat contact surface. The gel was then loaded in to the Ettan™ DALT*twelve* system which was programmed to perform electrophoresis with the Ettan™ DALT buffer kit controlled at 25°C and at constantly controlled power. This represented an approximate three-fold increase in scale and took a total of five hours for the single gel– 30 minutes at 2.5 W per gel and then 4 hours 30 minutes at 17 W per gel (other gels were present in the run which were not associated with this work). The gel was fixed and then stained in colloidal Coomassie overnight.

After rinsing the gel it was scanned using the Image Master scanner (GE Healthcare) and the spots were picked by electronically selecting the desired areas on the image and exporting the data to a pick list. Reference markers were also placed on the gel and defined in the pick list. By using the Ettan™ Spot Handler Workstation, the analysis of the spots by mass spectrometry was automated according to the manufacturers' instructions. Briefly, the gel was transferred to the Ettan™ spot picker, which used the pick list as input. The camera in the picker detected the reference markers and the robotics extracted gel plugs from the selected spots and deposited them in a 96 well plate. The Ettan™ Digester workstation was used for *in situ* tryptic digestion of the proteins followed by blending with MALDI matrix (α-Cyano-4-hydroxycinnamic acid) and then spotted on a MALDI plate using the Ettan Spotter.

3.2.3 Protein Identification by Mass Spectrometry

MALDI is a technique of producing ions for mass spectrometry analysis and stands for Matrix Assisted Laser Desorption Ionisation. It is described as a soft ionisation technique because the extent of molecular fragmentation is low relative to that seen in more conventional ionisation techniques such as electron ionisation or chemical ionisation. The core of the technique is the use of a matrix chemical which is added to the sample solution and then small sample spots are added to a polished steel plate. During evaporation the sample and the matrix co-precipitate in solid spots on the plate.

Sample plates are introduced to a high vacuum and the evaporation/ionisation is induced by the irradiation of the spot with a pulsed laser source, normally a nitrogen laser in the near UV. An important feature of the system is that the matrix absorbs at the wavelength of the incident light and, when it does, it produces heat and ionisation. Closely associated analyte molecules are ionised by transfer from the matrix ions and are said to "sputter" in to the gas phase.

After ionisation the ions enter the detector, which is normally a Time of Flight (TOF) instrument. A TOF detector is a relatively simple design with a long path length for the ions to pass through. The ions are all accelerated with the same potential and are then detected at the end of the ion path at a recorded time. Because the potential energy of all ions is approximately equal ($\frac{1}{2}mv^2$), and the velocity can be easily calculated from the defined ion path length and the recorded time for the ion to reach the detector after the laser desorption pulse, the mass of the ion can be determined. Often ion paths contain an ion-reflector, which doubles the available ion path in the instrument and significantly increase resolution. TOF detectors are particularly useful for biomolecules due to the wide linear mass range, so the combination with MALDI ionisation and TOF detection is widely used in proteomics. Each peptide produces a characteristic m/z spectral pattern that can be compared against theoretical predictions for peptide sequence (composition) identification.

Each spot was analysed by MALDI-TOF mass spectrometry (Voyager, Applied Biosystems) which was externally calibrated by applying a small sample of the calibration mix (Sequazyme™, Applied Biosystems) for peptides, and the signals were averaged over 150 laser shots. Spectra were analysed on Data Explore 4.5 with the only manipulation being a noise filter (0.7) and a baseline correction. Not all spots produced peptides that were detected but most of them did. The *Ranaspumin* sequences were subjected to *in silico* tryptic digest[100] and the predicted fragment masses were compared to the masses measured in each of the spots.

Because the digest was performed on fixed protein spots, and due to the probability that not all possible trypsin cleavage (C-terminal of an arginine or lysine residue) sites were successfully cut it was unlikely that all peptide fragments would to be observed. To account for this, protein spots were considered to share identity, or be strongly related to a particular *Ranaspumin* sequence, if three separate predicted tryptic fragments were detected. Due to intrinsic error in the mass spectrometry, a tolerance of ±1 m/z units was allowed between predicted and observed fragments.

All of the six *Ranaspumin* protein sequences were identified except for RSN-2, which was probably too small to be resolved on the gel. Some of the spots were found to contain human keratin, which is a common impurity from contamination in the DTT or other reagent or possibly due to manual handling during collection and separation of the foam. Three spots were identified as being related to RSN-1, one of which was at an unexpectedly high molecular weight and was likely to be an artefact from precipitation in the IEF strip followed by dissolution during SDS-PAGE. This occurs when the IEF step causes localised increases in concentration as each protein molecule migrates to the same strip in the gel. Although the preparation buffer is designed to limit protein precipitation it can still occur where a particular protein is present in high abundance and so this indicated that RSN-1 was present at relatively high levels. The other two spots identified as RSN-1 were at the expected molecular weight and were likely to be a post-translational modification or degradation isoforms of each other.

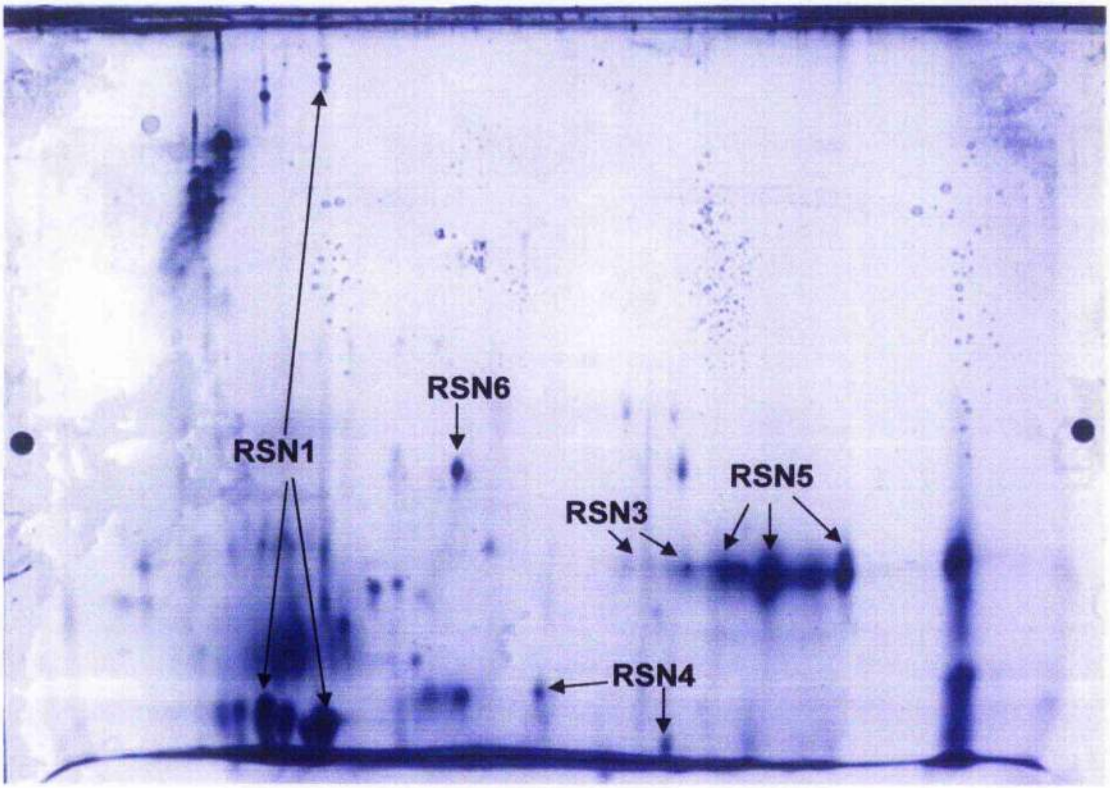


Figure 3.3. 2-DE of *P. pustulosus* foam fluid with pH 3-10 non-linear IEF strip (left to right) and 12.5% polyacrylamide precast SDS-PAGE. The spots identified by MALDI-TOF analysis of the tryptic fragments are indicated.

RSN-3 ran on the 2D gel at a slightly higher molecular weight than expected from the 1D SDS-PAGE and in two pH-dependant isoforms. RSN-4 showed the largest discrepancy relative to the 1D gels. Two spots were identified containing RSN-4 fragments, with differences of both pI and molecular mass observed between them and the spot with the lowest apparent molecular mass appearing to run even lower (less mass) than the spots observed for RSN-1 (11kDa), and the other spot running unexpectedly lower than RSN-3.

Analysis of the peptides showed that the fragments observed in both of the spots contained almost the same tryptic fragments.

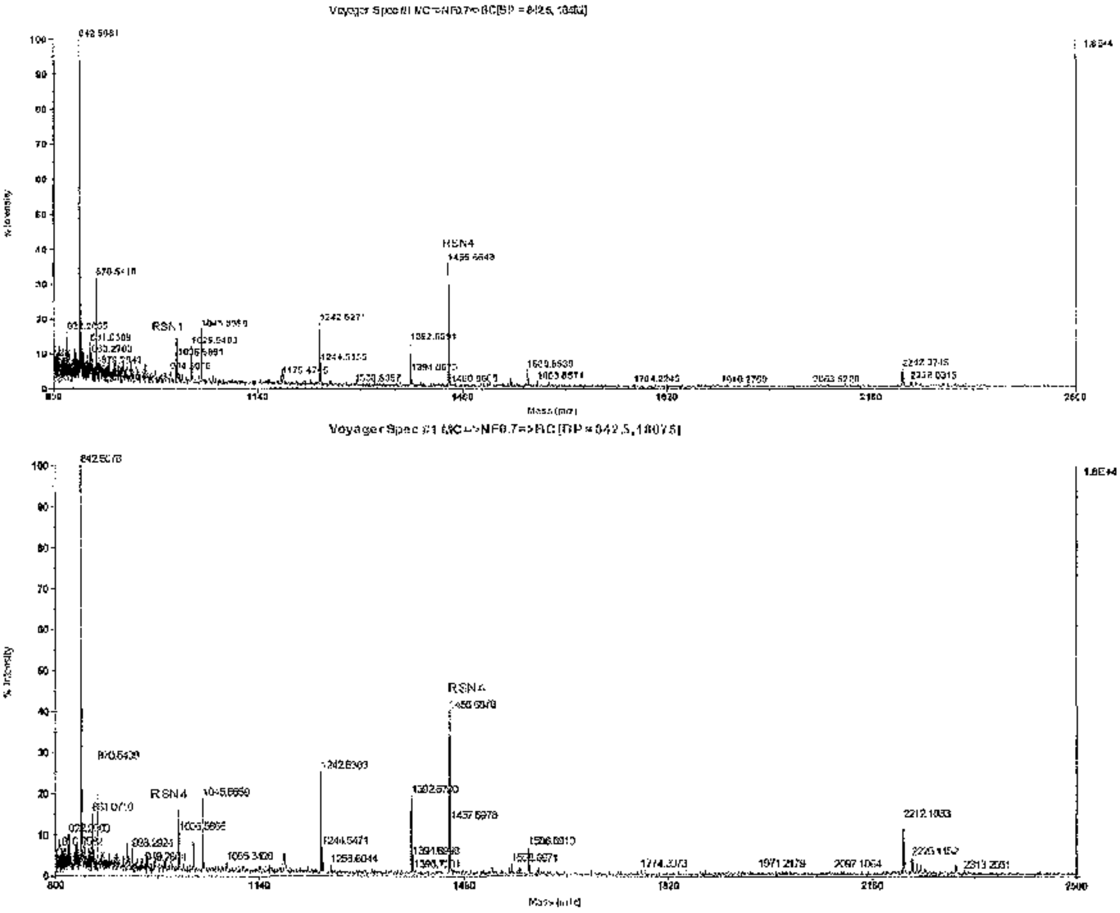


Figure 3.4. Comparison of the mass spectra from the tryptic digest of the two spots identified as RSN-4 related.

All three spots identified as RSN-5 related were of very similar apparent molecular masses but were distinguished by a narrow range of pI points. Despite some of the spots being observed at differing molecular masses from 1D SDS-PAGE, the observed pI points were nominally those expected from those predicted from the amino acid sequences[101], with RSN-1 appearing to be the most acidic (predicted pI of 5.1) and RSN-5 being the most basic (predicted pI of 7.7).

These results confirmed the hypothesis that the all of the major proteins had been sequenced and identified. A number of smaller spots were identified which were probably constituent foam proteins that merit further investigation and a number of others were from known contaminants. The cluster of spots with acidic pI and high molecular mass (top left of gel) that corresponded to none of the known *Ranaspumin* sequences were of particular interest. RSN-2 was not identified from this gel so an alternative strategy was proposed to resolve the RSN-2 spots.

3.2.4 Method Optimisation for RSN-2 Identification

In order to identify RSN-2, it was proposed to use a narrower pH range in the first dimension and to cast a higher percentage acrylamide gel for the second dimension. These changes were expected to give greater resolution in the first dimension where RSN-2 was predicted to migrate to pH 5.2 and to retain RSN-2 on the gel in the second dimension, rather that running in the gel front as previously observed. There was no requirement for the extra resolution afforded by the larger IEF strip and gel so, to save material and time, this experiment was performed on a small scale.

Sample preparation was the same as for the previous small-scale gel. In this case a pH 4-7 non-linear immobilised IPG strip was used for the iso-electric focussing, based on the assumption that RSN-2 would migrate between pH 4 and 7 and the observation on the large gel that the solvent front was particularly intense around pH 5-6. For the second dimension, the polyacrylamide gel was cast with a higher acrylamide content to a final of 15.5% throughout.

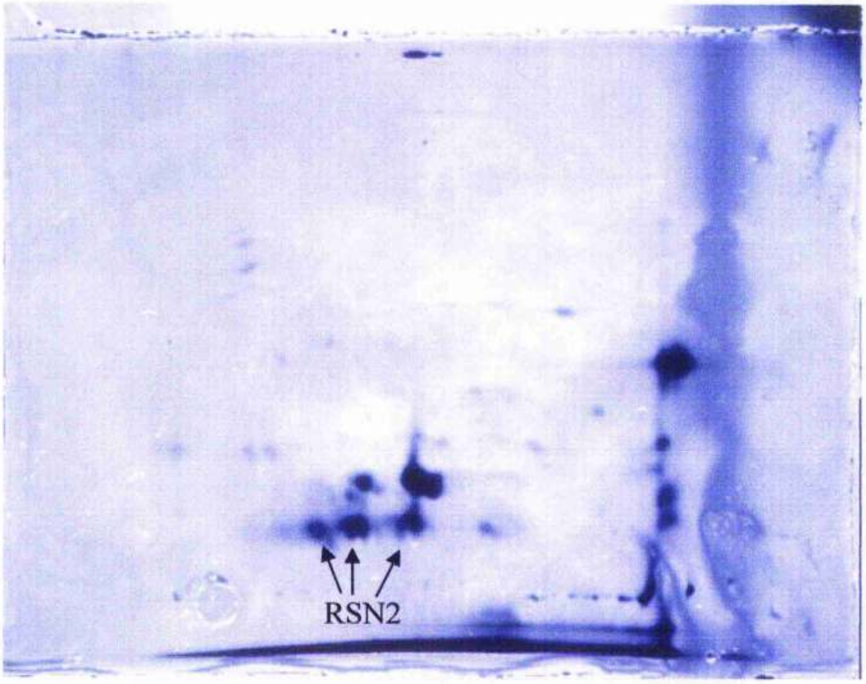


Figure 3.5. **2D Electrophoresis of *P. pustulosus* foam fluid. The pH range was pH 4-7 (left to right) and the acrylamide content was 15.5%.**

RSN-2 was expected to run at a similar pI point to RSN-1, and very slightly lower molecular mass. From the pH 4-7, 15.5% gel three spots were observed in this area. The spots were excised with a scalpel, digested with trypsin and analysed by MALDI-TOF. The results showed good correlation between all of the spots and RSN-2. Keratin was also present in the most acidic of the three spots. These observations confirmed the molecular mass and pI for RSN-2 and also suggested that it exists in at least three different isoforms.

3.2.5 Sample Reproducibility

All of the electrophoresis up to this point had been performed on a single foam nest sample that had been separated from the eggs, aliquoted and stored frozen before use. Two-dimensional electrophoresis was used to examine the possible differences between different samples from the same frozen batch and, more importantly, foam samples from different nests on different days and produced in slightly different environmental conditions.

Using the same conditions and preparations as already described, two different samples were prepared and analysed by 2D electrophoresis and visualised by colloidal Coomassie stain. Both gels were repeats of the proof of concept mini-gel with pH 3-10 IEF strip and 15% acrylamide gels. One sample was from the same foam sample as previously analysed and one from a different foam sample collected from a different location (but from within the Northern Range, Trinidad) on a different evening.

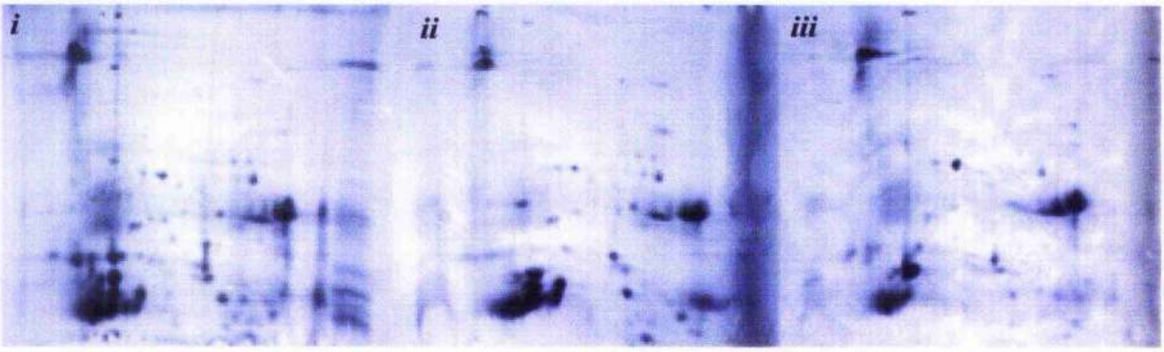


Figure 3.6. Comparison of 2D electrophoresis of *P. pustulosus* foam fluid from the same (*i* and *ii*) and different (*iii*) nest samples. There are obvious similarities between the 3 gels but they are not identical.

A comparison of the two gels, run under identical conditions with the samples from the same foam nest (Figure 3.6) showed that the proteomes were very similar with the predominant features from the original gel being found in the repeat, including the high molecular mass, low pI proteins spots. There were slightly fewer background spots in the repeat, indicating possible less degradation in this run. Comparing these gels to the gel of a different foam sample illustrated that the spots identified as *Ranaspumins* were again the main features, with much fewer other unknown protein spots being visible. This may have been due to relative sample loading (the protein concentrations were unknown and were likely to be variable as different nests retain different volumes of water) or due to a different level of contamination or degradation between different nests.

Two different nest samples were also prepared and separated on pH 4-7 strips and 15% acrlyamide mini-gels. Again, these gels showed (Figure 3.7) some variability in background protein spots and slight differences in the migration of some of the *Ranaspumin* spots, particularly RSN-3 and RSN-5. Curiously, in this comparison the high molecular weight, low pH protein peaks were not at all visible suggesting that they are not an important, functional part of the foam composition. They may be multimers or complexes that are not always present but their identification would be of interest for future work.

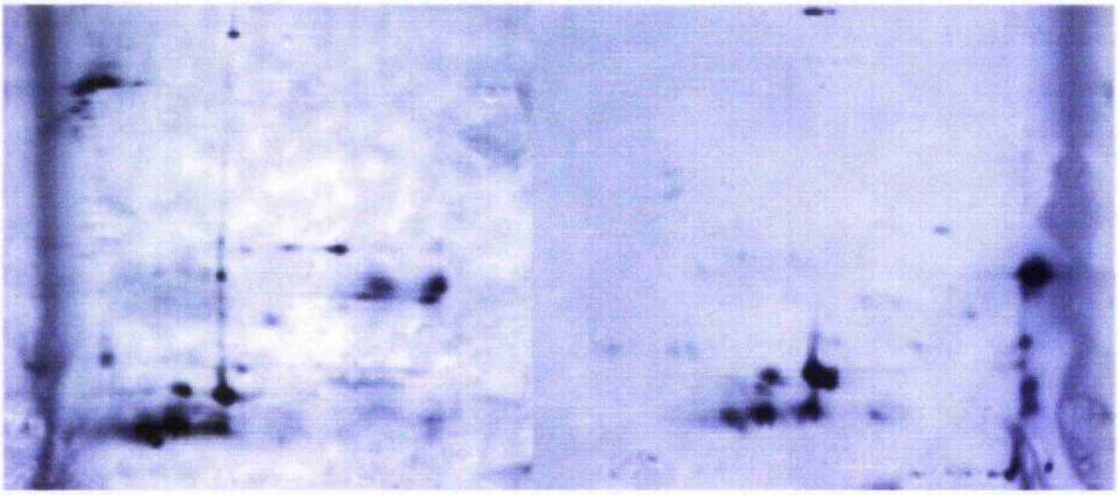


Figure 3.7. Comparison of two different foam nest samples from *P. pustulosus* on pH 4-7 IEF, 15% SDS-PAGE gels. The high molecular weight protein spots were conspicuously absent in the second sample.

3.3 Separation and Analysis of Foam by HPLC

3.3.1 Introduction to HPLC

High Performance Liquid Chromatography (HPLC) is a chromatographic method where a liquid mixture of dissolved components (mobile phase) is separated by an interaction with the solid support of the column (stationary phase). This form of liquid chromatography is described as high performance, because the separation takes place inside a column at high pressures (it is common for columns to be rated for operation up to 30 MPa) and as a result the components are separated with much greater resolution (less diffusion) relative to a low pressure column.

The widespread use of HPLC has led to constantly improving instrumentation which is divided into a number of components. The mobile phase is driven by pumps which are now available in many different configurations. Typically an HPLC instrument will contain four different channels for the introduction of the mobile phase and these can be driven by up to four separate pump head units. Mobile phases are drawn in the correct proportions from each reservoir and then co-mingled in a mixing chamber of various sizes, depending on the instrument. The sample is introduced by injection through a sampling valve at this point.

Before proceeding into the high pressure system the fluid flows through a degasser, which removes dissolved gases usually by vacuum. If this step was not performed, gas bubbles may emerge from solution during high-low pressure phases of the separation, and the bubbles could introduce spurious peaks in the detector system.

After degassing, the mobile phase proceeds through the column where the chromatography takes place. Columns can be of various types and there are a number of important parameters that are often referred to when describing the columns. At the most fundamental level, the column can operate in normal or reverse phase. Normal phase was the first to be developed and is most useful for polar molecules. The stationary phase is composed of silica particles of defined sizes, which are packed in the column to create a polar medium and the mobile phase is non-polar. Whilst providing some efficient separations normal phase HPLC is not widely used due to problems with reproducibility.

The converse separation system is reverse phase HPLC where a non-polar stationary phase is used to separate molecules in a polar mobile phase. The columns are packed with silica particles that are modified with aliphatic sidechains to create a hydrophobic surface area. Common mobile phases are C_4, C_8 and C_{18} chains where the number describes the number of carbon atoms in the aliphatic chains. This is the most common HPLC separation mode and is used in the analysis of small organic molecules, large biomolecules, polymers, lipids and carbohydrates amongst others.

Different modifications on the silica particles can produce stationary phases with different separation modes including size exclusion, ion-exchange and chiral interaction chromatography. Important parameters associated with the columns include particle size, porosity and column dimensions. Particle size describes the size of the individual particles which dictates resolution but also affects back-pressure and therefore separation times. Similarly, the porosity of the column determines pressure and flow kinetics but can also have an effect on resolution. Some modern columns use monolithic silica material where the stationary phase is made up of a polymeric matrix rather than a population of particles and these columns offer much higher porosity and therefore flow rates. The dimensions of the column are important to consider and relate to the results that are required. Narrower columns give more sensitive detection for analytical purposes whilst wider columns are more useful for preparative techniques. Longer columns provide better resolution than their shorter counterparts but increase back pressure and separation times.

After the separation step is complete the flow from the column proceeds to a detector which can take many forms, but the standard device is a fixed UV-visible absorption detector or a diode array detector which is capable of monitoring many different wavelengths at once. Other detectors can include conductivity, pH, fluorescence, refractive index or infrared. Specially designed valves have been developed to couple the eluant from an HPLC with a mass spectrometer. This requires the interface from atmospheric pressure to the vacuum chamber where ionisation takes place and produces an extremely powerful instrument.

3.3.2 Sample Preparation and Chromatographic Conditions

For efficient separation, foam fluid (1 mL) was placed in a 1.5 mL microcentrifuge tube and dried in a vacuum concentrator (Eppendorf 5301) at 30°C until there was no liquid visible. This took approximately four hours. The residual solid was dissolved in 500μL of 0.1% v/v TFA (trifluoroacetic acid) in 6 M guanidine hydrochloride. Guanidine was included as a denaturant as unfolded proteins were more likely to bind more reproducibly and with greater differentiation than when in the native state. The organic solvents used in HPLC denature the proteins anyway so denaturing the sample is of no consequence.

The HPLC system was a DX500 (Dionex) consisting of a chilled temperature controlled autosampler and injection port, a quaternary gradient pump, a thermostatically controlled column oven, diode-array detector and fraction collector. The system did not have an in-line degasser, so all mobile phase solutions were degassed with helium before use and lightly sparged throughout the runs. Mobile phase solutions were also 0.22μm filtered after preparation. The stationary phase was a 4.6 x 250mm 5μm particle size reversed phase column (Grace Vydac 214TP54) using C_4 modified silica and this was coupled to a corresponding guard column (Grace Vydac) which prevented fouling of the column from particulates which may have been present in the foam. All of the system components were controlled by the *Chromeleon* software on an attached PC that was also used to record the data.

3.3.3 Chromatography

As an exploratory experiment, a generic method was applied to the foam whereby the mobile phase was changed from 100% 0.1% v/v TFA in dH_2O to 100% 0.1% v/v in acetonitrile in a linear gradient over 40 minutes. TFA (trifluoroacetic acid) is a common ion-pair additive that controls the pH and interacts with both the sample and the stationary phase to alter (usually improve) separations. The injection volume was 100μl, the elution was monitored at 214 nm (for peptide bond) and 280nm (for protein aromatic sidechains) and the flow rate was 1 ml/min.

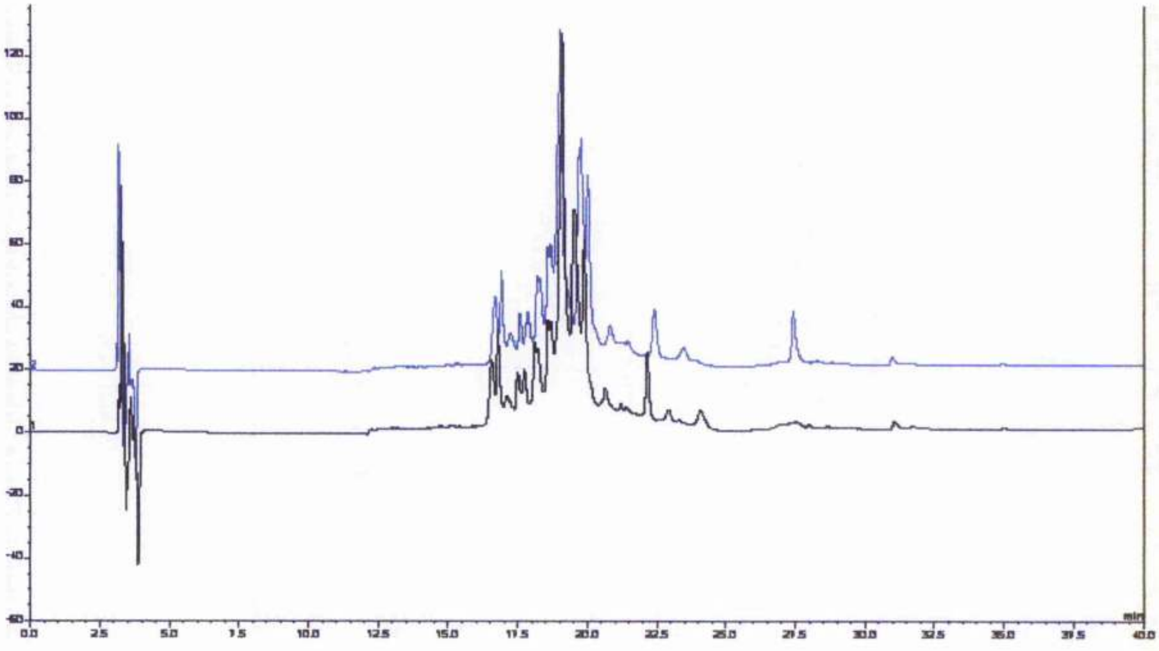


Figure 3.8. Chromatogram of frog foam fluid, separated by reverse phase HPLC showing A_{280} (mAU) versus time (mins). The two traces represent duplicate injections. In this case the water to acetonitrile gradient was 4% min^{-1}.

Applying this method to the foam fluid showed a variety of proteinaceous peaks separated mostly in the 25-50% ACN (acetonitrile) range. Most of the peaks were not well resolved, especially in the area with the most intense signals. Duplicate injections were made, one immediately after the other, from the same sample in the same vial. The vial was maintained at room temperature between injections. Comparison of the chromatograms of the two injections showed that some of the peaks were reliably reproduced but a small number had shifted in retention time or intensity or had disappeared all together. This observation questioned the robustness of the method and the sample but it was decided to be good enough to continue.

For mass spectrometry, it was important to separate the peaks as much as possible so that the analysis of a peak would not be compromised by the interaction between different protein or peptide species (e.g. ion suppression). To improve the separation the method was modified to make the mobile phase gradient more gradual in the region where components were eluted from the column.

Time (mins)	%ACN
0	0
10	0
15	20
70	50
100	100
105	0
115	0

Table 3.1. Conditions applied to RP-HPLC of denatured foam fluid. The table describes the modified mobile phase gradient applied to improve the resolution between protein/peptide peaks.

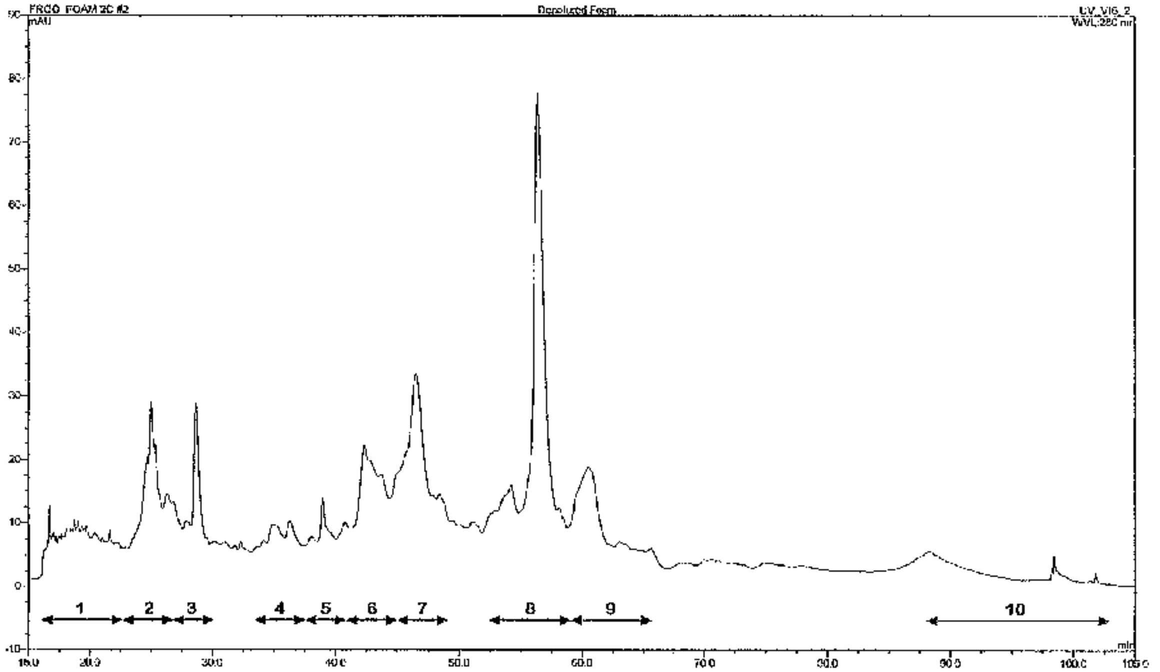
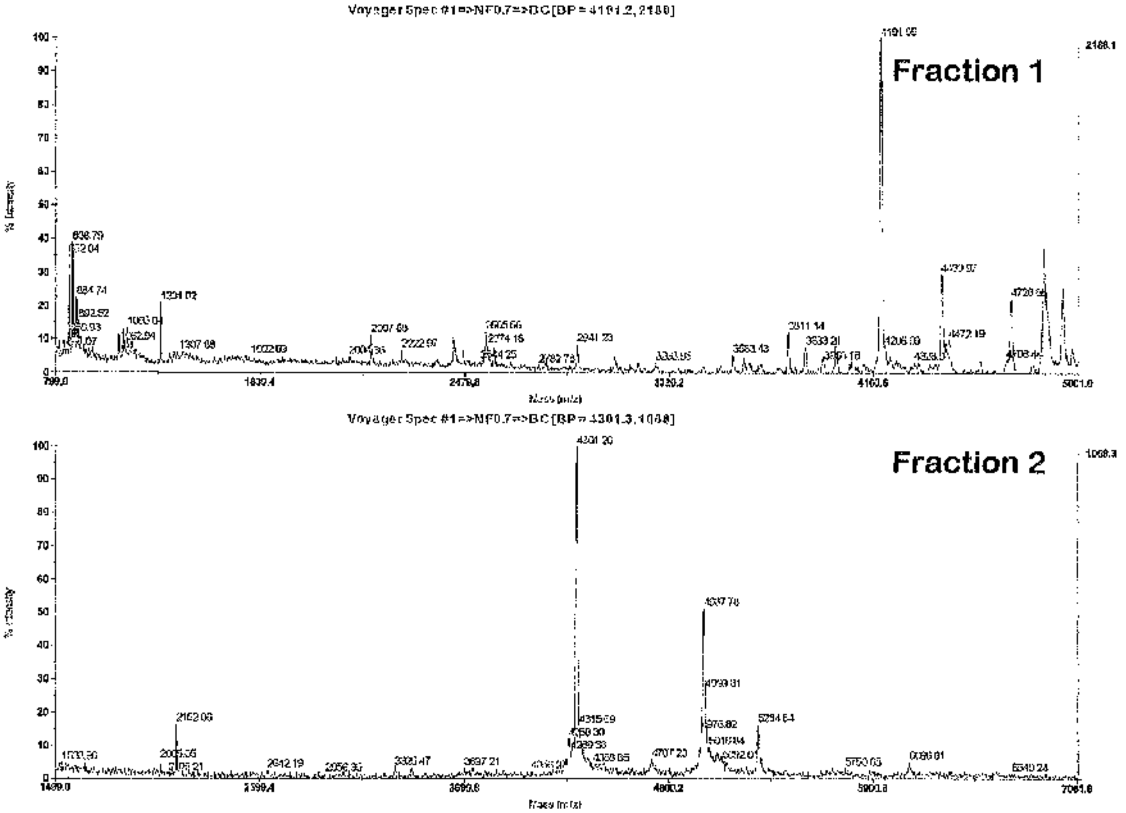


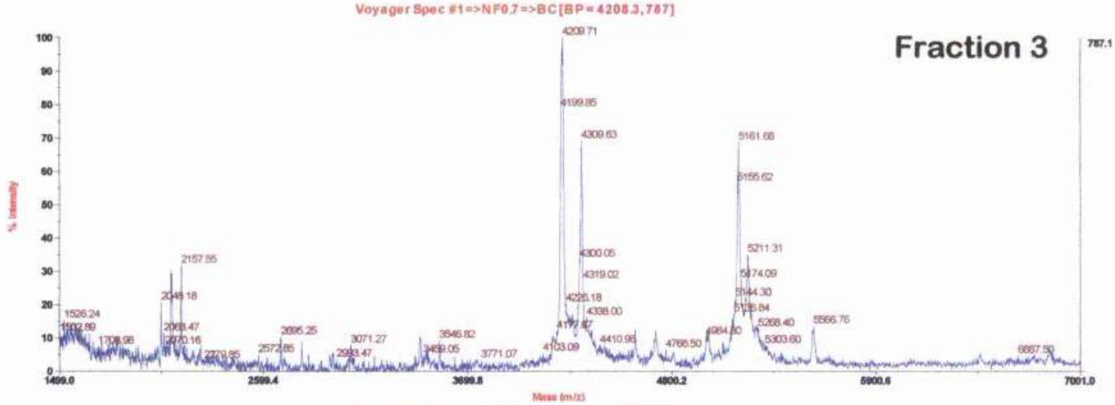
Figure 3.9. Chromatogram at A_{280} of denatured frog foam, separated by RP-HPLC with a relatively shallow water to acetonitrile gradient. The majority of the separations take place at a gradient of 0.5% min^{-1}. Collected fractions are indicated by the numbered arrows.

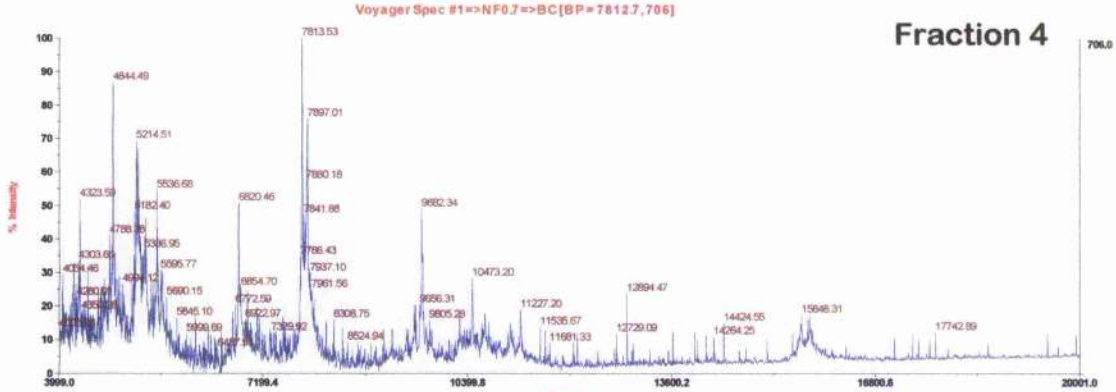
Although the HPLC system incorporated a fraction collector, it was found that more accurate collection could be facilitated by manually switching a short piece of tubing between different 15 mL falcon tubes. The different fractions collected are indicated in Figure 3.9. The modified method improved peak resolution between the foam components although many peaks were probably still mixtures. Many of the peaks were of very low intensity indicating that they were only present in low concentrations.

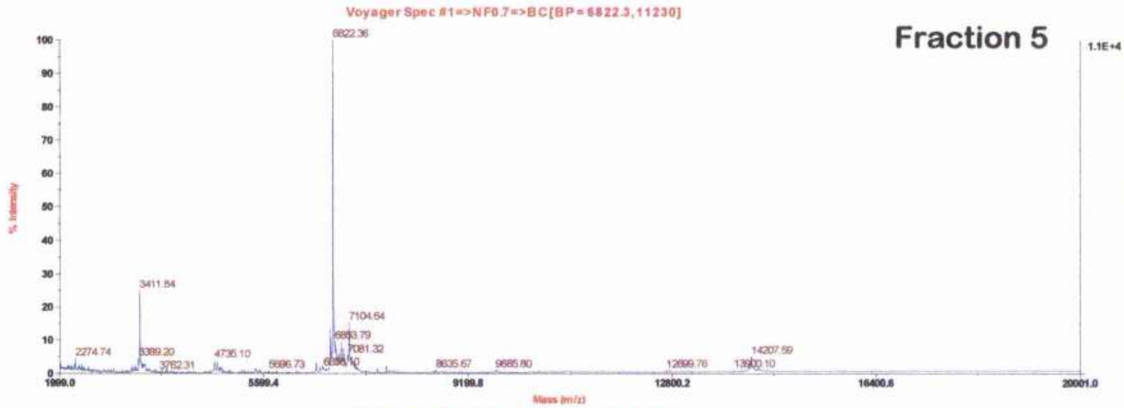
3.3.4 Analysis by Mass Spectrometry

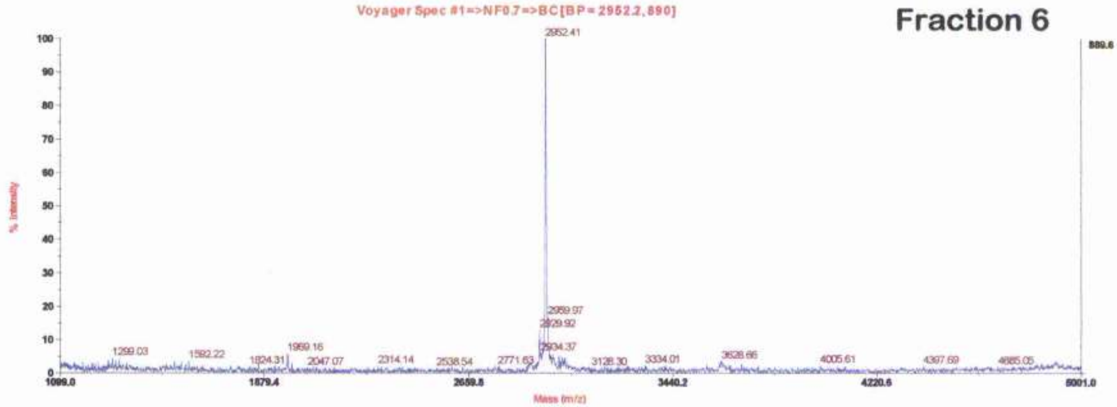
After the run was finished, each sample was transferred into 1 mL fractions in 1.5 mL microcentrifuge tubes and vacuum concentrated as before. Once a solid residue was obtained the samples were re-dissolved in MALDI matrix solution and spotted for analysis by mass spectrometry. Each spot was analysed separately at low (800 – 4000 Da), medium (4000 – 10000 Da) and high (10000 – 50000 Da) ranges and the resulting information was used to analyse each spot at an appropriate mass range. Each spot yielded information on different peptide species except for fraction 10, which gave no significant signal. Different areas of the spot from fraction 7 gave two different spectra, i and ii.




Voyager Spec #1=>NF0.7=>BC[BP=4208.3,787]
Fraction 3
787.1
% Intensity
Mass (m/z)
0
10
20
30
40
50
60
70
80
90
100
1499.0
2599.4
3699.8
4800.2
5900.6
7001.0
4209.71
4199.85
4309.63
5161.68
5155.62
4300.05
4319.02
4226.18
4338.00
4177.87
4103.09
4410.96
5211.31
5174.09
5144.30
5135.84
5268.40
5303.60
5566.76
4984.80
4766.50
2157.55
2048.18
2068.47
2070.16
2279.85
1526.24
1502.89
1708.96
2695.25
2572.85
3071.27
2983.47
3546.82
3469.05
3771.07
6667.50


Voyager Spec #1=>NF0.7=>BC[BP=7812.7,706]
Fraction 4
706.0
% Intensity
Mass (m/z)
0
10
20
30
40
50
60
70
80
90
100
3999.0
7199.4
10399.8
13600.2
16800.6
20001.0
7813.53
7897.01
7880.16
7841.88
7786.43
7937.10
7961.56
4844.49
5214.51
5536.68
5182.40
4323.59
4788.38
5336.95
4303.65
5595.77
4084.46
4994.12
5690.15
4280.61
5845.10
5999.69
6820.46
6854.70
6772.59
6922.97
7329.92
8308.75
8524.94
9682.34
9666.31
9805.28
10473.20
11227.20
11536.67
11681.33
12894.47
12729.09
14424.55
14264.25
15646.31
17742.89


Voyager Spec #1=>NF0.7=>BC[BP=6822.3,11230]
Fraction 5
1.1E+4
% Intensity
Mass (m/z)
0
10
20
30
40
50
60
70
80
90
100
1999.0
5599.4
9199.8
12800.2
16400.6
20001.0
6822.36
3411.84
7104.64
6883.79
7081.32
2274.74
3389.20
3782.31
4736.10
5696.73
8635.67
9685.80
12699.76
13900.10
14207.59


Voyager Spec #1=>NF0.7=>BC[BP=2952.2,890]
Fraction 6
889.6
% Intensity
Mass (m/z)
0
10
20
30
40
50
60
70
80
90
100
1099.0
1879.4
2659.8
3440.2
4220.6
5001.0
2952.41
2959.97
2929.92
2934.37
2771.63
1299.03
1592.22
1824.31
1969.16
2047.07
2314.14
2538.54
3128.30
3334.01
3628.66
4005.61
4397.69
4685.05

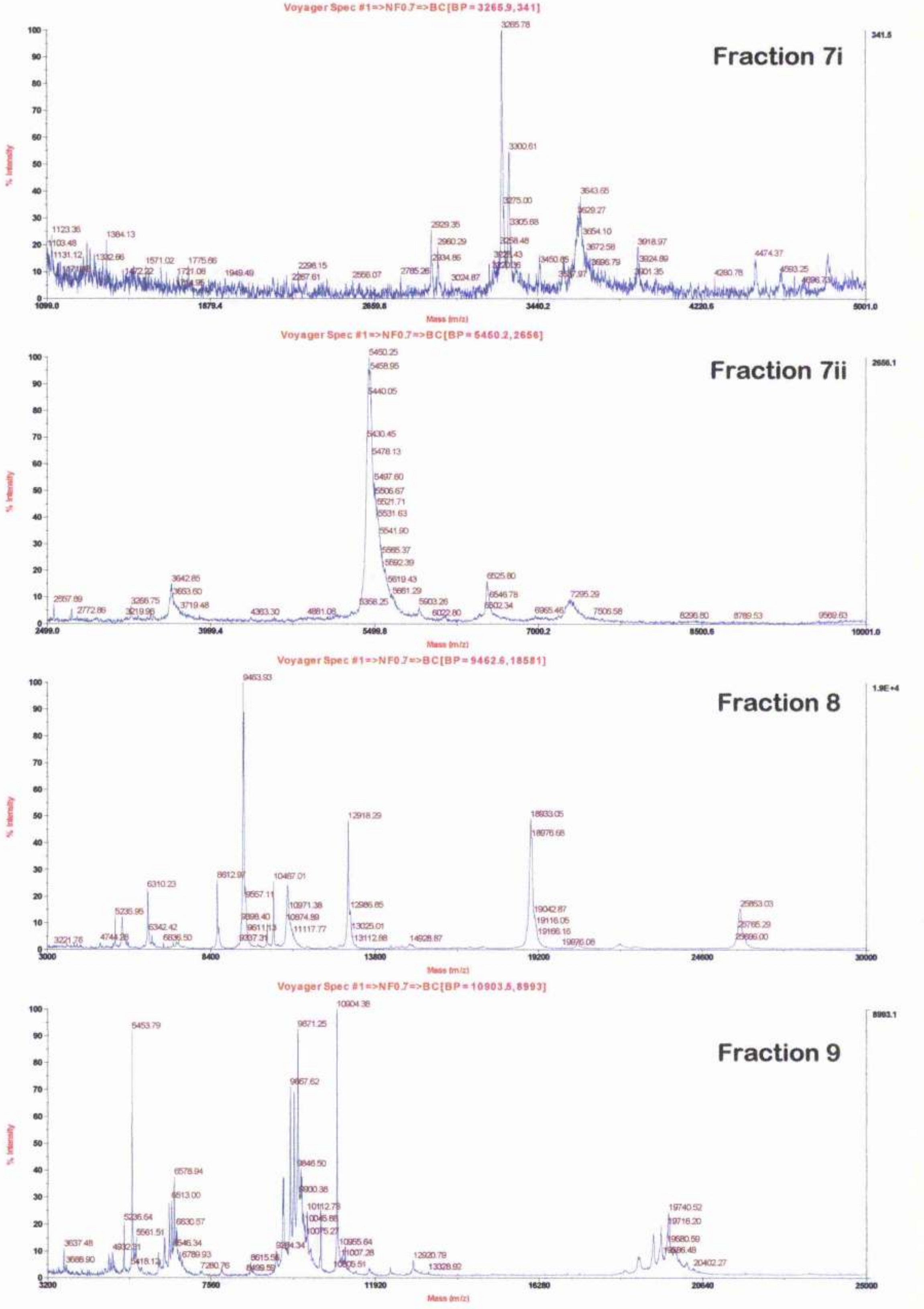


Figure 3.10. Mass spectra from each of the fractions separated by RP-HPLC of denatured frog foam fluid.

Both the HPLC chromatograms and the m/z peaks in the subsequent mass spectra suggest that the foam contains a large number of different peptide components, certainly a great deal more than the six *Ranaspumin* polypeptides previously identified. An important observation from the mass spectrometry results is that there are a large number of peptides present in the foam over a range of masses from small peptides of 10-20 amino acids and a number of large peptides or small proteins, likely to have tertiary structure.

Larger fragments were only found in fractions 8 and 9, and only fraction 8 contained proteins with masses corresponding to known *Ranaspumins*. Peaks for RSN-2, 3, 5 and 6 were observed in this fraction and it is likely that RSN-1 and 4 were also present in this fraction or fraction 9 but were masked in these spectra. Fractions 8 and 9 also gave m/z peaks for a number of small proteins that are, as yet, unidentified and were not observed in SDS-PAGE. These may be fragments from known proteins or multiply charged ions from smaller proteins and are not observed as particularly intense peaks anyway. There may be some merit in MS/MS sequencing of these peaks in the future for elucidation or confirmation of identity.

3.4 Summary

P. pustulosus foam nest fluid was separated by two different methods: two-dimensional electrophoresis and reverse phase HPLC. Subsequent analysis of each component by MALDI-TOF mass spectrometry gave new insights into the composition of the foam. The electrophoresis supported previous work that suggested that there were six major foam proteins (*Ranaspumins*). Further, it was shown that some of these proteins exist in two or more isoforms, fragments or states of degradation. It was also observed that there are subtle differences in the composition and the relative quantities of proteins found in different foam nests. The HPLC analysis indicated that, in addition to the proteins observed, a large fraction of the species in the foam are peptides or small proteins (<10kDa). An abundance of peptides secreted from frogs (usually through the skin) are known so this observation is not unexpected but does add to the battery of information known on the material. It is not yet known whether these additional components are intrinsic to the original foam material, or whether they arise from other sources such as sperm from the male, or frog skin peptides and other components produced inadvertently during the nesting process.

4 Solution NMR Assignment of RSN-2

4.1 Background

Given the unusual sequence and properties of RSN-2, it was decided to investigate its solution structure. Current methods of deriving protein structures are limited to X-ray crystallography or nuclear magnetic resonance (NMR) although other methods can give valuable information on secondary and tertiary structure. In the case of RSN-2, NMR was chosen as the method of analysis because of the extra information that can be supplied in terms of evaluation of dynamic properties in solution. These dynamic processes can be indicative of function and may help to clarify the properties of RSN-2 and function in the foam. Also, at 10.8 kDa RSN-2 is of an appropriate monomer size for structure determination by NMR, assuming that it is not highly aggregated in solution.

4.2 NMR Theory

In NMR experiments the nuclear spins of individual atoms are manipulated by magnetic fields to yield structurally relevant information. Each atomic nucleus has an associated quantum spin number and it is this property that is exploited by NMR. Specifically, the power of NMR is in the observation of the different sizes of quanta of electromagnetic radiation that are absorbed by nuclei in different chemical environments. Since the initial observation of the NMR phenomenon[102,103,104], technological advancements have driven its use in first small molecule structure elucidation and more recently the calculation of solution structures of biomacromolecules. The field is constantly growing and there are continuous advancements in many directions, including instrument hardware and software, novel pulse sequences and new applications. Whilst a certain amount of knowledge is required for the biochemist to understand and profitably use the technique, this text is only a broad overview of the field and the particular applications used. More detailed theoretical and practical overviews are available elsewhere[105,106].

Useful NMR experiments for probing protein structures were only available after the introduction of pulsed Fourier transform NMR and the development of multidimensional pulse sequences for the purpose. From the initial strategic developments more than twenty years ago[107], there are now more than five thousand NMR-derived structures deposited in the PDB. New and constantly evolving methodologies allow the study of larger proteins, instrumental advances mean that less and less sample is required, and progress has been made in the automation of structure elucidation from acquired data. All of this means that NMR is likely to continue and increase in use as a powerful tool in structural biology.

4.2.1 The Vector Model

A fundamental property of an atomic nucleus is the spin quantum number. According to quantum mechanics the spin number *I*, is defined by the sum of the protons and neutrons (fermions) in the nucleus, each of which can have +/- spin ½. Nuclei with an odd number of fermions will have half-integer spin, those with even numbers will have full-integer spin. In protein NMR nuclei which are spin ±½ are almost exclusively used, so only this spin number will be considered in the following discussion, although NMR theory can be extended to take other nuclei into account. Electrons have an associated intrinsic angular momentum. The spin angular momentum results in a magnetic moment (μ) which is given by $\mu = \gamma I$, where γ is the gyromagnetic ratio which is a nucleus-specific constant. From the equation it can be seen that nuclei with zero spin numbers will not have any magnetic moment and so will be NMR inactive.

Under an external magnetic field the magnetic nuclei can adopt one of two quantised energies, one aligned partially with the magnetic field and the other a higher energy level aligned against the field. Because of this small energy difference, there is a slight difference in populations between the two energy levels with slightly more nuclei aligning with the field (lower energy) and it is this energy difference that is recorded in an NMR spectrum.

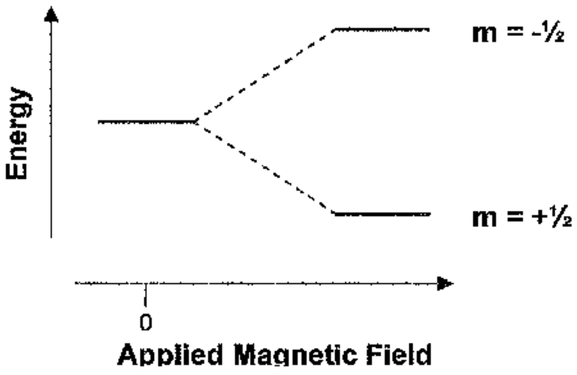


Figure 4.1. Energy splitting of a spin ½ nucleus under an applied magnetic field.

The difference in energies and therefore populations is very small and this is the reason why NMR is historically considered to be an insensitive technique. The energy difference is also related to the strength of the magnetic field, and therefore sensitivity, which is one reason why instruments with higher magnetic fields are being developed.

Nuclei absorb radiation equal to the energy difference between the opposing spins, which is in the RF (radio frequency) range. This is classically viewed in a vector model using Cartesian coordinates to represent the bulk magnetic moment of the collection of nuclei in a sample. Each of the nuclei precesses about the magnetic axis due to the torque created by the angular momentum and in a sample the bulk magnetic moment is created by the sum of each constituent nucleus. Because there is a slight surplus of nuclei aligned with the magnetic field, the bulk magnetic moment is aligned with the field and the sum of the vectors of the individual nuclei precessing about the axis is a vector in line with this axis. To aid this visualisation the coordinates are often transformed from the laboratory frame to the rotating frame, where the axes rotate about the field axis at the same rate as the magnetic moment in question. This rate of precession is called the *Larmor* frequency.

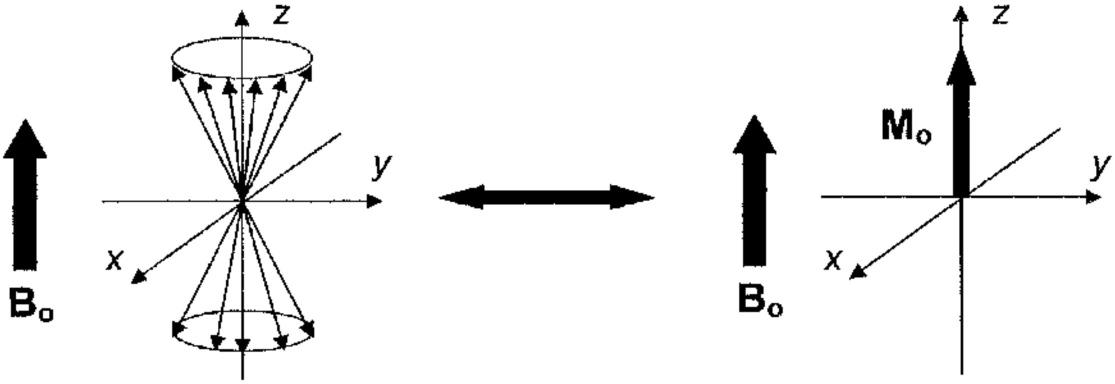


Figure 4.2. The summing of individual magnetic moments to create the bulk magnetic moment.

In an NMR experiment, the sample is placed in a static magnetic field to create the split energy levels and then an RF pulse is applied to the system perpendicular to the magnetic moment (along either the x or y axis). Absorption of the RF energy at a particular frequency (and therefore energy and quanta) inverts one of the aligned nuclear spins so that it is against the external magnetisation. In the rotating frame this is viewed as a rotation of the bulk magnetic moment about the axis of application and the rotation is increased as the pulse length is increased (as more spins are inverted). Pulses are often termed 90° or 180° with reference to this phenomenon.

After the pulse, the magnetisation gradually returns to equilibrium by relaxation mechanisms that are described in more detail in Chapter 6. The relaxation is recorded as a decay in amplitude of the magnetic moment in the x-y plane to give a decaying cosine curve. This time-domain data are transformed to the frequency domain by Fourier transformation so that a group of chemically identical nuclei decaying give a single Lorentzian absorption peak (the dispersion part of the peak is discarded) in the NMR spectrum.

The power of NMR lies in the fact that the static field is not just a result of the applied external field, but it also includes a component due to the local physical and chemical environment. Spectrometers are designed so that any local field inhomogeneities are minimised so that the only differences are due to the nuclei themselves and their environments. These local differences are due to the relative shielding effects of different electron densities around nuclei and so certain chemical structures have predictable resonance frequencies. These frequencies are usually referred to as chemical shifts (δ) which are relative to a reference frequency in parts per million. This system has the effect of normalising frequencies across a range of static fields.

In systems with many different nuclear spins the spectra are complicated by the splitting of the peaks due to interactions between the different nuclei, called J-coupling. These are due to the different combinations of spins that are possible in spin systems. For example in the ^{1}H (proton) spectrum of a CH_2 moiety there are two different proton nuclear spins with a possibility of four different combinations of alignments, each combining to give slightly different energies. So, any other protons that are covalently coupled to this moiety and are close enough to be affected due to relaxation interactions, are exposed to three (two of the environments are of the same energy) different shieldings resulting in a peak that is split into three (a triplet). The magnitude (in Hz) of the splitting is equal for two coupled protons so this can be used as a diagnostic tool for structure elucidation. This through-bond coupling is called scalar coupling.

4.2.2 Product Operators in NMR

The vector model is extremely useful for conceptualising NMR phenomenon but, when more complicated systems and pulse sequences are considered, it is not a powerful enough model to explain the behaviour of the spin magnetisations. Such systems can be fully described by quantum mechanics in the density matrix theory but this is an extremely unwieldy approach and so the product operator formalism is often used instead. Whilst this is based on quantum mechanics, it incorporates geometric parameters that can be applied to physical systems.

The basic concept is based on the treatment of an applied field to a magnetic vector. For example, an RF pulse applied about *x* (β_x) can be described in terms of the applied pulse and the component axial magnetisations: $M_z \xrightarrow{B_x} M_x\cos\beta + M_y\sin\beta$. Operators can be used to describe magnetisation in a coupled system and in a two-spin system there are 16 product operators. The alteration of magnetisation due to chemical shift evolution and scalar coupling can be treated independently and separately for each spin in the system that allows a complete description of the state of the system at any point. The combination of a number of pulses and delays results in extremely long and complicated expressions, but these can be condensed from knowledge of trigonometric identities and the removal of unobservable components.

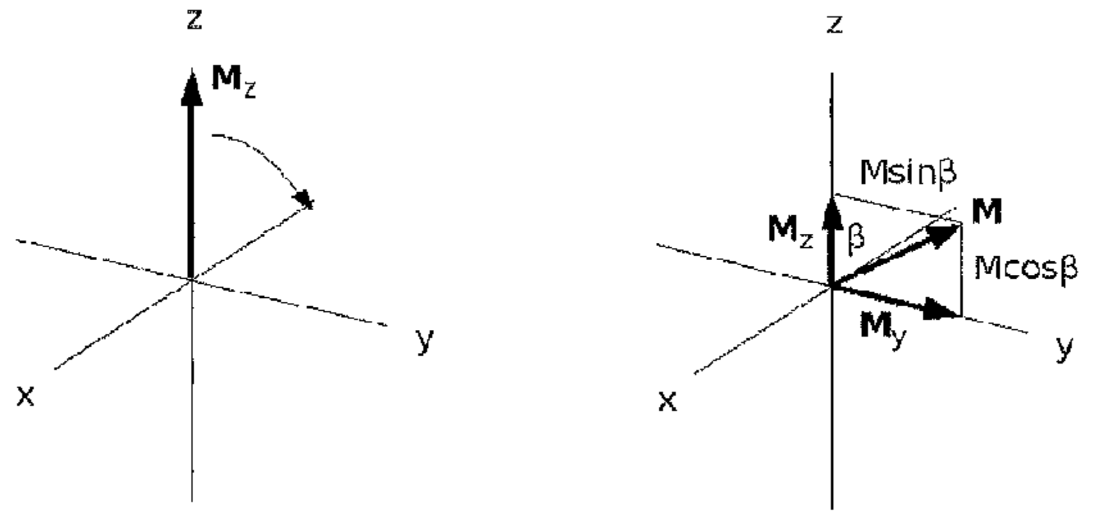


Figure 4.3. Splitting of an off-axis magnetisation vector produced by an applied RF pulse, using trigonometry.

A detailed discussion of the product operator formalism is beyond the scope of this work but one particular outcome of the treatment of spin states by this method should be noted. In particular, the product operator is able to describe quantum coherence and coherence transfer where the vector model cannot, and it is coherence transfer that allows multidimensional experiments. In the most simple form for an H-C coupled system, this is done using an INEPT (Insensitive Nuclear Enhancement by Polarisation Transfer) where the pulse sequence begins with a preparative 90° pulse on the proton channel. The system then evolves under the effect of chemical shift evolution, field inhomogeneities and heteronuclear coupling. A 180° pulse on both proton and carbon and subsequent evolution (same time as the first evolution), refocuses chemical shift and field inhomogeneities but the vector diffusion from scalar coupling remains. If the evolution delay is a quarter of the coupling constant then the two coupled vectors are antiphase to each other and a simultaneous 90° pulse on proton and carbon completes the transfer of the relatively large proton polarisation (spin population difference) from proton to carbon so that detection sensitivity is drastically improved. Variations on this concept and similar sequences for manipulating the spin populations and the use of the distinct and relatively invariable coupling constants found in proteins, complex multidimensional spectra can be recorded for atomic assignment.

4.2.3 Instrumentation

Modern NMR spectrometers are extremely sophisticated and expensive instruments. They are composed of several different groups of components and are almost all controlled by a computer workstation. At the core is a magnet which provides the static field. The strength of the magnet is measured in Tesla but is typically described by the *Larmor* frequency of a proton when subjected to the field. Conventional electro-magnets have been superseded by superconducting magnets composed of coiled metal alloys which are immersed in liquid helium to lower the temperature for zero resistance and these magnets are currently commercialised at strengths of up to 950 MHz (22.31 Tesla) and the first 1 GHz magnet is likely to be available in the near future. An outer chamber of liquid nitrogen and a series of vacuum chambers prevent excess helium evaporation but the liquid nitrogen requires regular refilling and the helium does need to be replenished periodically. Some modern magnets contain shielding to prevent excess stray magnetism from being present in the laboratory.

The liquid sample is placed in a glass tube and positioned at the centre of the magnetic field. In some cases the sample is spun at several hertz to normalise the irregularities in the sample tube but in protein NMR, high quality tubes are used and the sample remains stationary. Also, peaks from proteins are broader and so there is no advantage to "spinning out" minor inhomogeneities Compressed air is used to control sample height and also to regulate temperature in the sample chamber. The sample is flanked by a series of shim coils, which are electromagnets and can be varied by the user to ensure a homogeneous magnetic field across the sample.

When in place, the sample is seated in the coils of a probe, which is used to apply the radio frequency pulses that manipulate the atomic magnetisation. The same coils also detect the changing magnetic moments that provide an NMR signal. Probes are manufactured in a variety of different geometries and electronic configurations. Usually, a number of channels are included that can be tuned to the appropriate frequencies for different atomic nuclei: hydrogen, nitrogen, carbon and sometimes phosphorous, are included in probes for biomolecular analysis. Multi-channel probes were initially manufactured with the carbon channel coils being inside the proton channel so that the proximity to the sample could help compensate for the relatively low natural abundance of NMR active carbon (1.1% ^{13}C), but modern instruments are usually of inverted geometry with the proton channel being closest to the sample. This is the most efficient use of the system when using modern multidimensional pulse sequences and isotopic enrichment techniques.

Spectrometers use quadrature detection, which is the splitting of the signal in the receiver coils into two separate signals, one with an imposed 90° phase shift to the other. This allows for discrimination between positive and negative signals. Some coherent noise and other artefacts are removed by phase cycling. Phase cycling is the process by which the phases of the pulses in a sequence are systematically varied. The addition of the acquired spectra results in the cancellation of some of the undesired peaks. Phase cycling can also be used to select the magnetisation of interest and discard the unwanted signal (otherwise spectra would be extremely complicated). Some probes also contain gradient channels. Gradients are applied in NMR pulse sequences as an alternative or an addition to phase cycling as they can select specific coherences. High specification instruments now utilise special low temperature cryo-probes which operate at close to absolute zero. At these temperatures, electrical noise is drastically reduced relative to room temperature probes and so the sensitivity of the technique is significantly enhanced.

Outside of the main magnet a console contains all of the electronics required for application of the applied RF radiation, recording of the spectra, control of the experiments and all other required signals. The pulse program is interpreted to produce a series of pulses via signal generators and then amplifiers. Analogue to digital converters and digitisers are used to transfer the exciting frequencies or the measured signal, to the probe or the data storage device. To compensate for temporal inhomogeneity in the static field a lock signal is recorded and used to normalise the effective field. Practically this involves an internal system which has separate components of another spectrometer and is used to record the deuterium signal of a spiked deuterated solvent, which is usually 5-10% D_2O in protein samples.

4.2.4 Practical Considerations

In protein NMR spectroscopy, the first hurdle that needs to be overcome is the insensitivity of the technique, for naturally occurring nuclei. The most abundant hydrogen isotope (^1H)

is NMR active but, for multidimensional experiments, NMR-active nitrogen and carbon isotopes are required, and these only occur naturally at very low levels (1.1% for ^{13}C and 0.37% for ^{15}N). This problem can be circumvented by the incorporation of an enriched quantity of the active isotopes into proteins produced biosynthetically. This is achieved by growing the host cells in media where the sole sources of carbon and nitrogen nutrients are in the isotopically enriched form. In some cases cells are reluctant to grow in this medium and further development is required to develop an appropriate expression system. This pre-requisite, coupled with the insensitivity of NMR (even for enriched samples) means that protein samples are relatively expensive to produce for NMR structure determination.

Also due to the low sensitivity, NMR experiments can take anywhere from seconds to months. A simple pulse-record experiment takes less than a second but multiple scans are usually employed to increase signal strength although this must be traded against the increase in time required for multiple scan experiments with the knowledge that signal to (random) noise is proportional to the square root of the number of scans. Typically, isotopically enriched protein samples for NMR are prepared at approximately 1 mM concentration and experiments are performed with approximately four to sixteen scans in the direct dimension, depending on the particular experiment.

Using these techniques, NMR has emerged as a standard technique for generating protein structures but the remaining drawback is the size limitation. The size of protein that can be studied is limiting for two reasons: firstly because the increase in nuclei in the same chemical shift range produce extremely complicated and overlapping spectra and secondly, because larger molecules give broad NMR signals because as the tumbling time becomes slower and transverse relaxation is faster (the short FID makes frequency determination less precise). The issues relating to complicated spectra can be largely overcome by multi-dimensional (3D, 4D, 5D) experiments and increasing bioinformatic support towards automation although the higher order experiments do require significantly longer acquisition times, and are intrinsically less sensitive. The size limit is also being challenged by modern labelling and experimental techniques, so that large proteins or domains can be studied. For proteins above ~40 kDa, TROSY (Transverse Relaxation Optimised Spectroscopy), experiments are used which select single coherences which are less sensitive to transverse relaxation and this results in much higher resolution[108] and complex systems up to ~900 kDa have been successfully studied[109].

Chemical labelling with selected isotopes also helps to sharpen and simplify complex spectra. Deuteration or partial deuteration reduces the effect of ^{1}H-^{1}H dipolar coupling on the transverse relaxation and allows higher resolution. Other labelling techniques using cell free expression such as SAIL (Stereo Array Isotope Labeling)[110] can reduce the complicated spectra so that they contain just the information required for probing structures. With more advances being made and high throughput systems being brought closer to fruition it is likely that NMR will continue to provide important information about complex biological systems.

4.3 Sample Preparation

Recombinant RSN-2 was produced from plasmid stocks of the previously prepared construct. For preparation of isotopically enriched proteins, the growth media was M9 minimal media based on the recipe of Maniatis[111], with the sole source of nitrogen being ammonium chloride ($^{15}NH_4Cl$ for enriched) and the carbon source was glucose (^{13}C-D-glucose for enriched). Enriched ammonium chloride and glucose were purchased from Spectra Stable Isotopes. The M9 salts solutions were prepared as 500 mL aliquots in 2L culture flasks and were autoclaved prior to use. Glucose and ampicillin solutions were sterile filtered into the media immediately prior to use. Due to the expense of isotopically enriched materials, small scale culture trials were made to confirm that the *E. coli* would grow in the relatively nutrient-poor media, and to find what the optimum glucose concentrations were. The cells were found to grow more slowly in the minimal media compared to the LB media, with lower glucose concentrations resulting in longer lag times at the start of the growth. However, a lower glucose concentration (2 g/L) was found to give the same RSN-2 levels as the standard concentration (3 g/L) with a lower background protein expression so the lower concentration was used for larger scale production.

20 mL LB cultures of RSN-2 were grown overnight and used to inoculate each 500mL flask of minimal media. Growth typically took five hours until IPTG induction, after which the cells were allowed to grow for a further four hours before harvesting. RSN-2 was purified from the cells as described previously.

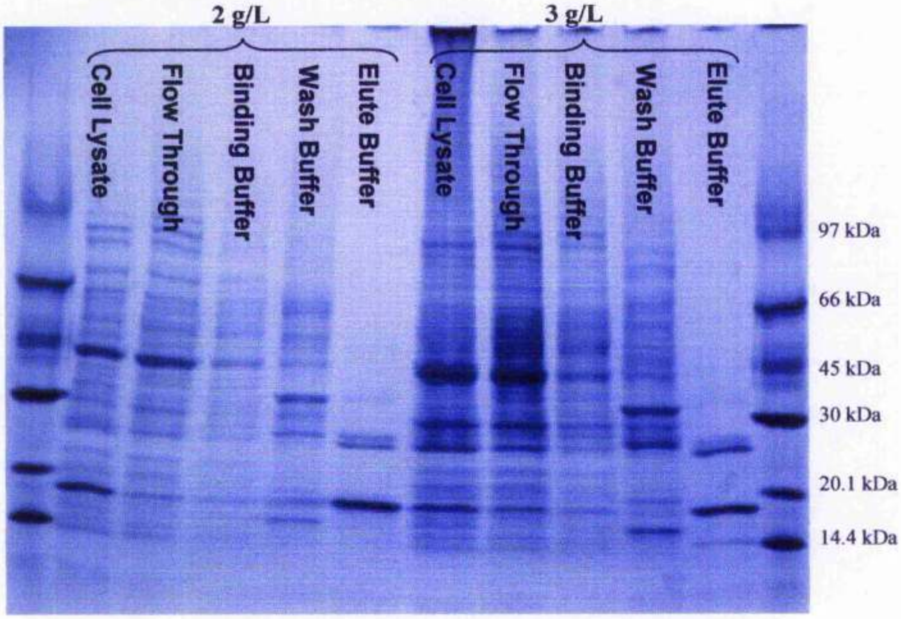


Figure 4.4. Comparison of 200 mL expressions with 2 g/L glucose (lanes 2-6) and 3 g/L glucose (lanes 7-11) minimal media. Lanes 6 and 11 are the affinity purified RSN-2 samples at the same loading. Densitometry analysis indicated virtually identical protein expression. Lanes 1 and 12 are molecular weight standards.

After gel filtration chromatography, RSN-2 fractions were pooled and centrifugal ultrafiltration was used to concentrate the protein solution to ~600μL. Prior to NMR experiments, 540μL of RSN-2 solution was mixed with 60μL of D_2O (Sigma) and placed in an NMR tube (Wilmad 535-PP). For all NMR experiments the buffer was 10% D_2O, 90% 20 mM sodium phosphate 50 mM sodium chloride 0.02% sodium azide pH 7.0. In all, two ^1H, ^{14}N, ^{12}C, samples were prepared (designated H1 and H2), three ^1H, ^{15}N, ^{12}C, samples (N1 – N3) and two ^1H, ^{15}N, ^{13}C, samples (CN1 and CN2). It was necessary to prepare multiple samples because protein degradation was observed after prolonged exposure to elevated experimental temperatures; this was not observed when samples were stored refrigerated.

4.4 NMR Experimentation and Data Collection

All spectra were processed using the *Azara* suite of programs written by Dr. Wayne Boucher, University of Cambridge. Initial processing scripts were created from the Bruker data files via the program *ReferenceB* (written by Dr. Krystyna Bromek, University of Glasgow) and were adapted for specific experiments. In summary:

- the data were zero-filled to improve digital resolution
- a weighting function was applied to alter resolution/sensitivity

- Fourier Transform to the frequency domain (Maximum Entropy algorithms for some indirect dimensions)
- the resulting spectrum was phased correctly (absorption peak)
- repeated for all dimensions

Scripts were executed by the *process* program, the output of which was a *plot2* input file. Typically, processed data were viewed in *plot2* and then the scripts were altered, reprocessed and viewed again until the spectra were optimised.

Once the experiments were processed satisfactorily, they were input into the CCPN (Collaborative Computing Project for NMR) *Analysis*[112] software for spectrum display and assignment. During the use of this software a variety of versions were used, ranging from 1.0.3 to 1.0.13.

4.4.1 Initial Optimisation of Conditions

All NMR experiments were conducted on a Bruker Avance II 600 MHz (actually 600.13 MHz) spectrometer equipped with triple resonance cryoprobe and gradient channel, courtesy of Dr. Brian Smith, University of Glasgow biomolecular NMR facility. Initially, the H1 sample was used to design the experimental strategy. A 1D experiment showed reasonably dispersed signals indicative of a folded protein of this size. Despite this, it was observed that several signals were rather broad and it was decided that an increased temperature (decrease viscosity and T2) would be required in order to improve the resolution.

In these experiments, and all others described here, specific pulse sequences were used to reduce the contribution of the ^{1}H signal from the water solvent. In the simplest form this was achieved by performing a water presaturation experiment where the water signal is irradiated with a low power RF pulse - the low power pulse is selective for the particular frequency whereas the high power pulse normally used for excitation irradiates a wide frequency range encompassing all proton frequencies. Irradiation of the water signal eventually produces equalised populations of spin states and at this point the typical 1D experiment is performed and because the water signal is saturated, the majority of the peak is removed from the spectrum.

The use of pulsed field gradients has allowed for even greater levels of solvent suppression with fewer artefacts. Throughout this work, most experiments were conducted with solvent suppression using the WATERGATE pulse sequence[113]. In this sequence, gradient pulses are applied symmetrically about a selective 180° pulse – the transverse magnetisation is all perturbed by the initial gradient and then the selective pulse inverts all proton spins other than water. The second gradient pulse refocuses all of the proton magnetisation except those in water, which are destroyed. This is a versatile sequence which can be applied to the transverse magnetisation at the end of multidimensional pulse programs to remove water signals.

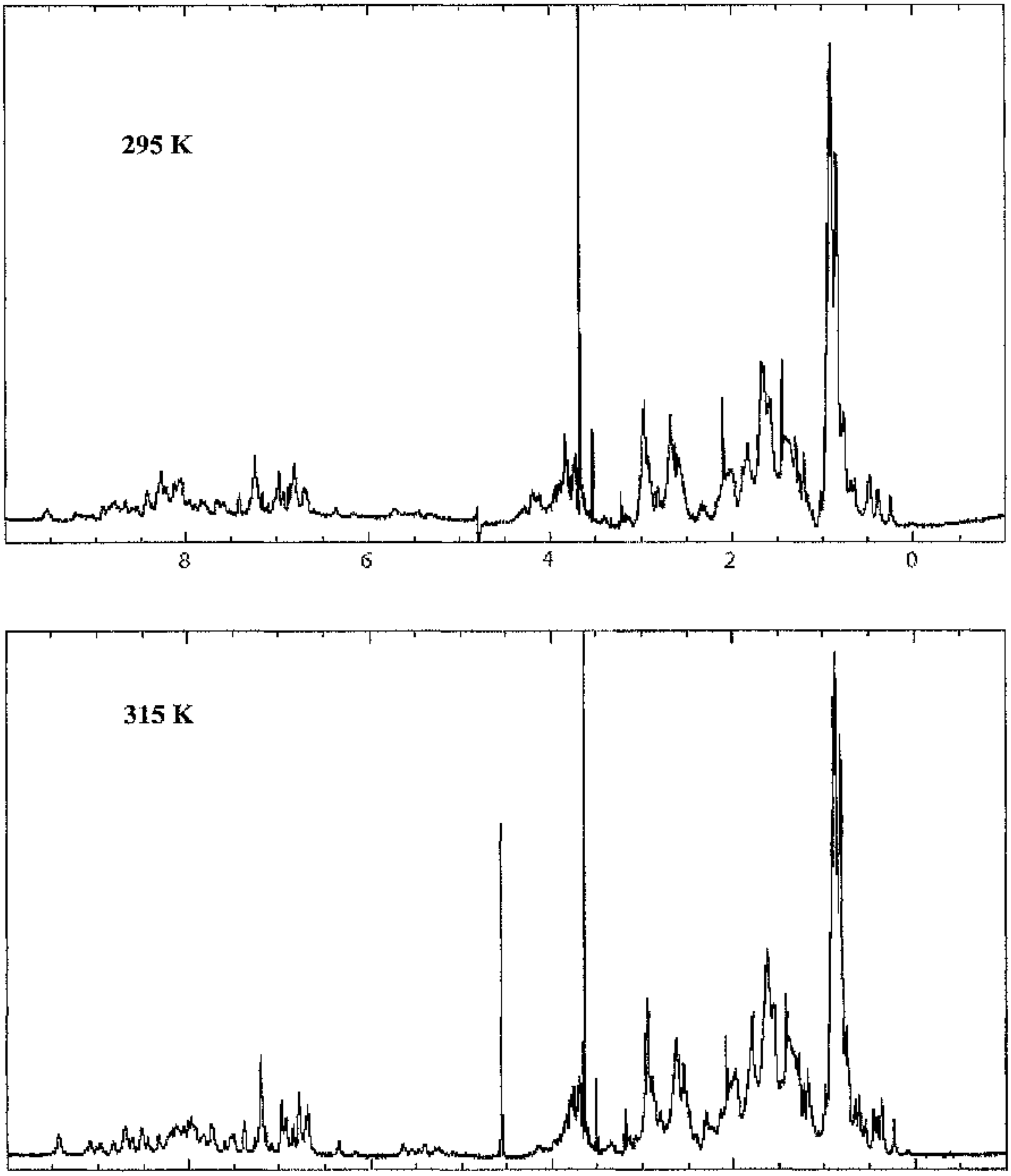


Figure 4.5. **^{1}H NMR spectra of RSN-2 at 295 and 315 K. The increase in resolution with temperature was especially evident amongst the amide protons. Spectra displayed using *plot1*. The x-axis is ^{1}H chemical shift.**

A ^{15}N enriched sample (N1) was prepared and used to optimize the HSQC (Heteronuclear Single Quantum Correlation) experiment, upon which most of the other experiments would be based. An HSQC is a 2D experiment where the first dimension is the proton nuclei and the second dimension is the one-bond scalar coupled heteroatom (nitrogen or carbon). Magnetisation is transferred from the proton to the heteronuclei using an INEPT-type sequence and the coherence is then heteroatom-frequency labelled before being transferred back to the proton for detection of quantum filtered coherance.

In a ^{1}H^{15}N-HSQC experiment, all peaks were observed to fall within the ^{1}H range of 6-10 ppm and a ^{15}N range of 104-132 ppm. Prior to embarking on 3 dimensional experiments, the optimised ^{1}H^{15}N-HSQC experiment was performed at a variety of temperatures from 308 K to 278 K to find the ideal temperature for dispersion and resolution of amide signals (Figure 4.6). Because of the temperature changes, some of the chemical shifts changed slightly due to changes in exchange rates, hydrogen-bond lengths etc. It was clear that the increased temperature caused a decrease in peak width so a temperature of 308K was chosen for future experiments (all temperatures were calibrated for the spectrometer via glycerol peak reference calculations).

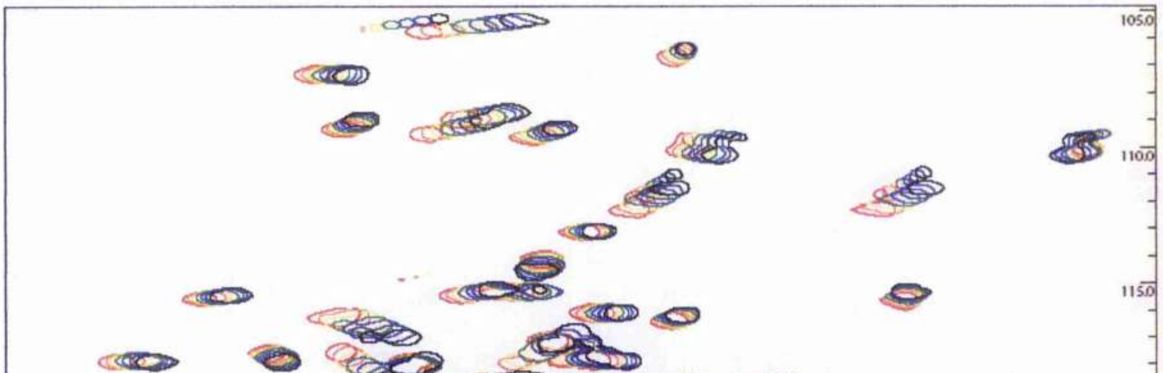


Figure 4.6. **^{1}H-^{15}N HSQC of RSN-2 at temperatures from 308K (black) to 278K (red). The x and y-axes are ^{1}H and ^{15}N chemical shifts respectively.**

4.4.2 Assignment Strategy

To adequately describe the protein structure, each of the amide resonances must be mapped to specific residues in the protein sequence. This was achieved by using triple-resonance (^{1}H, ^{15}N and ^{13}C) experiments to measure the chemical shifts of the associated alpha, beta and carbonyl carbons and protons. The chemical shifts fall within a range and are indicative of particular residues given known chemical shifts from databases.

The important feature of the experimental strategy is the acquisition of complementary experiments that describe the chemical shifts of parent residues and preceeding residues. For example, the 3D HNCACB experiment correlates the amide chemical shifts with those from the alpha and beta carbons for both the attached and the preceding residue, whereas the CBCA(CO)NH experiment correlates the same chemical shifts for just the preceding residue. By comparison of the two experiments, amide links can be inferred and a complete chemical shift sequence can be generated without any link to the actual chemical structure. In practice, the known amino acid was used to help build up the assignments, to help distinguish between overlapping residues and to join fragmented sequences where resonances were not observed (proline residues, for example).

With the complete assignment of the amide resonances in the ^{1}H^{15}N-HSQC, the knowledge of the backbone carbon and proton resonances could be used to assign carbon and proton resonances in the residue sidechains to give a fully assigned ^{1}H^{13}C-HSQC which essentially completes the atomic assignments.

4.4.3 Backbone Assignment

Analysis of the ^{15}N HSQC spectrum indicated that most or all of the amide nitrogen and proton resonances were represented. Peaks were picked, assigned to resonance objects, and sequential links made by analysis of the carbon chemical shifts from the CBCA(CO)NH, HNCACB, HNCO and HNCACO[114]. Using the chemical shifts of the C_α and C_β and the carbonyl carbons some provisional assignments were made of short stretches of sequence. Subsequently, use of the CC(CO)NH experiment allowed assignments to be made with greater certainty and lengths of sequence of up to approximately 20 residues were successfully assigned. Other resonances were linked and treated as unknown residues until further information was gained.

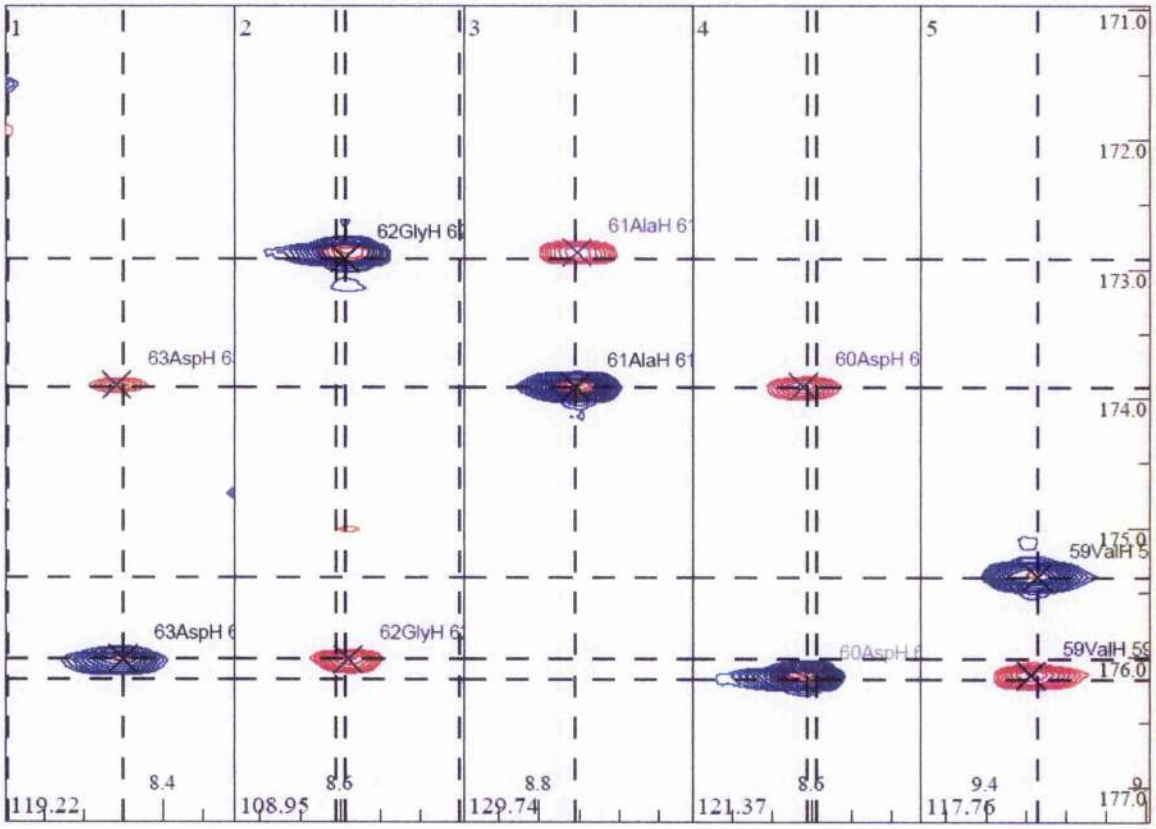


Figure 4.7. **Example of "walk" along RSN-2 backbone. Complementary experiments were used to identify the links between carbonyl resonances of residues 59-63. The x and y axes are the ^{1}H and ^{13}C chemical shifts respectively and the ^{15}N chemical shift is indicated in the bottom left-hand corner of each plane.**

By performing ^{1}H^{15}N-HSQC experiments before and after each experiment, protein degradation was quickly observed by the appearance of sharp peaks where previously there were none. To retard this degradation all future experiments were performed at 303K; a comparison of peaks at 303K and 308K showed that chemical shifts were similar enough to directly transfer assignments. The temperature decrease was found to slow down protein degradation as anticipated. It is postulated that the degradation was hydrolysis at the amino-terminus, due to the previous observations made in mass spectrometry analysis.

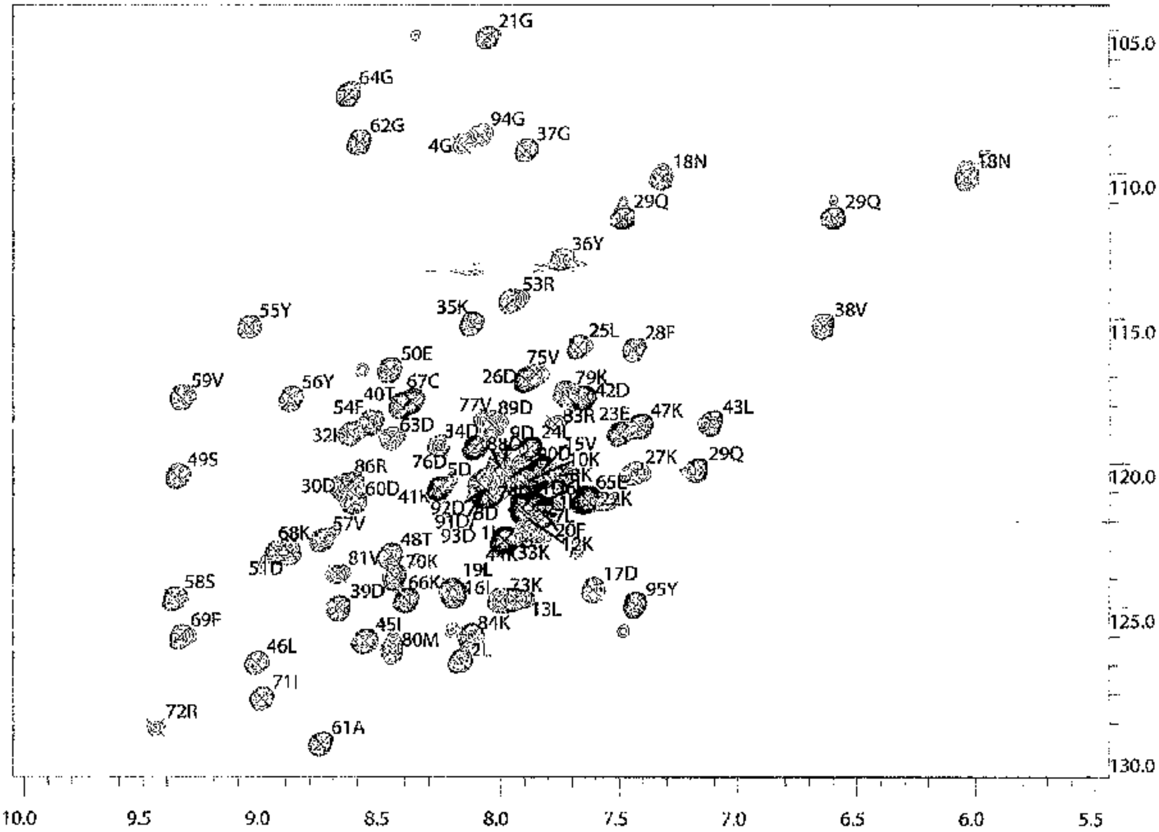


Figure 4.8. **^1H-^{15}N HSQC of RSN-2 annotated with residue identifier. The x and y-axes represent ^1H and ^{15}N chemical shifts respectively.**

After the assignment of the amide proton and nitrogen, the carbonyl carbon and the sidechain carbons, the proton analogues of the heteronuclear backbone experiments were performed (HBHACONH and HNHAHB) to assign proton resonances. Approximately 70% of the amino acid residues were assigned in this way.

4.4.4 Sidechain Assignment

For sidechain assignments, the reference spectrum was a 2D $^1H^{13}C$-HSQC shown in Figure 4.9. For 3D experiments, minimising the spectral width of the indirect dimensions is important so that short acquisition times are balanced with spectral resolution and spectral folding was optimised to accommodate these requirements. An initial HSQC with large spectral widths was used to find ideal points to fold the spectra where no resonances were split and overlap was avoided. In RSN-2 there are only a small number of aromatic residues – four tyrosine and four phenylalanine. The characteristically high chemical shifts for aromatic protons means that separate experiments (HBCBCGCDHD and HBCBCGCEHE) could be used to distinguish the associated resonances by association with the known C_β.

Finally, the HCCH-TOCSY experiment[115] was used to assign all visible proton and carbon resonances. This was a particularly successful experiment with excellent sensitivity. Many peaks were observed which were not present in preceding experiments and assignments could be inferred for the residues that remained unassigned. After this process, all residues were assigned except for three residues at the N-terminus. A small number of expected resonances were not observed (probably due to solvent exposure/exchange although they did show some proteolytic degradation by mass spectrometry). The result of this was the complete assignment of the $^{1}H^{13}C$-HSQC.

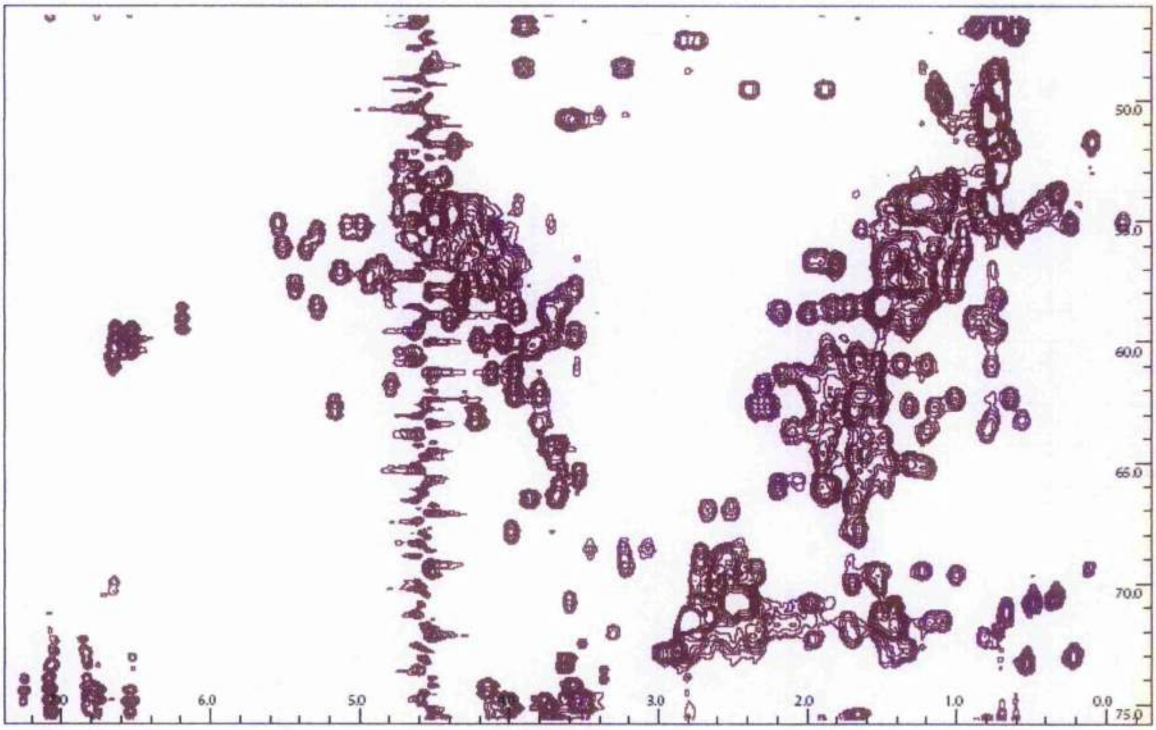


Figure 4.9. **^{1}H-^{13}C HSQC of RSN-2 at 303 K. The x and y-axes represent the ^{1}H and the ^{13}C chemical shifts respectively. The spectral width of the indirect dimension was arranged to give optimal folding and therefore reduce experiment time and disk space (increase resolution) for the 3D experiments.**

4.5 Summary

Multidimensional NMR pulse sequences were used to assign chemical shifts to almost all of the atoms in the RSN-2 sequence. Whilst this information does not provide biological structure in itself, it is a necessary prerequisite for structure calculations. Manual assignment meant that a great deal of familiarity with individual residues was gained and allowed insights into local chemical environments through considerations of individual peak intensities and chemical shifts.

5 Structure Calculation of RSN-2

5.1 Introduction

Having assigned all detected residues in RSN-2, the information could be used to calculate the solution structure. Modern methods have been developed to minimise the manual effort required to transfer atomic assignments to solution structures but some specialist knowledge and systems are still required to complete this task.

5.2 Experimental Strategy

Structural information can be obtained by measuring the interaction between nuclei that are close in space as well as those that are covalently attached. Recently, residual dipolar couplings (RDCs) have been used to deduce long-range orientational alignments between atoms, in anisotropically tumbling proteins[116]. NMR can provide this spatial information, because nuclei are found to couple via dipolar through-space couplings as well as scalar through-bond coupling. Through-space dipolar couplings are physically manifested by the spin relaxation between non-equilibrium populations of nuclei, when the energy difference for spin inversion of both nuclei matches the frequency of the effective field fluctuations experienced by the neighbouring nuclei. The cross-relaxation is called the Nuclear Overhauser Effect (NOE), and is observed in spectra as a change in peak intensity (relative to a spectrum where NOE effects are minimised) due to the altered population differences.

This dipolar mechanism is a major part of the spin-lattice relaxation (longitudinal) mechanism of a perturbed nucleus and is proportional to the inverse of the sixth power of the interatomic distance. The result of this is that only nuclei which are close in space (≤5 Å) to each other can interact efficiently in this way. It is also important to note that the NOE intensity is inversely related to distance for a molecule undergoing isotropic tumbling in solution and because of this, NOE peaks can be classified by approximate distance. In a protein structure the NOEs created can be divided into groups based on peak intensity, and range from intense peaks indicating atoms close in space which are usually scalar coupled (but those that are not close in sequence strongly influence the derived structure), to medium-range interactions which can help to define the secondary structure and finally long-range interactions which help to characterise the global fold. The absolute accuracy of the peak intensity versus interatomic distance is less important than the number of NOE peaks, because the structure should be almost completely specified by the matrix of atomic contacts. This information is used in computational methods to calculate the solution structure.

Inputs to the structure calculations were based on three experiments – ^{15}N and ^{13}C edited 3D NOESYs and inferred knowledge of strong hydrogen bonds from amide exchange. As well as this, the starting point for structure calculations is the input amino acid sequence, which is sufficient to derive the empirical restraints that define the covalent structure of the molecule.

5.2.1 NOESY Experiments

The basic NOESY experiment is a series of three 90° proton pulse with two different delays between them. The first delay (t_1) is the incremental delay used to confer chemical shift data in the indirect dimension and the second delay is the chosen NOESY mixing time (t_m) during which time, the cross-relaxation occurs.

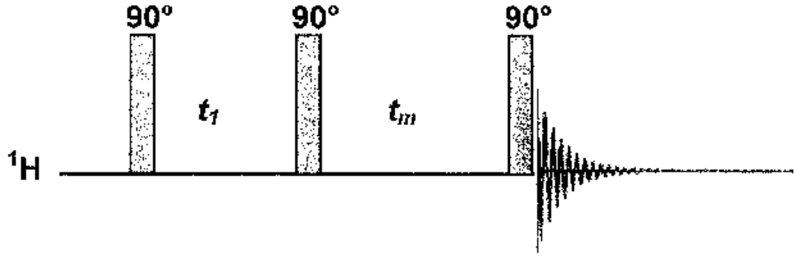


Figure 5.1. NOESY pulse sequence.

Cross-relaxation is most efficient at a similar mixing time to the spin-lattice relaxation but in proteins, the mixing times chosen are shorter to minimise the errors from spin-diffusion (dipole-dipole relaxation via an intermediary nucleus). Chemical exchange can also give crosspeaks in this experiment and is unavoidable -- the same sequence can be deliberately used to probe exchange (EXSY). The first pulse transfers equilibrium from equilibrium to the transverse plane where it evolves under the effects of chemical shift and relaxation for the first delay. The second pulse puts magnetisation onto the longitudinal axis except for the component lost during the first delay so there is z-magnetisation with a frequency label from spin 1. During the mixing time, cross-relaxation occurs between spin 1 and spin 2 which transfers some of the magnetisation to spin 2. The final pulse places the magnetisation vectors in the transverse plane again where it is detected with spin 1 and 2 frequency (chemical shift) signals.

By adding an HSQC sequence to the end of the typical NOESY pulse sequence, 3D experiments can be created that separate the NOESY experiments according to the attached heteronucleus. ^{15}N- and ^{13}C-edited NOESY experiments[117,118] were performed with a mixing time of 120 ms and then the data was processed to give positive NOE peaks. Peaks were manually picked but not assigned.

Traditionally NOESY experiments have been assigned, at least partially, and the peaks converted into an interatomic matrix of peak intensities (distances) which could be used as input to structure calculations. This creates some difficulties because overlapping peaks are sometimes impossible to assign and the possibility of a single peak being composed of more than one NOE contact is not considered. This can lead to errors in structure. Additionally, the assignment process is very time-consuming. In this work an alternative strategy of ambiguous distance restraints was used.

5.2.2 Hydrogen Bonds

To investigate the number, extent and residue specificity of hydrogen bonds donated from the amide proton, the N3 sample was lyophilised (Christ Alpha 1-4) and then dissolved in 600 µL of D_2O (Sigma). Immediately after dissolution, the sample was transferred into an NMR tube and prepared for experiments (tuned, matched, locked and shimmed) which took approximately five minutes. A total of eight ^{1}H^{15}N-HSQC experiments were executed in series, for a total of two hours after which, all amide protons had exchanged with D_2O.

The presence of hydrogen bonds was indicated by the visibility of crosspeaks in the HSQC caused by the retarded proton - deuterium exchange. Initially, 17 of the residues were observed representing a reasonably strong hydrogen bonded network and these were seen to disappear at a similar rate for all residues confirming that all areas of the protein are solvent accessible (through structure or conformational fluctuations).

Although the amide donor nucleus could be identified from this method, the acceptor peak could not be explicitly established and so the data could not be used until a reasonably accurate solution structure was obtained from NOE data alone. Once the solution structure was characterised, the donor protons were examined to find if they were arranged in proximity with a potential donor. This was found to be the case for all of the observed H-bond HSQC crosspeaks and so a table of unambiguous restraints could be created.

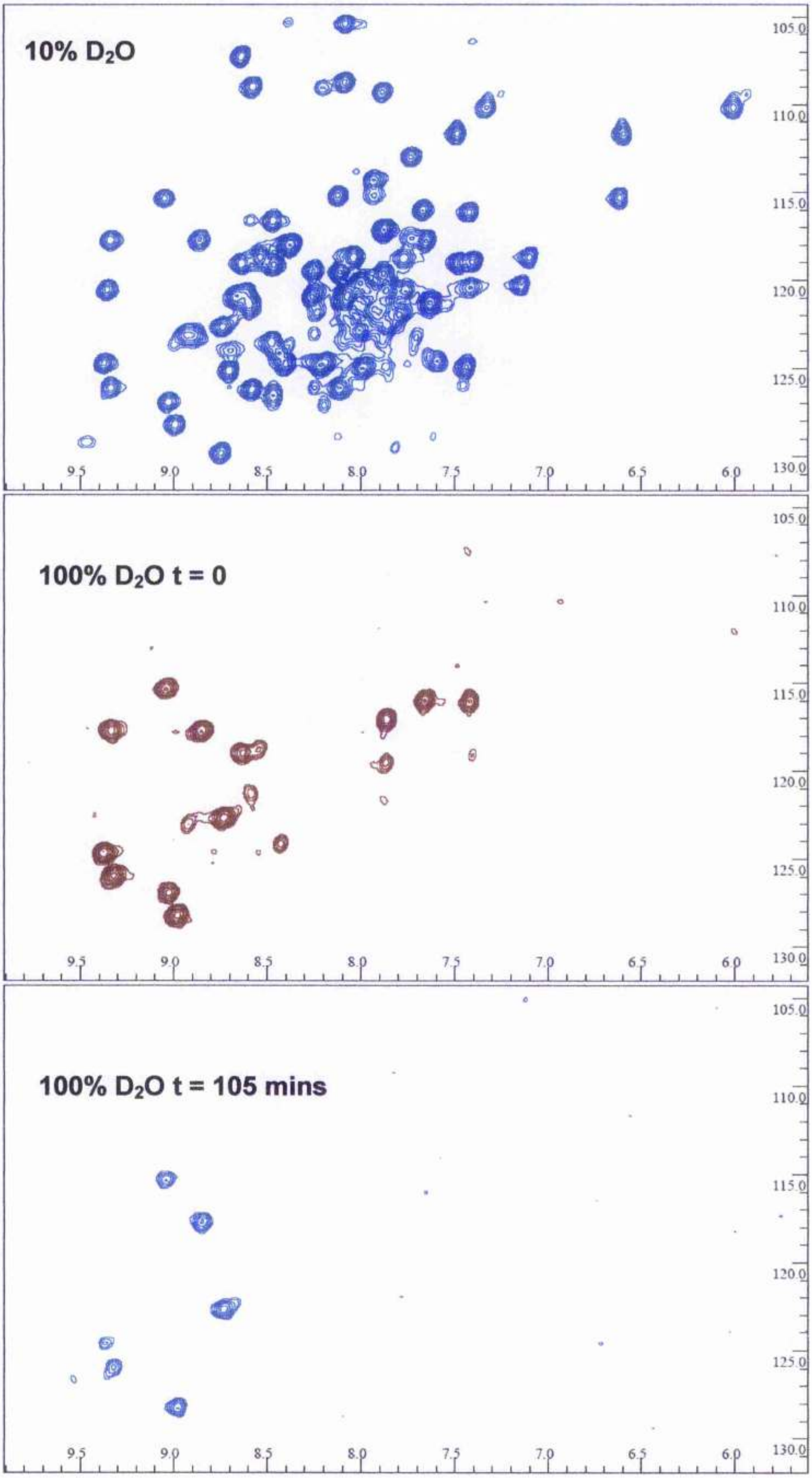


Figure 5.2. **$^1H^{15}N$-HSQC spectra of RSN-2 undergoing deuterium exchange. *i)* RSN-2 in 90% H_2O/10% D_2O, *ii)* Immediately after dissolution with D_2O, *iii)* 105 minutes after dissolution.**

5.3 Generation of Experimental Restraints

After completion of resonance assignment, the "shift match distance constraints" function of *Analysis* was used to produce a list of ambiguous restraints from the NOE data for input to structure calculation software. Ambiguous restraints interpret an NOE crosspeak as a restraint between all possible pairs of resonances that are within a defined chemical shift range of the observed chemical shifts[119]. In many cases, this will be entirely a contribution from a single dipole – dipole interaction but it allows for the possibility of the peak being composed of more than one NOE interaction. A complete set of protein NOE peaks can be converted to ambiguous restraints, which can be used to calculate structures without any initial bias introduced from assignments. Clearly, the quality of the final structure will still depend on the accurate resonance assignments used for the ambiguous restraints, and the quality of the peak picking in the NOE spectra which are prone to artefact peaks.

The ambiguous distance restraints took the form of single NOE peak with a number of options for the both of the protons involved in the dipolar coupling, where the protons were selected from their assigned chemical shifts which lay between an allowed range from the peak. For example, a peak picked from a ^{13}C-edited NOESY experiment with centre at 2.656 ppm (^{1}H direct), 8.110 ppm (^{1}H indirect) and 71.257 ppm (^{13}C) was given the following ambiguous assignment:

```
assign ( resid 27 and name HE2
         or resid 34 and name HB2
         or resid 84 and name HE2
         )
       ( resid 5 and name H
         or resid 21 and name H
         or resid 34 and name H
         or resid 35 and name H
         or resid 84 and name H
         or resid 93 and name H
         or resid 94 and name H
         ) 2.8 2.8 0.0 volume=2.419 peak=59 ppm1=2.656 ppm2=8.110
```

This data indicates that there were three different residues that were within the given range for the H-C atoms and at least one of these was coupled to at least one of a possible seven protons in the indirect dimension. In addition to this, the software has assigned the normalised peak volume of 2.419 as equating to a target H-H distance of less than 2.8Å (an intense peak).

In this way, each peak in the ^{15}N- and ^{13}C-edited NOESY spectra was ambiguously assigned. Lists of ambiguous restraints were exported from *Analysis* for input to CNS by using the "export ARIA list" function.

5.4 Conversion of Restraints into Structure

5.4.1 Software and Methods

In this work, structures were calculated using restrained molecular dynamics with simulated annealing (MD-SA) protocol described by *Nilges*[119]. Data manipulation and calculation was performed using the Crystallography and NMR System[120] (CNS version 1.1) with a version of the PARALLHDG 5.3 forcefield[121] which was modified (by Dr. Brian Smith) to use the IUPAC atom nomenclature. Calculations were carried out on clustered nodes (IBM X Series 330, 340 and 370), access courtesy of the University of Glasgow, Department of Computing Science.

Unlike some other methods, the MD-SA calculations work entirely in Cartesian space. The molecular dynamics part of the protocol involves the minimisation of a potential energy function for the protein which is composed of individual energy components:

$$E_{total} = E_{bond} + E_{angle} + E_{impropers} + E_{coup} + E_{dihedral} + E_{vdw} + E_{electr} + E_{NOE} + E_{H\text{-}Bonds}$$

Actual, experimentally-defined restraints only represent two of the parts of this sum in these calculations further inputs from experiments, such as dihedral angles or residual dipolar couplings are possible). Because of this, an important part of the process is the initial generation of a random structure based on protein sequence and using known physical attributes for bond lengths, angles, van der Waals radii etc. The random structure is then minimised under the influence of the applied functions. The actual form of the functions used in this work was simple repulsive functions.

Constant temperature molecular dynamics cannot effectively minimise a protein structure from a random distribution to a global fold because structures will almost certainly get stuck in local minima. This problem is rectified by simulating a temperature rise which provides the energy required to overcome local minima. In the programs used in this work, the structure is heated to a theoretical 2000 K where the high kinetic energy allows an extensive sampling of conformational space. The potential energy is minimised by molecular dynamics and then the system is cooled slowly. This continues in a large number of small steps, with energy minimisations and cooling, with temperature-dependant weights being placed on the various energy terms. At low temperature, a final round of molecular dynamics yields the energy minimised structure. This calculation is performed a number of times from different randomly generated starting structures and a range of different minimised structures are obtained. This group of structures is known as an NMR structure ensemble.

5.4.2 Initial Calculations

For the first round of structure calculations, only strong and well defined (definitely not noise/artefacts) NOE peaks were used as input. Despite the observed change in hydrodynamic profile observed after chemical reduction, and the presence of only two cysteine residues, the inferred disulphide bridge was not input as a calculation prerequisite. As already stated, the deuterium exchange data was not used for initial structure calculations. In the first round, 26 individual structures were calculated.

Structures created from the initial calculations were analysed for restraint violations using the ARIA[122] *analyse* tool within CNS. From the initial ensemble, consistent violations (especially high energy violations) were investigated and found to be either assignment errors (two identified), peak picking issues (not centred or in the wrong plane) or water exchange peaks (the majority of error peaks). Resolution of these violations was an iterative process whereby violations were analysed and peaks restraints were removed or adjusted followed by another round of structure calculations.

After ten iterations, an ensemble of structures was obtained with a consistently low NOE energy and reasonably high structural homology. During the refinement process, the initial 2800 restraints were reduced to 2267 restraints from the removal of noise and artefact peaks, and from discarding the symmetry-related peaks using the *check* tool (see section 5.4.3).

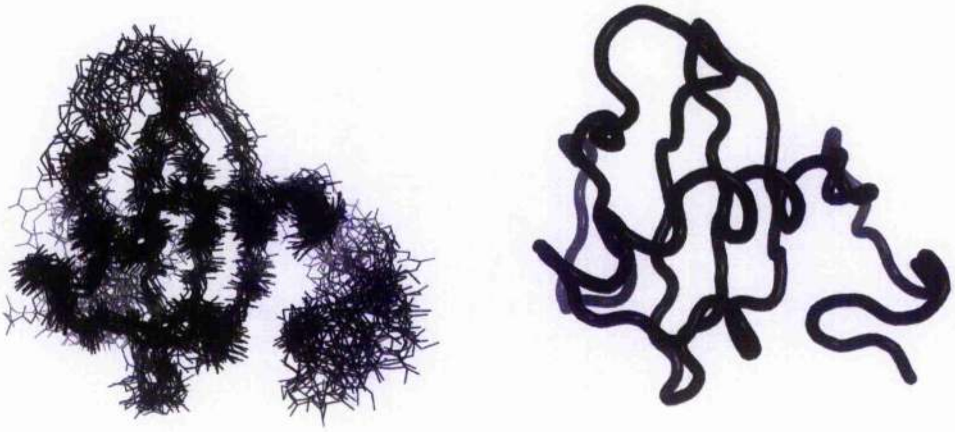


Figure 5.3. Structures from the first round of structure calculations. The superimposed structures are shown on the left and the average structure from the maximum likelihood superposition see Chapter 7) is shown on the right.

5.4.3 Secondary Calculations

Examination of the initial structures (shown in Figure 5.3) revealed that there were areas of the structures which were well defined and showed little differences within the ensemble. In general, all the structures consisted of four strands making a curved β-sheet structure with a bent helix running perpendicular over the sheet. The helix and strands 1, 2 and 3 were well defined with the 4th strand and the N-terminus being only loosely defined. Amongst the well defined area, it was clear that the two cysteine residues were close in space to each other and their input structure as discrete oxidised thiol sidechains appeared to result in a false kink in the strands and led to consistent violations. Based on this evidence it was decided to treat these residues as a disulphide bridge for subsequent calculations by modifying the covalent structure definition to include a 2.8 Å S-S bond.

Initial structures were also examined with reference to the hydrogen bond experiments and it was observed that the well defined areas overlapped with the H-bond donors that had been deduced from deuterium exchange experiments. It is probably unsurprising that those areas that had the strongest and most consistent NOE data also showed the highest degree of order in terms of hydrogen bonds, with both sets of data being indicative of a relatively rigid and immobile central core. H-bonds were introduced into structure calculations by providing a list of input restraints for hydrogen bond donor (proton atoms) and acceptor (nitrogen atoms) with a range of bond lengths being allowed.

Before proceeding to further structure calculations, the NOESY spectra were re-examined and more peaks were picked. These extra peaks were of much lower intensity and less well resolved than in the first round but were chosen to improve the accuracy of the global fold. Because these peaks were of low intensities, they were unlikely to change the general topology so could be introduced without fear of causing improper folds, but incorrectly picked peaks could be identified by local energy violations. In total (^{13}C- and ^{15}N-edited NOESY), 2847 peaks were picked to create ambiguous restraints.

Using all of these inputs, a round of 26 structures were calculated without any reference to the previous calculations. These calculations were analysed for consistent violations and a number of the newly picked NOE peaks were identified as false and were removed. Removal in *Analysis* is properly achieved by setting the peak merit to zero rather than deleting the peak. In this way, peaks would not be re-picked but the figure of merit could be used to remove the peak from restraint generation. As expected, the hydrogen bond data and the inclusion of the disulphide bridge did not significantly affect the calculated structures, or the energies thereof.

After appropriate examination (and sometimes removal) of the consistently violating restraints using the *analyse* tool, the structures were refined using additional ARIA tools. The *filter* tool reduces the ambiguity of the restraints based on the contacts made in the previous round of structures, to create restraints that are most likely to represent the information contained in the crosspeaks. This allows the structures to be defined more precisely since the disambiguated restraints are less "soft". Any likelihood of this method leading to incorrect restraints being selected (local minima rather than global) was minimised by analysis of consistent violations throughout the ensembles, as described previously.

Efficiency of CPU time was increased by running the *check* tool which reduced duplicate NOE data to a single restraint i.e. H' with NOE restraint to H'' also often has an associated restraint for H'' to H'. By transforming these two restraints into a single restraint, there were less total restraints for the structures to conform to and the weight placed on such a restraint was normalised versus others observed. This procedure also made one NOE restraint table from the two originally input.

Restraint filtering was introduced iteratively by reducing the filtering by 0.01 (1%) at each round from 1.00 to a final value of 0.95. The protocol used was:

- Input restraints
- Calculate structures

- Check for consistent violations using *analyse* – alter restraints and restart structure calculations, if necessary
- Filter the restraints
- Remove duplicate restraints using *check*
- Use new restraints for next round of analysis

This protocol was used until the filtering had reached 0.95 (95%) using a total of 2358 restraints. At this point the NOE data was calibrated, in terms of normalised NOE peak intensity versus inter-atomic distance, using the *calibrate* tool, and the structures were calculated again. Comparison of the uncalibrated versus the calibrated structures revealed a marginal increase in NOE energies and total energies in the calibrated structures but less violations were evident as reported by the *analyse* tool. In terms of structure, the only significant difference between the two sets of structures, was the position of the amino-terminus (Figure 5.4) which was not defined by experimental data in any case.

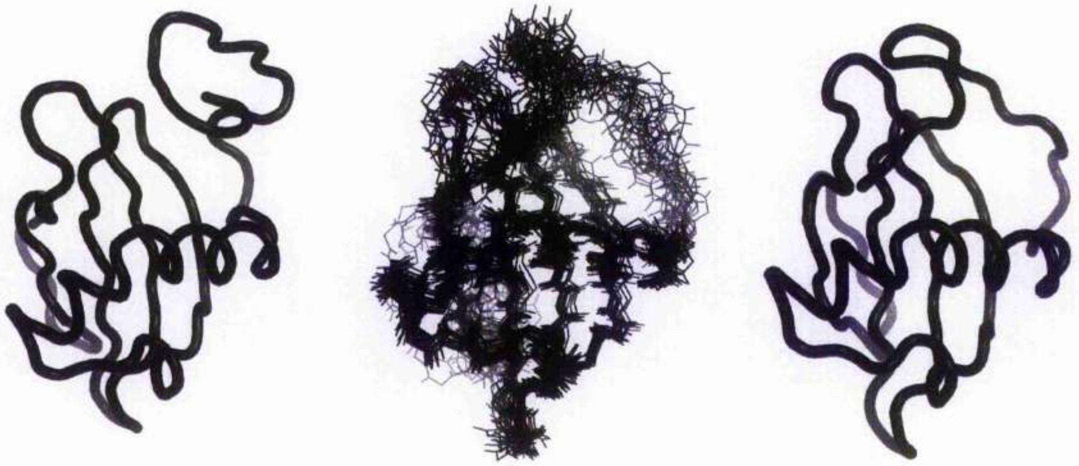


Figure 5.4 **Structures from the second round of calculations. The structure on the left hand side is the most likely superposition of the structures calculated without NOE peak calibration, the middle and the right hand side pictures are the superimposed and most likely superposition structures respectively for the structures calculates with NOE peak calibration.**

5.4.4 Final Structure Refinements

After the final round of structure calculations had been made from the filter/analyse/check/calibrate protocol a final step was performed to increase the quality of the structure. This step acknowledged the fact that the standard molecular dynamics force fields can represent the nonbonding (electrostatic, van der Waals) interactions poorly which can manifest itself in inadequate atomic packing and unfavourable Ramachandran plots.

To overcome this problem, a final structure refinement was performed using simulated annealing of the protein which is virtually encapsulated by a thin layer of solvent (water) molecules. The final group of refined structures were used as a template for treatment with the *water_refine* protocol, based on the published procedure[12] and adjusted for implementation within CNS by Dr. Brian Smith.

An ensemble of 100 structures was created in this way with 95% filtering, removal of violations and explicit solvent filtering. No consistent violations of >0.5Å were observed in these structures but a small number (4%) of the structures were found to have much greater NOE energies (by a factor of 5-10 times) than the others. It is likely that these structures were stuck in local energy minima during the structural annealing procedure and never managed to overcome the energy barrier to further minimisation to the global minimum. Accordingly these structures were removed and replaced with random (but checked for high energy rejection) structures selected from a set of twenty structures which were calculated in exactly the same way as described above.

5.5 Summary

Through-space atomic interactions were probed for all RSN-2 protons that were bonded to nitrogen or carbon atoms. The RSN-2 sequence and the matrix of proton-proton contacts, were used as inputs for structure calculations using a simulated annealing with molecular dynamics protocol. Initial structures were augmented with inputs from a deduced disulphide bridge and hydrogen bond inferred from deuterium exchange experiments. ARIA tools were used to improve the precision of structures and remove erroneous input restraints. Finally, a set of one hundred final structures were created which took into account the effect of electrostatic solvent interactions and internal non-bonding interactions more correctly. The results of the structure calculations are discussed in Chapter 7.

6 Solution Dynamics of RSN-2

6.1 *Introduction*

It is unlikely that a single structural arrangement could result in an RSN-2 structure that is both water soluble and surface active because the polarity requirements for these properties are conflicting. It was proposed that examination of the mobility in the protein and individual structural motifs and residues might better explain the structure to function relationship. All polypeptides are mobile in solution, to some degree, and the inherent mobility is an integral component of many proteins' function, and the ability to study these motions by NMR is a distinct advantage of the technique over X-ray crystallography. Protein motions can occur on the picosecond scale for relatively fast and short-range motions to second scale for concerted motions of large domains. Recent advances in instrumentation and labelling techniques have allowed the examination of dynamics of extremely large proteins such as the 670 kDa 20S proteasome[123].

Excited magnetic spins return to the equilibrium state over time by processes that are collectively known as relaxation. Because of the frequencies involved, the relaxation processes are almost entirely non-radiative. NMR is able to analyse nucleus-specific relaxations, and relaxation processes are directly related to the physical environments and the movements of the atoms in this environment, so information of protein dynamics can be inferred from the study of individual and combinations of relaxation mechanisms.

Typically, and in this work, relaxation studies are performed by measuring the magnetic relaxation of ^{15}N amide nitrogens. Although not a complete analysis of every atom in the protein, this method gives a generalised overview of concerted movements of residues or regions within the protein. The analysis is of three separate relaxation mechanisms discussed separately below: longitudinal relaxation, transverse relaxation and heteronuclear NOE cross-relaxation.

6.1.1 Relaxation Mechanisms

Longitudinal relaxation is also known as spin-lattice relaxation because it is the process by which non-equilibrium magnetism returns to the equilibrium i.e. net magnetic moment along the z axis. Transverse magnetism and therefore non-equilibrium populations are lost by interactions such as with the sample environment (lattice) through heat exchange with molecular vibrations and rotations. The relaxation time is called T_1 and is often described in terms of the relaxation rate R_1 which is the inverse of T_1.

Transverse relaxation describes the decay of coherences or the magnetisation components that are in the xy plane. The relaxation time and rate is T_2 and R_2 respectively. T_2 relaxation occurs through interactions of the excited nuclei with other spins – spin-spin relaxation. Although this is essentially a swapping of excited states, the net effect is relaxation because the average time that a magnetic vector spends in the excited state is reduced. T_2 is the same or less than T_1 (never more than). Cross-relaxation by the nuclear Overhauser effect, has already been described elsewhere in this work.

Figure 6.1 illustrates the relationship between T_1, T_2 and NOE with the rotational correlation time. T_1 is fastest at frequencies close to the *Larmor* frequency, and so reaches a minimum, whereas T_2 decreases to zero. The NOE ratio is the most sensitive parameter[124] within the nanosecond timescale covered in this work, starting from positive NOE intensities for larger molecules which tumble slowly, and moving steeply to negative intensities for small molecules. It should be noted that all of these parameters are field dependant.

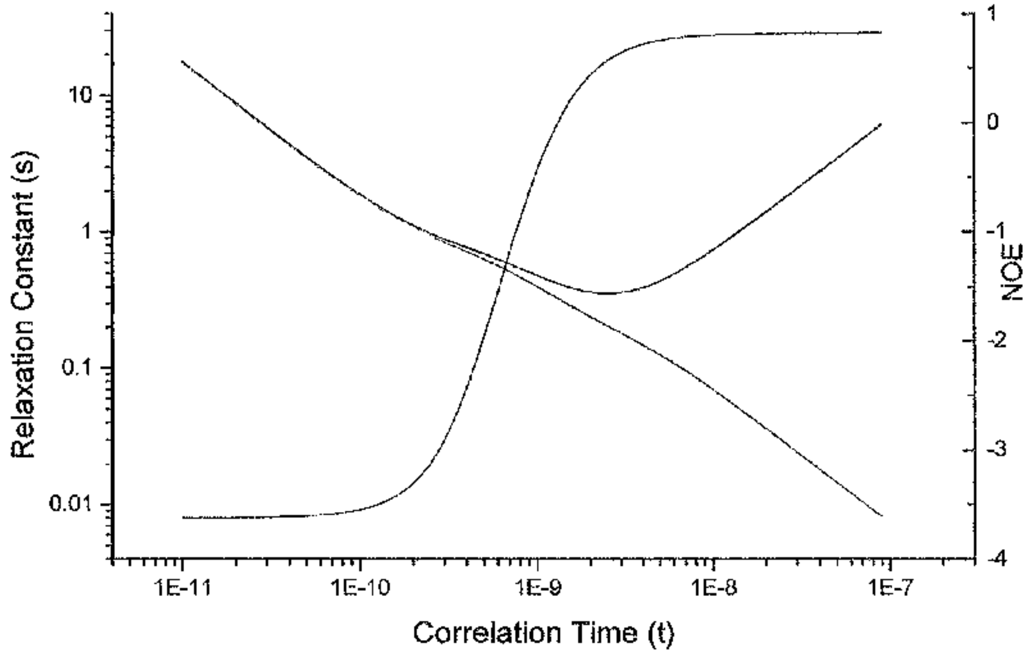


Figure 6.1. Simulation of T_1, T_2 and NOE at 14.1 Tesla. Data provided by Dr. Brian Smith.

A magnetic spin, in isotropic solution, experiences magnetic fluctuations which vary depending on the particular molecule in question. For a spin experiencing rapid fluctuations, the time it takes before it experiences a different magnetic environment is short compared to a spin that experiences relatively slow fluctuations. The collective average magnetic fluctuations experienced by the ensemble of magnetic vectors can be described, in terms of the time taken to experience change from the initial magnetic environment, in the form of an autocorrelation function. Such a function takes the form of an exponential decay where the decay rate constant is called the rotational correlation time, τ_c. A small molecule will tumble in solution faster than a large molecule, and therefore experience faster magnetic fluctuations represented by a faster rotational correlation time. Because of this relationship, the rotational correlation time can be used as a diagnostic of protein size, although it is also a function of shape, temperature and solution viscosity.

Small molecules that are tumbling very quickly are able to sample a large number of different frequencies compared to large molecules that are tumbling slowly and only able to sample a limited number of frequencies. This is represented by the spectral density function, which is the Fourier transformation of the autocorrelation function, analogous to FT-NMR processing from the time to the frequency domain. Also analogous to the FT processing is the observation that magnetic vectors with short correlation times broaden the spectral density function, similar to rapidly relaxing spins broadening NMR peaks. These phenomenon are shown in Figure 6.2, the frequency is represented by ω.

$$G(t) = \frac{1}{5}\exp\left(\frac{t}{\tau_c}\right) \xrightarrow{\text{Fourier Transformation}} J(\omega) = \frac{2}{5}\left(\frac{\tau_c}{1+\omega^2\tau_c{}^2}\right)$$

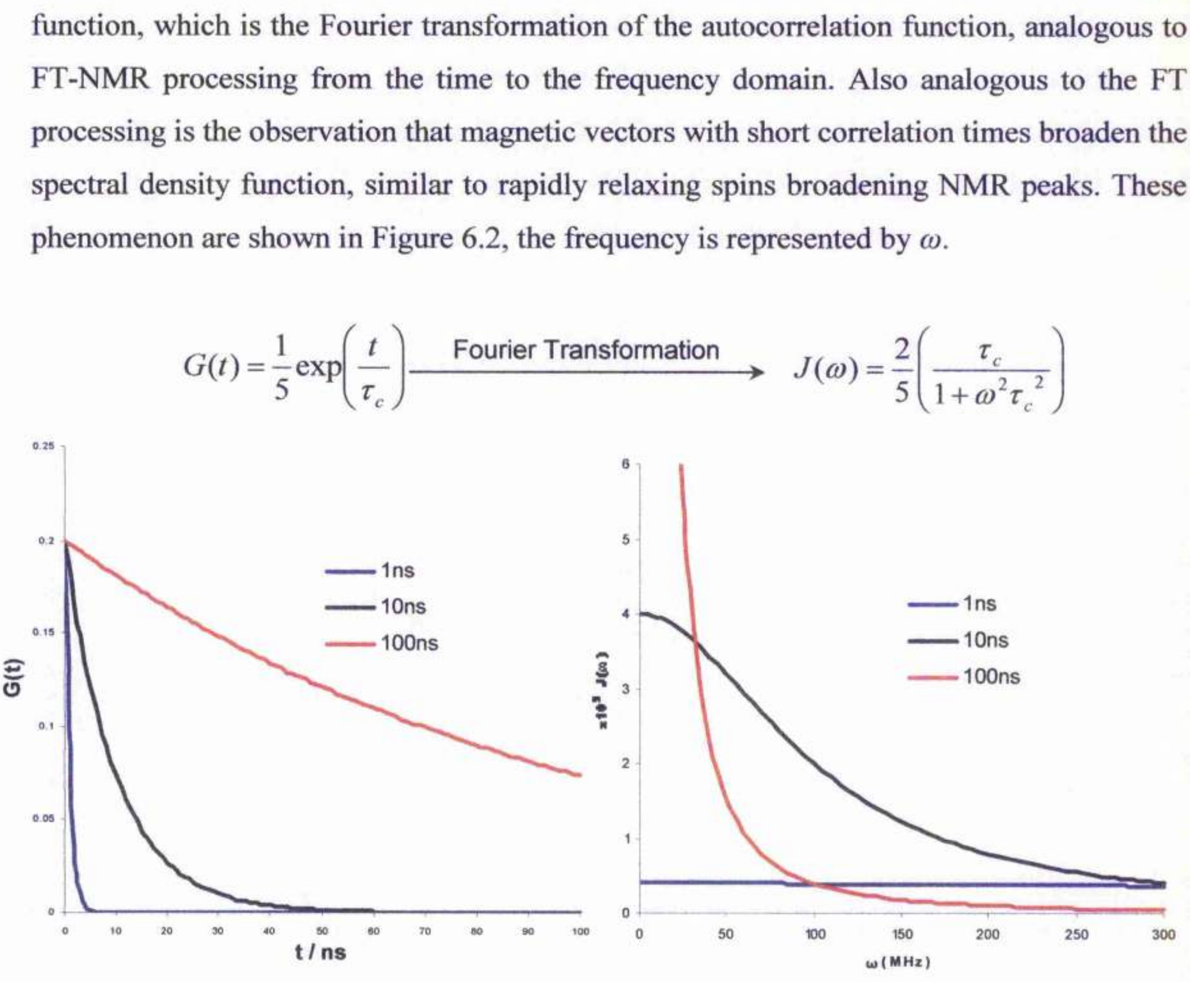


Figure 6.2. The autocorrelation function and spectral density functions for magnetic vectors with τ_c of 1, 10 and 100 ns.

These mathematical descriptions hold true for idealised spherical objects but do not adequately describe the situation in proteins. In these complex systems a number of different contributions can be made to relaxation, resulting in a variety of relaxations being present for different nuclei, which experience movements with a range of dynamic processes being possible.

Due to these complexities, relaxation of individual nuclei in proteins are very difficult to predict. This is overcome, to some degree, by the selection of backbone nitrogen nuclei for monitoring relaxation processes. The amide nitrogen relaxation is dominated by only two processes – the dipolar relaxation with the attached proton already described, and chemical shift anisotropy (CSA). CSA describes the field inhomogeneities experienced by the magnetic vector as it reorientates in solution which can be an efficient relaxation mechanism. The amide nitrogen is also only able to move in a limited range and so some limited assumptions can be made about the relaxation, and the process can be fitted to experimental data to derive specific relaxation parameters. Other advantages of using the amide nitrogen as a probe is that information can be gained from every residue so relaxation mechanisms can be mapped across the structure and used to help explain biological functions. It should always be noted though, that the backbone atoms are relatively rigid and do not represent the full protein dynamics absolutely.

6.2 Experimental Analysis

6.2.1 Strategy

As described above, the analysis of amide nitrogen relaxation provides a relatively easy way to examine dynamic processes across the protein, without running a large number of experiments and using elaborate models to fit the experimental data. In this way, the RSN-2 backbone dynamics could be reasonably well defined and concerted regions of movement could be observed without analysing individual sidechain atomic dynamics.

From the spectral density function described, it can be seen that to provide an adequate description of relaxation, the function would need to be "mapped" over a wide range of different frequencies. To overcome this, the Lipari-Szabo model-free formalism is used to calculate internal motions from T_1, T_2 and NOE data at a single or a small number of magnetic strengths. The model-free formalism is so-called because it does not assume any particular model of motion prior to analysis. The analysis uses known equations relating T_1, T_2 and NOE to a modified spectral density and correlation function which includes a term for internal motion of the N-H vector which is not dependant on the overall tumbling of the molecule. The revised autocorrelation function can be Fourier transformed into a spectral density function which includes a generalised order parameter S and an effective correlation time τ_e of the amide bond motion[125,126]. The order parameter describes the internal motion with the extreme values of 1 and 0 describing a completely fixed and a completely free amide vector, respectively.

$$C(\tau) = S^2 + (1-S^2)e^{-\frac{\tau}{\tau_e}} \longrightarrow J(\omega) = \frac{2}{5}\left(\frac{S^2\tau_c}{1+\omega^2\tau_c^{\;2}} + \frac{(1-S^2)\tau}{1+\omega^2\tau^2}\right) \quad \text{, where } \frac{1}{\tau} = \frac{1}{\tau_c} + \frac{1}{\tau_e}$$

This strategy allows for simple experiments, all taking the form of a pseudo-3D $^1H^{15}N$-HSQC experiment, where each 2D plane represents a different delay time. These can then be analysed to calculate individual relaxation rates. It was hoped that the backbone dynamics would give an accurate enough description of the overall protein dynamics and provide some indication as to the structural basis for surfactant activity.

6.2.2 Experiments for R_1, R_2 and NOE

R_1 describes the rate of re-appearance of net magnetisation along the axis of the fixed field (conventionally the z axis). To record this, magnetisation is perturbed to a non-equilibrium state and then allowed to relax for a defined delay. After this, the z component is flipped onto the transverse plane where its intensity can be recorded. By measuring the peak intensity at several delays, delay time versus intensity can be plotted and the R_1 value can be found from the exponential decay constant. The sensitivity of the experiment is increased by bracketing the pulse sequence with INEPT and reverse INEPT sequences for proton-nitrogen magnetisation transfer. This sequence yielded a series of peaks with increasing intensities with delay time, creating a curve that could be fit to an exponential constant proportional to give T_1. Delay times used were 43.1, 253.1, 505.1, 575.1 and 1009.1 ms. The experiments with 43.1 and 253.1 ms delays were repeated to assist with error estimation.

R_2 is the rate of disappearance of the transverse components of the bulk magnetic moment and so the experiment measures the remaining transverse magnetism after a variable delay time. Proton magnetisation is transferred to nitrogen and developed in the transverse plane. During the delay time, all other magnetisation perturbations from chemical shifts, scalar coupling, longitudinal relaxation and field inhomogenieties are refocused by the application of multiples of the CPMG (Carr-Purcell-Meiboom-Gill) pulse sequence.

Increasing delay times resulted in an exponentially decaying curve with a constant, T_2. Delays of 16.96, 33.92, 50.88, 84.8, 101.76, 118.72, 135.68 and 169.6 ms were used with the 33.92 ms delay being repeated to indicate experimental error.

To quantify the degree of relaxation that was a provided by NOE cross-relaxation, two different experiments were compared. The first experiment used presaturation of the amide protons prior to measuring the nitrogen peak intensity, so that a difference in populations due to NOE could be observed in the peak intensities. The second experiment was not presaturated and so acted as a reference. For a nucleus that relaxes rapidly through the nuclear Overhauser effect, the relative decrease in peak volume in the saturated experiment can be diagnostic. Although this experiment is can be particularly sensitive to nanosecond scale dynamics, the saturation experiment is prone to error and can lead to misleading data for relaxation analysis. The reason for these errors is the long longitudinal relaxation time for water protons which is often less than the recycle time, leading to water saturation and subsequent saturation transfer to rapidly exchanging protons in the protein. In this work, the experiment used was designed to perturb the water protons to a minimal degree so that the longitudinal relaxation was no longer a problem[127]. Both of the spectra, which were also of the form of a $^1H^{15}N$-HSQC, were recorded twice so that an estimate of experimental error could be made.

6.2.3 Analysis of Data

Each of the data sets was processed using *Azara*, as described previously. These spectra were imported into *Analysis* and the peaks were picked and assigned based on the conventional HSQC spectrum. For the R_1 and R_2 experiments, each set of spectra were entered as an experimental series with corresponding relaxation delays. Given the relaxation delays and the associated peak heights and volumes for each residue detected, the program was able to analyse the data to produce an associated relaxation rate (or relaxation time) for each residue as well as an error value based on the fit of the curve. *Analysis* was also able to perform the relatively trivial task of producing a NOE ratio value for each residue with an error value based on signal to noise values.

Subsequent analysis of this data revealed that the error estimates in the relaxation data were very small. Perseverance with this approach could not alter the erroneous data and so the relaxation data fitting within *Analysis* was abandoned in favour of more established methods. The reason for the poor estimation of error has since been discovered as a mistake in the software code and has been remedied for current software releases.

Peak height versus delay time data was exported from Analysis as tab-separated files and files were adjusted for input to *Curvefit* (written by A. G. Palmer III, Columbia University), using the scripts supplied by the author and modified by Dr. Brian Smith, as well as some manual manipulations. Data curves were fitted and rate constants were derived in batch analysis mode. The output relaxation rate value and errors (via least squares analysis) were used as the definitive values for further work as they were more in-line with expected errors than observed in *Analysis*.

6.2.4 Derived Relaxation Rates and NOE Cross-Relaxation

The relaxation rates (R1, R2 and NOE ratios) generated from *Curvefit*, were plotted as a function of amino acid sequence.

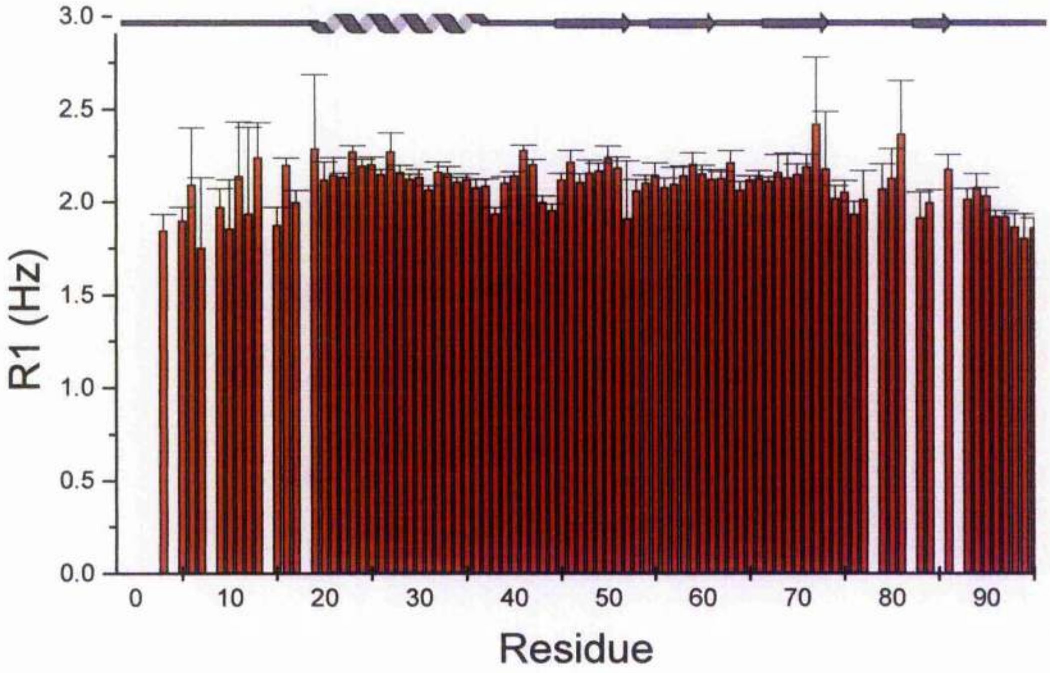


Figure 6.3. Longitudinal relaxation rates for RSN-2 residues.

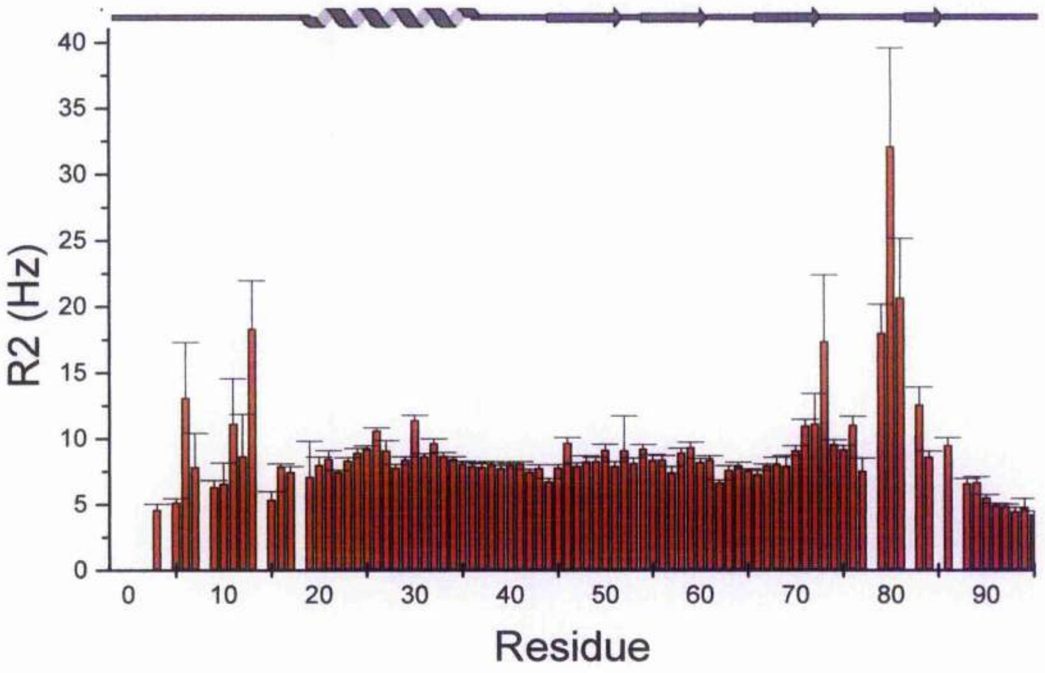


Figure 6.4. **Transverse relaxation rates for RSN-2 residues.**

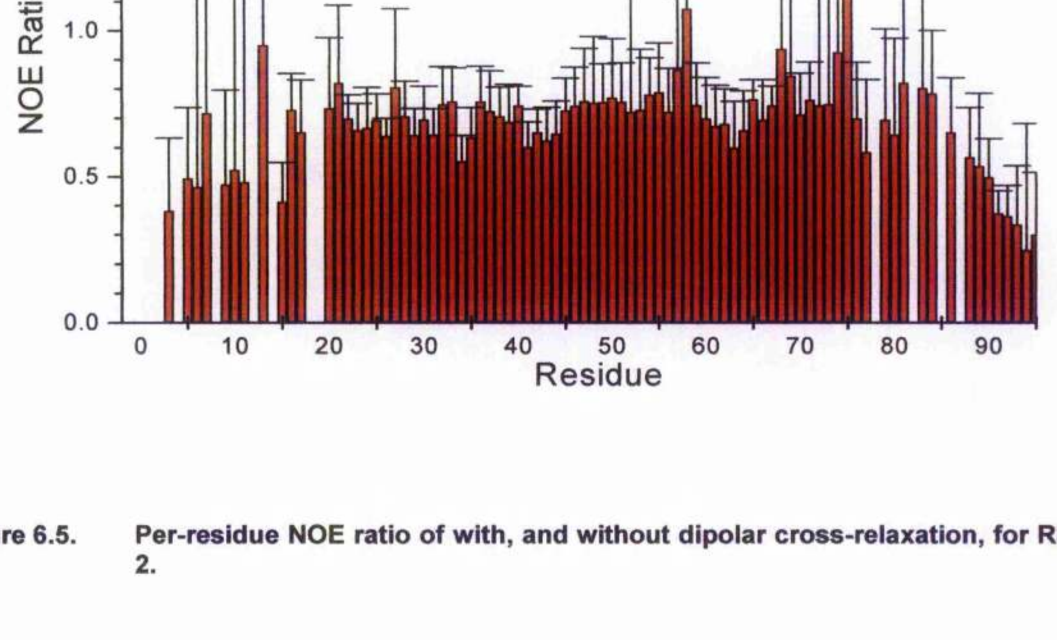


Figure 6.5. **Per-residue NOE ratio of with, and without dipolar cross-relaxation, for RSN-2.**

It is immediately apparent from these plots are the relatively high experimental errors observed in some of these fittings and ratios. Data for the NOE ratios were particularly imprecise towards the N-terminus and analysis of the data showed that this was primarily due to the low peak intensity in these areas. Whilst the low signal intensity was possibly indicative of molecular motion in themselves (e.g. fast exchange of the amide proton with water), the NOE values incorporated such high error values that the absolute values had to be treated with caution.

A qualitative analysis of the calculated data shows that there are some areas with relative variability in relaxation parameters. There are some regions where the decreased NOE ratio and lengthened T_1 times (reduced R_1 rates) suggest fast internal motion. These correspond to the termini and an area from approximately residue 75-85 where NMR signals were also observed to be weak suggesting that some dynamic process is occurring here. In the middle of this group, there are three residues (79-81) which exhibit depressed T_2 times, which is indicative of dynamic processes on the relatively slow millisecond timescale. The decreased T_2 times were also manifested by broad peaks in the RSN-2 spectra.

6.2.5 Model-Free Analysis

For quantitative analysis of the dynamic properties of each amide residue the model-free approach described previously was used. The original method described was augmented with further models[128] and further developed[129] to include a statistical protocol for model selection between the five different models describing relaxation. Each model can describe relaxation best described by one or a combination of parameters for an order parameter (general, slow or fast), an internal correlation time and a parameter accounting for chemical exchange (R_{ex}).

As an initial step, an estimate for the rotational correlation time of the molecule (τ_c) was made. This was based on the 10% trimmed mean R_2/R_1 ratios for relatively fixed residues in the rigid core of the polypeptide; residues 20-70 in this case. Given this ratio, the estimated correlation time was fixed at 5 ns via the *R2R1diffusion* program (based on reference 130, implemented and distributed by A. G. Palmer III, Columbia University) assuming an isotropic tumbling molecule. This was a reasonable assumption based on the relative principle components of the moment of inertia, calculated from the solution structure described in chapter 7 (1.00:0.94:0.68), by the program *pdbinertia* (Palmer). A value of 5 ns is entirely typical for a globular protein of 11 kDa, and strengthens the assumption that RSN-2 is a monomer in solution at biologically relevant concentrations, temperatures and pH's.

For model-free analysis the output from *Curvefit* was transformed into input for the *Modelfree* program version 4.15 (Palmer). The procedure followed was based on that of Mandel *et. al.*[129], using a statistical evaluation of each model and providing a protocol for model selection. Generally, a model was selected if the fitted data was close enough to 500 *Monte Carlo* simulations (synthetic data created randomly on a Gaussian distribution with maximum at the estimated τ_c and a width related to the estimated error that was input). The closeness of fit between real and simulated data was described by the sum-squared error (*Γ*) and a model was deemed appropriate if *Γ* was less than a critical value, and an F-test between this, and another model, was less than another critical value. The values chosen followed that shown in Figure 6.6.

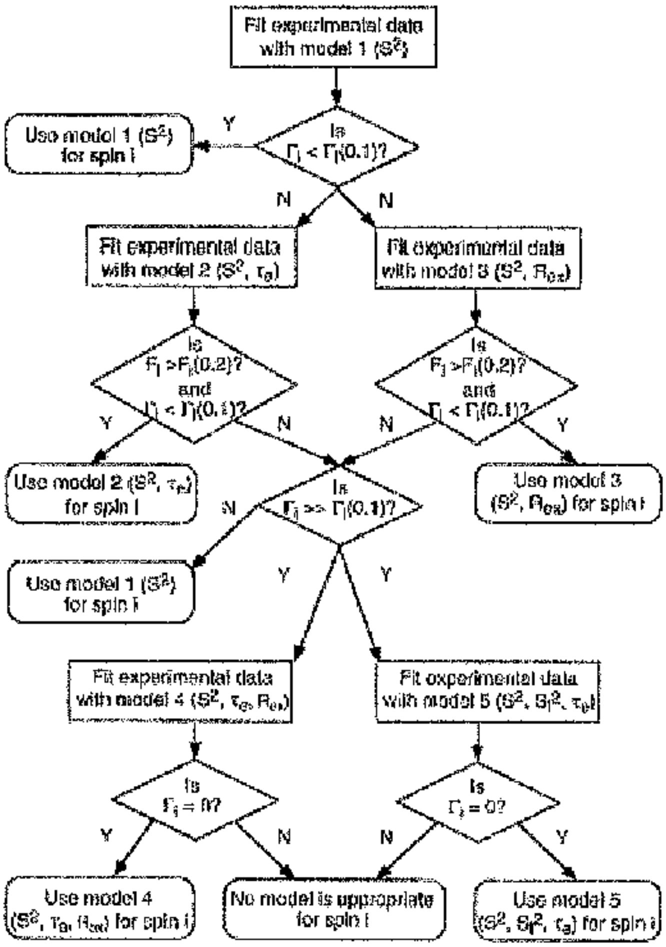

Fit experimental data with model 1 (S^2)
Is $\Gamma_i < \Gamma_i(0.1)$?
Y
Use model 1 (S^2) for spin i
N
N
Fit experimental data with model 2 (S^2, τ_e)
Fit experimental data with model 3 (S^2, R_{ex})
Is $F_i > F_i(0.2)$? and $\Gamma_i < \Gamma_i(0.1)$?
Is $F_i > F_i(0.2)$? and $\Gamma_i < \Gamma_i(0.1)$?
Y
N
N
Y
Use model 2 (S^2, τ_e) for spin i
Is $\Gamma_i >> \Gamma_i(0.1)$?
Use model 3 (S^2, R_{ex}) for spin i
N
Use model 1 (S^2) for spin i
Y
Y
Fit experimental data with model 4 (S^2, τ_e, R_{ex})
Fit experimental data with model 5 (S^2, S_f^2, τ_e)
Is $\Gamma_i = 0$?
Is $\Gamma_i = 0$?
Y
N
N
Y
Use model 4 (S^2, τ_e, R_{ex}) for spin i
No model is appropriate for spin i
Use model 5 (S^2, S_f^2, τ_e) for spin i

Figure 6.6. Flow-chart describing the protocol for model evaluation and selection. This diagram was reproduced from reference 129.

A number of analyses were performed to determine the appropriate model for each residue prior to a global optimisation. After model selection, the data was examined, and those peaks which were not well enough resolved to assure a quantitative analysis were removed from the final optimisation. After this filtering step, 54 spins were selected for analysis in the final Brent optimisation, and 52 spins were fitted to model 1 (just the order parameter *S*) and the remaining two (Lys66 and Asp89) were fitted to model 2 (the order parameter *S*, and an internal effective correlation time τ_e), but these were both fitted to the upper bound limit for this correlation time (400 ps) suggesting that the model was not a good approximation for the data.

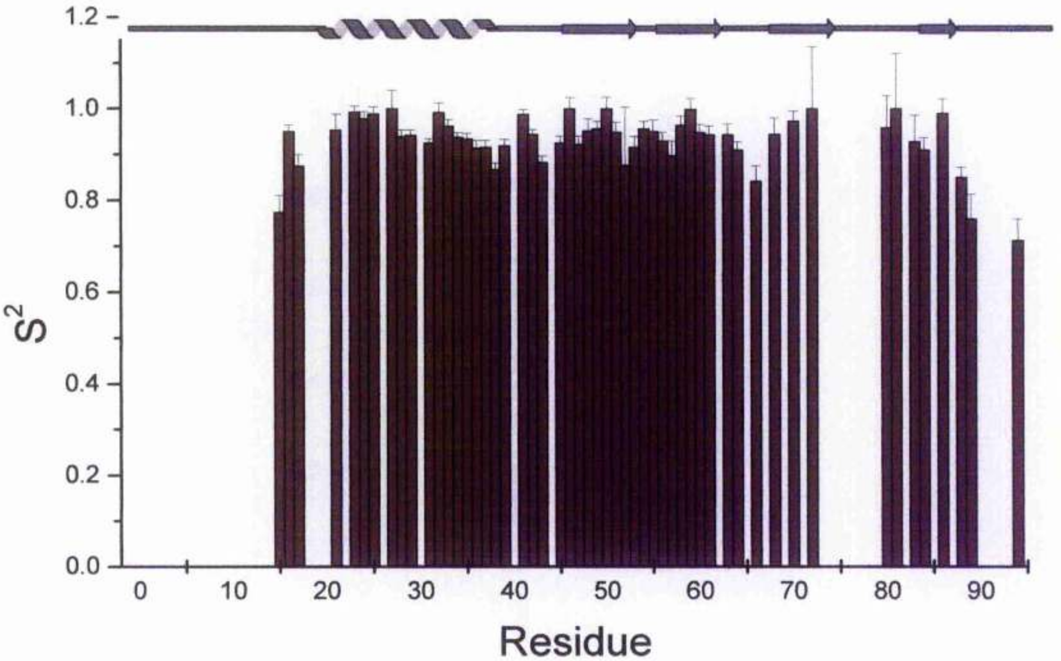


Figure 6.7. Plot of per-residue order parameter for RSN-2. The parameter was optimised using *Modelfree* model 1.

From Figure 6.7, it can be seen that the majority of the N-terminus was excluded from the final optimisation because of the quality of the data obtained. However, the generally low order parameter at this end of the molecule suggests flexibility, and the weak intensity of the peaks reinforces the concept of a rapidly moving amino-terminus. Similarly, the carboxy-terminus displays a distinctly lowered order parameter and is likely to be undergoing rapid dynamic processes.

There is some variability in the order parameter plot and the gaps in fitted data means that some possible trends were not well enough defined to draw any conclusions for many of the specific regions. Overall, it appears that the core of the structure is relatively rigid. However, residues 70-80 gave NMR peaks that were relatively weak and broad. This reduced the quality of the relaxation data and so many of the peaks could not be fitted to relaxation models, but the nature of the peaks suggests that short T_2 times (slow movement) or, more likely, rapid chemical exchange was taking place in this region.

6.3 Summary

Experimental methods were used to collect information on longitudinal and transverse relaxation rates, as well as NOE cross-relaxation effects, for amide nitrogen resonances in RSN-2. Computational methods were used to fit this data to models describing internal motion on the nanosecond timescale and statistical analysis of the fits demonstrated the validity.

Estimation of the rotational correlation time underpinned the assumption that RSN-2 is essentially monomeric. Per-residue analysis of the order parameter showed that both of the termini fluctuate relatively rapidly and a number of other areas are participating in dynamic processes. Qualitative descriptions of the relaxation data and the spectra, suggested that a number of residues at residues 70-80 were undergoing complex movements, with some being involved in rapid solvent exchange and others appearing to move on a relatively slow micro-millisecond timescale. These observations will be related to structure in the following chapter.

7 Assessment of the Calculated RSN-2 Structures

7.1 *Quality of Calculated Structures*

7.1.1 Energies and Violations

Final calculations of the RSN-2 structure were inspected for violations as part of the refinement protocol. It was observed that a small number of violations were evident at the end of the process but these were referenced back to the NOESY spectra and evaluated as real NOESY contacts. Calculation of the final round of structures produced a small number (6 out of 100) of structures that were of much higher energies than the others and were probably evidence of local energy minima traps along the simulated annealing pathway. These were replaced with six structures from a subsequent calculation of ten structures (from the same input) which were found to be of comparable energies to the other 94, and these 100 structures were used as templates for the *water_refine* protocol.

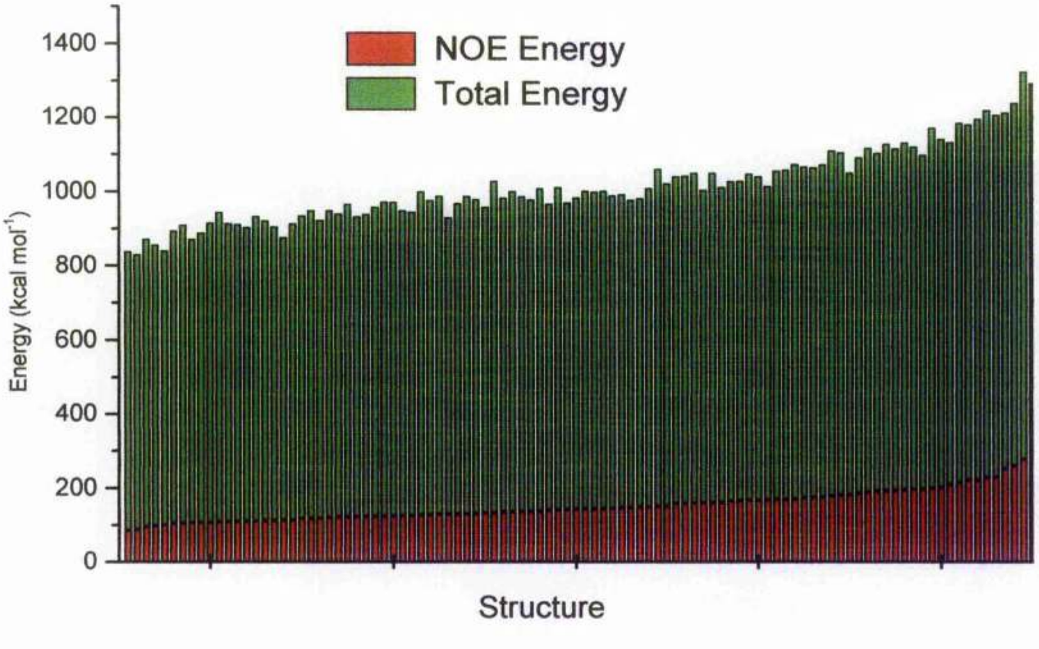


Figure 7.1. **Total and NOE energies of refined structures (prior to explicit solvent refinement). The total and NOE energies ranged from ~800-1000 and ~100-300 kcal mol^{-1}.**

After refinement in simulated water, the 100 calculated structures showed a slight increase in NOE violations (no change in H-bond analysis) relative to the structures derived without explicit solvent refinement, but the related energies were likely to be a result of the altered force-field rather than due to any structural errors. The concept of refinement in explicit solvent slightly adjusts the energy-minimised structure based on introduced external effects so a rise in NOE violations was expected. Overall energies of the structures were observed to reduce dramatically compared to those refined with undefined solvent, but this was probably mostly due to the favourable electrostatic attractions between charged residues on the protein surface and the solvent layer, and partly due to the use of a *Lennard-Jones* potential to describe the Van der Waals interactions. NOE energies were seen to increase slightly (~180-430 kcal mol^{-1}) due to the energy violations described.

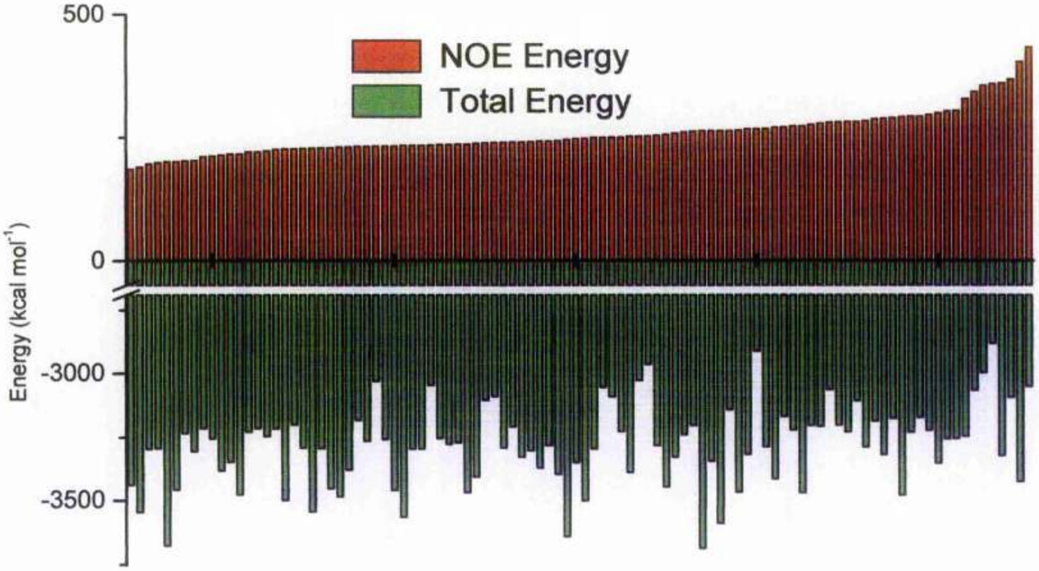


Figure 7.2. Total and NOE energies of calculated structures after refinement in specified water solvent.

All of these structures were analysed for associated NOE violations. Of the 2322 energy violations, 21 to 40 violations were observed in each of the structures (only reported if the difference between the inter-atomic distance in the calculated structure was > 0.3Å from the distance estimated from NOE intensities). Those violations that were consistently observed were investigated and found to be true contacts, almost always due to intra-residue NOE cross-relaxation. No violations from hydrogen-bond restraints were consistently observed.

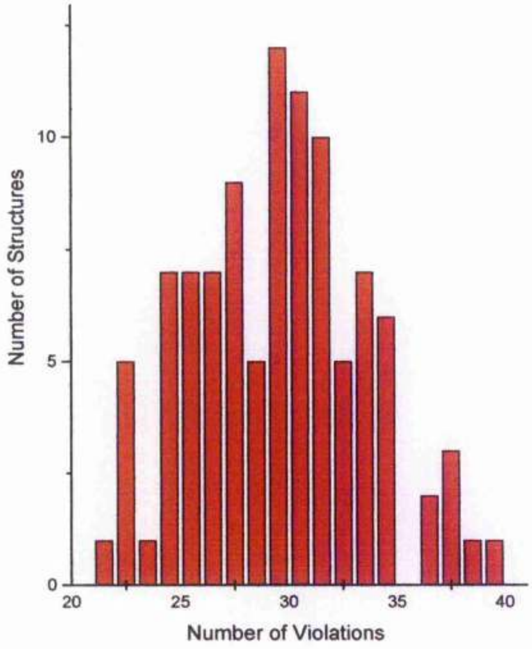


Figure 7.3. Histogram of the number of violations observed in the 100 structures in the final round of structure refinement.

Overall, the structures were judged to be acceptable in terms of *in silico* potential energies and associated violations. The close clustering of the energy data implies good precision but does not guarantee an accurate structure.

7.1.2 Structure Validation by Database Comparisons

A number of software packages are available for comparison of a calculated structure ensemble against database structures or idealised polypeptides and these are useful in assessing the quality of the structures and for diagnosing any errors. This process is now a routine supplement to NMR structure determination given the historical problems with NMR structures deposited in the Protein Data Bank (PDB) that have been documented[37]. More recently, advances in high-throughput structural genomics have highlighted these difficulties by the duplication of errors through comparative modelling[131]. In this work, structure validation was performed using the two most common platforms – WHATCHECK[132] and PROCHECK[133].

PROCHECK Results	
Most Favoured	66.1%
Allowed	26.8%
Generously Allowed	4.4%
Disallowed	2.8%
WHAT CHECK Z-scores	
2nd Generation Packing Quality	-0.881
Ramachandran Plot Appearance	-5.302
χ^1 - χ^2 Rotamer Normality	-3.427
Backbone Conformation	-9.316

Table 7.1. **Structure validation from two software packages. PROCHECK analysis was performed on an ensemble of the first sixty structures. WHAT CHECK analysis was performed on the structure with lowest total energy (number 22).**

The PROCHECK results illustrate an ensemble of structures that are reasonably compliant with preferred phi-psi angles. Examination of the residue specific results (Appendix D) confirm that the well defined structured areas of helix and strand have favourable dihedral angles and those residues which were found to create unfavourable angles were generally in more flexible areas of the structure and were observed inconsistently throughout the structure ensemble. These observations are unsurprising, given the nature of RSN-2 and all proteins, and the results were judged to be indicative of a largely accurate structure and gave further evidence for the dynamic nature of certain areas of the protein structure.

WHAT CHECK is a module of the WHAT IF[134] suite and provides a wealth of information based on statistical analysis of the given PDB input structure. The Z-score for each factor is a measure of the observed deviation in the test structure versus an average from a database of high-resolution crystal structures where a score of 0 is average and the value is expressed in units of standard deviations from the mean. A higher absolute score indicates a much larger distribution about the mean. In general, a score below -3 is considered poor and below -4 is considered bad. Examination of the results for RSN-2 show generally poor Z-scores although these results are comparable to those often observed in NMR derived structures[135].

Close analysis of the results show similar properties to the PROCHECK analysis, in that those residues producing errors and low Z-scores are those in the mobile, relatively unstructured regions. An area of particular note is the region between residues 72 and 80, which gave particularly poor backbone conformations. This may be indicative of a local structural inaccuracy brought about by lack of NOE contacts. These residues were found to produce broad and insensitive peaks, and the relaxation analysis suggested that there was a complex dynamic process occurring, possibly a combination of rapid solvent exchange and relatively slow intra-molecular movements. The per-residue Ramachandran plots confirm that the well-structured areas of the protein generally have allowed phi-psi angles whereas those parts of the structure that were only loosely defined from the NMR data had less favourable geometry. This was expected from the lack of NOE contacts and serves to illustrate those parts of the spectrum which are not well characterised from the experimental data. Overall, the global accuracy of the structure was shown to be accurate when treated as an ensemble.

WHAT CHECK RMS Z-scores	
Bond Lengths	0.323
Bond Angles	0.431
Omega Angle Restraints	0.998
Side Chain Planarity	0.424
Improper Dihedral Distribution	0.397
Inside/Outside Distribution	1.159

Table 7.2. WHAT CHECK RMS Z-scores for RSN-2 low-energy structure. These scores relate to a more local structure analysis than the Z-scores and are relative to a reference score of 1.

The WHAT CHECK analysis also produced RMS Z-scores which describe the local geometry of the protein structure. In these measures a score of 1 is the reference point and is equal to a Gaussian distribution about the mean; a score less than 1 represents a tighter distribution and a score of greater than 1 represents looser definition. For all appropriate measures (bond lengths and angles are only appropriate for x-ray crystal structures) all scores were reported to be within normal ranges.

7.1.3 Precision of Ensemble

The 100 structures produced in the final refinement in explicit solvent were all clustered tightly. The fact that the high energy structures were replaced in the previous iteration of the refinement makes this unsurprising but in any case, the original outliers represented only a small percentage which could reasonably have been expected to be carried through water refinement without further significant alterations.

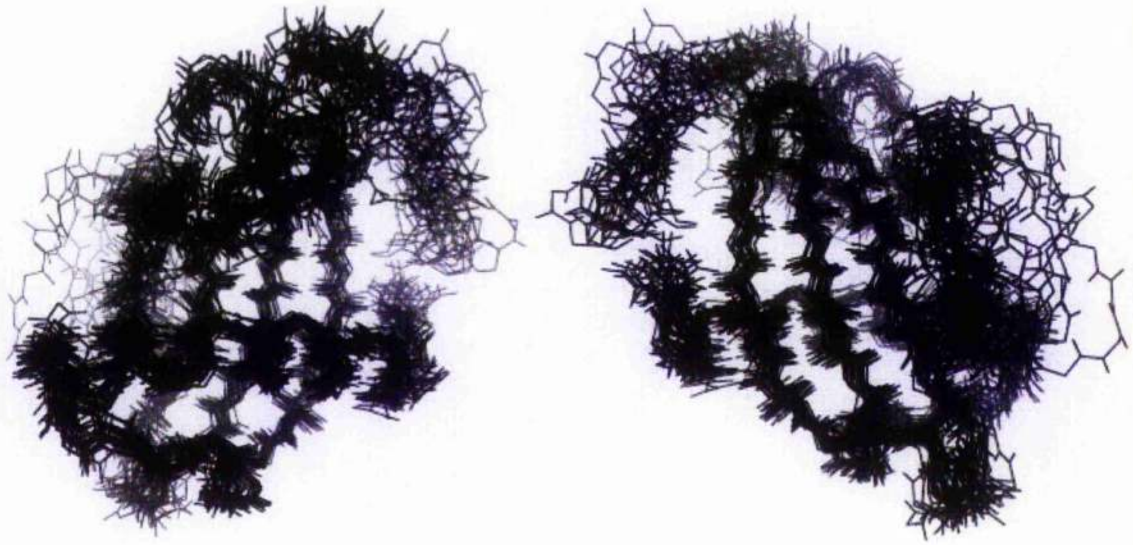


Figure 7.4. Twenty of the final 100 RSN-2 structures, superimposed on each other and viewed with 180° rotation.

As a final measure of variability within the final ensemble, the actual structures were reviewed for clustering of atomic co-ordinates. Structures were viewed using PyMol™ and superimposed using the *fit* function. Figure 7.4 shows the first 20 of the 100 calculated structures superimposed on each other and viewed on a 180° single-axis rotation relative to each other. For simplicity, only the polypeptide backbone was observed. Examination of the structures revealed that the core of the each protein, where secondary structure can clearly be viewed were very similar and superimpose almost exactly. Conversely to this, there are other regions where the backbones are dispersed quite widely through Cartesian space, with these regions being observed primarily at the termini and between regions of secondary structure as predicted from previous observations of protein dynamics.

For a quantitative comparison of all of the structures, the *UWMN*[136] program was used to generate RMSD (root mean square deviation) values for each atom and an overall RMSD for the structure. This program uses matrix algebra to calculate an average inter-atomic distance between all pairs of atoms and then transforms this back into three-dimensional space (with some manipulation to prevent structural irregularities) to create a truly average structure without lending any more bias to one structure over another i.e. *UnWeightedMeaN*. From the 100 structures in the final ensemble all atoms had RMSD values calculated excluding those associated with the first two residues (*UWMN* cannot accommodate negative residue identifiers), but it is reasonable to assume that these residues would possess a high degree of variability as they were not detected in NMR experiments and therefore had no associated restraints in structure calculations. Accordingly, the results for the first two residues would be largely meaningless anyway.

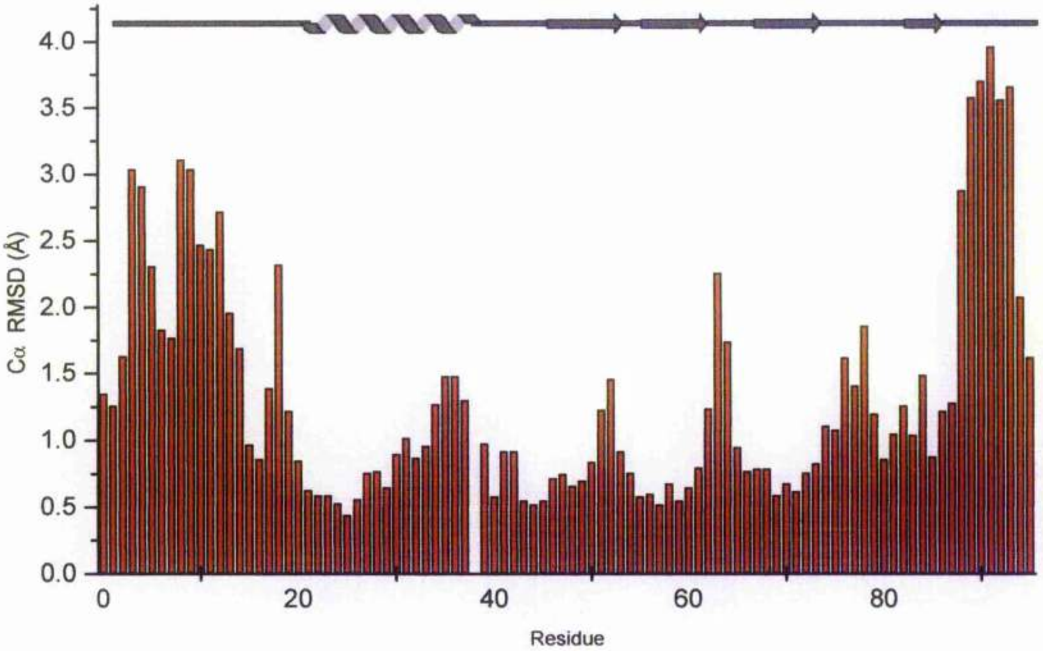


Figure 7.5. RMSD over 100 structures of the C_α atoms of each residue of RSN-2.

Per-residue RMSD values for each C_α showed a generally low level of variability but with a definite increase around the termini and at other less well defined areas of the protein structure. It was also found that a large degree of variability was seen for the poly-Asp_6 sequence which is likely to be a result of the lack of distance restraints, due to overlapping NMR peaks. A summary of the RMSD values across the final 100 structures is given in Table 7.3 and shows that the removal of the mobile N-terminus from consideration, improves the RMSD value considerably. The final overall RMSD of 2.1Å was judged to be acceptable given the number of dynamic atoms included in the comparison.

Atoms	Residues	RMSD (Å)
All	0 → 95	2.1
C_α	0 → 95	1.6
All	16 → 95	1.8
C_α	16 → 95	1.4

Table 7.3. Root mean square deviations for the 100 RSN-2 structures with and without the relatively unstructured N-terminus and with all atoms or just the C_α atoms.

A program, called *Theseus* has been released recently that aims to provide an accurate superposition and measure of precision of structural ensembles without making the same (unrealistic) assumptions made in LS (least squares) calculations[137,138]. The maximum likelihood (ML) technique assumes a correlation between atoms where they are covalently bound (whereas LS does not) and also places less weight on variable regions. As well as performing the superposition, it computes statistical comparison data to evaluate the adequacy of the superposition.

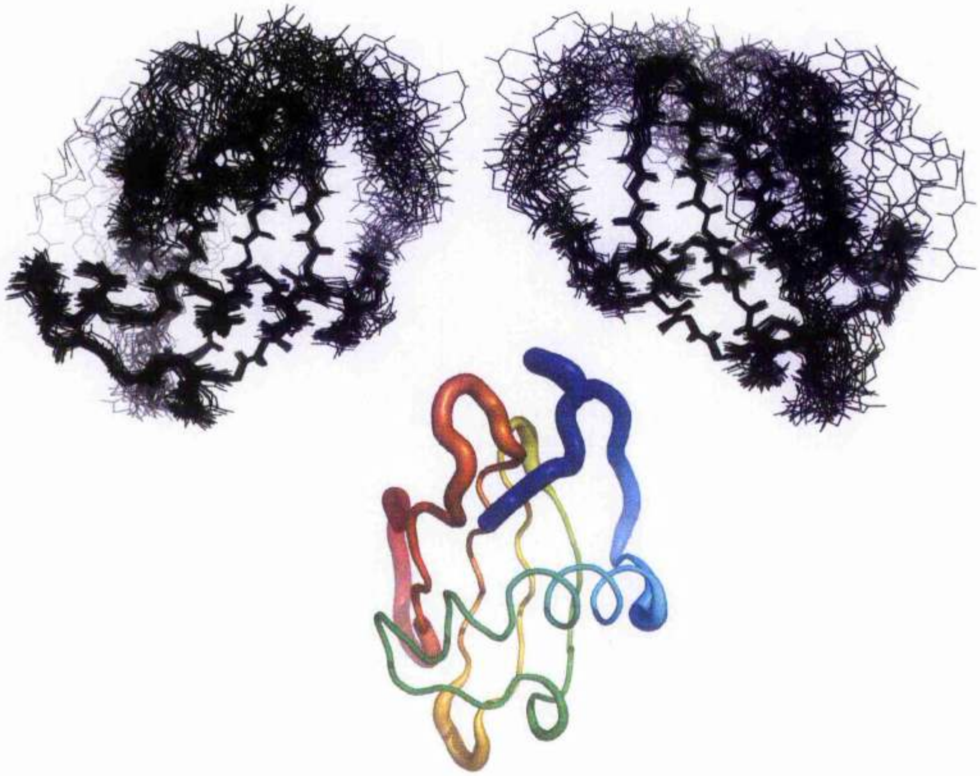


Figure 7.6. Superposition of 24 of the final RSN-2 structures using *Theseus*. The top two structures are the different views of the superposition, mimicking that in Figure 7.4. The bottom structure is the maximum likelihood mean, with higher variability indicated by thicker tubes.

Figure 7.6 shows that the ML superposition is very similar to that found using PyMol. The mean structure confirms that the maximum variability is observed at the termini and the turns. In the *Theseus* calculation, only backbone atoms were evaluated, and the statistical evaluation suggests that the ML superposition is a better model than the LS fit. The program was also used to provide a RMSD_{ML} of 0.36 Å versus RMSD_{LS} of 0.88 Å for the tightly defined regions of structure (the four strands and the helix).

Theseus Statistic	Value
LS RMSD	2.4 Å
Maximum Likelihood (σ)	0.27
AIC (Akaike Information Criterion)	-5579
Omnibus χ^2	2.1

Table 7.4. Statistical evaluation of the *Theseus* superposition.

7.2 Topology of the Calculated RSN-2 Solution Structure

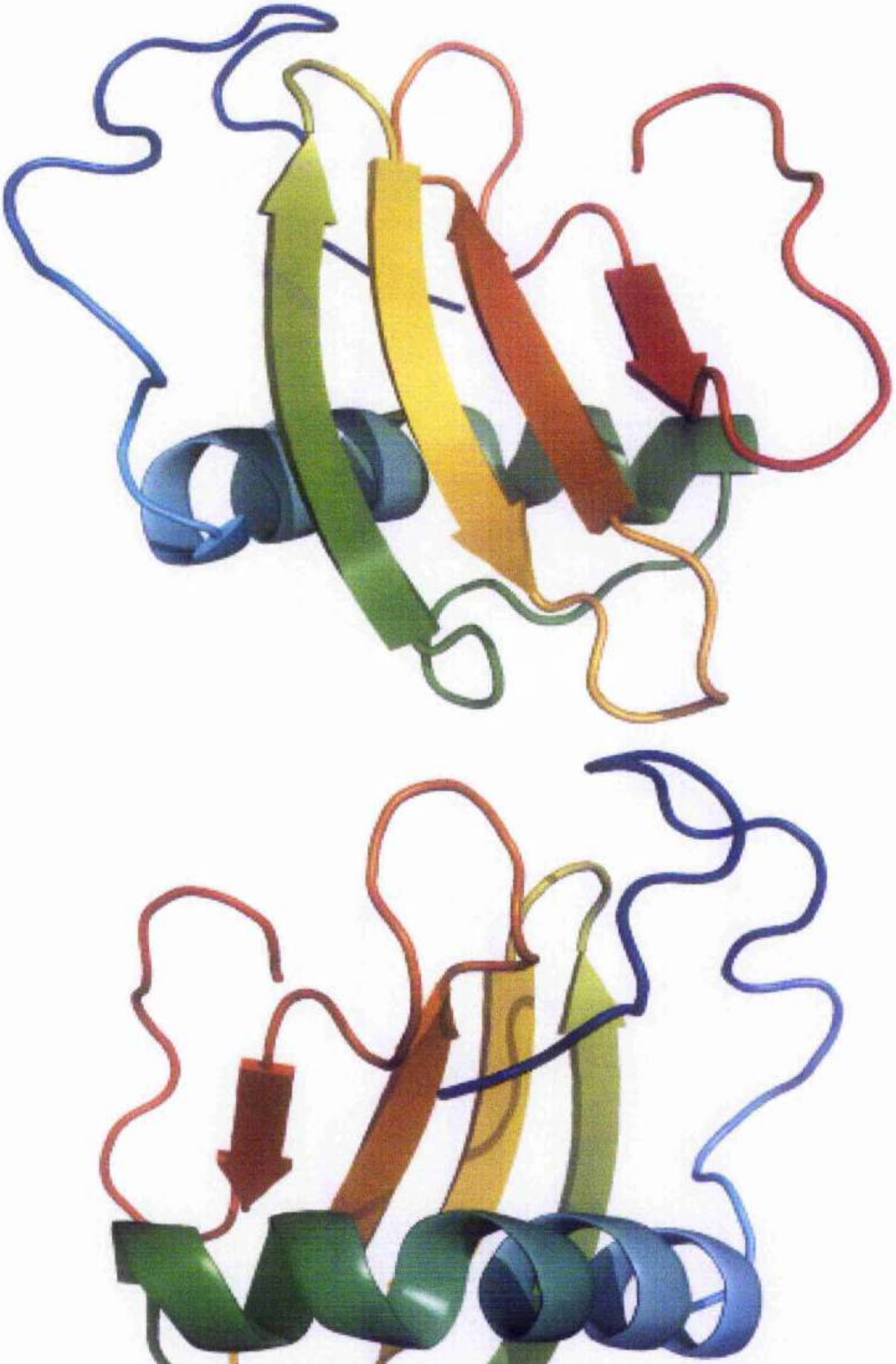


Figure 7.7. Cartoon diagrams of the average RSN-2 structure from the final ensemble of 100 structures. The sequence direction is from blue (N-terminus) to red (C-terminus). The bottom view is a 180° single-axis rotation of the top view.

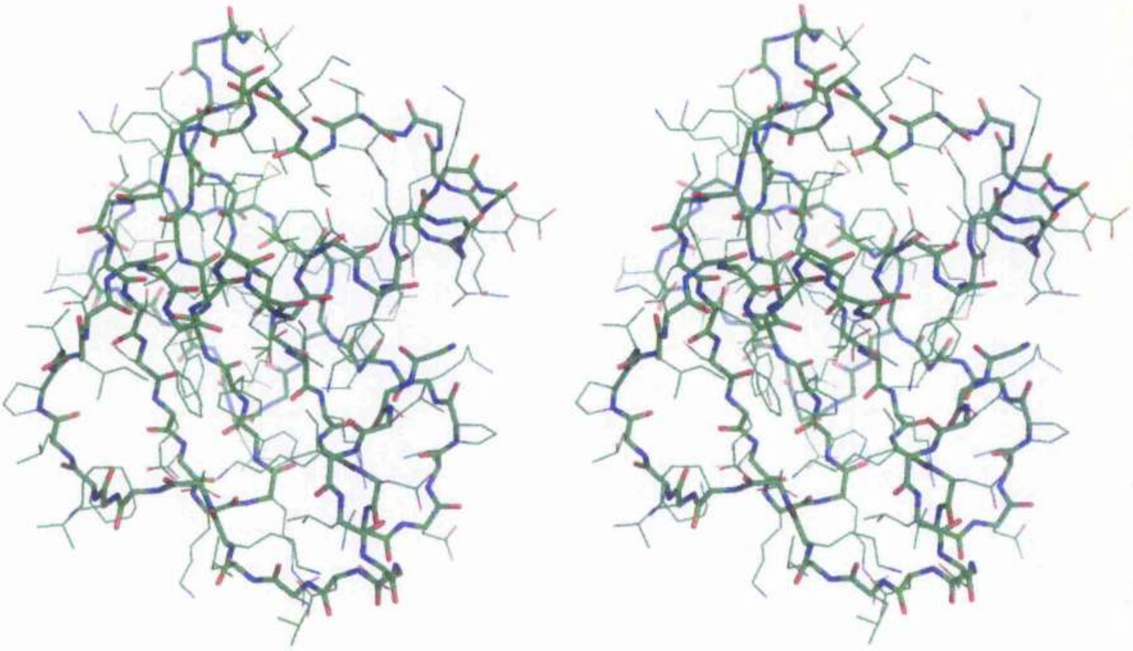


Figure 7.8. **Heavy atom stereo view of the average RSN-2 structure. The backbone is traced with a heavier line.**

Using *UWMN* an RSN-2 structure that was closest to the mean of the ensemble was determined, and was a good representation of the final ensemble. Scrutiny of this structure allows for a detailed description of the topology of the RSN-2 solution structure. The distinguishing feature of the protein is a core of four anti-parallel β-strands forming a twisted sheet with a kinked α-helix resting perpendicular across it. Analysis with SCOP[139] revealed that RSN-2 fits firmly within the Cystatin/Monellin superfamily and has a cystatin-like fold.

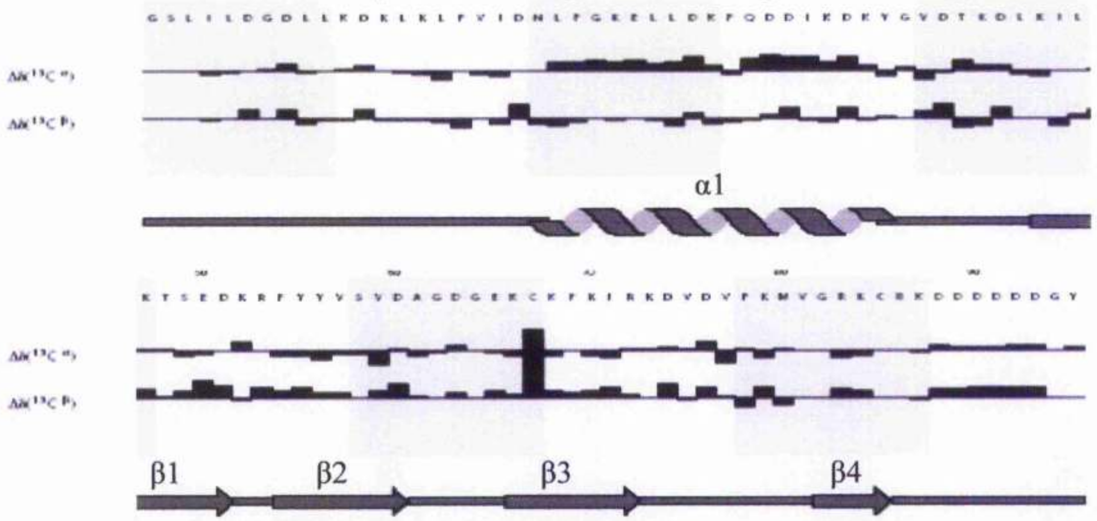


Figure 7.9. **Secondary structure diagram of the RSN-2 solution structure, displayed with amino acid sequence and differences in C_α and C_β chemical shifts relative to reference data. This figure was generated within *CCPN Analysis*.**

At the N-terminus, a long tail with a high degree of mobility (containing 3 residues as a cloning artefact) shows some attraction to β1. The single helix is initiated by Asn18 and appears to be a rigid structure of 18 residues and 5 complete turns. The linker arm between the helix and β1 also appears to be rigid and three β-strands pack well in a typical anti-parallel sheet fashion. Between β3 and β4, the turn lacks the definition seen elsewhere in the molecule. Relaxation data and RMSDs confirm this observation, as does the fact that only very few (and often weak) NMR signals were observed from the residues in this turn suggesting relatively rapid water exchange for the amide protons.

Residues Pro78 to Met80 create an unexpected deviation in the fourth strand which is highlighted in the poor WHAT CHECK backbone Z-score in this area. Although this is undoubtedly influenced by the mobility described previously, there is also the possibility that some related isomerisation occurs in Pro78. The applied forcefield assumes a trans conformation and the NOE crosspeaks for the surrounding amino acids suggest that this is the case, but there may be some slow isomerisation that was not observed due to the low intensity of the NMR peaks in this region. The C_β-C_γ chemical shifts for Pro78 is 5.27 ppm (32.4 – 27.7) and a 77ValH_α→Pro78H_δ NOE cross peak was observed which is indicative of a higher probability of a trans conformation[140]. The proximity of the Pro78C_α to the water peak prevented the investigation of any diagnostic minor cis-proline NOE crosspeaks. Other unusual phi-psi angles were observed further along β4 on Val81 and Cys85. Because of the lack of NMR crosspeaks observed, this area of the structure would benefit from further investigation by complimentary techniques.

The C-terminus consists of a loop followed by a poly-Asp_6 and a terminal tyrosine anchored against the outside of the β-sheet. The diffuse nature of the aspartic acid hexamer chain is most likely to be to the lack of resolution in the NMR data, but this very fact would suggest that there were very few interactions to be observed - there was little dispersion in the NMR peaks so the residues were unlikely to be involved in defined secondary structure elements. The terminal tyrosine contact was investigated further as its unusual nature was potentially of functional importance. The possibility of an error arising from ambiguous restraints was discounted as the ^{13}C-NOESY showed definitive contact between the tyrosine H_ε and LYS68 and LYS 70 sidechains.

Figure 7.9 summarises the secondary structure of RSN-2 and also shows the calculated per-residue sidechain C chemical shift δ values versus a database of reference chemical shifts. The only residue seen to deviate significantly from the mean database values was

Cys67 C_β which, upon examination, can be attributed to the atom lying adjacent to the electron aromatic ring on Phe69 which exerts a ring current shift on the Cys67 C_β.

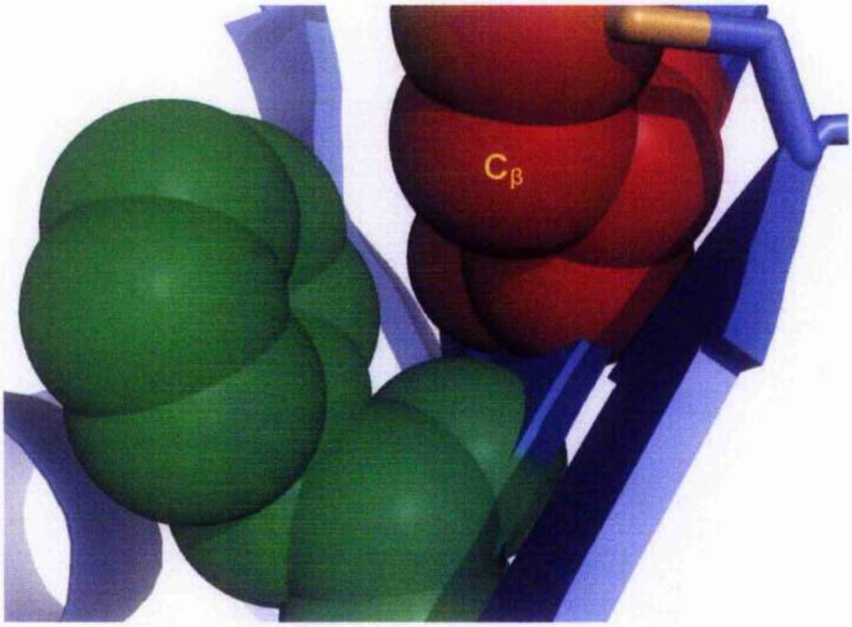


Figure 7.10. Atomic sphere representation illustrating the proximity of the CYS67 C_β (red) to the phenyl ring of PHE69.

Another observation made during NMR data analysis was that a small number of residues had small satellite peaks associated with them in the $^1H^{15}N$-HSQC. These were not the same as the degradation peaks observed from other residues (they were immediately apparent and did not change significantly over time), but could be seen in 3D experiments to connect to other amino acid residues, and were only distinguishable from the parent peak through decreased intensity (approximately 1:6) and slightly altered chemical shifts. With the availability of structural data these residues could be examined and were shown to have special significance. Those residues which could unambiguously be assigned as conforming to this regime were: Gly21, Lys22, Thr48, Glu50 and Val57. From Figure 7.11, it can be seen that these six residues are part of three distinct regions of secondary structure (α1, β1 and β2) which all come together within the protein fold to form a cleft between the helix and sheet. By mapping the heterogeneity in amide signals to this area, it is reasonable to assume that some structural heterogeneity follows.

From this data, it is unfeasible to predict the nature of the structural change but because it can be seen as slow on the NMR timescale, because it affects only a small number of residues in the proximity and because it is of significantly lower intensity (therefore population) than the main peaks, it is likely that this is only a short but concerted movement of the structure in this vicinity, resulting in the presence of two slightly different structures with different populations being observed in the NMR experiments. It is suggested here that the proximity of Pro14 to the region may be the reason for the structural changes, due to cis-trans isomerisation. Again, the $\delta(C_\delta\text{-}C_\beta)$ for Pro14 is 4.6 ppm and a Leu13$H_\alpha \rightarrow$Pro14H_δ was observed but the proximity of Pro14H_α to the water peak prevented the detection of any minor cis-isomer population.

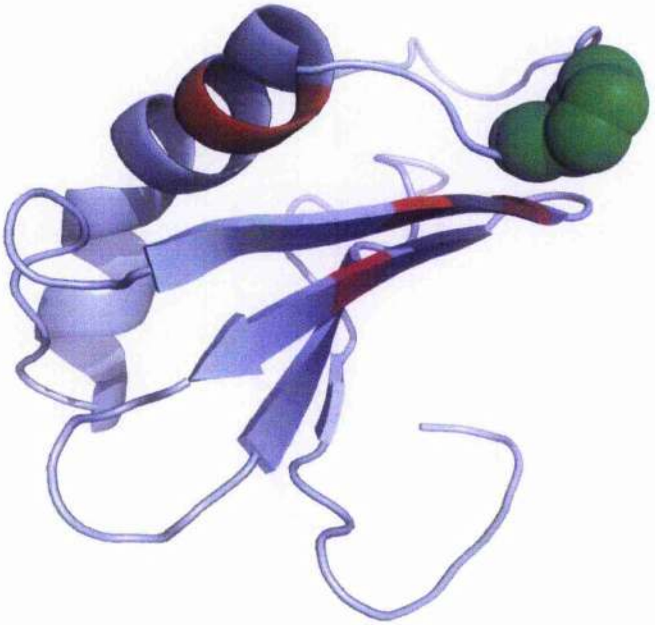


Figure 7.11. Cartoon diagram of RSN-2 showing residues in red where structural heterogeneity was inferred from the presence of amide satellite peaks. The Pro14 residue is indicated by green spheres.

7.3 Electrostatic Surface

The original hypothesis that RSN-2 is responsible for the surfactant nature of the foam fluid and that the highly charged sequence might be responsible for this, can be partly tested by examining the charge distribution about the calculated protein structure. Proof of the surfactant nature of recombinant RSN-2 is described elsewhere (in Figure 8.8). In this work, the electrostatic surface was calculated using APBS[141] (Adaptive Poisson-Boltzmann Solver) and visualised in PyMol. APBS numerically solves the Poisson-Boltzmann equation which describes electrostatic interactions in aqueous media.

Assessment of the solution structure showed that many of the charged hydrophilic residues were directed to the surface or solvent accessible areas of the structure, whereas the uncharged hydrophobic residues were directed into an unexposed core. The helix forms a biphasic structure, with hydrophilic and hydrophobic residues being largely divided on opposite sides of the helix, which is best visualised in a helix wheel, as in Figure 7.12.

Similarly, many of the hydrophobic residues on the β sheet point out onto the opposite surface from many of the hydrophilic residues to form a hydrophobic interior and a hydrophilic exterior.

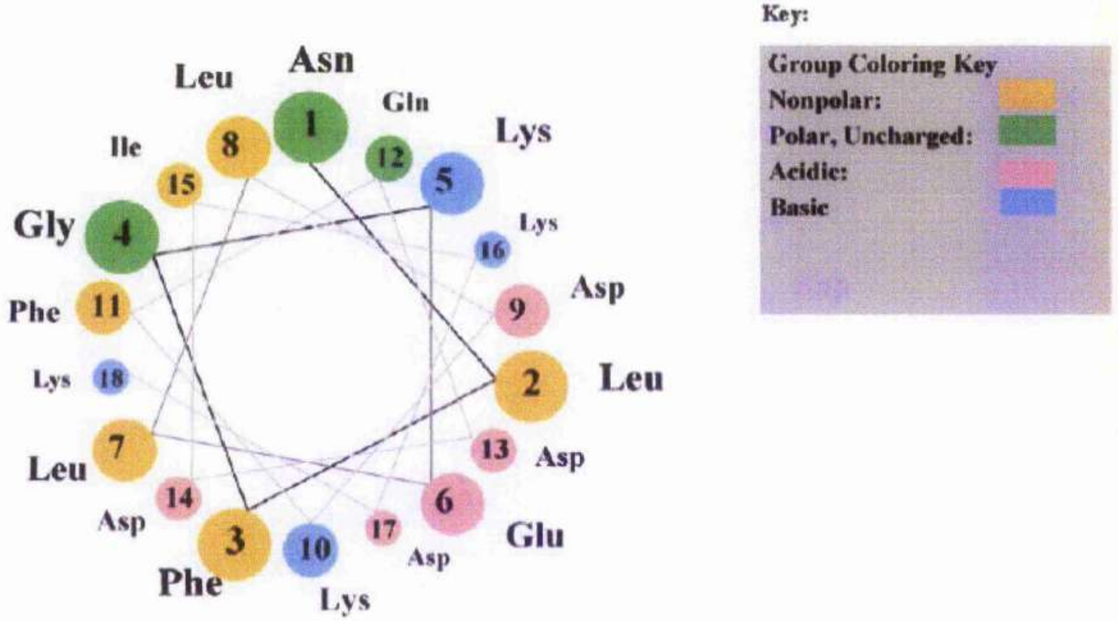


Figure 7.12. A helix wheel representation of RSN-2 α1 using a Java applet written by E. K. O'Neil and C. M. Grisham, *University of Virginia*[142].

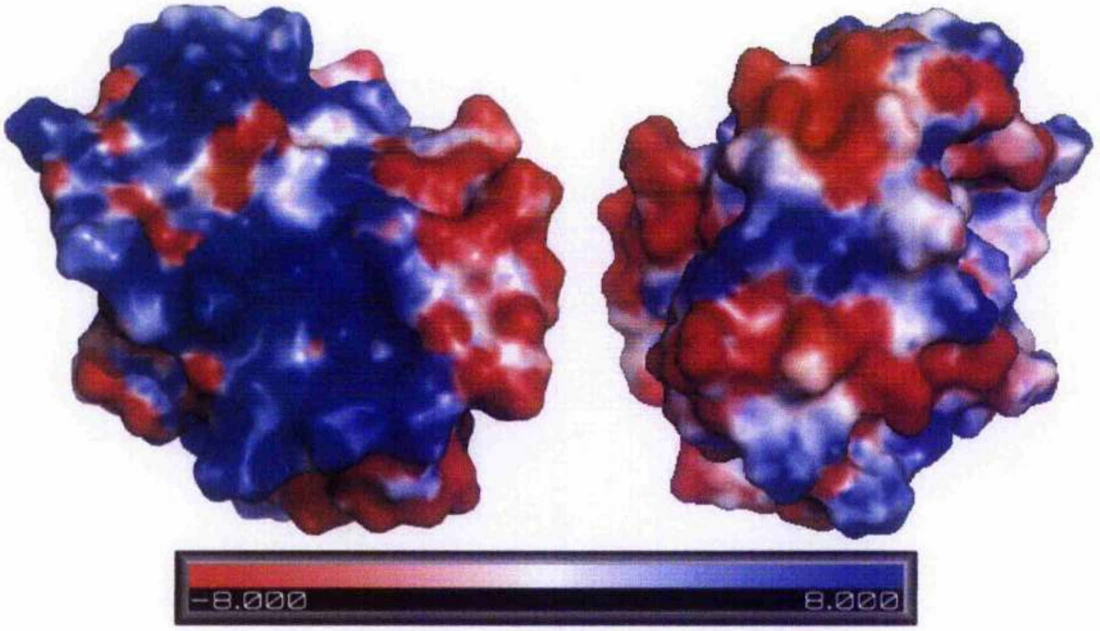


Figure 7.13. Electrostatic surface of RSN-2 where the left and right images mimic the top and bottom views from Figure 7.7. The scale is in units of *kT* where k is Boltzmann's constant and T is the system temperature in K (303).

The APBS solution for surface potential is shown in Figure 7.13. As expected, the highly charged primary sequence results in a structure with an extremely charged surface. Interesting regions include the highly positive potential of the underside of the β sheet and the negative ridge created by the poly-Asp_6 region. Overall, the charge is dissipated fairly evenly around the surface with no obviously uncharged regions. This observation undermines the hypothesis that the structure imitates a classical "charged head, un-charged tail" surfactant, but is perhaps unsurprising given the propensity for proteins to fold with minimum exposure of non-polar atoms in aqueous environment to form the most energy-favourable structure i.e. the hydrophobic effect.

Given the structural data described, it is difficult to speculate on a relationship between the three dimensional structure of RSN-2 and the surfactant activity. From Figure 8.10, it can be seen that the protein molecules have an extremely high affinity for monolayer formation at the air-water interface and yet the protein forms a tightly packed solution structure with high water solubility. Therefore, it seems likely that some other process is taking place that makes exposure to air much more energy favourable for the molecule than can be seen from the solution structure. With the knowledge that the majority of the hydrophobic residues are buried in the centre of the protein it can be seen that a partial unfolding might create an air-seeking patch and the amide satellite peaks discussed may be indicative of this. In addition, a potential hinge region between α1 and β1 showed a slightly depressed order parameter suggesting a relative increase in mobility, but no more so than the other regions without defined secondary structure (turns and termini).

Although these observations suggest that RSN-2 could be amenable to unfolding, there is no structural reason to suggest that this would be more driven in RSN-2 than any other protein with buried hydrophobic regions. Another possibility is that some other foam component may be interacting with RSN-2 to assist in surfactant activity, but the fact that recombinant RSN-2 displays such high activity would make this unlikely and would mean that the component would have to be common to all the buffers and water used in recombinant protein production, such as a trace metal. Despite some evidence that the poly-His_6 motif may form metal-binding pocket, no ion cavity is obvious from the solution structure.

7.4 Significance of the RSN-2 Structure

Despite the unique and highly unusual primary sequence of RSN-2, the derived solution structure represents a fold that has been seen in many structures previously. As already discussed, a SCOP analysis shows greatest similarity to the structure of Monellin[143,144], a sweet tasting protein isolated from the berries from *Dioscoreophyllum cumminsii.* Monellin is composed of two subunits, artificially joined together for structural studies to form a 95 residue polypeptide. Structurally, the protein consists of very similar secondary structure to RSN-2, whilst lacking some of the same unusual details at the termini and loops. A rudimentary taste experiment of 1 mg/mL recombinant RSN-2 in PBS did not suggest any sweet flavour, probably due to the lack of the Asp and Tyr residues conserved in Monellin and other sweet proteins.

Also within the same family of proteins are the cystatins. These are a range of proteins found in animals and plants where they inhibit cysteine proteinases. Due to some interest in human cystatin-C as a disease marker, the cystatins are covered widely in the literature with some recent reviews being especially comprehensive[145,146,147,148,149,150].

Cystatins have been divided into three broad sub-families. The Type 1 cystatins are of similar size to RSN-2 at ~100 amino acids and have been found in the cells of humans and many other mammals. Members of this family type lack any disulphide bridge, are non-glycosylated and have a number of conserved residues. They are often termed *Stefins* and solved structures include human cystatin A[151, 152] and B[153]. Structurally, RSN-2 and Stefin A are similar with a helix running approximately perpendicular to a right-hand twisted sheet. Some confusion has arisen in the past, as to the exact nature of the helix where the slight curvature seen in one structure was seen as a distinct kink in another. Indeed, the helix was seen as two short lengths rather than one continuous length in the latter structure. Further NMR analysis[154] of Stefin A and the P25S mutant has revealed that the structure is probably more of a curved helix than a kinked one, which compares well to that derived for RSN-2.

The structural rationale for the inhibitory action is well understood and is related to the conserved regions in three areas of the protein. The three areas responsible, are the loop between β1 and β2, the loop between β3 and β4 and a short strand near the N-terminus. These regions form the sharp edge of a wedge shape that insert and bind irreversibly to papain-like cysteine proteinases[155]. The C-terminus of Stefin A hooks back against the underside of the β-sheet in the same fashion as in RSN-2 but no mention is made in the literature of any functional significance.

Other studies of Stefin A have included the observation of domain-swapped dimers[156], formed from the extension of β1 and thought to be a kinetic trap in the folding pathway.

Dimerisation can be induced through heat, chemical denaturant or point mutations. No amyloid formation was observed in this work, nor was it observed in the native foam[63]. Cystatin B was found to bind copper tightly although the functionality of this remains unclear[157].

Type 2 cystatins are slightly larger than type 1, with ~120 amino acids, and are usually secreted and released from cells via a 20-30 residue signal peptide. They contain two conserved disulphide bridges and are sometimes phosphorylated. Structural examples of type 2 cystatins are the Chicken Egg White (CEW) cystatin (which is easily available and therefore well-studied[158, 159]) and human cystatin D[160]. These structures are similar to type 1 cystatins but with curtailed termini, elongated β1 and β2 and a much longer loop between β2 and β3. For CEW, this loop was observed to contain and extra β-strand in the X-ray structure compared with the NMR derived structure and is now thought to be an artefact from crystal packing. Another type 2 cystatin of interest is human cystatin F which has recently been shown to form a disulphide-mediated dimer[161]. Type 3 cystatins are much larger, multi-domain proteins, mainly consisting of the kininogen family. These proteins are glycosylated and have multiple disulphide bridges with regions of sequence similar to the type 2 cystatins. They also contain the common cystatin structural motifs.

Other cystatins that don't fit into these classifications include those plant-derived examples named *phytostatins*. As yet, the only structure from this family is Oryzacystatin-I, from the rice *Oryza sativa* L. japonica and it is functional as a potent inhibitor[162]. This protein is structurally similar to the type 2 cystatins but lacks the characteristic disulphide bonds. As with many of the type 1 and type 2 cystatins, Oryzacystatin contains a fifth β-strand at the N-terminus which is absent in RSN-2 although the backbone does follow a similar spatial trajectory. Although no frog proteins have been found to be cystatins, a protein isolated from the African Puff Adder has been shown to be a strong inhibitor[163].

Although lacking cystatin activity (see Chapter 9), some other proteins such as RSN-2 and Monellin have been found to adapt the same fold, defined by the MEROPS[164] database as family I25. Both X-ray[165] and NMR[166] structures are available for the precursor to Protegrin-3, a 101 residue conserved sequence which is shared amongst a large number of mammalian antimicrobial peptides. Like RSN-2, the protegrin-3 precursor has been shown to have very little sequence identity to the cystatins and shows no activity as an inhibitor of papain, but shares in the same overall fold. Similarly, the structure for mouse Latexin, the only known mammalian carboxypeptidase inhibitor, has been solved[167] and shown to consist of two cystatin-like domains linked through a helix forming a pseudo 2-fold axis of symmetry.

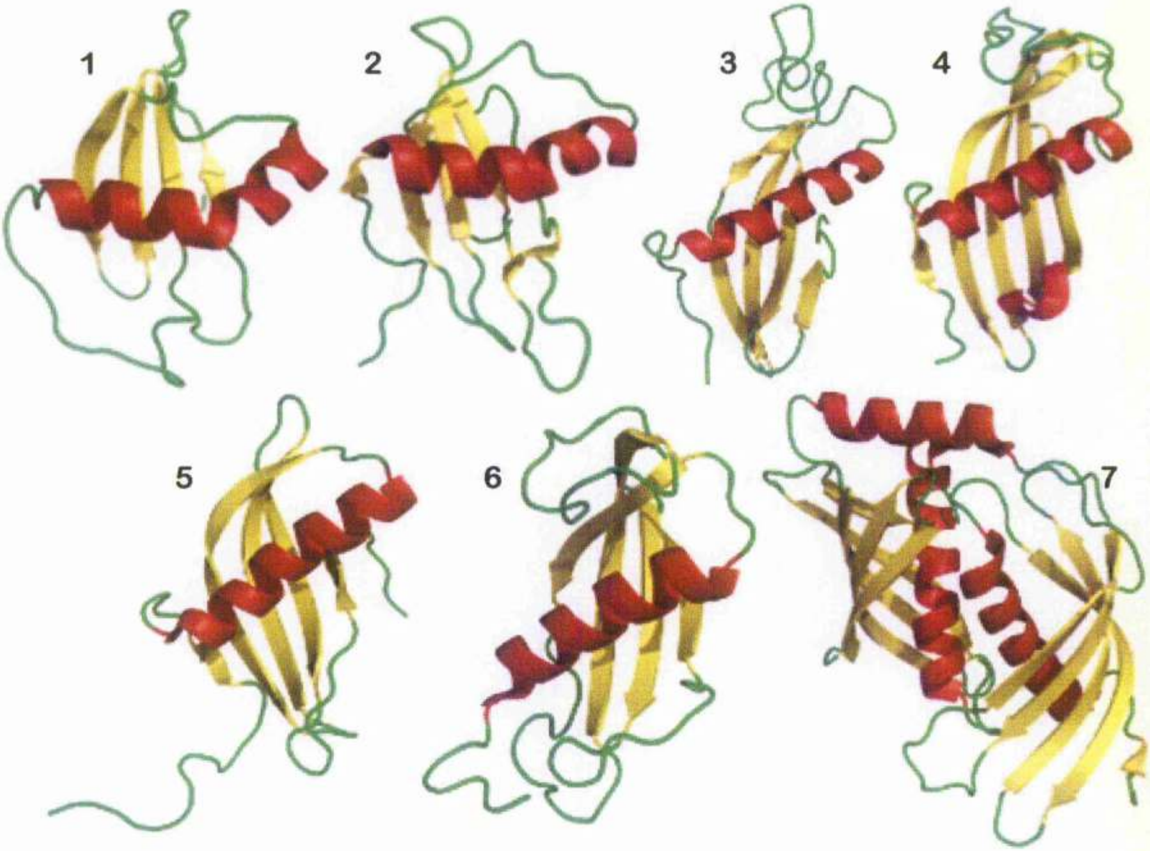


Figure 7.14. Comparison of RSN-2 with structural family members from the PDB[168]. 1) UWMN average of 100 RSN-2 structures compared against 2) Stefin A (1GD3), 3) Chicken egg white cystatin (1A67), 4) Cystatin D (1RN7), 5) Oryzacystatin (1EQK), 6) Protegrin precursor (1N5H) and 7) Mouse latexin (1WNH).

7.5 Summary

Examination of the calculated RSN-2 structures suggested that the highly structured core of the molecule has been well described and conforms well to known tests for precision and accuracy. The less structured areas show much less structural homogeneity and give correspondingly low scores when analysed by structure integrity tools. These low scores are expected for less structured mobile regions and, while they are good indicators for regions where care must be taken in drawing conclusions, they do not necessarily reflect errors in the calculated structures.

The solution structure consists of a compact globular fold, composed of a single helix over a four-stranded sheet. Both termini were observed to be mobile but the carboxy terminus appeared to be anchored, with the poly-Asp$_6$ sequence forming a solvent-exposed ridge. The length of peptide between the helix and the first strand was well-defined but had increased mobility and it is proposed that it may be a hinge point for surface unfolding. In addition, the third turn has some unusual structural features and may be a combination of slow proline isomerisation and faster motions. Amide satellite peaks in the NMR spectra could be mapped back to the solution structure to reveal a region where slow conformational exchange was observed, and suggested some alternative, less populated local structure at an exterior cleft, most probably due to cis-trans isomerisation of the spatially close proline residue.

The calculated structure of RSN-2 shows that the highly unusual primary sequence adopts a global fold that has been seen in a range of proteins from a range of different species, classes and even kingdoms. RSN-2 is another addition to this protein superfamily and is the first amphibian representative. In terms of evolutionary biology, the growth of this collection of proteins, many of which are related to antimicrobial activity, suggest that the particular fold is a significantly mature protein which has progressed through numerous evolutionary steps creating altered functionality despite maintaining a broadly similar structure. By adding surfactant activity to the list of functions, RSN-2 widens the potential scope of the cystatin family further.

8 Biophysical Studies

8.1 Solution Analysis

8.1.1 Ultraviolet-Visible Spectroscopy

Ultraviolet-visible (UV-Vis) spectroscopy was used extensively in this work, primarily as a convenient analytical tool. In proteins, the peptide bond absorbs electromagnetic radiation at a wavelength maximum of 214 nm but this phenomenon cannot discriminate between different proteins and is at a position in the light spectrum where many potential impurities could interfere. Instead, the absorption of light ~280nm by aromatic sidechains and cysteine can be used to quantify proteins in solution. In the case of RSN-2, it can also be used to distinguish the target from other proteins.

Of the three aromatic sidechains, the most useful is tryptophan because it absorbs very strongly with a local maximum at 280 nm. Phenylalanine absorbs much less strongly with a λ_{max} at 257 nm and accordingly, it doesn't contribute strongly to measurements at 280 nm. Tyrosine has an intermediate absorbance at 280 nm with a λ_{max} at 276 nm. Free cysteine does not absorb at 280nm, but when oxidised as cystine, the disulphide form does absorb weakly. In RSN-2, there are no tryptophans present, so when measuring absorbance at 280nm the only contributing components were tyrosine and possibly the two cysteines. Correspondingly the spectrum appears slightly different from tryptophan-containing proteins, with a lower λ_{max} at approximately 278 nm. Also, unlike other strongly absorbing proteins, dilutions were rarely necessary. It has been shown[169] that examination of the primary sequence and knowledge of the appropriate molar extinction coefficients can allow predictions of the molar absorption coefficients of proteins by a simple additive formula with precision of ±5%. For RSN-2, the extinction coefficient value (ε) was calculated[170] as 5960 (mol dm^{-3})$^{-1}$ cm^{-1} when it is assumed that the two cysteine residues are in the oxidised form.

Solutions were diluted if appropriate, to avoid errors from high concentrations and the stray light effect. Samples were always blanked against the pure buffer, either before being measured (single beam instrument) or concurrently (dual beam instrument). Quartz cuvettes with a 1 cm path length were used and the absorbance at 280 nm was recorded on one of a number of Shimadzu spectrophotometers available. Using the Beer-Lambert law and the calculated extinction coefficient, concentrations could be estimated with some confidence although a larger error is sometimes associated with proteins containing no tryptophans[170].

8.1.2 Circular Dichroism

Circular dichroism (CD) is a spectroscopic technique that takes advantage of the chiral nature of proteins. Optically active molecules respond differently to the two (left and right) circular components of plane polarised light and this difference is measured in CD. Because proteins are chiral at both the amino acid and the secondary structure levels, CD can be used to estimate the secondary structure content of proteins by comparison of an experimental spectrum to a database of spectra from known structures.

In this work far-UV CD was used to estimate secondary structure in uncleaved (fusion tag still attached) and cleaved recombinant RSN-2. These experiments were performed with the assistance of Dr. S. Kelly and Mr. T. Jess at the BBSCR Scottish Circular Dichroism Facility at the University of Glasgow. Samples were supplied ~1 mg/mL in 20mM sodium phosphate buffer at pH 7.0 and the instrument was a Jasco J-810 with a 0.02cm pathlength. Spectra were averaged over 8 scans.

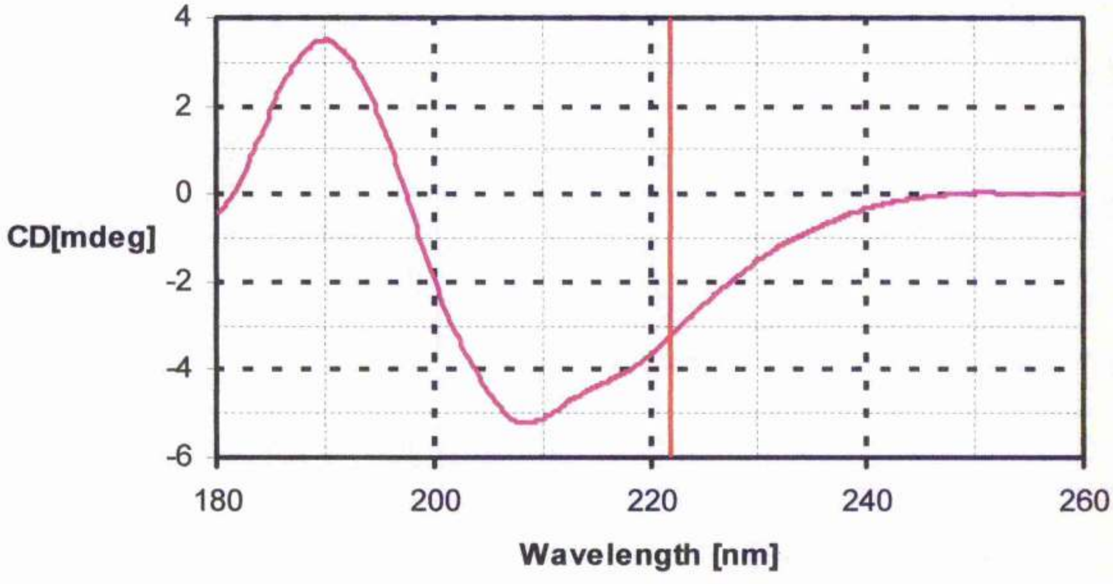


Figure 8.1. CD spectrum of recombinant RSN-2 with fusion tag.

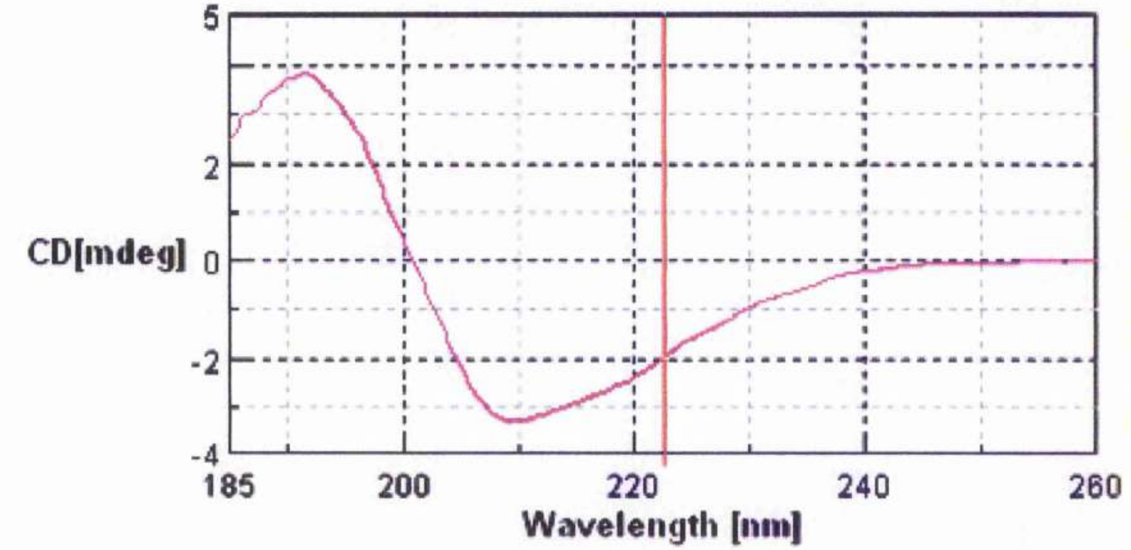


Figure 8.2. CD spectrum of recombinant RSN-2 without fusion tag.

Comparison of Figure 8.1 and Figure 8.2 suggests that there were very few conformational changes observed from removal of the fusion tag as both spectra appeared very similar in profile. Absolute magnitudes of the peaks changed slightly between samples but this was likely to be due to differences in sample concentrations

The spectrum of the uncleaved protein was analysed on the *Dichroweb*[171, 172] server against two databases (Contin[173] and Selcon-3[174]) for secondary structure content, and the results showed a minor level of helix content and a larger strand content which conforms well to the derived solution structure. Indeed, the results for the two different RSN-2 forms are so close that they may point towards the fusion tag adopting some secondary structure, or some of the less ordered parts of the derived solution structure actually forming a more ordered structure. These observations are supported by a $^1H^1H$-TOCSY NMR experiment performed on uncleaved RSN-2 (performed by Dr. Brian Smith, data not shown).

	Contin	Selcon 3	Solution Structure
Helix	11%	17%	14%
Strand	36%	35%	35%
Turn	21%	19%	19%
Unordered	32%	29%	32%

Table 8.1. Summary of results from RSN-2 secondary structure analysis by far-UV CD and the NMR-derived solution structure.

8.1.3 Fluorescence

Fluorescence is the inelastic luminescent phenomenon whereby a molecule in an excited state decays to the ground state with the emission of a photon. Typically, light in the ultraviolet or visible blue region is used to excite the molecule to a higher electronic state, and some proportion (described by the quantum yield) of these states will decay with the emission of a photon in the visible range. The energy difference is normally lost through vibrations/rotations etc. and produces heat rather than light.

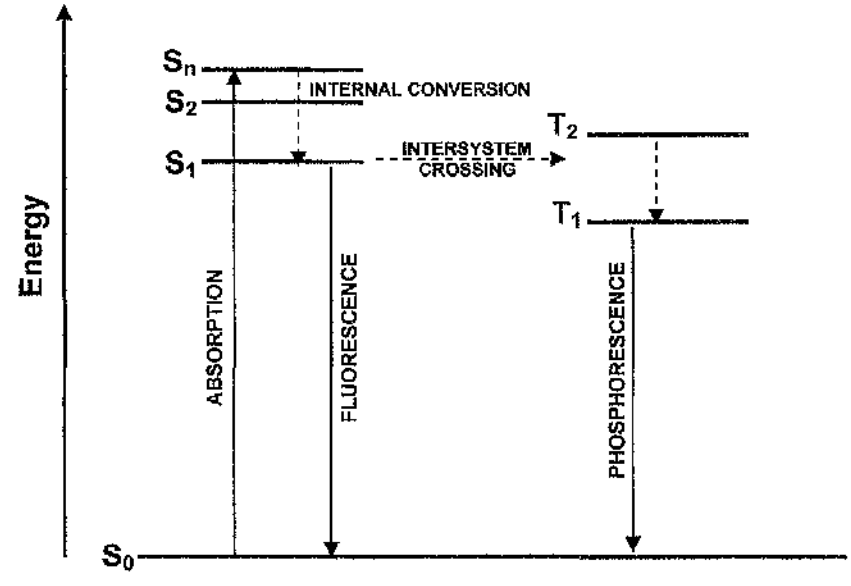


Figure 8.3. A simplified Jablonski diagram, depicting fluorescence from an energy level model.

Because of the lack of tryptophan residues in RSN-2, intrinsic fluorescence was unlikely to be informative. However, by adding a fluorescent probe, the state of the protein in solution could be further investigated. 1-anilinonaphthalene-8-sulphonic acid (ANS) is a dye that binds to hydrophobic patches on proteins with a characteristic intensity and wavelength shift in the emission spectra[175]. This phenomenon has been used for studying protein folding in solution, with the fluorescence changes being descriptive of the changes in exposure of the hydrophobic core[176]. The emission intensity of ANS (10 μM) solution was scanned with and without RSN-2 (approximately 5 μM) using standard 1 cm quartz fluorescence cuvettes in a Spex Fluoromax II spectrofluorometer. The excitation wavelength was 390 nm and slit widths were 5 nm.

Figure 8.4 demonstrates ANS binding to RSN-2 suggesting that there are some significantly exposed or accessible hydrophobic areas despite the solution structure suggesting a highly charged interface. ANS binding is probably a result of solvent accessibility at the protein core (known from deuterium exchange NMR experiments) but it may also be indicative of some dynamic property which means that the protein can adopt some other transient conformation in solution. It should be noted that this is not an entirely sturdy result as the ANS may not exclusively bind to hydrophobic patches and some studies have shown that protein structures can change to accommodate the dye[177].

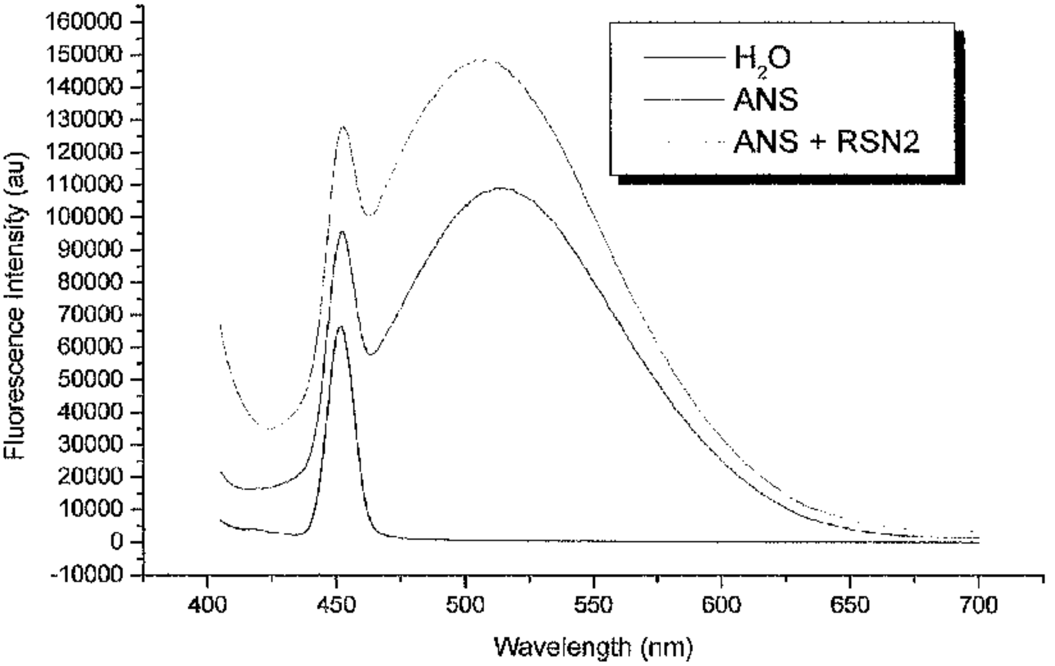


Figure 8.4. Fluorescence emission spectra of ANS with and without RSN-2. The increase in intensity and blue-shift indicates ANS binding (λ_{excite} = 390 nm).

8.1.4 Differential Scanning Calorimetry

A fundamental property of globular proteins is their propensity to denature at elevated temperatures. Differential Scanning Calorimetry (DSC) is an analytical technique that measures the energy employed by the protein to unfold[178]. The instrumentation consists of two cells containing a sample and reference solution, and a number of heaters with associated electronics. The cells are gradually heated and the temperature is monitored, and independently compensated for, by feedback heaters to maintain the same temperature in each cell. When thermal unfolding occurs in the protein, some energy is used and the rate of temperature increase slows relative to the reference cell. By measuring the extra heat required to keep temperatures equal in both cells the endothermic properties of the protein-specific transition can be examined.

Both cleaved and uncleaved recombinant RSN-2 were analysed by DSC at approximately 5 μM referenced against phosphate buffer at pH 7 which was also the solute. In each case the cells were heated to 110°C and then allowed to cool before being scanned again. Both samples showed a similar profile with the re-scan displaying significant residual activity which suggests that the thermal denaturing does not necessarily destroy ability for the protein to re-fold i.e. a large (>50%) of the protein re-folds into the native (or very similar) conformation when cooled back down. The temperature where the transition took place, T_m, was higher in the cleaved protein compared to the uncleaved protein (~81°C versus ~73°C), which adds weight to the premise that the cleaved protein is a more stable structure and that the fusion tag is likely to be a less ordered fold.

Interestingly, this transition temperature is strikingly similar to that of the ubiquitous human cystatin C (82°C), lower than that of cystatin A (95°C) and CEW cystatin (>115°C), and higher than cystatin B (66°C) all of which are structural analogues[179]. Whilst it might be expected that the inclusion or otherwise of disulphide bridges would be the primary reason for altering resistance to thermal unfolding from these structurally similar proteins, the relatively high T_m for cystatin A contradicts this pattern. The nature of the curve is a smooth transition suggesting that RSN-2 goes through a simple two-state (folded to unfolded) transition without any oligimerisation or amyloidogenesis seen in other similar structures. This is supported by the similarities in the magnitude of the calorimetric (ΔH) and van't Hoff (ΔH_v) estimated from the DSC data.

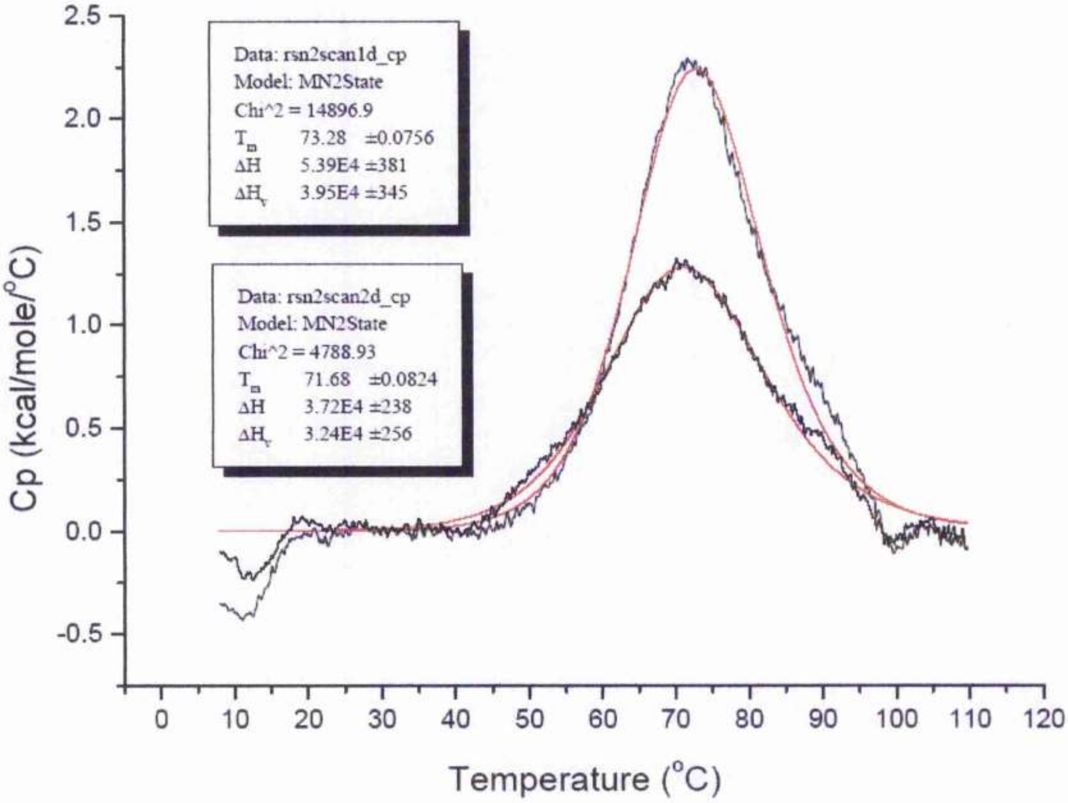


Figure 8.5. **DSC thermal unfolding transition and re-scan (lower trace) of RSN-2 prior to fusion tag cleavage.**

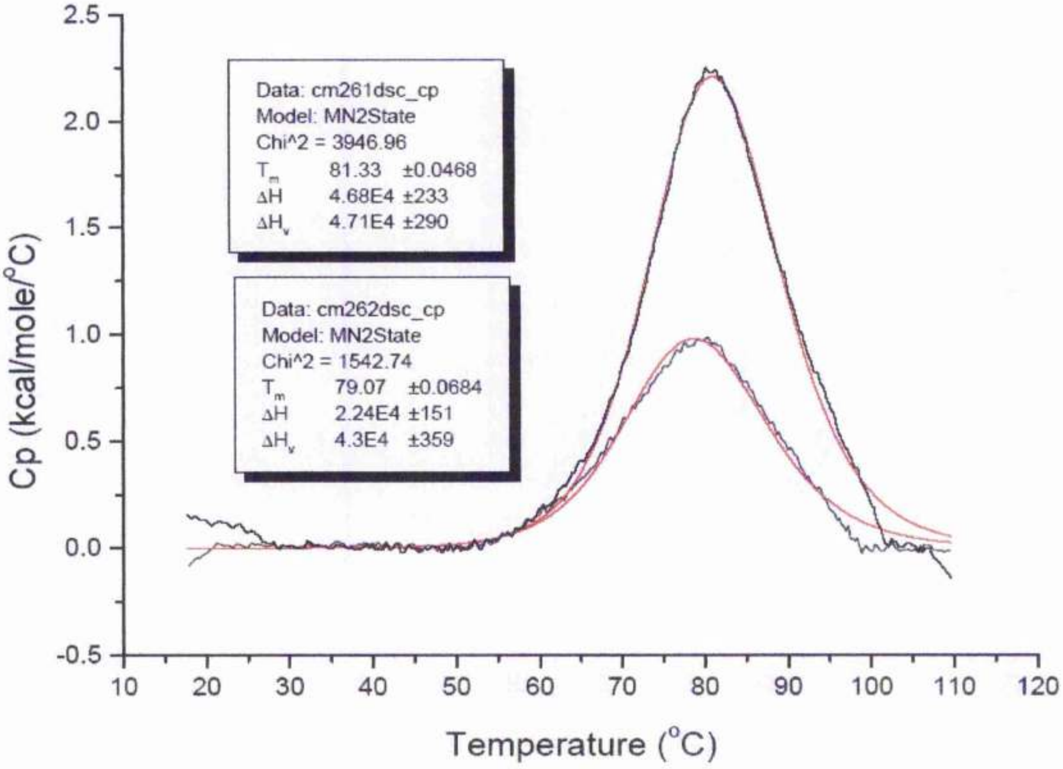


Figure 8.6. **DSC thermal unfolding transition and re-scan (bottom) of RSN-2 following fusion tag cleavage.**

8.2 Interfacial Analysis

8.2.1 Surfactant Activity

Surface tension in a liquid describes the physical force that prevents a liquid from spreading. This force is caused by the relative stability of molecules or atoms in the solution phase relative to the air-water interface. This can be summarised as the difference between the adhesive and cohesive forces at a gas-liquid interface. In aqueous solutions, surface tension is provided by the preference of water molecules to remain in a fully hydrated, hydrogen-bonded network rather than have contact with the air interface where the molecule cannot take part in any bonding.

Surface tension is affected by chemical composition and also temperature – surface tension decreases with increasing temperature. A higher intermolecular affinity in solution will result in a higher surface tension, therefore mercury which has strong metallic bonding has extremely high surface tension of 435.5 mNm^{-1}, ethanol which is a relatively non-polar molecule, has a surface tension of 22.1 mNm^{-1} (at 20°C) and water is between these, at 72.8 mNm^{-1} [180].

Surfactants (a name devised from Surface Active Agents) generally have the effect of decreasing surface tension, except for some ionic salts that have been observed to increase surface tension for reasons that are not completely understood[181]. Because a decrease in surface tension also decreases free energy of the surface, molecules that lower surface tension preferentially concentrate at the surface, and *vice versa* for molecules that increase surface tension.

In this work, a Langmuir trough was used for the study of Langmuir Blodgett films. This term describes interfacial monolayers and derives from the work of Irving Langmuir and Katherine Blodgett, who investigated the properties of surfaces and extended the work on thermodynamics of three dimensional systems to the two dimensional counterparts seen at surfaces[182]. This is work that was awarded the Nobel prize in chemistry for Langmuir in 1932. As an aside, this was a formalisation of the observations originally made by Benjamin Franklin, who observed the calming influence of a teaspoon full of oil on half an acre of water on a pond in Clapham common, London, and was reported in 1773[183].

The bulk of the studies performed in this area relate to the study of oils or lipids at the air water interface. These solutions oppose each other very strongly and form immiscible biphasic mixtures, so surface layers are easily achieved and studied. After forming an oil-on-water system, the surface is swept by barriers, thereby compressing the oil molecules arranged at the surface layer and, by monitoring physical and chemical parameters of the surface, various kinds of structural information can be inferred.

In its simplest form, the apparatus can be used to monitor the change in surface pressure of a hydrophobic surface over an aqueous solution. Surface pressure (π) is defined as the difference between the surface tension of pure water and the measured surface tension ($\pi = \gamma_0 - \gamma$). As the surface layer is compressed, the hydrophobic surface layer is also compressed, and the surface area over which it acts is decreased. By measuring the increase in surface pressure with a decrease in surface area, pressure area isotherms can be produced which are analogous to pressure volume isotherms for gaseous systems.

P. pustulosus foam fluid has been shown[62] to be surface active as would be anticipated from a foam-derived solution. At the simplest level a decrease in surface tension has the observable result of increasing the wetting ability of aqueous solutions. Figure 8.7 shows a photograph of a 0.9 mg/mL recombinant RSN-2 solution in 50 mM sodium phosphate buffer at pH 7.0 and the pure buffer for comparison. The solutions were 50 μL drops on Nescofilm®, a highly hydrophobic surface, mounted on a glass slide. In a qualitative analysis RSN-2 appeared to cause a decrease in contact angle and an increase in spreading and wettability of the solution, and therefore must have decreased surface tension.

Following the completion of this experiment, it was noted that the RSN-2 solution appeared to evaporate at a faster rate than the non-proteinaceous solutions which was possibly just an effect of the increased surface area as a result of the spreading. It was also noted that the RSN-2 solution left a significant residue on the film surface after being poured off for disposal. Further investigation (by Prof A. Cooper, personal communication) showed that the residue bound a much higher level of Coomassie blue stain than other sites where control (non-RSN) protein solutions had been sited. These observations lend weight to the hypothesis that RSN-2 binds strongly to hydrophobic surfaces or interfaces whilst maintaining good solubility in aqueous solutions.

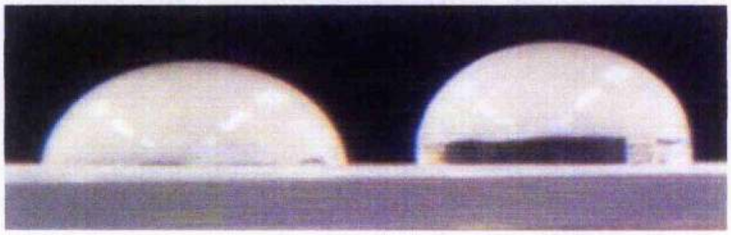


Figure 8.7. Photograph of recombinant RSN-2 solution (left) compared to buffer (right).

To quantify the surface activity of RSN-2 a classical torsion balance (White Electrical Instrument Company) with a platinum Du Noüy ring was used to measure the surface tension of RSN-2 over a range of concentrations. The ring was cleaned in detergent solution and warm water followed by exhaustive rinsing with deionised water. Finally the ring was rinsed with ethanol (HPLC grade, Fisher Scientific) and then flamed with a Bunsen. The clean ring was placed on the torsion balance and 1mL drops of sample solution were placed on a glass evaporating dish and allowed to equilibrate for five minutes before measurement. Measurements were taken according to the manufacturer's instructions which involved zeroing the balance in air then lowering the ring to the surface of the solution and then raising it until it broke free of the surface. The force exerted at this time was measured in triplicate and recorded. RSN-2 concentration for the most concentrated sample was measured by A_{280} as described previously in this work, and subsequent concentrations were calculated from dilutions in water. Experimental error for the surface tension was estimated by the standard deviation from the mean for each data point.

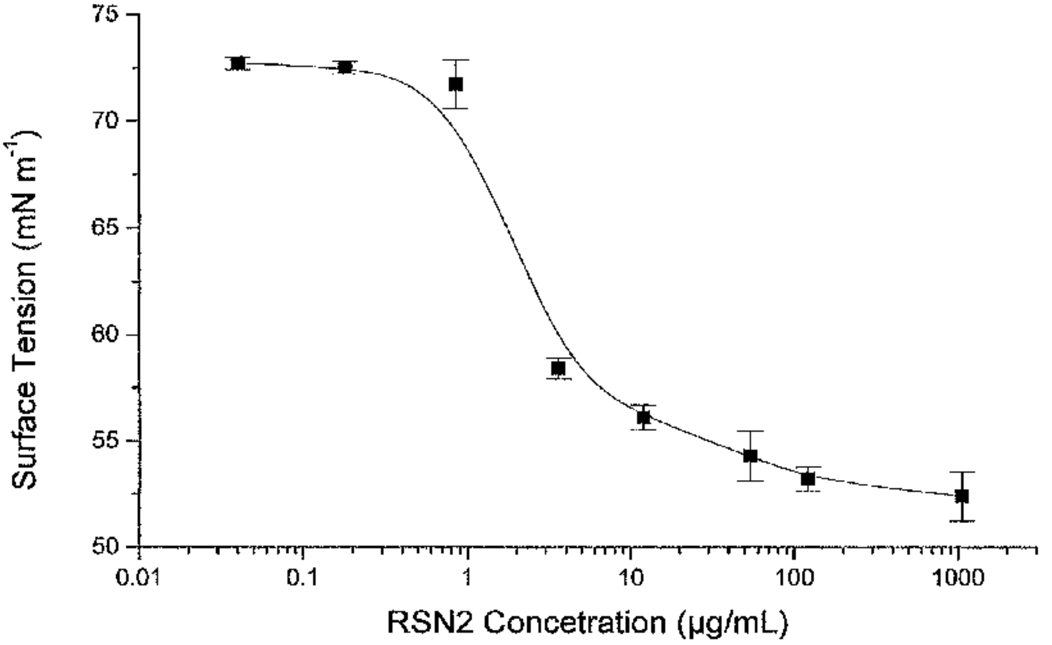


Figure 8.8. Plot of surface tension versus the RSN-2 concentration (on a logarithmic scale) measured using the *Du Noüy ring* method. The curve is a best fit line, used as a guide only.

These results indicate that RSN-2 is surface active at very low concentrations (<10µg/mL) and reduces the surface tension down to less than 60 mNm^{-1}. This is not as much of a reduction in surface tension than has been observed for some other surface active proteins (~40 mNm^{-1} for Latherin[184], 27 mNm^{-1} for some fungal hydrophobins[185]) and small molecules such as SDS[186], but is still a greater reduction than that seen in some other proteins that are known to foam (63 mNm^{-1} for BSA, for example).

A more sophisticated measurement device was available for characterisation of surface activity in the form of a Langmuir trough (Kibron Inc.). This instrument uses computer control and recording for measurement of surface tension/pressure in a variety of formats – the micro-trough well formats allowed for multiple analyses of small volumes of sample, the standard Teflon trough allowed for measurements of surface-area versus surface-pressure isotherms, and a depth-modified trough allowed for deposition of *Langmuir-Blodgett* layers. Surface tension measurements used a wire probe mounted on a sensitive electronic balance and was based on the Wilhelmy plate method[187]. The system was installed and housed in a thermostatically controlled room as per the manufacturer's instructions.

In the first instance, the micro-trough was used to measure the surface tension across a concentration range in the same way as for the tensiometer. During use of this instrument it was found that it was very important to maintain a high degree of cleanliness and to only use limited contact materials during preparation of surfaces and solutions. Before use, the trough was thoroughly cleaned with a soft brush (any rough abrasion would affect surface properties of the Teflon), warm water and detergent. After this, it was extremely important to rinse all of the detergent away so that no residue could affect measurements. Water of the highest quality available was used as any constituent impurity could affect readings. It was also found that any solution contact with disposable plastic containers such as falcon tubes released surface active agents in small quantities, consequently only glass containers were used. After washing, the plate was allowed to dry (covered to minimise dust contact) and then mounted on the trough apparatus. The probe was rinsed in purified water and then high-purity ethanol and flamed in a Bunsen.

The probe was calibrated against deionised, sterile-filtered water. Solutions were applied, normally in 500μL drops, and the probe was lowered to within the solution to ensure complete wetting, before being raised to the air-water interface where measurement was initiated. This generally took less than a minute. The instrument was allowed to record data for several minutes but it was found that after longer periods (more than an hour) evaporation began to interfere with measurements and would sometimes mean that the probe was no longer in contact with the solution.

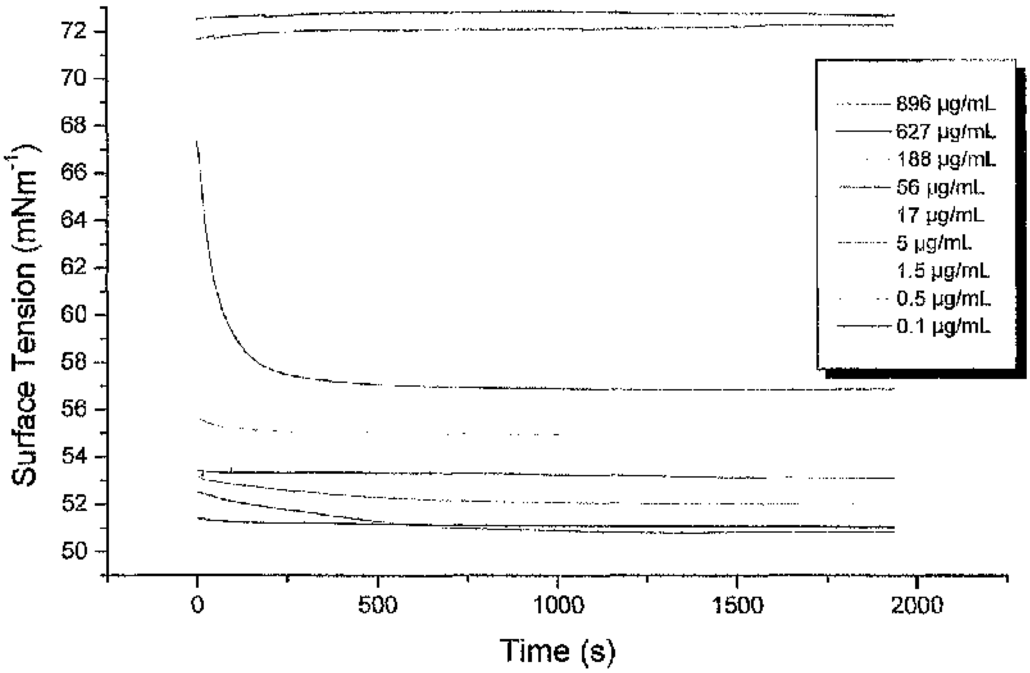


Figure 8.9. Plot of surface tension versus time for RSN-2 at a range of concentrations.

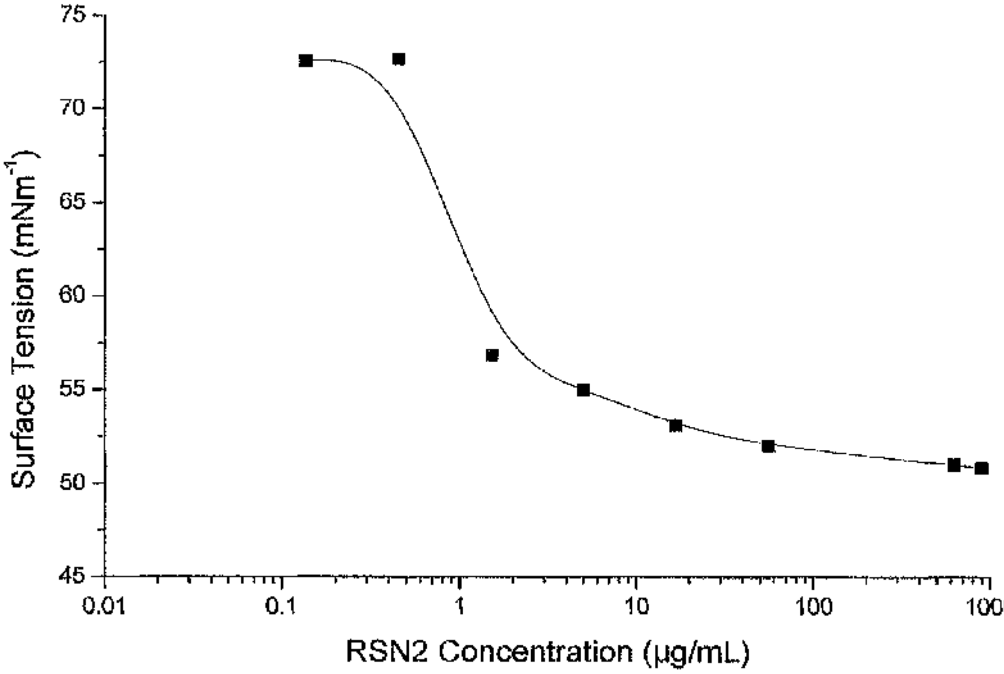


Figure 8.10. Plot of Surface tension versus concentration of RSN-2 (Langmuir trough method) where the surface tension is the plateau value reached in Figure 8.9.

It was observed that the concentrated solutions initially showed a marked decrease in surface tension that drifted downwards for some time before reaching a plateau value. The time taken to reach this limit varied slightly between solutions but was generally of the order of 300-500 seconds which is in good agreement with measurements of the foam solution and has been shown to have much faster kinetics than other surfactant proteins[62]. On the molecular level there are no clear reasons as to why RSN-2 self-associates at the air-water interface so quickly. If there are any structural rearrangements related to the surface association then they must happen rapidly, suggesting a low energy barrier. From the elucidated solution structure, there is no obvious unfolding mechanism that could occur without a probable high energy barrier (due to the loss of energetically favourable internal non-bonding attractions).

In this work RSN-2 has been shown to reduce surface tension to an almost identical degree as seen for the foam fluid but at a lower concentration (1.5 μg/mL compared to ~13 μg/mL)[63]. Proteomic analysis (Chapter 3) has also demonstrated that RSN-2 is a significant component in the foam matrix. This supports the case for RSN-2 being the surfactant component in the foam fluid. The fact that RSN-2 affects a surface tension quickly, but to a degree that is not as large as observed in some surfactant solutions, fits very well with its considered biological function. When excreted in the foam fluid it creates a surface tension decrement that is great enough to facilitate foam formation and it does this quickly while the male is beating his legs. Furthermore it can associate and dissociate with the monolayer upon agitation, which may be an important property in the formation of the foam nests.

Surface tension analysis from the two different techniques was in good agreement. At 296 K, RSN-2 was observed to reduce the surface tension to ~55 mNm^{-1} ($\pi \approx 18$ mNm^{-1}) at a concentration of ~1-5 μg/mL in 500 μL drops. One of these drops was measured as having a diameter of 11mm which equates to a surface area of $9.5x10^{-5}$ m^2 (ignoring the curvature affects). If one assumes a minimum active concentration of 3 μg/mL RSN-2 this equates to 1.5 μg in the 500 μL drop, which is the same as $8.0x10^{13}$ protein molecules, the surface area per molecule can be calculated as $1.2x10^{-18}$ m^2. From the calculated structure, arbitrary x, y and z molecular dimensions were calculated as 3.1x2.9x2.5 nm using the PyMol *distance* function. Using this data, and assuming that the RSN-2 structure does not alter conformation at the surface its surface area can be calculated as $7\text{-}9x10^{-18}$ m^2. Even given the error associated with these calculations, it is clear that RSN-2 partitions into the surface active layer at very low concentrations. Indeed, it seems likely that it preferentially forms a monolayer when in solution but that after this, the remainder stays in solution or perhaps additional layers do not create any further significant reduction in surface tension.

Before making quantitative measurements of RSN-2, a model fatty-acid was used to confirm satisfactory operation of the trough. A 1 mg/mL solution of stearic acid (Sigma) in chloroform was prepared. The glass trough with Teflon mount was filled with deionised water and allowed to settle for a number of minutes. The clean probe was calibrated and zeroed and then the surface was aspirated to remove any contaminant surface active molecules and dust. A compression was initiated and the surface pressure was monitored for any increase which identified contamination. If contamination was observed the surface was further aspirated and the process was repeated until cleanliness was observed, at which point the probe was calibrated and zeroed again. A clean glass syringe (Hamilton) was used to deposit 50 μL of stearic acid solution onto the surface of the water, taking care not to disturb the surface more than necessary by applying the drop from a level just above the surface. An increase in surface pressure was observed after deposition and this was seen to return to zero as the solvent layer evaporated. Once the solvent was fully evaporated, the compression was initiated and the apparatus recorded the temperature-compensated pressure-area isotherm which was in good agreement with published data[188]. This confirmed that the trough was fit for further use.

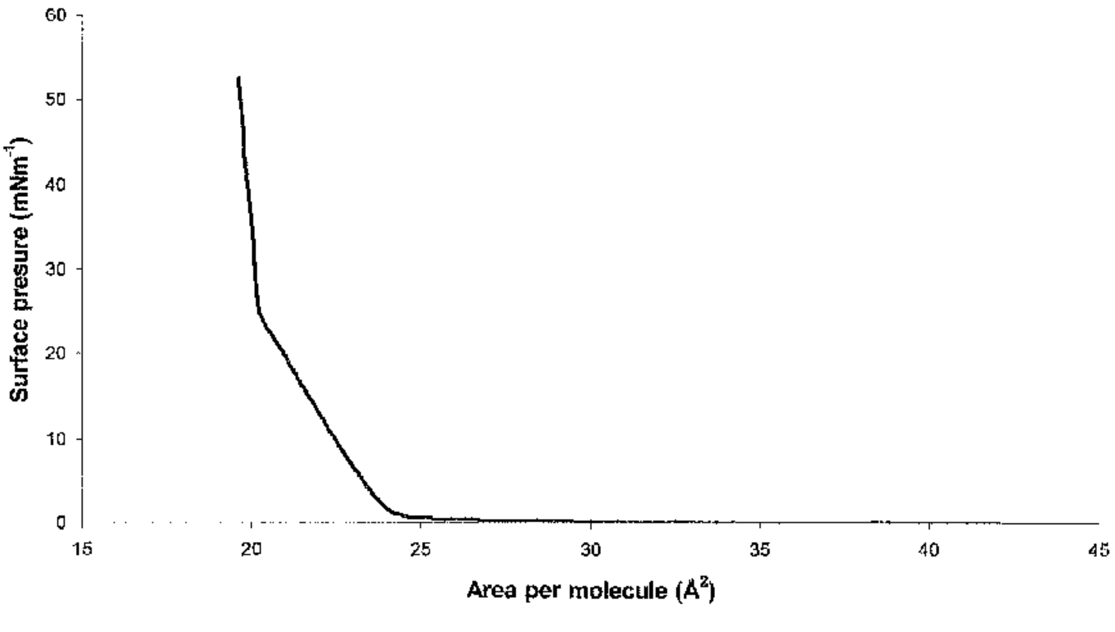


Figure 8.11. Surface pressure-are isotherm of stearic acid.

In order to examine proteins some of the procedures used in fatty acid studies had to be altered. Most importantly, solvents could not be used to carry and apply the monolayer as they would have denatured the protein. Instead RSN-2 was applied in buffer solution and allowed to equilibrate for 30 minutes prior to any compression. A blank run was performed with the compression of just the buffer solution to confirm that the buffer did not contribute towards surface activity. A concentrated solution (~6 mg/mL) of RSN-2 in 20mM sodium phosphate pH 7.0 was diluted to 1 mg/mL with deionised water and confirmed by A_{280}. 12 μL of this solution was applied to the subphase of the cleaned water in the trough using a glass syringe. Compression was initiated at 50 cm^2 min^{-1} and was allowed to continue until the probe mounting prevented any further movement.

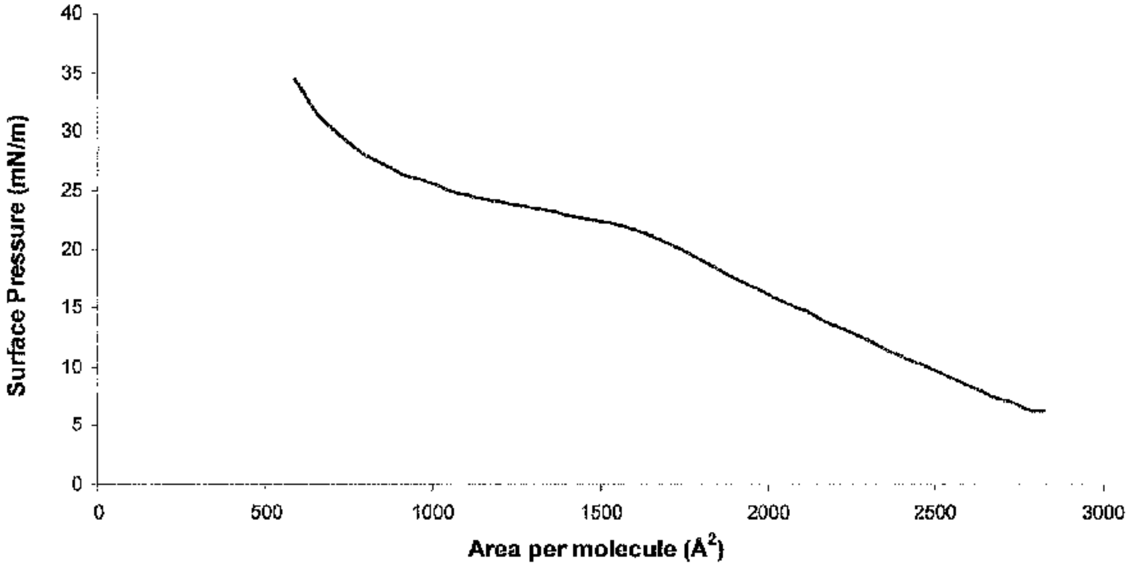


Figure 8.12. Surface pressure – area isotherm of RSN-2.

Figure 8.12 shows the pressure area isotherm for RSN-2. The general shape of the plot is similar to that seen for proteins in other studies[189,190,191] and a number of points can be made. Although the x-axis is conventionally represented as area per molecule (the total surface area divided by the number of sample molecules applied), this is meaningless for RSN-2, where the molecule is able to exist at both the air-water interface and in the bulk solution. Unlike fatty acids, there is no collapse pressure – that is a surface area that is reached where the molecules can no longer all be arranged at the surface and the monolayer fragments and collapses. A collapse pressure does not occur in the RSN-2 solution, consistent with the expectation that protein is constantly in equilibrium with the aqueous sub-phase.

It was also observed that at any point in the compression cycle, if the compression was halted, the surface pressure did not remain constant, but rather relaxed back to a surface pressure expected from the static surface pressure versus concentration observations seen previously. It is speculated that this is a dynamic effect where the molecules reorganise themselves into a surface concentration that is ideal. Repeated analysis of the same solution showed that the same solution gave largely reproducible data after a compress-relax cycle that would indicate that any structural reorganisation in the protein was fully reversible.

8.2.2 Two-Photon Excitation Fluorescence Microscopy

As described in section 8.1.3, RSN-2 increases the fluorescence from ANS, and this observation was used for structural imaging of RSN-2 foams by two-photon excitation fluorescence microscopy. This optical technique takes advantage of second-order fluorescence emission events that are stimulated by absorption of two photons of half the energy of the excited state. Accordingly, excitation with light at 810 nm can be absorbed twice to emit a photon of >405 nm (the same as a fluorescence emission after excitation with light of 405nm). By using an excitation wavelength of 810 nm, the scattering is less than that of visible light. Also, because the excitation is non-linear, the emission is much more confined to the focal point, than in conventional fluorescence. The net effect is that a microscope can be used to focus the light and probe to a greater depth into a scattering sample than is normally feasible in visible microscopy[192]. This work was carried out at The Centre for Biophotonics, University of Strathclyde, with the assistance of Prof. A. Cooper and Dr. D. Wokosin (who performed all of the microscopy). The procedures were similar to those previously used for natural *P. pustulosus* foam[62].

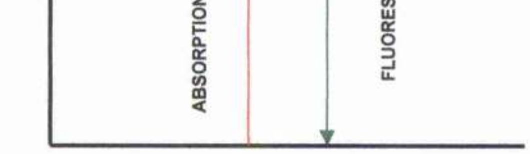


Figure 8.13. Energy level diagram of two-photon fluorescence. The dashed line represents a virtual energy level - absorption is actually simultaneous.

An RSN-2 solution (4mL of ~0.6 mg/mL in PBS) was mixed with 100 µL of 10 µg/mL solution of ANS in ethanol and whipped to a foam using a commercially available cappuccino frother. The foam was of a much wetter consistency than the native *P.*

pustulosus foam nest and was observed to drain relatively quickly and then collapse which may be a result of the difference in concentration between native foam and recombinant RSN-2 solution, but is more likely to be due to the lack of other foam constituents that may help to stabilise the foam structure. Microscopy of the foam structure was acquired on a Radiance 2000MP microscope (Bio-Rad) with pulsed laser excitation at 810 nm and detection at 440 – 500 nm. The images were viewed using the Picviewer software (written by Dr. J. Dempster, The University of Strathclyde). It was found to be difficult to obtain clear pictures of the bubble interfaces of the RSN-2 foam, as the bubbles were continually draining and collapsing. Only one clear picture was taken before halting analysis.

Figure 8.14 shows the image obtained, displayed with maximum contrast. The circles represent enclosed bubbles within the wet foam and fluorescence intensity is represented by darkness. From the image, it is clear that there is an accumulation of ANS at the air-water interface, due to the increased fluorescence intensity in these regions. There are two possible explanations for this – either there is a very large concentration difference of ANS-bound protein at the interface relative to the bulk, or the RSN-2 molecules at the interface have adopted a different conformation than those in the bulk, accompanied by increased exposure of hydrophobic patches (therefore binding more ANS). From the surface tension experiments, it is known that RSN-2 does accumulate at the surface, but at 0.6 mg/mL the majority of the protein will still be present in the bulk solution. This would suggest that although a concentration difference between the surface and the bulk solution does exist, it may not be large enough to account for the diversity of fluorescence intensity seen between these two areas of the foam. Comparison with data from the whole foam[63] suggests that there is much less fluorescence intensity in the bulk solution in the RSN-2 foam. This information adds weight to the hypothesis that RSN-2 undergoes a rearrangement at the air-water interface, but is not definitive.

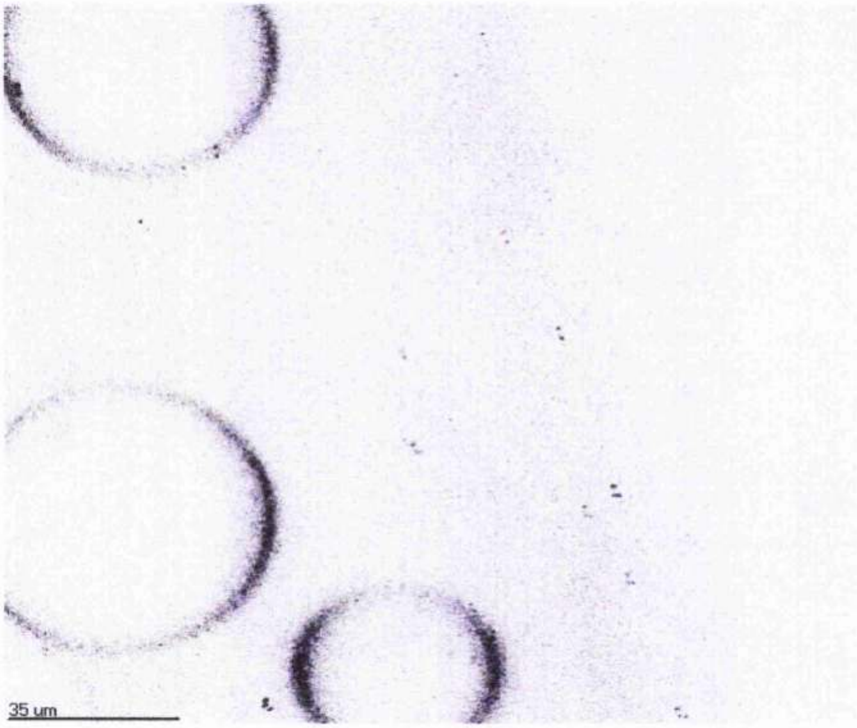


Figure 8.14. Two-photon excitation fluorescence microscopy image of a plane through an artificial recombinant RSN-2 foam. Fluorescence intensity increases are represented by darkening on the grey-scale.

8.2.3 Atomic Force Microscopy

To investigate the possibility of a structural change in RSN-2 at the surface it was necessary to try to find a way to analyse static, surface-bound properties. Whereas protein immobilisation would normally be a straightforward procedure, there are difficulties associated with RSN-2 because an immobilised structure should be designed to mimic the native state of the protein as much as possible. Because nothing is known about the relative surface orientation of RSN-2, covalent binding to the surface would be unsatisfactory. Instead the protein must be bound to a surface that closely mimics the aqueous environment. Clearly, there are problems in trying to create a surface where the solid-air interface is similar to the water-air interface and even more so, there are difficulties in trying to bind RSN-2 to such a surface when it already exists in an aqueous environment, i.e. to bind to the surface properly, through non-covalent interactions, the solid must be more like water than water.

Langmuir-Blodgett depositions are immobilised mono or multilayers of a substrate on solid surfaces and the field has been surveyed in depth[182]. The previously described Langmuir trough was used with a deposition accessory to dip solid surfaces into aqueous solutions containing RSN-2 with associated monolayer. In the first instance, silicon wafer was used as a solid substrate due to its ready availability and the ability to chemically alter its surface. An electronics grade silicon wafer was cleaved to create small oblong sheets for preparing RSN-2 surfaces with approximate dimensions of 5 x 20 mm. These surfaces were cleaned and prepared by sonicating in Piranha solution (75% sulphuric acid, 25% hydrogen peroxide (30% solution)) for 20 minutes to remove any organic residues and create a hydrophilic surface. These surfaces were rinsed in deionised water and then mounted on the deposition device.

The process for each deposition was similar to the preparation of monolayers: the solid substrate was inserted into the clean water in the trough and then the RSN-2 was added to the surface of the water. It was important to perform the operation in this way as any true monolayer deposition must only be allowed to pass through the interface once or a bilayer could be formed. Three different surfaces were prepared. The first used a 10 μg/mL RSN-2 solution and the wafer was dragged through the interface once at a slow rate (approximately 1 mm/min), whereas the second was created by dipping the wafer a further three times therefore creating a possible seven layers. The third deposition was created by simply immersing a wafer sample in a relatively concentrated RSN-2 solution (0.9 mg/mL), stirring and incubating at room temperature for thirty minutes.

For analysis of the surfaces, AFM (Atomic Force Microscopy) was used. This technique uses a laser light to monitor movements in a cantilever that is dragged or tapped across a surface[193]. By monitoring the changes in depth, a three dimensional image of the surface can be created. The cantilever (and tip) can be operated in a number of different modes but in this work it was operated exclusively in the dynamic oscillation or tapping mode where the cantilever oscillates vertically, making contact with the surface each time but with not enough force not to stick to the surface. In this mode, complementary information can be gained by looking at both the amplitude and the phase data. All atomic force microscopy described here was performed with the assistance of Dr. Nikolaj Gadegaard (Centre for Cell Engineering, University of Glasgow).

Surfaces were cleaved to the correct dimension squares and then mounted in an AFM sample disc (15mm diameter) for analysis. Analysis was performed on a Nanoscope Dimension 3000 AFM (Veeco) in air. On the first and second surface (mono and multi-layers), there was no obvious pattern observed. There were no visible features and software analysis of the general roughness were of the order of 1 Å and were almost certainly a result of the surface roughness of the wafer[194] and were far too small to be related to protein deposition. No regular features were observed after Fourier transformation of the distance data. The sample that had been immersed in concentrated RSN-2 solution appeared quite different microscopically; a variety of surface features were observed including evidence for ordered multilayers, shown in Figure 8.15 as amplitude and phase images, processed and rendered using *WSxM 2.2* software (Nanotec Electrónica).

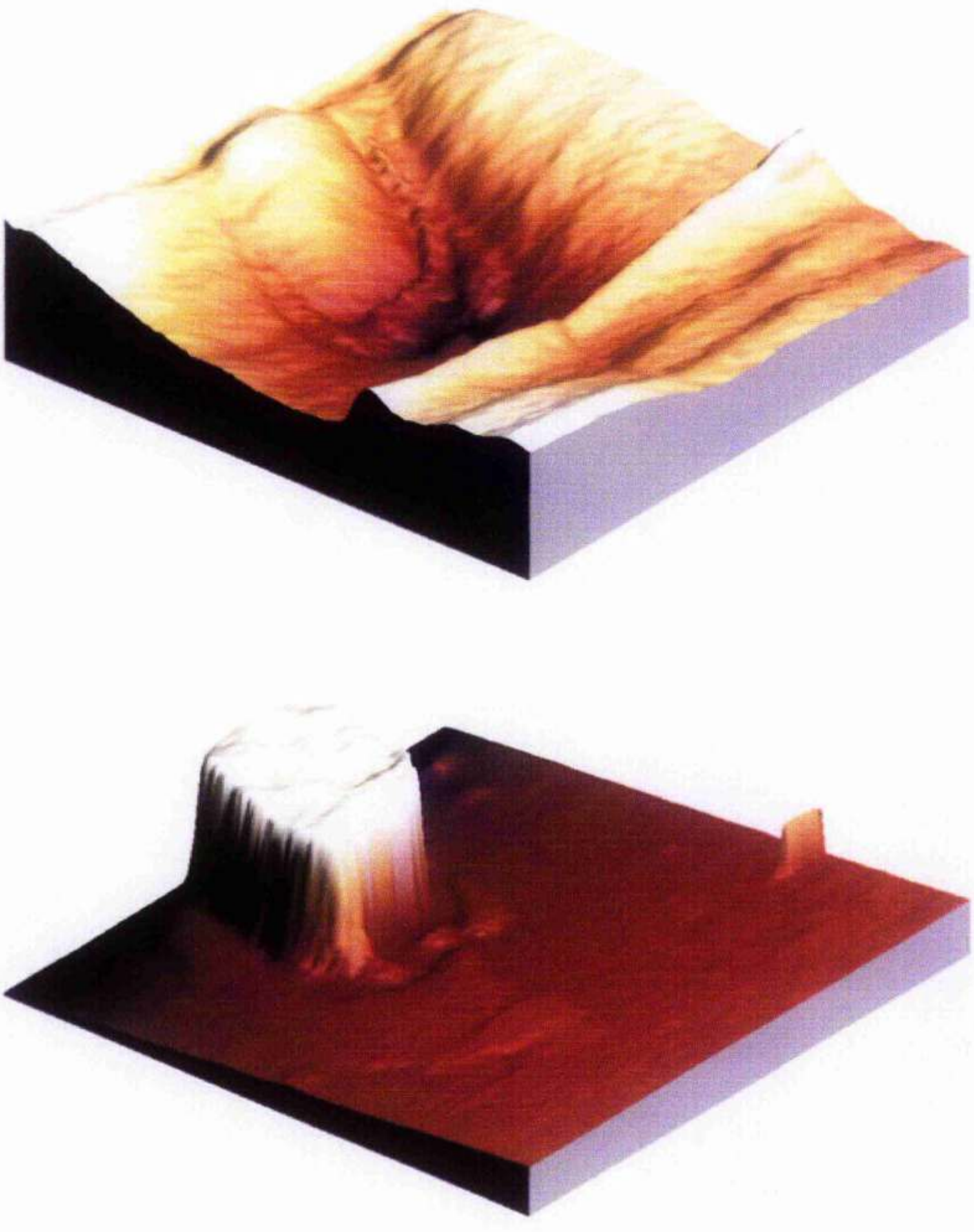


Figure 8.15. Amplitude (top) and phase (bottom) images from AFM of RSN-2 deposition on silicon wafer. The squares measure 200 x 200 nm.

Although no organised monolayer deposition was observed, the features observed on the immersed surface indicated that there may be some potential for further experiments. Because RSN-2 hadn't adhered to the silicon surface with enough strength to assemble and image, it was postulated that a more hydrophilic surface would be more likely to provide monolayers. To this end, Langmuir Blodgett techniques were used to attempt RSN-2 deposition on mica surfaces. The mica (Structure Probe Inc) surface was muscovite V-4 which is optically very flat and the mineral structure provides exposed surface hydroxyl groups to create a polar surface. Mono and multilayer depositions were prepared as previously described and are currently awaiting analysis.

It was hoped that AFM analysis might help to describe some regular structure formed when RSN-2 self assembles at the air water interface as the resolution of AFM would be unlikely to allow single protein morphology to be observed. As a complementary surface analysis recombinant RSN-2 was prepared and submitted for neutron reflection (Prof. J. Lu, University of Manchester) and are awaiting analysis. This technique uses the detection of reflected neutron particles against monolayers on mixed H_2O/D_2O surfaces to identify dimensions of various surface components. It is hoped that neutron reflection results will help to characterise the surface layer dimensions of RSN-2 at the air-water interface and, knowing the molecular dimensions of the protein in solution, this data may provide some information of the interfacial structure.

8.3 Summary

Spectroscopic techniques were used to demonstrate that the secondary structure of recombinant RSN-2 is not significantly altered upon cleavage of the fusion sequence. Additionally, calorimetry was used to demonstrate that the cleavage of the fusion tag results in a more stable protein structure. Solution fluorescence and fluorescence microscopy illustrated protein binding, partitioning in solution and supported the hypothesis of molecular reorganisation at the air-water interface.

Analysis of the surfaces of RSN-2 solutions proved that the recombinant protein is surface active at extremely low levels and basic arithmetic was used to suggest that the protein is active at, or close to, monolayer levels. Examination of the dynamics of RSN-2 surface association indicated that RSN-2 actively maintains a dynamic equilibrium of soluble protein in the bulk phase with monolayer coverage at the interface. Comparison with published data for the intact foam implied that RSN-2 is probably the surface-active component in the foam, but laboratory observations suggested that other foam components were required to stabilise the otherwise short-lived foam structure.

Initial AFM studies were promising and it is hoped that follow-up experiments with this, and neutron reflection techniques, will yield structurally significant insights to RSN-2 at the surface.

9 Anti-Microbial Activity

9.1 Introduction

P. pustulosus foam nests were observed to be free of parasitic or pathogenic attack over a period of several days in their natural environment, which is hot, humid and constantly exposed to bacteria and other microbes. Indeed, the nests appear to be free of any visible degradation or infection throughout their natural lifetime (the time taken for the eggs to move into the water). Eventually, it is the process of the eggs being released that creates instability in the foam structure. After a number of days, the eggs can be seen to be encapsulated in an expanding liquid layer that allows the eggs to move through the foam and eventually into the water below. It is postulated that this is more likely to occur in wet weather while the nest is in standing water, and the fluid is less likely to evaporate. In this way, survival of the tadpoles is more likely. The constituents of the fluid are unknown but samples have been taken and frozen for future analysis. It is proposed that when the have reached an advanced development stage and they are ready to move into water, they release a specific enzyme that breaks down the structural components of the foam (protein and/or carbohydrate) to allow their release. Analysis of this material may yield further information on the biochemical complexities of the system.

As previously stated, the sequence similarity searches of the six known *Ranaspumin* sequences indicated that RSN-1 had a high similarity score with cystatins, and RSN-2 was shown to have some (but low) degree of similarity to the same proteins when analysed using the 3D-PSSM program. From the elucidated solution structure this can now be seen to be an accurate prediction, and a remarkable one given the low degree of primary sequence similarity. These links to proteinase inhibitors and the observed anti-microbial action of the foam, suggest that these proteins may be acting as cystatins in the foam to prevent microbial attack via cysteinyl proteinases.

9.2 Cystatin Assays

9.2.1 Initial Testing

As an initial inhibition activity study, the foam nest fluid and solutions of recombinant RSN-1 and RSN-2 (both still with fusion tags) were prepared by Dr. R. Fleming and sent to a collaborator with expertise in cystatin testing (Dr. D. Knox, Moredun Research

Institute). The exact assay that was used is unknown but is thought to be similar to that used in published data[195].

These initial results gave a strong indication that the frog foam fluid was an inhibitor of papain-like proteinases and suggested that RSN-1 and RSN-2 were also inhibitors. In the same assay, the same solutions were also found to be inhibitors of a similar cysteine proteinase, Cathepsin B (an intracellular cysteinyl proteinase), but to a lesser extent (data not shown).

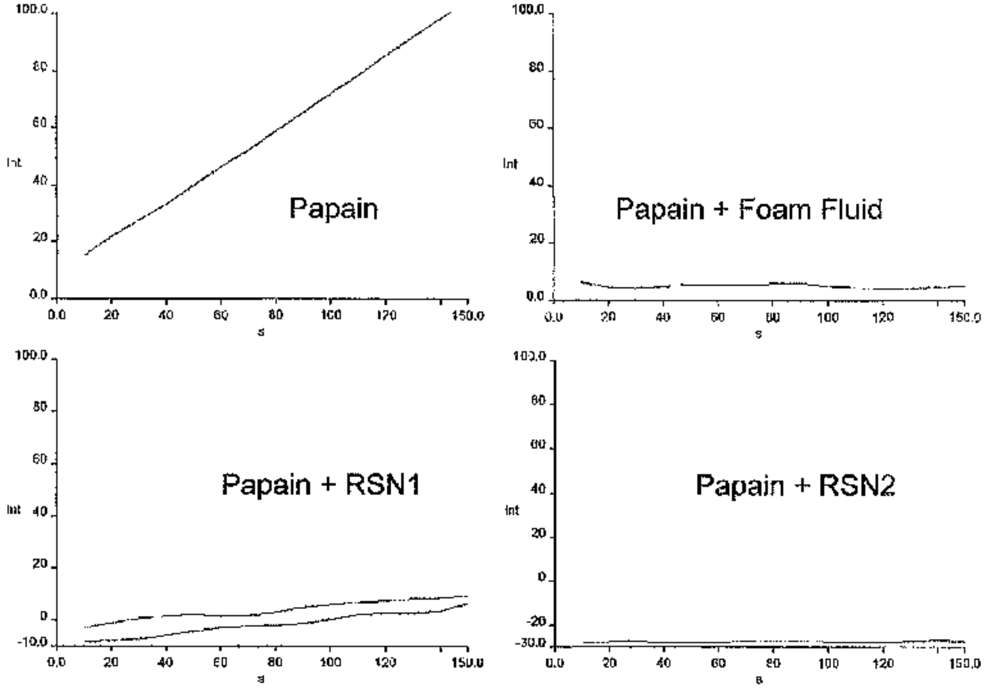


Figure 9.1. Fluorescence intensity against time for papain and papain with foam components. The rise in intensity indicates proteinase activity.

That the foam fluid was indicated as an inhibitor, was not surprising, but the activity shown from the recombinant proteins was more so. This was particularly true for RSN-2, because the sequence showed such little similarity to cystatins (although it has now been shown to be structurally similar). These results were treated with great importance as they helped to add to the evidence of anti-microbial function in particular proteinaceous components. It also suggested that RSN-2 acted as an inhibitor, despite not containing the highly conserved residues found in other inhibitors, raising the possibility of an unexplored functional mechanism.

For confirmation and to provide further data, fresh samples of foam and recombinant RSN-1 and 2 were prepared and submitted for analysis as before. The results from the repeat experiments were less conclusive, showing some possible inhibitory activity but not to the same extent as previously shown. It was also noted that when these fresh samples were compared to thawed samples from the previous analysis, there was an unsatisfactory degree of variability between samples. It was decided to move the cystatin analysis in-house for investigation of the variability of the results and confirmation of their surprising nature.

9.2.2 Assay Development and Foam Activity

Detection of the cystatin activity was based, wherever possible, on the same procedure as that used in the initial analysis, with the only differences being the manufacturers of the chemicals and the instrumentation. The assay was based on the detection of a fluorogenic substrate which, in its unaltered state self-quenches, but when a proteinase is present the internal peptide bond is cleaved and the 7-amino-4-methylcoumarin fragment can be detected sensitively by fluorescence emission[196]. In this work, the fluorogenic substrate was Z-Phe-Arg 7-amido-4-methylcoumarin (Sigma) which was prepared by dissolution of the hydrochloride salt form in methanol to a stock concentration of 1 mM.

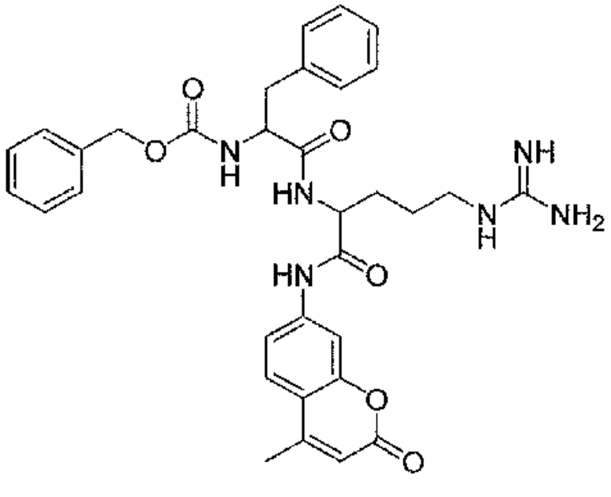


Figure 9.2. Z-Phe-Arg 7-amido-4-methylcoumarin - the fluorogenic substrate used to detect proteinase activity or inhibition.

Papain activity was detected by an increase in fluorescence intensity over time due to the cleavage of the peptide. Fluorescence emission was recorded using the Spex Fluoromax II spectrofluorometer described previously, operated in the time-resolved mode with λ_{ex} = 370 nm and λ_{em} = 470 nm and 5 nm slit width. The assay buffer was 100 mM sodium acetate pH 5 with 5 mM DTT (papain is most active in a slightly acidic environment). The active fold of papain contains three disulphide bridges and a free sulphydryl at the active site (Cys25)[197], which binds the substrate in the open cleft (Figure 9.3). A reducing environment may reduce the disulphide bridges but does not appear to alter activity, whereas oxidation of the free sulphydryl would make the enzyme inactive, so a slightly reducing environment is preferred. Papain solutions were prepared immediately before use by dissolution of lyophilised papain (from *Papaya* latex, Sigma) in assay buffer, with several minutes of vortexing, to a final concentration of 0.5 Uml^{-1}. Assay solutions were mixed in a fluorescence cuvette with a defined order: 1230 μL of assay buffer followed by 75 μL of papain solution and then 150 μL of test sample. This solution was gently mixed and allowed to equilibrate at room temperature for 3 minutes (to allow any potential proteinase-inhibitor complex binding to form) before addition of 45 μL of the substrate. Fluorescence was recorded immediately and monitored for 300 seconds.

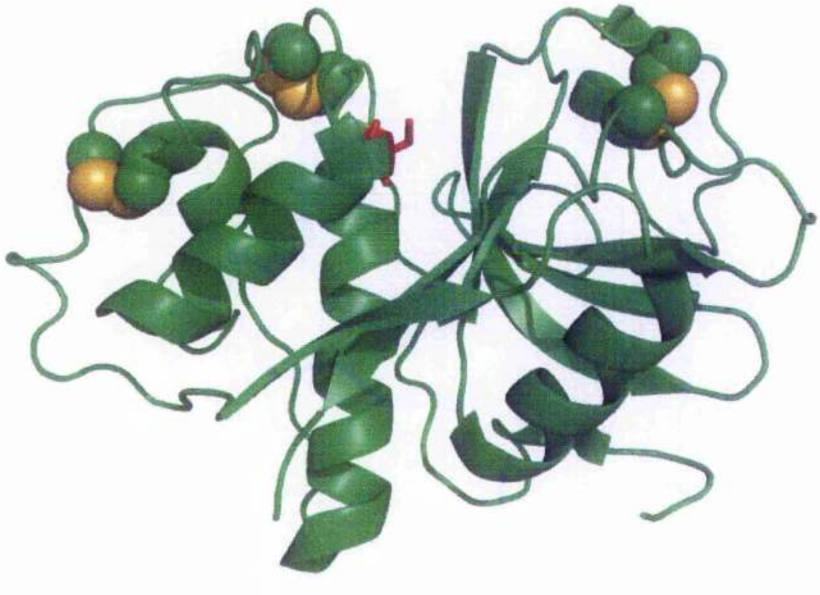


Figure 9.3. Crystal structure of activated papain[197]. The disulphide bridges are represented by gold spheres and the free thiol is shaded red within the binding cleft.

As a test sample, foam fluid was used for the initial experiment. As with all surfactant manipulation, care had to be taken to prevent any bubble formation in the cuvette during pipetting or mixing. Pure deionised water was used as the negative control. A BSA (Sigma) sample was prepared in assay buffer at a relatively high concentration (3 mg/mL) to test whether the availability of protein at high levels could act as an alternative substrate, and therefore produce a false positive inhibition.

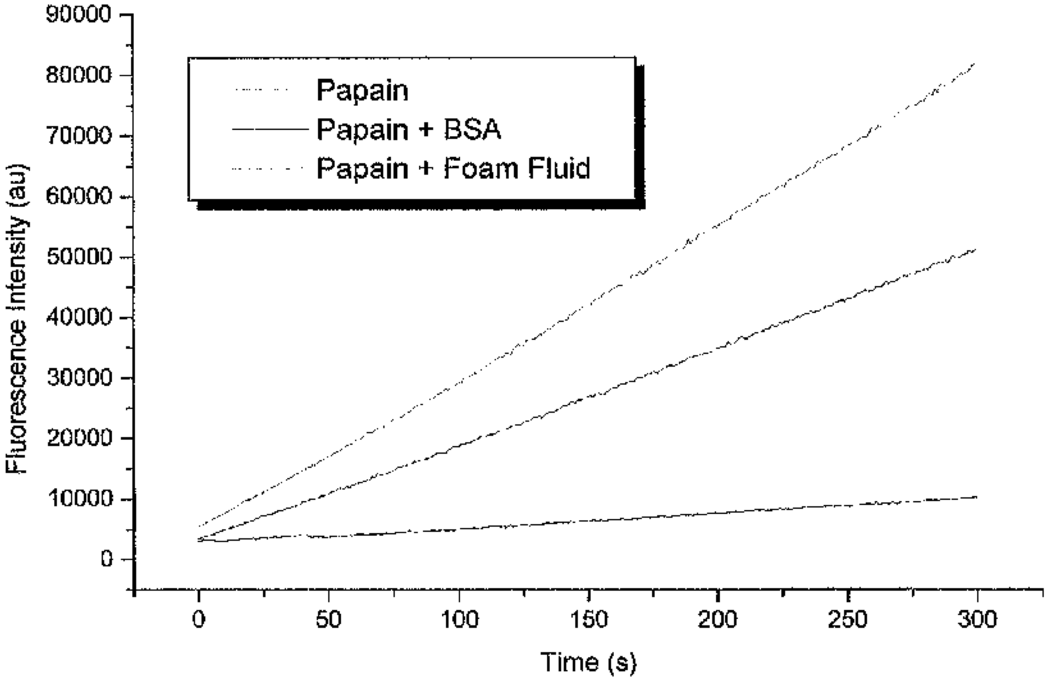


Figure 9.4. Fluorescence intensity versus time for papain with concentrated BSA and dilute foam fluid.

Although the high concentration of BSA did reduce fluorescence intensity slightly, the gradient change was relatively small and could not account for the activity seen for the foam fluid, which was seen to be active down to volumes of only 30 μL (20% of that described in the other experiments here).

9.2.3 Synthesis of RSN-1 and Cystatin Assay of Recombinant Ranaspumins

RSN-1 was produced following a very similar procedure to the production of RSN-2 (without the thrombin tag). The RSN-1 gene[63] (transcribed sequence shown in Appendix B) was supplied (prepared by Dr. R. Fleming) in the T7 NT-TOPO plasmid shown in Figure 2.5, transformed into *E. coli* BL21 cells and cultured as described previously. Following this, the over-expressed RSN-1 was purified by gravity chromatography through Ni NTA resin. Compared to the reliable production of RSN-2, the expression of RSN-1 was found to work very unreliably with a success rate of only approximately one in five and even then, only in low concentrations (~5 mg from 2L broth). Presence of overexpressed RSN-1 was evaluated by SDS-PAGE initially, followed by confirmation by externally calibrated MALDI-TOF mass spectrometry.

A concentrated sample of RSN-1 (0.5-1 mM) was subjected to a simple 1D proton NMR analysis, using the gradient water suppression described previously. The resulting spectrum suggested that the protein was folded, and the dispersion of the H_α peaks indicated a high strand content. This experiment showed that RSN-1 is a good candidate for structure elucidation by NMR, if a reliable expression system can be developed.

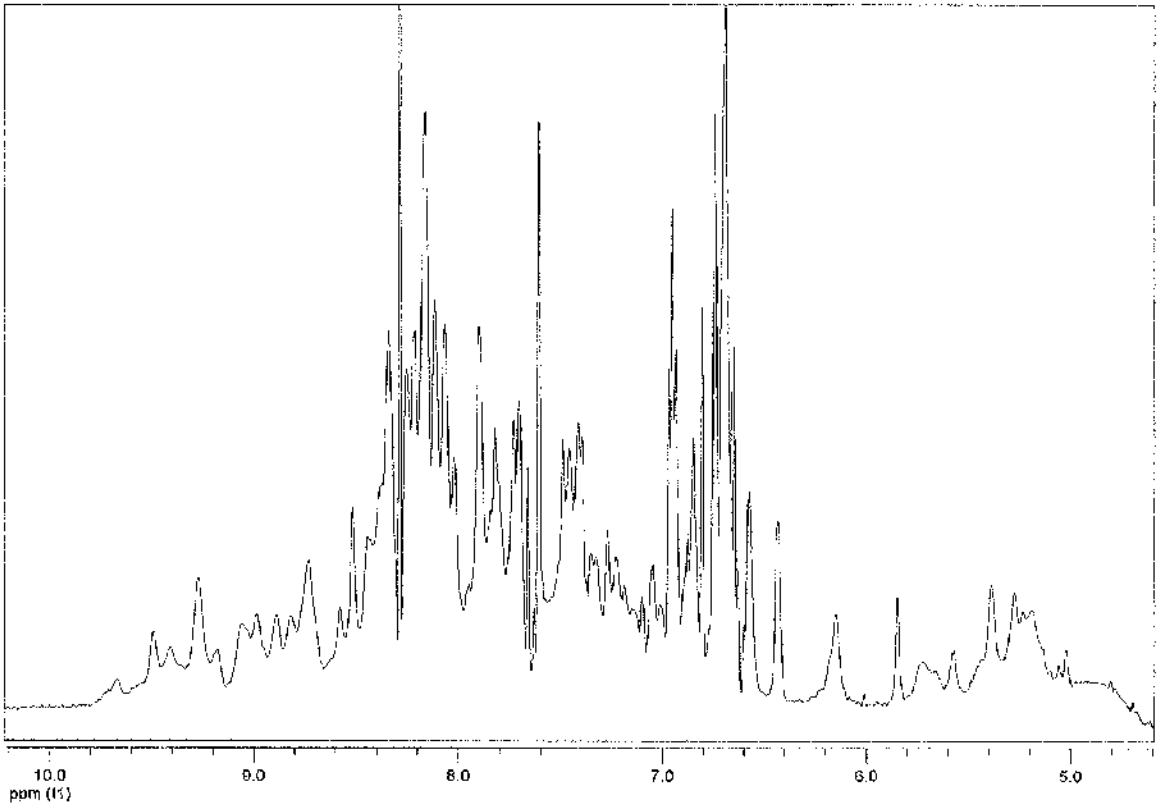


Figure 9.5. The amide region of the RSN-1 proton NMR spectrum. The sharp peaks are due to residual small molecule contamination.

As was the case for the original cystatin assays, RSN-1 was tested for activity without cleavage of the fusion tag. Recombinant RSN-2 was tested in its full, uncleaved form (His_6-Ek-RSN-2), with the fusion tag and thrombin cleavage site (His_6-Throm-rRSN-2) and the cleaved form (rRSN-2). Recombinant proteins were buffer exchanged into PBS and normalised to 0.5 mg/mL (by A_{280} determination) prior to analysis.

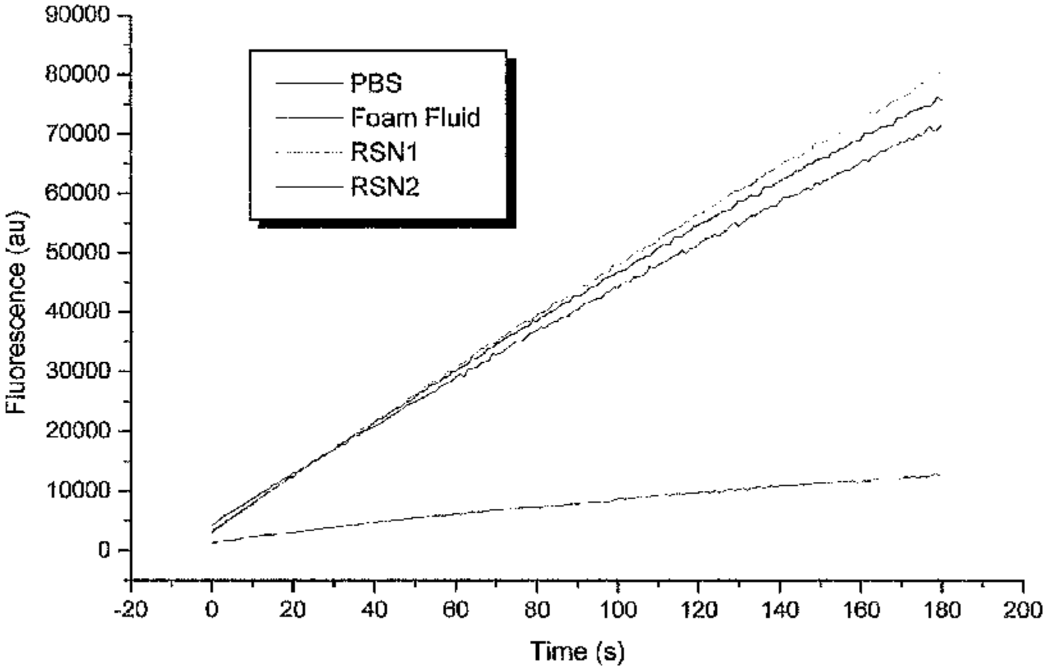


Figure 9.6. Fluorescence intensity over time for papain cleavage of labile substrate, incubated with potential inhibitors.

Examination of the fluorescence versus time data did not emulate the results found externally for RSN-1 or RSN-2. Conversely, the data for the bulk foam fluid confirmed that it was a potent inhibitor of papain proteinase activity. A great deal of time was spent changing the solution environment, in terms of relative concentrations and incubation times, to reproduce the previous results but this was never achieved. The easy availability of recombinant RSN-2 meant that it could be tested numerous times to a point where some confidence in the negative result could be achieved, but the relative scarcity of recombinant RSN-1 meant that the negative results were not a certainty.

The conflicting results were difficult to interpret because all variables between external and internal experiments were kept to a minimum where possible, and the variables that had to be changed could not account for the differences observed. BSA and water were effective negative controls and foam fluid proved to be an extremely potent positive control so the efficacy of the assay was demonstrated, and the 1.5 mL sample size provided much more stable fluorescence signals than seen in the external testing, where much smaller sample volumes were used.

A difference that may have caused the observed discrepancies was changes in inhibitor concentrations – the external testing simply added 10 µL of inhibitor to a final volume of 250 µL, without correcting for concentration. Is it possible that extremely high protein concentration may have acted as a competitive substrate, although demonstration of a negative result with BSA would make this unlikely. Another possibility may be that the proteins could be in different states (folds, aggregation state, complexes) due to preparations, storage and transport conditions, but several attempts to alter these properties did not result in increased activity. Other explanations for these observations could involve the differences in the implementation of the actual assays: that is that the expertise of the external lab may mean that something was performed differently without being transferred to the internal assay, or the heightened sensitivity of the specialist equipment used in the external assay. The remaining possibility is that the high number of cystatins that are tested in the external lab could have caused contamination, although the external assay performed adequately with a negative control. During the course of this work, the conflicting results were never fully resolved, but it must be assumed for the purposes of reporting results, that neither RSN-1 nor RSN-2 showed inhibitory activity against papain-like cysteinyl proteinases.

In a similar assay (same substrate) RSN-2 was tested for serpin activity by Dr. Rhona MacDonald and Dr. Rosalind Tan. Serpins are the serine protease functional analogues of cystatins, and may also be related to anti-microbial activity. In a single assay RSN-2 was shown to have no inhibitory activity against alpha-1 antitrypsin, which is unsurprising given the effective use of thrombin in the purification process. The serpins are particularly interesting here because the family has been found to contain members that are structurally similar to each other, but some of them do not inhibit proteases. The best characterised of this is ovalbumin from hen egg white, which belongs to the serpin superfamily but shows no inhibition of serine proteases[198,199] and shows reversible association properties at the air water interface[200]. As such, ovalbumin appears to be a hold a similar relationship to the serpin family as RSN-2 does to the cystatins, and the parallels drawn may prove to be significant.

9.2.4 Inhibitory Activity of Native Foam Components

Despite the difficulties associated with recording the inhibitory activity of the recombinant proteins, it was confirmed that the foam itself was a powerful inhibitor. It was decided to investigate the source of this inhibition by separating the foam into its constituent parts and recording the activity of each part. In this way, it was hoped that the source of the activity may be identified and that this may help to interpret the inconsistent results observed for the recombinant proteins.

To investigate this, a project was initiated with a postgraduate student, Mr. William MacEwan, and all of the work described in this section was carried out in collaboration, and it should be recognised that approximately half of the practical work was performed by him.

As a very simple separation, the foam fluid was separated according to size by gel filtration chromatography using assay buffer (without DTT) as the mobile phase. Chromatography was performed as described previously (Section 2.4.2) and the resulting chromatogram showed no discernable differences, suggesting that buffer composition and pH did not appreciably affect the foam composition. For all retention times with an associated UV response above the baseline, 1 mL fractions were automatically collected for testing for cystatin activity. From the 1 mL injection volume twelve 1 mL fractions were collected meaning that the total sample was diluted by more than 10-fold but, if it is assumed that the cystatin activity was contained in a single component, then virtually no dilution should have occurred. Accordingly, all fractions were tested for activity without any pre-concentration.

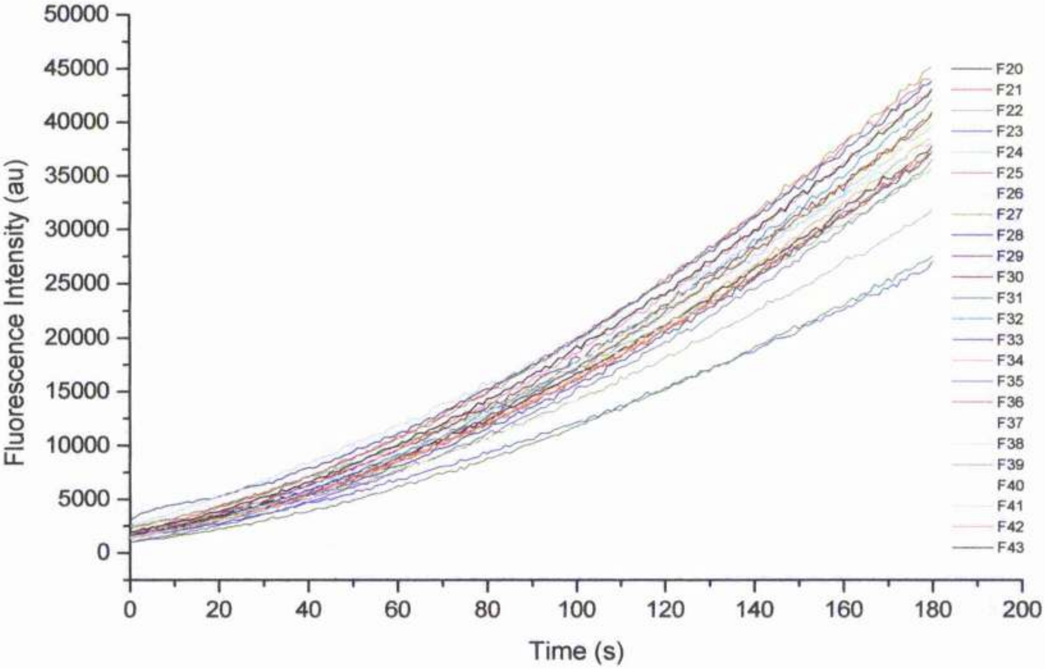


Figure 9.7. Fluorometric assay of papain inhibition by foam fractions separated by gel filtration chromatography.

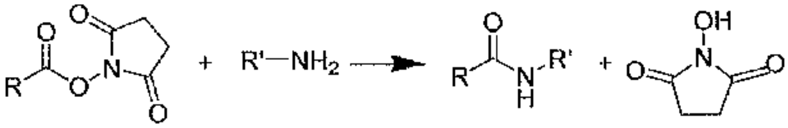
Although some variability was observed between fractions, this was within the range observed in the natural variability of the assay. From this data, there was no clear fraction or fractions that had significant inhibitory activity, and there was no observed correlation between extent of activity and fraction number, meaning that no particular size fraction of the component could be assigned as cystatin-active.

Due to the ambiguous result obtained through fractionation by gel filtration chromatography, a much coarser size-based fractionation regime was implemented. Foam fluid was filtered by ultrafiltration using a centrifuge and various membranes with different molecular weight cut offs (Vivaspin, Sartorious and Centricon, Millipore). Although these experiments overlap with those performed in section 2.4.1, the experiments were performed with slightly different conditions and were entirely independent. To prepare suitable fractions, a relatively large volume (~15 mL) of foam fluid was loaded onto a 300 kDa MWCO (molecular weight cut-off) spin concentrator and centrifuged until a volume of approximately 2 mL was retained at which point the filtrate was transferred to the top of a 100 kDa MWCO spin concentrator and the process was repeated. After this and in turn, the filtrates were filtered through 50 kDa, 30 kDa and 10 kDa MWCO concentrators and then each of the retentates from the five devices were rinsed thoroughly with deionised water.

Each of the retentate samples was prepared and spotted for MALDI-TOF mass spectrometry, and then analysed as described in section 3.2.3. Mass analysis proved that, surprisingly, virtually none of the foam components appeared to pass through the filter membrane, even at the highest pore size. Whereas, some circumstances could be envisaged where the concentration would preferentially concentrate large molecules and effectively dehydrate the solutions, in this case a large volume was passed through each membrane including a water rinse prior to analysis and yet small peptides ~1-5 kDa were retained in a spin concentrator with 300 kDa MWCO. Although it was observed that recombinant RSN-2 behaved with a larger hydrodynamic radius than predicted from the molecular mass, the unusual observations seen here were not replicated in the recombinant material and rRSN-2 passed through a 300 kDa MWCO filter with apparently little resistance.

This extremely unusual behaviour suggests that there are complicated cohesive forces acting within the foam fluid. This fact is obvious when trying to manipulate the whole foam, which has a sticky flexible consistency, but the fluid is much less viscous and is relatively easy to separate and pour. During ultrafiltration, the trans-membrane pressure is large enough to provide a pure water flux of the order of 1 mL/min. Centrifugal concentration produced a brown-coloured region at the membrane surface, consistent with highly concentrated protein solutions, so the proteins were associated with the membrane but apparently failed to pass through. Experiments were conducted both with PES and cellulose-based membranes to minimise the possibility of membrane binding clogging the pores. Reasons for the observed behaviour were not confirmed during this work but it seems clear that some cohesive or associative forces were preventing the movement of peptides across the membrane, and the observed behaviour supports the idea that the foam components are bound in a biomolecular matrix that supports and stabilises the integrity of the foam structure in aqueous environments.

As an alternative strategy, a papain-affinity column was prepared for selective retention of strong papain binders (stronger than the cohesion forces within the foam fluid) from the solution. It was intended that this system would allow for effective separation of active proteins and peptides from the matrix by relying on strong inter-molecular attractions, rather than the relatively weak forces applied in centrifugal filtration. An affinity column was prepared by binding papain to a functionalised polypropylene substrate for covalent conjugation (NHS-activated Hi-Trap 5 mL column, GE Healthcare). The column was prepared and operated using a syringe pump (Harvard Apparatus) and plastic, disposable syringes of various sizes. All solutions were passed through a 0.45 μm filter before being passed through the column to prevent contamination of the column and loss of performance.



Scheme 1. The reduction of an NHS-ester by a primary amine

In the pre-packed column, the resin is substituted via a ten-atom spacer arm with N-hydroxy succinimide ester groups which, at neutral or slightly basic pll, are readily reduced by primary amines such as those found on the N-terminus or lysine-ε of proteins, to yield stable, immobilised protein resins. Therefore, protein immobilisation is a relatively straightforward procedure but care must be taken to avoid the competing NHS hydrolysis reaction as much as possible.

Papain was bound according to the manufacturer's instructions: briefly, the column was supplied in an inert environment (isopropanol) so it was washed with water and equilibrated with 0.2 M $NaHCO_3$, 0.5 M NaCl, pH 8.3 (coupling buffer). One column volume (5 mL) of the papain solution (5 mg/mL in coupling buffer) was introduced and the column was incubated for 30 minutes at room temperature. The excess ligand was washed away with 0.1 M sodium acetate, 0.5 M NaCl, pH 4, and any remaining active NIIS-ester groups were deactivated with 0.5 M ethanolamine, 0.5 M NaCl, pH 8.3. Coupling efficiency was calculated by measuring the papain content in the ligand solution before addition to the column and after elution from the column by A_{280} - contaminating NHS groups were removed by a gravity PD10 desalting column (GE Healthcare). The yield was calculated as 62%.

Following papain binding, the column was used almost immediately. It was equilibrated with 20mM sodium phosphate, 150mM NaCl pH 7.5 (binding buffer) and then 3 mL of the foam fluid was passed through the column at a rate of 0.5 mL/min. The column was washed with binding buffer and then any papain-bound adducts were eluted with 100mM glycine/HCl, 400mM NaCl pH 3.0. The low pH dissociates the papain-inhibitor complex. Following neutralisation with 1M Tris pH 9.2, each of the solutions were promptly analysed using the fluorogenic cystatin assay.

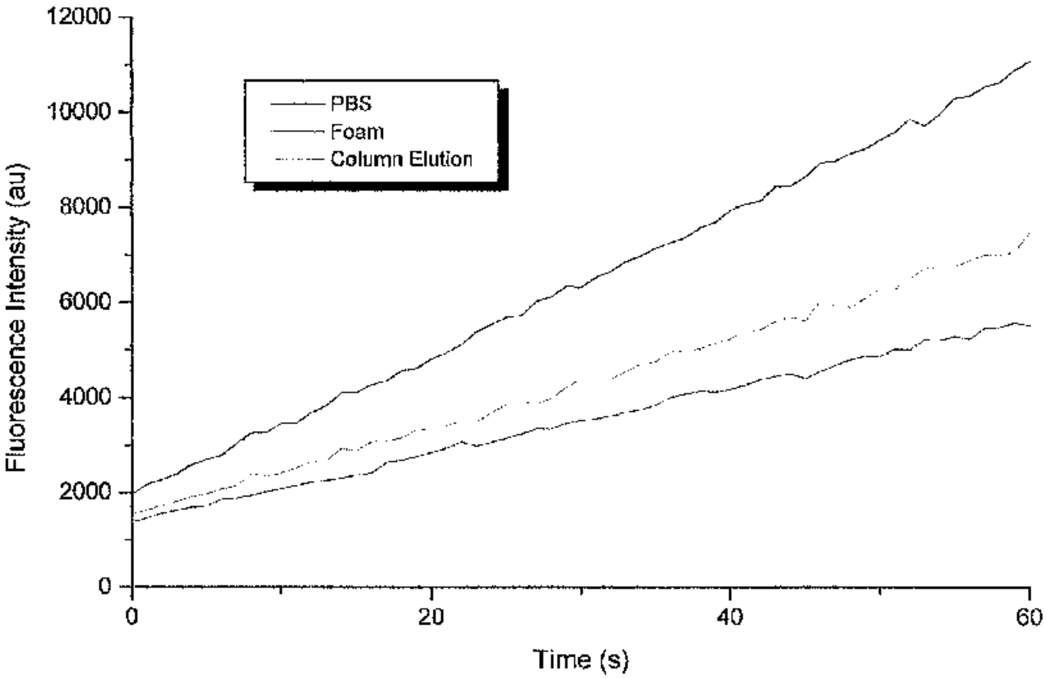


Figure 9.8. Cystatin activity of the elution from the papain affinity column using a fluorometric assay.

Although not conclusive, the results were indicative of some papain inhibition in the elution fraction. Accordingly, the composition was investigated by buffer exchanging into pure water using a 2 kDa MWCO dialysis cassette (Pierce Biotechnology), and analysed by MALDI-TOF for protein or peptide components.

An initial spectra revealed a ~8kDa polypeptide present in apparent isolation, suggesting that the cystatin activity had been purified completely (albeit in a weak concentration). The discovery of the peptide seen in the mass spectrum warranted further investigation so the sample was subjected to a tryptic digest, and the amino-acid sequences of some of the peptides were derived by *de novo* sequencing using tandem mass spectrometry. The sequencing was performed with the assistance of Dr. Carlos Bloch Jr. Samples were mixed with matrix and spotted in much the same way as previously described for MALDI-TOF, but were introduced to an Applied Biosystems 4700 TOF-TOF. By incorporating two independent time-of flight detectors, this instrument is able to target the intact peptide ions, fragment them by high energy collisions with carrier gas, and then perform a TOF analysis of the amino acid fragments. By careful analysis of the peaks, an iterative assignment process can be used to sequence the peptide from either (or both) termini[201]. Derived peptide sequences were as follows:

- NDVEPYFGRPR
- RNDVEPYFGRPR
- FEEPIPISK
- DQFVEQSDVTVELFVAFR

Database searches suggested that these peptides were previously undescribed, which was not surprising given the variety of peptides secreted from frogs of all species. To confirm this discovery and to sequence some of the other low-intensity peptides found in the digest mixture, the affinity column procedure was repeated with a larger (10 mL) volume of foam fluid, to create more concentrated cystatin elution. This time the cystatin assay showed the eluted fraction to be an extremely strong inhibitor, with potency similar to that seen for the intact foam fluid i.e. apparently complete inhibition. A mass spectrum was obtained for the column elution, and the signals were observed to be of a different mass as observed in the previous elution. Instead of a single 8 kDa polypeptide, two species were observed with m/z values of 12.3 and 5.9 kDa. This sample was lyophilised and sent to Dr. Carlos Bloch Jr. for TOF/TOF analysis at Embrapa Genetic Resources and Biotechnology, Brasília, but upon arrival and aqueous reconstitution, the 12.3 kDa signal was no longer observed and so no further characterisation was performed.

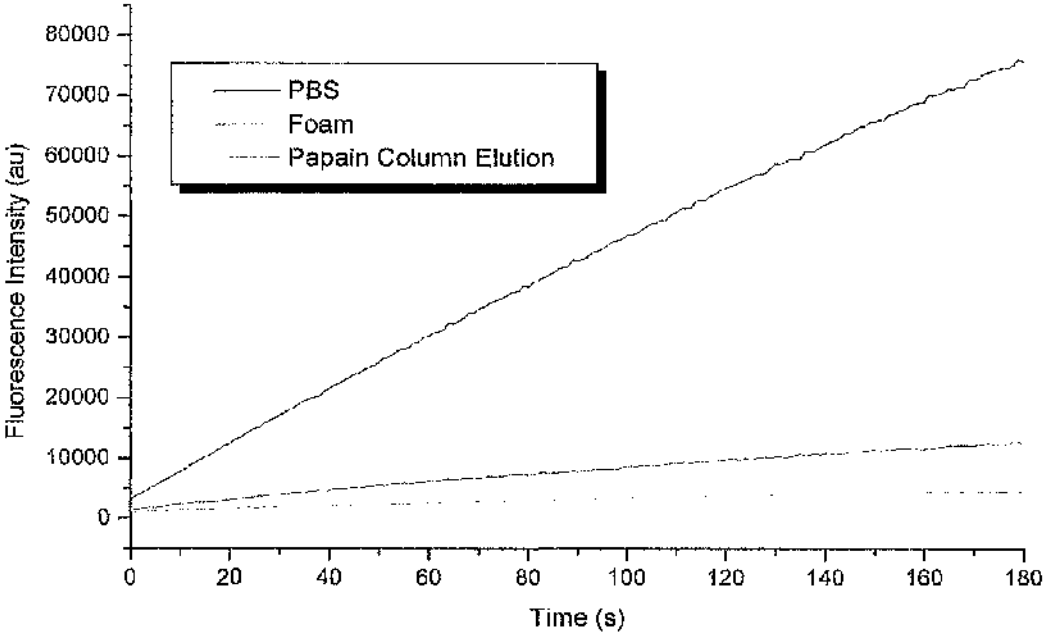


Figure 9.9. Cystatin activity of elution from papain affinity column repeat, via a fluorometric assay.

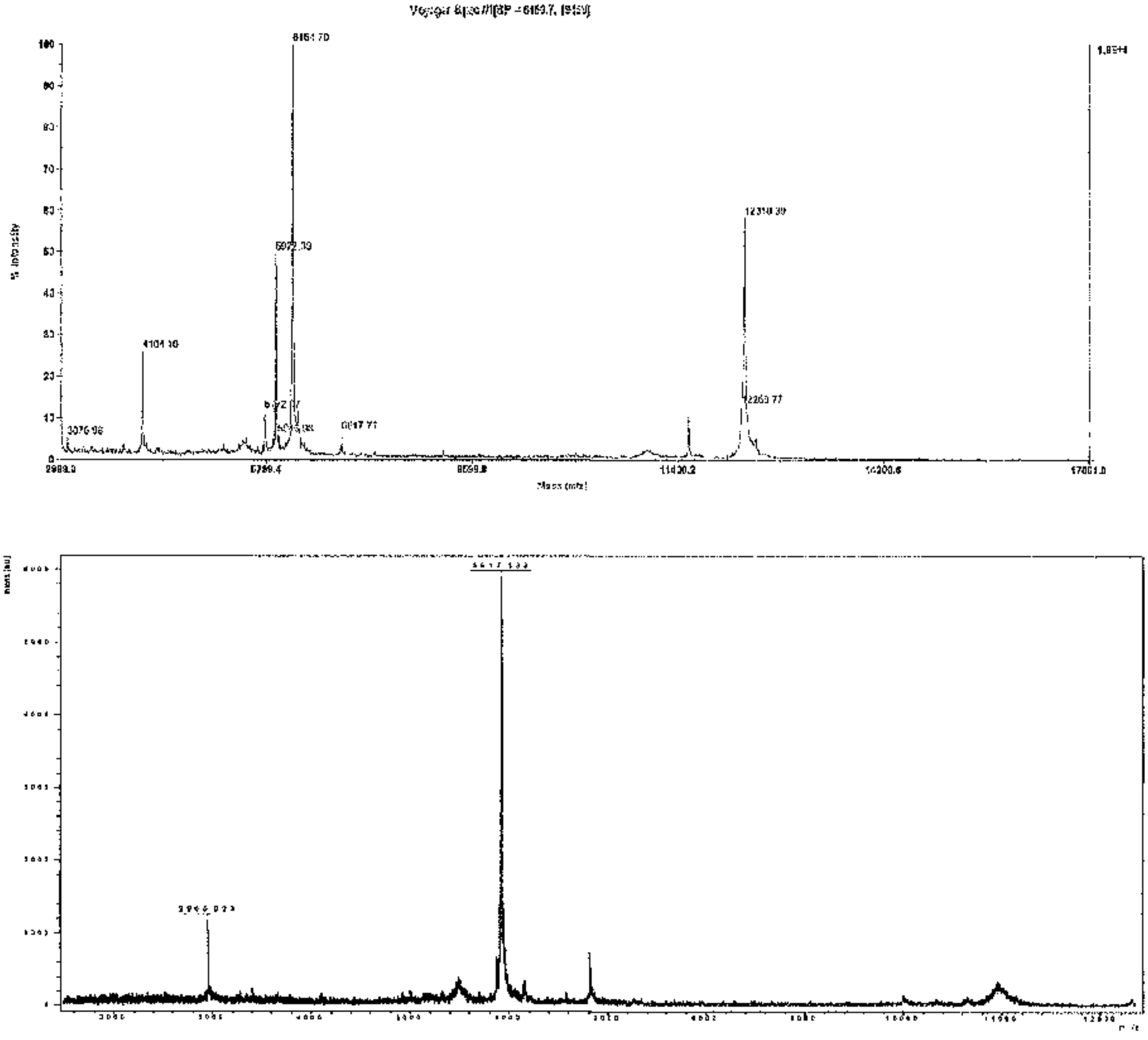


Figure 9.10. MALDI-TOF mass spectra of papain-binding component before (top) and after (bottom) lyophilisation and shipment to Brazil.

Papain-affinity chromatography was found to be an effective way of selectively isolating inhibitory peptides but could not be followed by a complete analysis in this case because of problems with reproducibility and lack of instrument time. However, there is great potential for this method, especially given the possibility of labelling an affinity column with practically any protein as a binding partner. The likelihood is that there are some peptide or polypeptide components produced in the foam that are active inhibitors of papain and cysteinyl proteinases, but they are probably targeted to inhibit a more environmentally relevant target. The investigation and identification of this target molecule, or organism(s), must be performed to reveal the true biological function.

9.3 Summary

P. pustulosus foam fluid was shown to be a potent inhibitor of papain. Although there was some indication that recombinant RSN-1 and RSN-2 may have been partially responsible for this action, this was not proven. From the results here, it appears that RSN-2 is not an inhibitor of papain, although it may be an inhibitor of some other enzyme found in the natural foam environment.

Analysis of individual foam components, separated by size, did not show any inhibitory activity. However, the preparation and use of a papain-affinity column allowed for the successful separation of some papain-inhibitory foam components. Some peptide fragments of these components were sequenced by mass spectrometry, but these components were not found to be stable enough for complete structure elucidation here.

10 Conclusions and Future Work

10.1 Conclusions

Proteins from the tropical frog *P. pustulosus*, were successfully separated and analysed in a context of their probable biological functions. A proteomic and chromatographic separation and identification strategy was used to identify a large number of peptide and polypeptide fragments that were present in various abundances, which dramatically add to the six proteins previously sequenced and the known carbohydrate analysis. The large number of peptides is typical of amphibian secretions and they are almost certainly functional to some degree, possibly anti-microbial in nature although this was not investigated here. These results prove the fantastic complexity of the visually simple foam material.

RSN-2, a particularly intriguing protein, was successfully expressed in a bacterial system as a recombinant fusion protein, reliably and in research quantities. The ease with which the protein was expressed in bacteria suggests that there are no complicated modifications or genetic manipulations seen in some other amphibian proteins. The recombinant fusion protein was shown to be susceptible to degradation by enteropeptidase which provoked an alteration to the expression and purification protocols, but was otherwise found to be stable at neutral pH and low temperature for a number of months. By successfully isolating the recombinant protein in a pure form, a detailed analysis on its structure and function was possible.

Recombinant RSN-2 was found to be highly surface active and was shown to reduce the surface tension of water by 20-30 mNm^{-1}. Biophysical analysis of the protein monolayers suggested that RSN-2 existed in equilibrium between the bulk solution and the air-water interface, whereby it was extremely favourable to have monolayer coverage of protein molecules with the remainder being favoured in the bulk. Although no definitive surface structures were obtained through AFM analysis, the ability for concentrated RSN-2 solutions to provide Langmuir-Blodgett films and depositions on solid, hydrophilic substrates was demonstrated. The attraction of RSN-2 to the interface and the relative hydrophobicity of the protein at the interface compared with the bulk, was visualised by employing a fluorescent probe in two-photon fluorescence microscopy, which implies a two-state structural model.

NMR was used to calculate the solution structures of RSN-2 and to probe its dynamic activities at neutral pH and the temperatures associated with the natural tropical environment. The solution structure was shown to consist of a core of a kinked 5-turn helix cradled by a 4-strand sheet, and these observations were in good agreement with circular dichroism studies. The central core of the structure appeared to be relatively rigid in solution, except for the fourth strand and the turn between strand three and four. These regions were poorly defined from the NMR data and proved problematic in fitting to models for amide nitrogen magnetic relaxation. Where cloning artefacts were present at the N-terminus, a large degree of flexibility was observed, with a length of polypeptide of approximately twenty residues showing little fixed secondary structure and a high level of solvent exposure, inferred by loose structural definition and enhanced mobility. At the caboxy-terminus the poly-Asp$_6$ motif was found to participate in an undefined loop structure which is anchored to the rest of the protein by the terminal tyrosine. These Asp residues are surface exposed and create a highly negatively charged ridge around a corner of the protein.

Some heterogeneity was observed in certain NMR signals which were proposed to be associated with residues in a slow conformational exchange in the vicinity, which was not exposed in the relaxation data. It is hypothesised that these observations are a result of the cis-trans (approximately 1:6) isomerisation of Pro14. Amide nitrogen relaxation analysis did suggest that there were some complicated dynamic processes occurring in the residue that made up the loop between the third and fourth strand. It is postulated that these motions include rapid solvent exchange and fast motions that are seen in other turns, but also a relatively slow underlying dynamic, possibly resulting from cis-trans isomerisation of the Pro78 residue. The inclusion of NOE contacts that were potentially from both forms may have been the reason for the unusual kink in the fourth strand.

Whilst some areas of poor definition were observed, the core of the protein was calculated as a tightly grouped ensemble of structures, corrected for interaction with explicitly defined aqueous solvent. Examination of the electrostatic surface, calculated by Poisson-Boltzmann solutions, suggested that the protein folded in a conventional manner with a compact hydrophobic centre and a highly charged solvent-exposed shell. These observations did not give any insight to a mechanical basis for surface activity, with the structure bearing no resemblance to the classical charged-head and oily-tail structure. Instead, the structure lent weight to the hypothesis that RSN-2 undergoes some reversible unfolding or rearrangement when assembled, or prior to assembly at the air-water interface. It is proposed that the link between the helix and the first strand may act as a hinge for unfolding, but there is no clear driving force to make such a mechanism energetically favourable.

Despite the lack of any similar primary sequences found in the public databases, RSN-2 was observed to adopt a conventional fold and almost certainly falls into the MEROPS I25 family. Unlike most other members of this family, RSN-2 does not inhibit the peptidase activity of papain and is readily cleaved by enteropeptidase and thrombin. However, the structure of RSN-2 showed such similarities to a wide variety of different peptidase inhibitors that it is quite possible that it is active in the foam against a specific degrading enzyme, found in the natural environment or secreted from parasitic intruders.

Within the overall foam structure, RSN-2 appears to provide the surface activity required to produce the bubbles in the structure but was not able to singularly produce the same stability observed in the native material. Accordingly, it is proposed that the data fits well with the structure previously postulated[62], where the RSN-2 monolayer film at the interface is supported by an underlying matrix, perhaps from lectin-carbohydrate interactions. Furthermore, RSN-2 possibly adds to the anti-microbial cocktail found in the mixture of proteins and peptides diffused around the foam interior. In this model, RSN-2 holds an extremely important role and can be seen as not just a simple protein surfactant, but as a protein with multiple purposes and abilities, most especially its ability to interact cohesively with other components.

10.2 Future Work

From the discoveries in this body of work, much has been revealed on the structure and function of RSN-2. There are two areas which still require attention to complete the understanding of such a complex protein; these are the structural rearrangements between the solution and surface structure and the potential of an inhibition target in the native environment.

Whilst some effort has already been made in examining the structure at the air water interface and some results from this are still pending, it has proved to be a challenging area. Whereas the field of examining surface films is reasonably well developed, proteins pose particular problems as they are soluble in the bulk phase as well as attracted to the interface, and are less inclined to adhere to solid surfaces whilst the possibility of being dissolved in solution remains a possibility. The difficulty of finding a substrate hydrophilic enough to simulate water in an aqueous environment, means that producing solid-bound monolayer depositions may require a change of tact. It may be possible to produce a deposition on a hydrophilic surface by simple evaporation but this is difficult to control and unlikely to result in a realistic monolayer. Further developments in this area may allow controlled depositions (e.g. protein printing) but it is more likely that RSN-2 will have to be deposited on a hydrophobic surface so that it can be imaged with the air mimicking the water at the air water interface. If some preliminary information can be used to suggest relative orientation at the surface (by neutron reflection, perhaps) then it may be possible to covalently immobilise RSN-2 on a solid surface.

Additional studies could also be made on the foam structure by raising antibodies against RSN-2 (and the same applies for any other protein). These antibodies could be easily conjugated to fluorescent, luminescent or some other detectable probe, for detailed analysis of protein localisation. For example, a primary anti-RSN-2 antibody could be conjugated with FITC (fluorescein isothiocyanate) and imaged with confocal microscopy. By employing anti-mouse (or whatever was appropriate) secondary antibodies a huge range of imaging or affinity investigations would be possible.

From the NMR performed here, the solution structure is well defined but further experiments could tighten the areas of loose definition, or provide further information on residues with overlapping NMR signals. Dihedral bond angle data and residual dipolar coupling information could be added to the structure calculations, as well as NOESY spectra with higher signal to noise ratios. Due to the well folded structure of RSN-2, these additions would probably be unnecessary without further reason, but if it became necessary for some further structural information to be gained, a crystal structure would probably be more useful. The unusual poly-Asp motif may be more easily visualised using a crystal structure as the overlapping signals would no longer be a problem. However, the lack of dispersal of the NMR peaks, is evidence in itself that these residues are not involved in a highly ordered local structure.

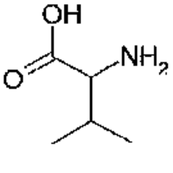
From the data shown here, it is clear that some part or parts of the foam fluid inhibit cysteine proteinases and are generally antimicrobial, but the active components were not unambiguously identified. Affinity chromatography showed some potential for resolving these issues and it could be an interesting avenue to investigate. If the pathogen, parasite or foreign object was selected from the environment, the inhibitor could be relatively easily found. Alternatively, this complex system may be particularly amenable to high throughput microfluidic technology where a library of potential binding partners could be immobilised and any interaction could be observed, such as Biacore where surface plasmon resonance is used to detect the refractive index change brought about by binding. If particular components could be found to bind to individual proteins then some clue would be provided as to biological functions.

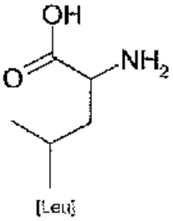
In addition to this, the five other *Ranaspumins* could be produced recombinantly and studied in detail individually, as described here for RSN-2. These studies could be initiated reasonably easily due to the existence of clones, but an efficient and stable expression system would need to be found as none of the other five proteins have been found to be suitable for use with the plasmid used for RSN-2. If these proteins were studied, and individual functions derived, then the understanding of the foam material would be greatly advanced. The remaining work would involve characterisation of the peptide and carbohydrate components and then, once individual components were well understood, their interactions with each other could be characterised. It is likely that the complete understanding of such a complex and unusual system would yield further interesting results from both a curiosity and application driven position.

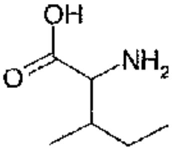
11 Appendices

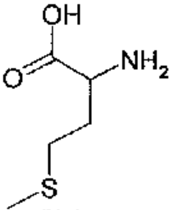
11.1 Appendix A

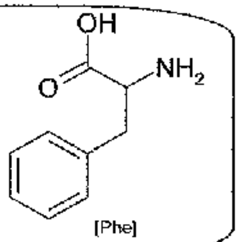
HYDROPHOBIC


[Val]

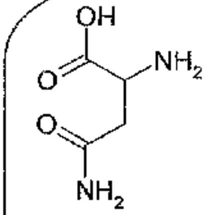

[Leu]

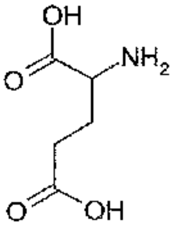

[Ile]

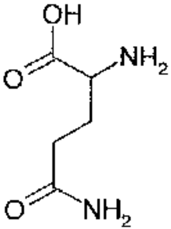

[Met]

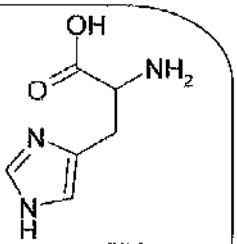

[Phe]

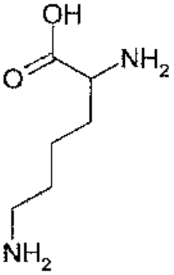
HYDROPHILIC

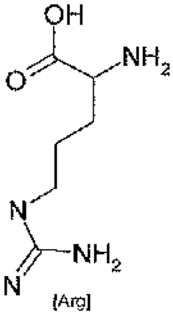

[Asn]

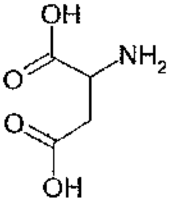

[Glu]


[Gln]

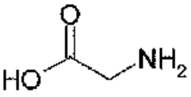

[His]

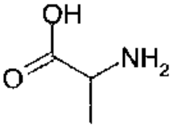

[Lys]

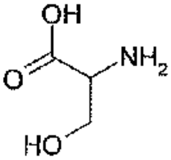

[Arg]

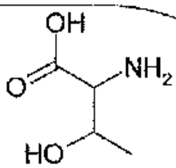

[Asp]

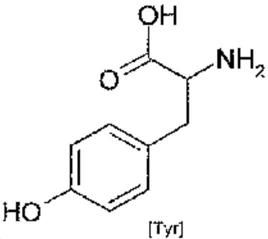
NEUTRAL HYDROPATHICITY

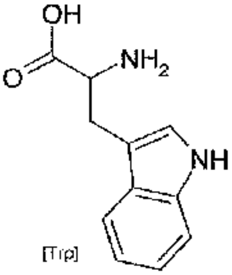

[Gly]

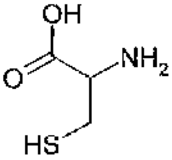

[Ala]

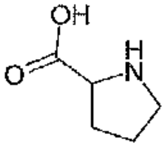

[Ser]


[Thr]


[Tyr]


[Trp]


[Cys]


[Pro]

11.2 Appendix B

Amino acid sequences of the six *Ranaspumin* proteins from *P. pustulosus*.

RSN-1
GGGNIGGGAKLGPEKPATPGIQDLLKSLLSVLNLSPPAIPE
DAEAVSYRDAKNGKFRLIKIHLGGELYCHVKQIAGPILALP
IVSDVVEVTGKECGKTEDDPLEDFPIP

RSN-2
LDGDLLKDKLKLPVIDNLFGKELLDKFQDDVKDKYGVDT
KDLKILKTSEDKRFYYVSVDAGDGEKCKFKIRKDVDVPKM
VGRKCRKDDDDDDGY

RSN-3
IDPTGLVQILLLEQVVHKIPPGNINLARTGIATQDSDYTAS
AVPSEARLAIDGNRNSDFNQKSCSHTGGNEPAWWRLELKK
KSKISVVVIAIRSDCCMDRFKGAELRIGNSQDATVNPICGK
VSAVKGSNYLFCCDGMEGKYISVVIPDRHEFLSLCEVEVY
GAKPIEGTHCK

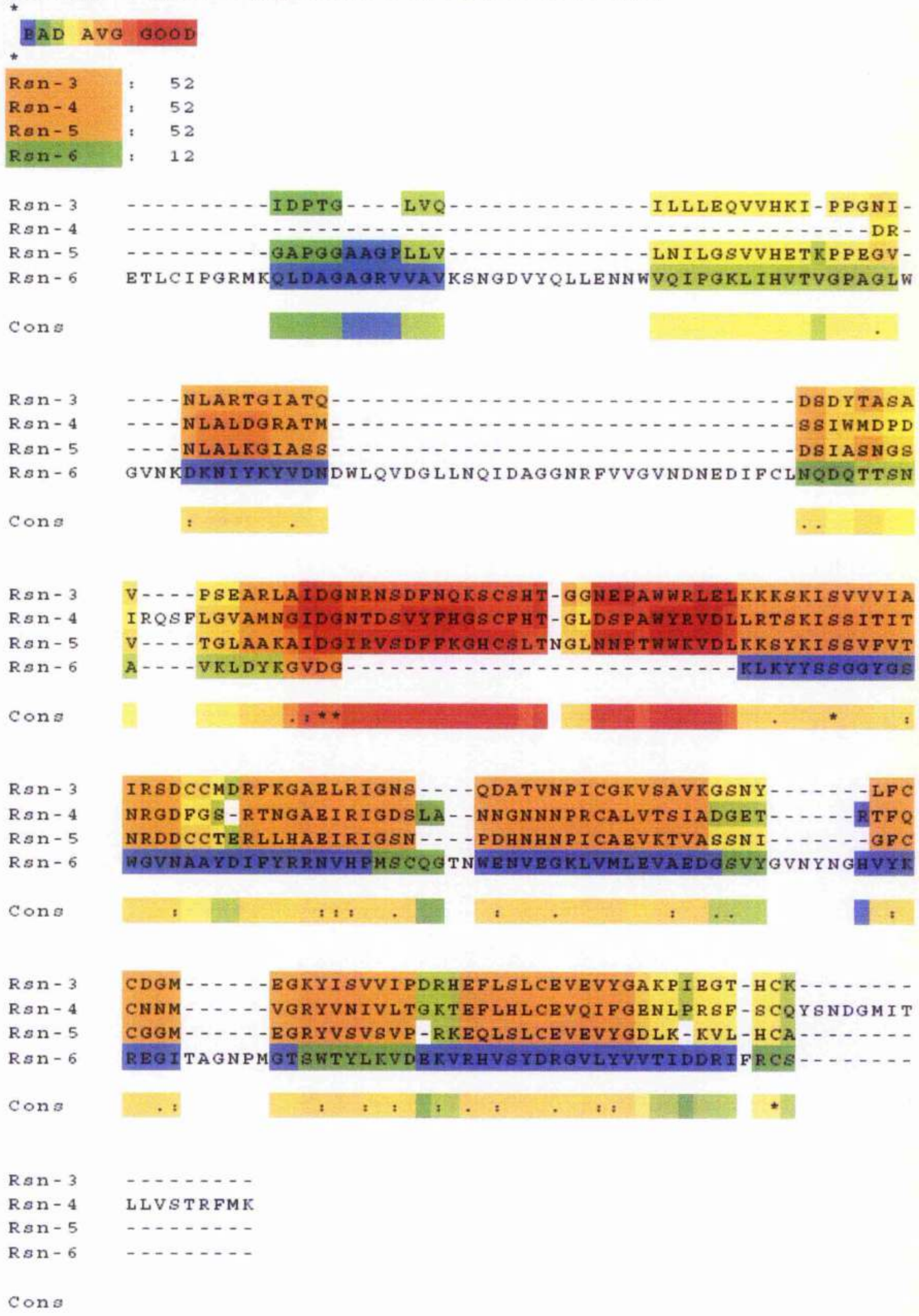
RSN-4
DRNLALDGRATMSSIWMDPDIRQSFLGVAMNGIDGNTDSV
YFHGSCFHTGLDSPAWYRVDLLRTSKISSITITNRGDFGSR
TNGAEIRIGDSLANNGNNNPRCALVTSIADGETRTFQCNN
MVGRYVNIVLTGKTEFLHLCEVQIFGENLPRSFSCQYSND
GMITLLVSTRFMK

RSN-5
GAPGGAAGPLLVLNILGSVVHETKPPEGVNLALKGIASSDS
IASNGSVTGLAAKAIDGIRVSDFFKGHCSLTNGLNNPTWW
KVDLKKSYKISSVFVTNRDDCCTERLLHAEIRIGSNPDHNH
NPICAEVKTVASSNIGFCCGGMEGRYVSVSVPRKEQLSLCE
VEVYGDLKKVLHCA

RSN-6
ETLCIPGRMKQLDAGAGRVVAVKSNGDVYQLLENNWVQI
PGKLIHVTVGPAGLWGVNKDKNIYKYVDNDWLQVDGLLN
QIDAGGNRFVVGVNDNEDIFCLNQDQTTSNAVKLDYKGV
DGKLKYYSSGGYGSWGVNAAYDIFYRRNVHPMSCQGTNW
ENVEGKLVMLEVAEDGSVYGVNYNGHVYKREGITAGNPM
GTSWTYLKVDEKVRHVSYDRGVLYVVTIDDRIFRCS

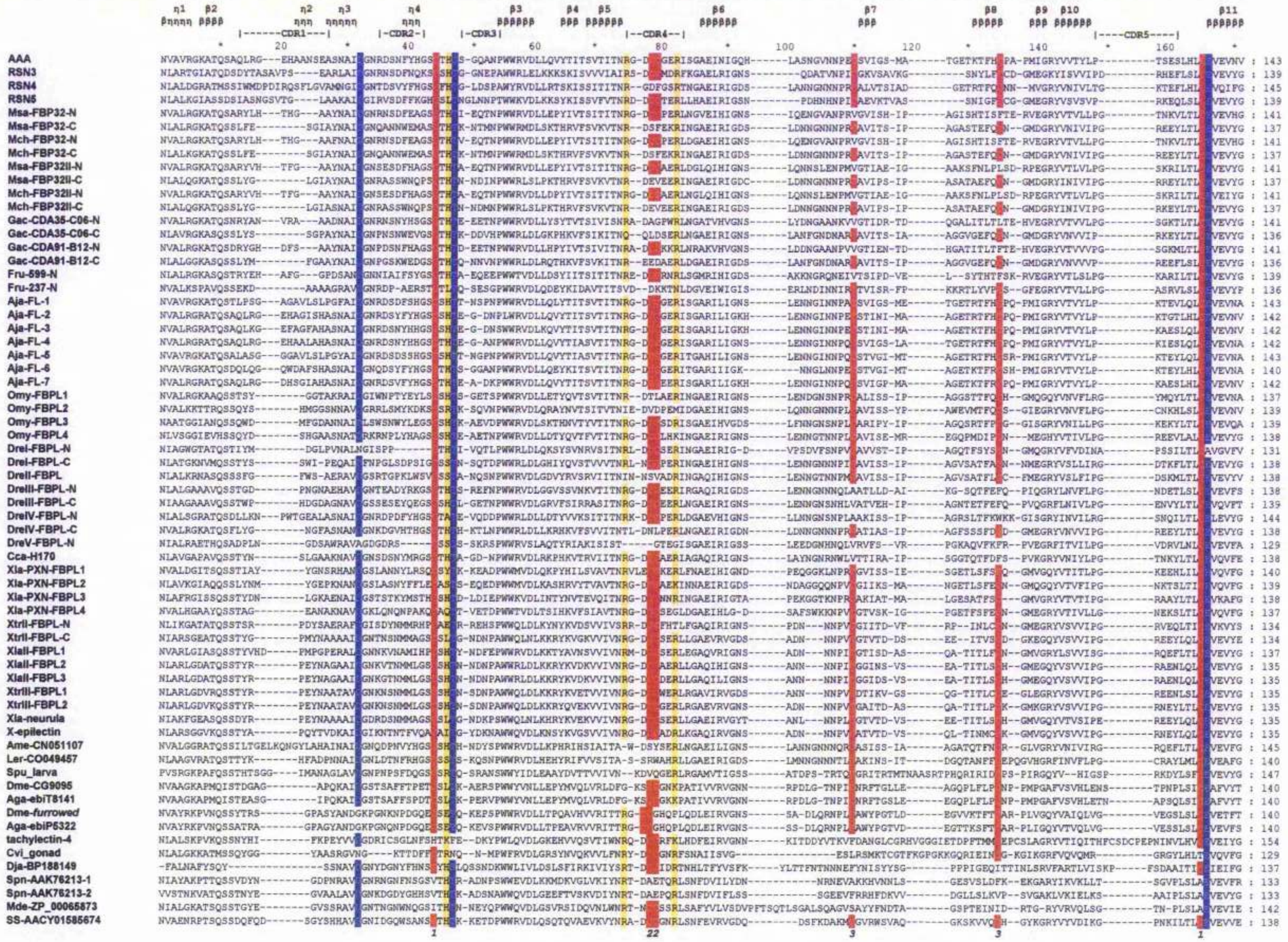
11.3 Appendix C

A multiple sequence alignment of the four foam lectins.

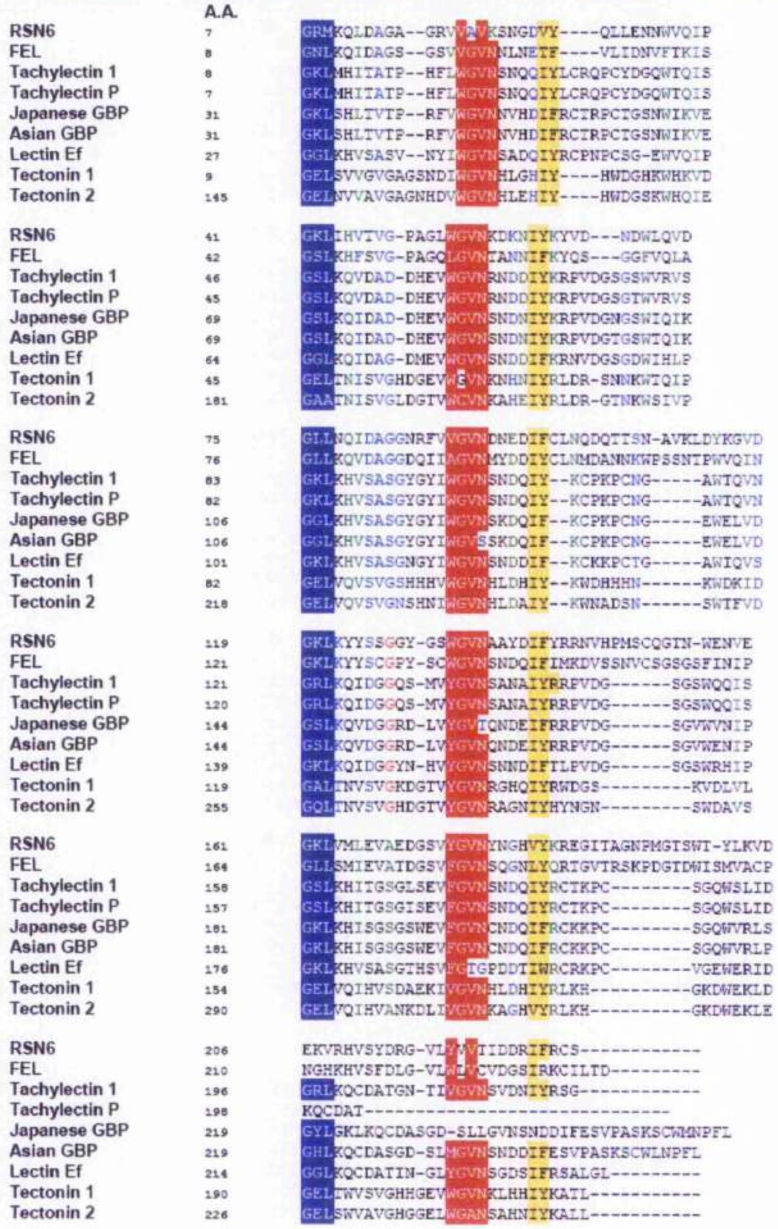


```
*
BAD AVG GOOD
*
Rsn-3   :  52
Rsn-4   :  52
Rsn-5   :  52
Rsn-6   :  12

Rsn-3   ----------IDPTG----LVQ--------------ILLLEQVVHKI-PPGNI-
Rsn-4   ---------------------------------------------------DR-
Rsn-5   ----------GAPGGAAGPLLV--------------LNILGSVVHETKPPEGV-
Rsn-6   ETLCIPGRMKQLDAGAGRVVAVKSNGDVYQLLENNWVQIPGKLIHVTVGPAGLW

Cons                                                       .

Rsn-3   ----NLARTGIATQ--------------------------------DSDYTASA
Rsn-4   ----NLALDGRATM--------------------------------SSIWMDPD
Rsn-5   ----NLALKGIASS--------------------------------DSIASNGS
Rsn-6   GVNKDKNIYKYVDNDWLQVDGLLNQIDAGGNRFVVGVNDNEDIFCLNQDQTTSN

Cons        :      .                                  ..

Rsn-3   V----PSEARLAIDGNRNSDFNQKSCSHT-GGNEPAWWRLELKKKSKISVVVIA
Rsn-4   IRQSFLGVAMNGIDGNTDSVYFHGSCFHT-GLDSPAWYRVDLLRTSKISSITIT
Rsn-5   V----TGLAAKAIDGIRVSDFFKGHCSLTNGLNNPTWWKVDLKKSYKISSVFVT
Rsn-6   A----VKLDYKGVDG---------------------------KLKYYSSGGYGS

Cons               .:**                             .   *    :

Rsn-3   IRSDCCMDRFKGAELRIGNS----QDATVNPICGKVSAVKGSNY-------LFC
Rsn-4   NRGDFGS-RTNGAEIRIGDSLA--NNGNNNPRCALVTSIADGET------RTFQ
Rsn-5   NRDDCCTERLLHAEIRIGSN----PDHNHNPICAEVKTVASSNI-------GFC
Rsn-6   WGVNAAYDIFYRRNVHPMSCQGTNWENVEGKLVMLEVAEDGSVYGVNYNGHVYK

Cons       :         :::  .      :   .       :  ..          :

Rsn-3   CDGM------EGKYISVVIPDRHEFLSLCEVEVYGAKPIEGT-HCK--------
Rsn-4   CNNM------VGRYVNIVLTGKTEFLHLCEVQIFGENLPRSF-SCQYSNDGMIT
Rsn-5   CGGM------EGRYVSVSVP-RKEQLSLCEVEVYGDLK-KVL-HCA--------
Rsn-6   REGITAGNPMGTSWTYLKVDEKVRHVSYDRGVLYVVTIDDRIFRCS--------

Cons      .:         :  : :  : . :   .  ::          *

Rsn-3   ---------
Rsn-4   LLVSTRFMK
Rsn-5   ---------
Rsn-6   ---------

Cons
```

A multiple sequence alignment of the F-type lectins.

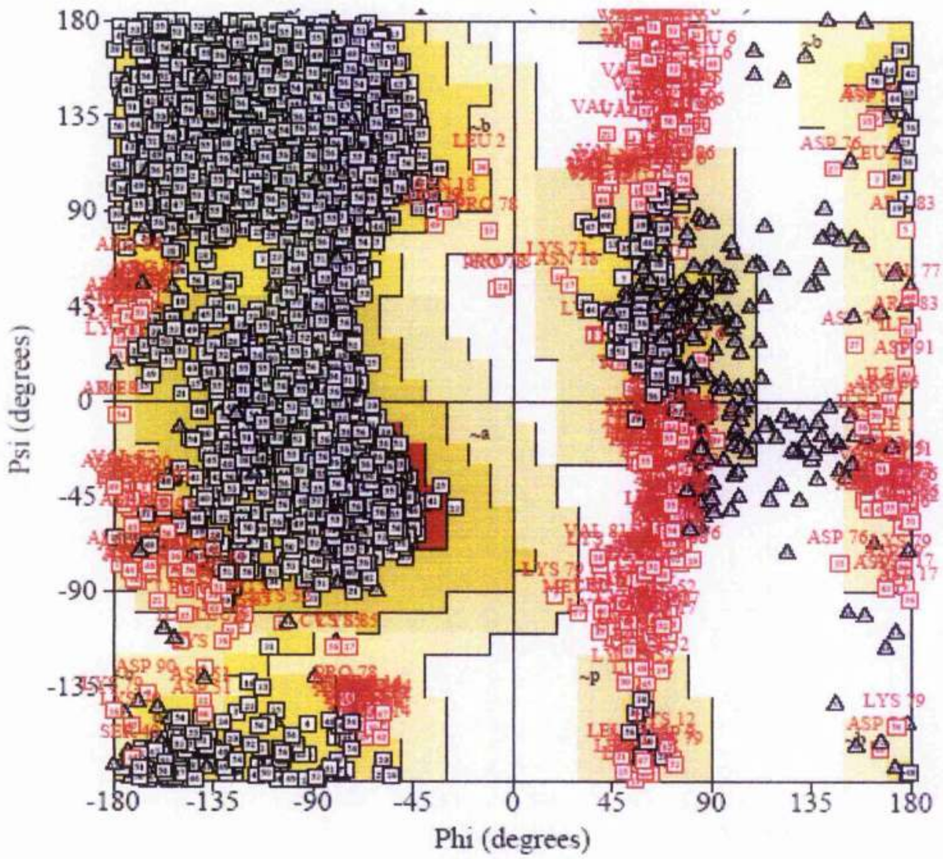


A multiple sequence alignment of the C-type lectins.



```
                 A.A.
RSN6             7     GRMKQLDAGA--GRVVAVKSNGDVY----QLLENNWVQIP
FEL              8     GNLKQIDAGS--GSVVGVNNLNETF----VLIDNVFTKIS
Tachylectin 1    8     GKLMHITATP--HFLWGVNSNQQIYLCRQPCYDGQWTQIS
Tachylectin P    7     GKLMHITATP--HFLWGVNSNQQIYLCRQPCYDGQWTQIS
Japanese GBP     31    GKLSHLTVTP--RFVWGVNNVHDIFRCTRPCTGSNWIKVE
Asian GBP        31    GKLSHLTVTP--RFVWGVNNVHDIFRCTRPCTGSNWIKVE
Lectin Ef        27    GGLKHVSASV--NYIWGVNSADQIYRCPNPCSG-EWVQIP
Tectonin 1       9     GELSVVGVGAGSNDIWGVNHLGHIY----HWDGHKWHKVD
Tectonin 2       145   GELNVVAVGAGNHDVWGVNHLEHIY----HWDGSKWHQIE

RSN6             41    GKLIHVTVG-PAGLWGVNKDKNIYKYVD---NDWLQVD
FEL              42    GSLKHFSVG-PAGQLGVNTANNIFKYQS---GGFVQLA
Tachylectin 1    46    GSLKQVDAD-DHEVWGVNRNDDIYKRPVDGSGSWVRVS
Tachylectin P    45    GSLKQVDAD-DHEVWGVNRNDDIYKRPVDGSGTWVRVS
Japanese GBP     69    GSLKQIDAD-DHEVWGVNSNDNIYKRPVDGNGSWIQIK
Asian GBP        69    GSLKQIDAD-DHEVWGVNSNDNIYKRPVDGTGSWTQIK
Lectin Ef        64    GGLKQIDAG-DMEVWGVNSNDDIFKRNVDGSGDWIHLP
Tectonin 1       45    GELTNISVGHDGEVWGVNKNHNIYRLDR-SNNKWTQIP
Tectonin 2       181   GAATNISVGLDGTVWCVNKAHEIYRLDR-GTNKWSIVP

RSN6             75    GLLNQIDAGGNRFVVGVNDNEDIFCLNQDQTTSN-AVKLDYKGVD
FEL              76    GLLKQVDAGGDQIIAGVNMYDDIYCLNMDANNKWPSSNTPWVQIN
Tachylectin 1    83    GKLKHVSASGYGYIWGVNSNDQIY--KCPKPCNG-----AWTQVN
Tachylectin P    82    GKLKHVSASGYGYIWGVNSNDQIY--KCPKPCNG-----AWTQVN
Japanese GBP     106   GGLKHVSASGYGYIWGVNSKDQIF--KCPKPCNG-----EWELVD
Asian GBP        106   GGLKHVSASGYGYIWGVSSKDQIF--KCPKPCNG-----EWELVD
Lectin Ef        101   GKLKHVSASGNGYIWGVNSNDDIF--KCKKPCTG-----AWIQVS
Tectonin 1       82    GELVQVSVGSHHHVWGVNHLDHIY--KWDHHHN------KWDKID
Tectonin 2       218   GELVQVSVGNSHNIWGVNHLDAIY--KWNADSN------SWTFVD

RSN6             119   GKLKYYSSGGY-GSWGVNAAYDIFYRRNVHPMSCQGTN-WENVE
FEL              121   GKLKYYSCGPY-SCWGVNSNDQIFIMKDVSSNVCSGSGSFINIP
Tachylectin 1    121   GRLKQIDGGQS-MVYGVNSANAIYRRPVDG------SGSWQQIS
Tachylectin P    120   GRLKQIDGGQS-MVYGVNSANAIYRRPVDG------SGSWQQIS
Japanese GBP     144   GSLKQVDGGRD-LVYGVTQNDEIFRRPVDG------SGVWVNIP
Asian GBP        144   GSLKQVDGGRD-LVYGVNQNDEIYRRPVDG------SGVWENIP
Lectin Ef        139   GKLKQIDGGYN-HVYGVNSNNDIFTLPVDG------SGSWRHIP
Tectonin 1       119   GALTNVSVGKDGTVYGVNRGHQIYRWDGS---------KVDLVL
Tectonin 2       255   GQLTNVSVGHDGTVYGVNRAGNIYHYNGN---------SWDAVS

RSN6             161   GKLVMLEVAEDGSVYGVNYNGHVYKREGITAGNPMGTSWT-YLKVD
FEL              164   GLLSMIEVATDGSVFGVNSQGNLYQRTGVTRSKPDGTDWISMVACP
Tachylectin 1    158   GSLKHITGSGLSEVFGVNSNDQIYRCTKPC--------SGQWSLID
Tachylectin P    157   GSLKHITGSGISEVFGVNSNDQIYRCTKPC--------SGQWSLID
Japanese GBP     181   GKLKHISGSGSWEVFGVNCNDQIFRCKKPC--------SGQWVRLS
Asian GBP        181   GKLKHISGSGSWEVFGVNCNDQIFRCKKPC--------SGQWVRLP
Lectin Ef        176   GKLKHVSASGTHSVFGTGPDDTIWRCRKPC--------VGEWERID
Tectonin 1       154   GELVQIHVSDAEKIVGVNHLDHIYRLKH----------GKDWEKLD
Tectonin 2       290   GELVQIHVANKDLIVGVNKAGHVYRLKH----------GKDWEKLE

RSN6             206   EKVRHVSYDRG-VLYMVTIDDRIFRCS------------
FEL              210   NGHKHVSFDLG-VLWLVCVDGSIRKCILTD---------
Tachylectin 1    196   GRLKQCDATGN-TIVGVNSVDNIYRSG------------
Tachylectin P    198   KQCDAT------------------------------
Japanese GBP     219   GYLGKLKQCDASGD-SLLGVNSNDDIFESVPASKSCWMNPFL
Asian GBP        219   GHLKQCDASGD-SLMGVNSNDDIFESVPASKSCWLNPFL
Lectin Ef        214   GGLKQCDATIN-GLYGVNSGDSIFRSALGL---------
Tectonin 1       190   GELTWVSVGHHGEVWGVNKLHHIYKATL-----------
Tectonin 2       226   GELSWVAVGHGGELWGANSAHNIYKALL-----------
```

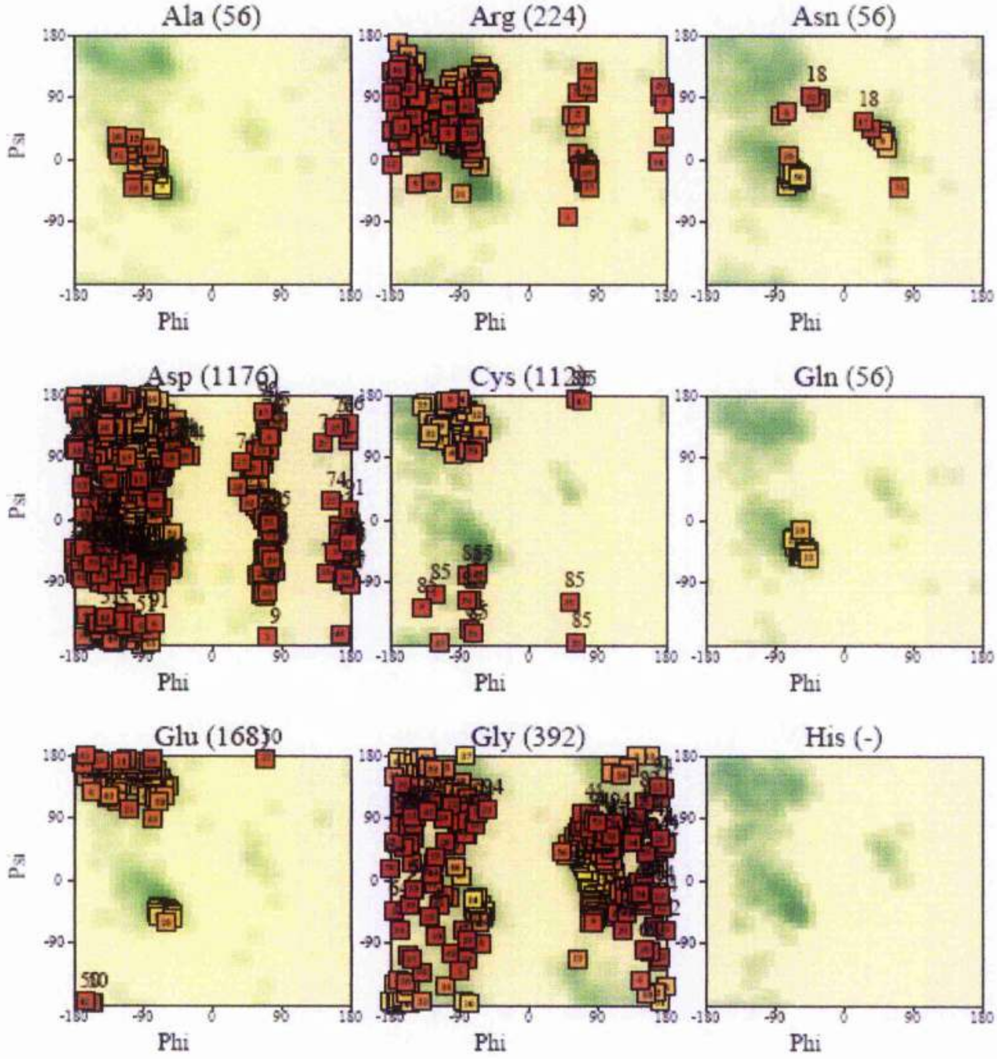
11.4 Appendix D



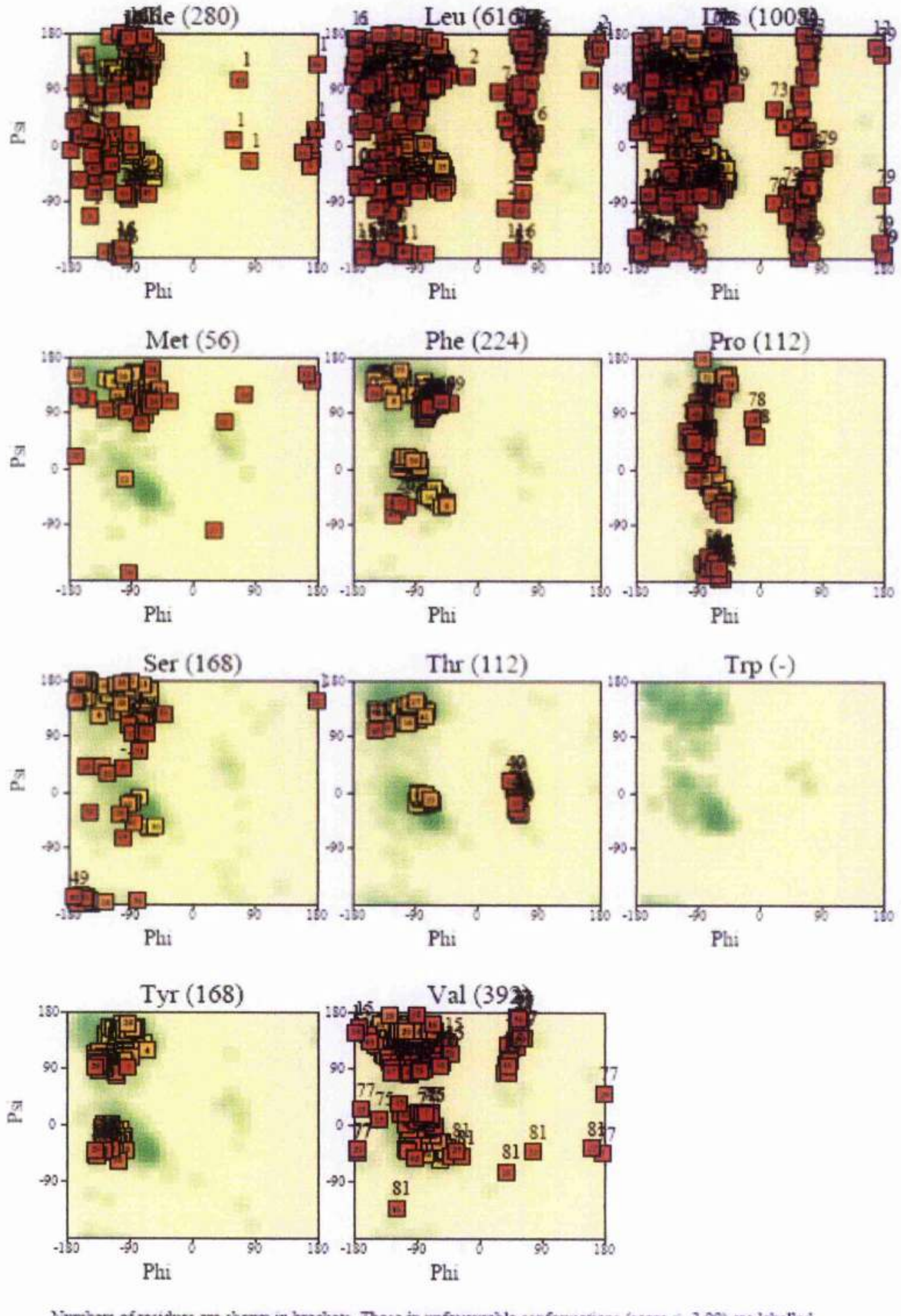
Plot statistics

Residues in most favoured regions [A,B,L]	3219	66.1%
Residues in additional allowed regions [a,b,l,p]	1304	26.8%
Residues in generously allowed regions [~a,~b,~l,~p]	215	4.4%
Residues in disallowed regions	134	2.8%
Number of non-glycine and non-proline residues	4872	100.0%
Number of end-residues (excl. Gly and Pro)	56	
Number of glycine residues (shown as triangles)	448	
Number of proline residues	112	
Total number of residues	5488	

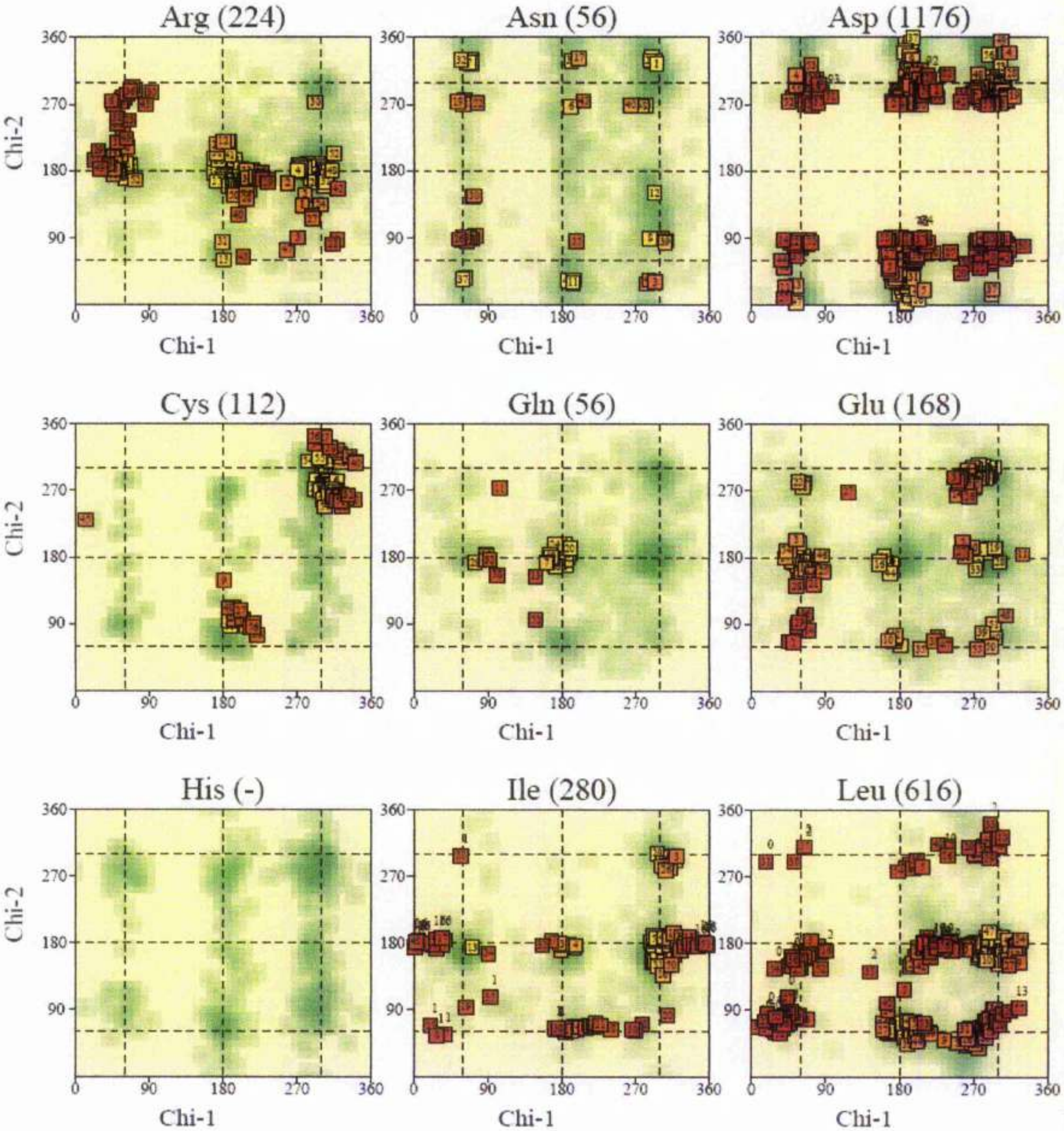
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.
Model numbers shown inside each data point.

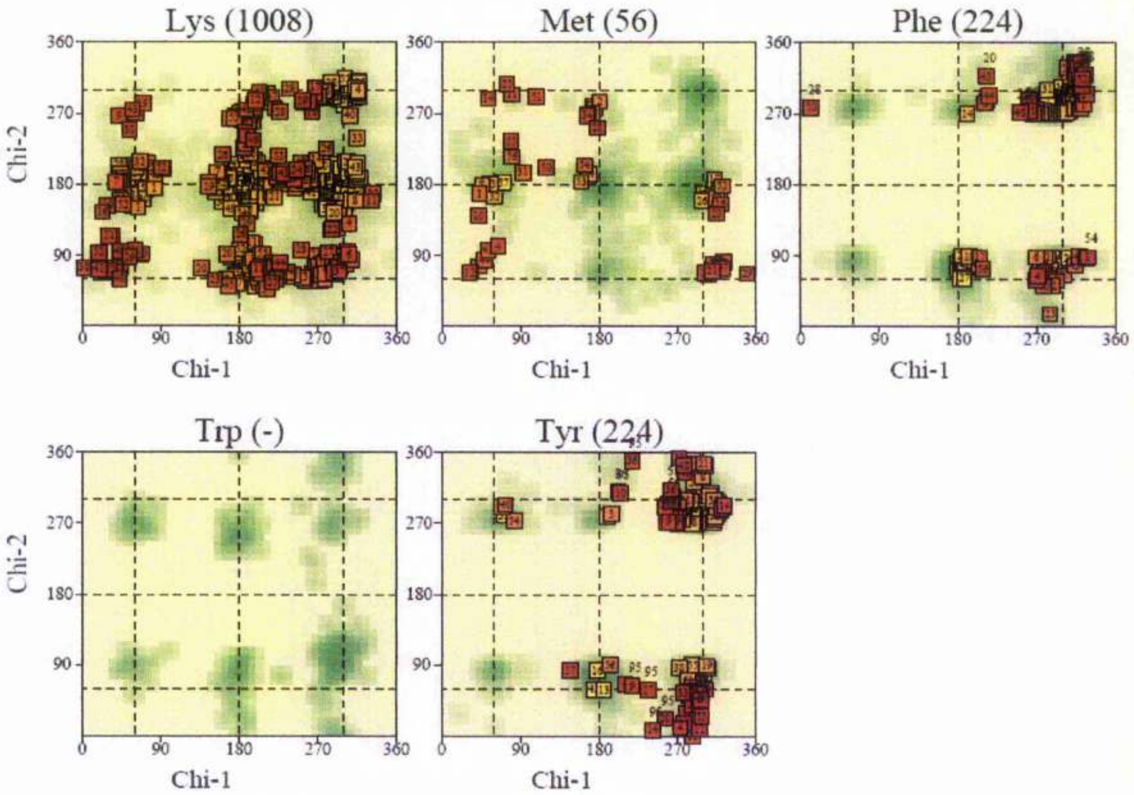


Numbers of residues are shown in brackets. Those in unfavourable conformations (score < -3.00) are labelled.
Shading shows favourable conformations as obtained from an analysis of 163 structures at resolution 2.0A or better.
Model numbers shown inside each data point.

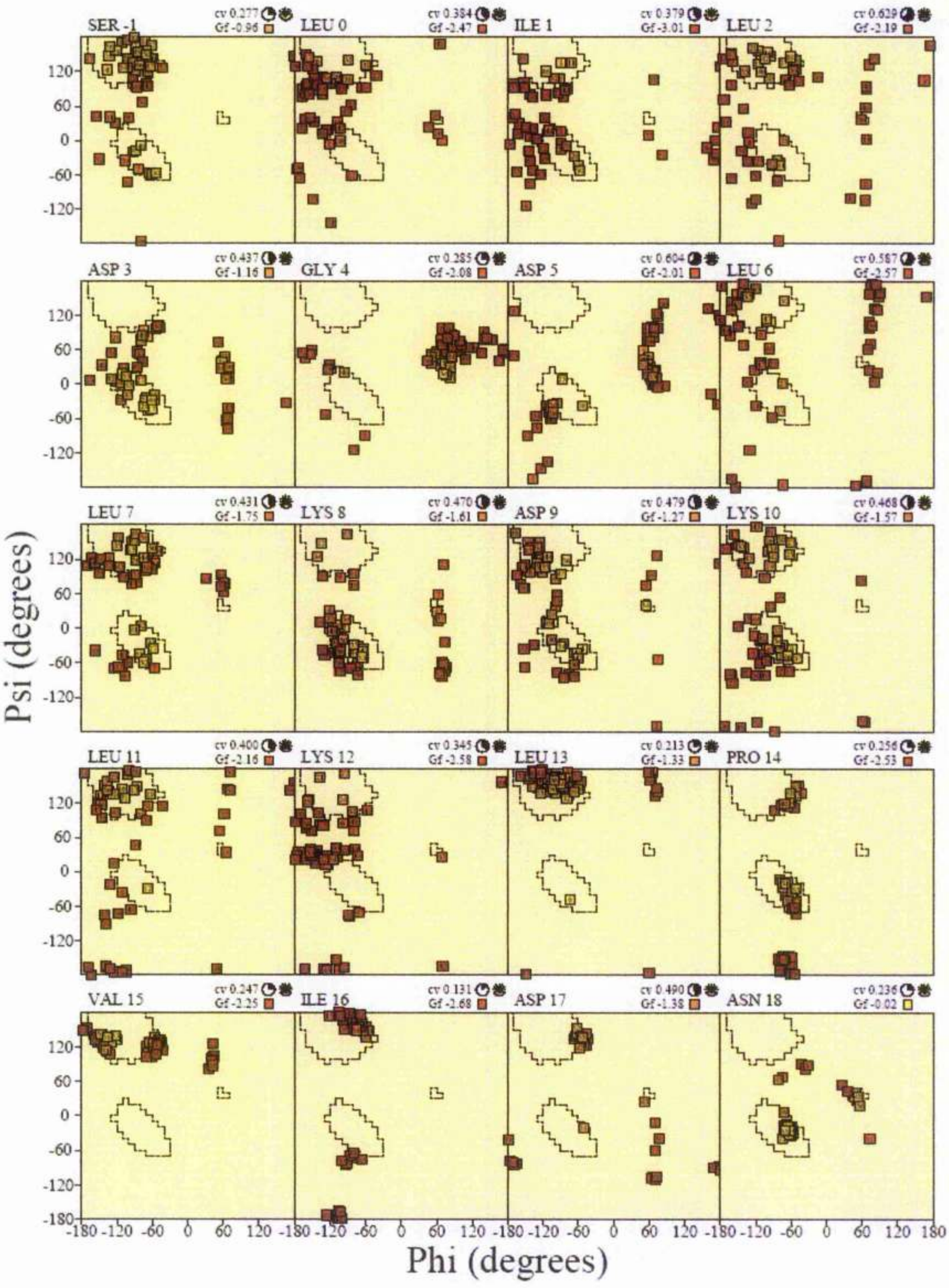


Numbers of residues are shown in brackets. Those in unfavourable conformations (score < -3.00) are labelled. Shading shows favourable conformations as obtained from an analysis of 163 structures at resolution 2.0A or better. Model numbers shown inside each data point.

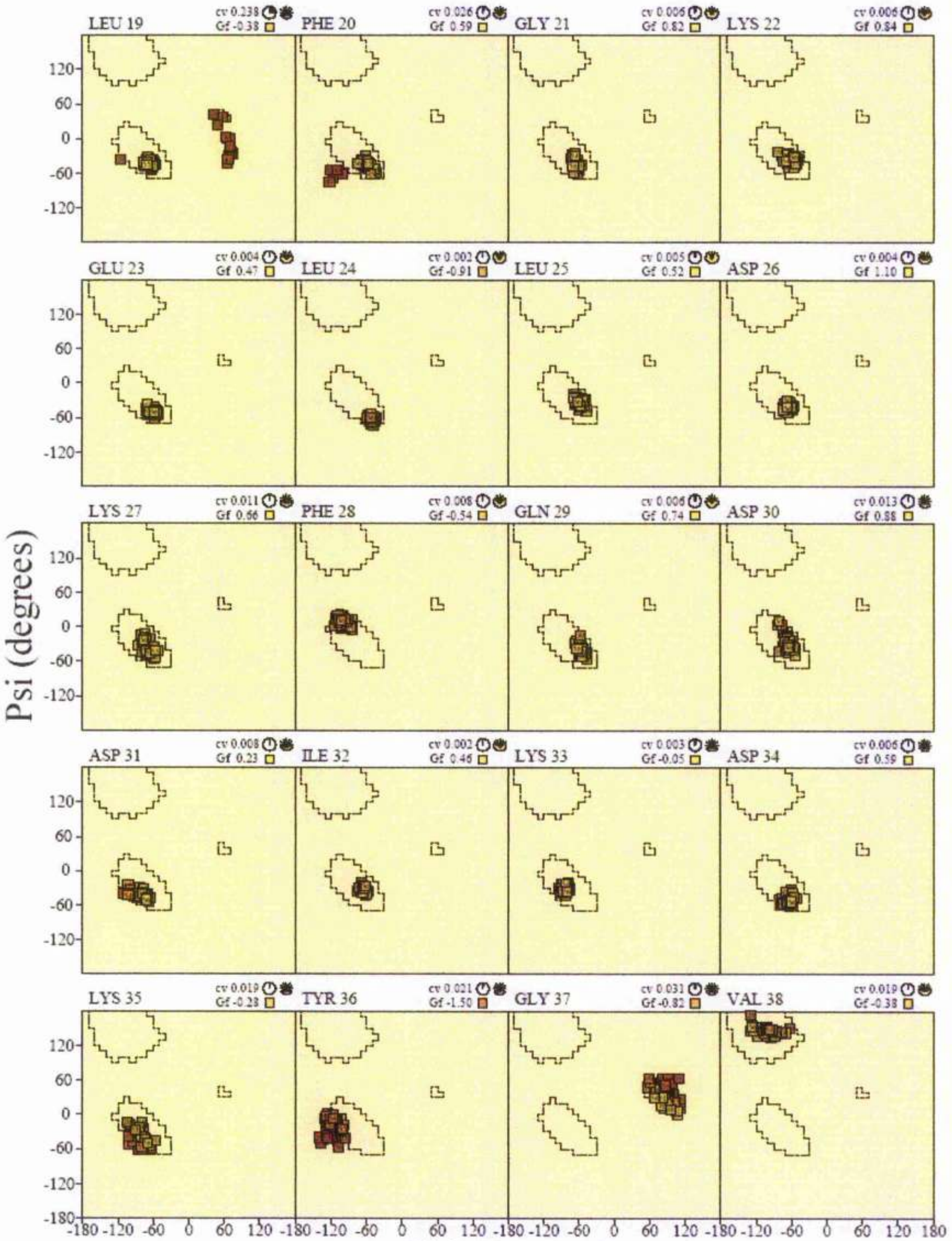

Arg (224)
Asn (56)
Asp (1176)
Cys (112)
Gln (56)
Glu (168)
His (-)
Ile (280)
Leu (616)
Chi-2
Chi-1
0
90
180
270
360



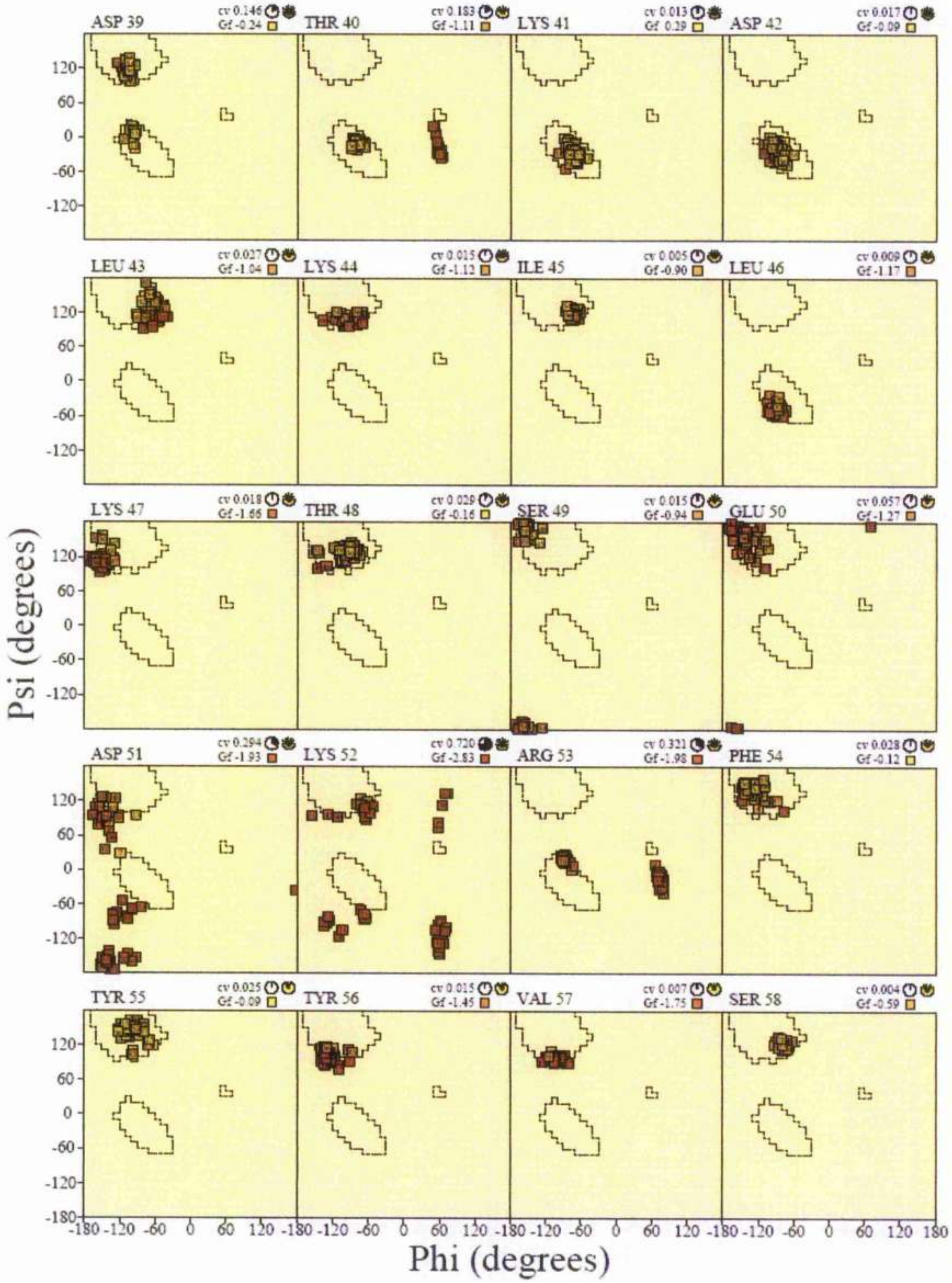
Numbers of residues are shown in brackets. Those in unfavourable conformations (score < -3.00) are labelled. Shading shows favourable conformations as obtained from an analysis of 163 structures at resolution 2.0A or better. Model numbers shown inside each data point.



cv = Circular Variance (low values signify high clustering of the data points). ✹ Accessible ◔ Buried
Gf = Average G-factor for the residue (the higher the value the more favourable the conformations) based on analysis of high-res. Xtal structures
Data points coloured according to G-factor: Favourable Unfavourable



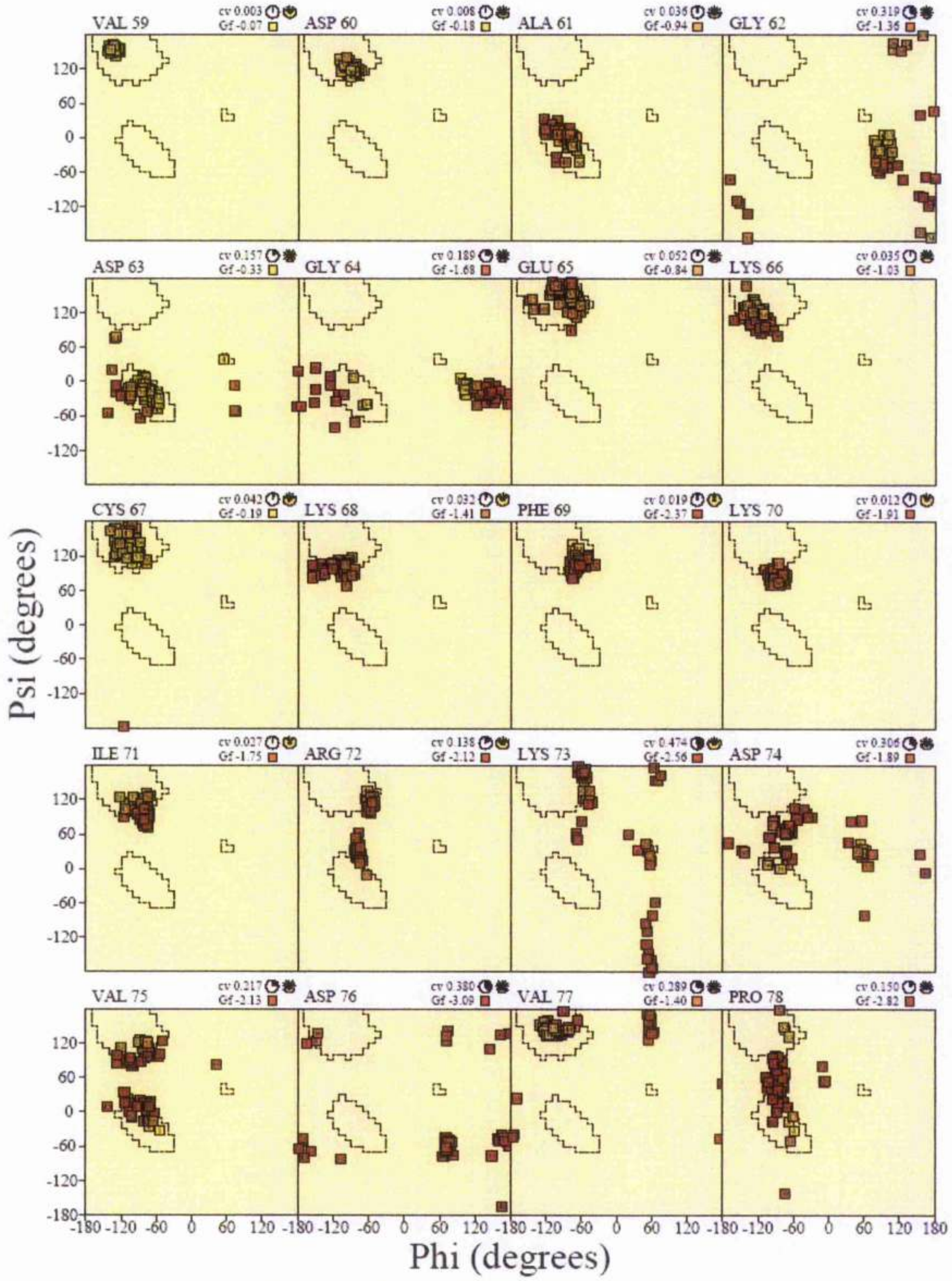
cv = Circular Variance (low values signify high clustering of the data points). ✱ Accessible ◔ Buried
Gf = Average G-factor for the residue (the higher the value the more favourable the conformations) based on analysis of high-res. Xstal structures
Data points coloured according to G-factor: Favourable Unfavourable



cv = Circular Variance (low values signify high clustering of the data points). Accessible Buried

Gf = Average G-factor for the residue (the higher the value the more favourable the conformations) based on analysis of high-res. Xstal structures

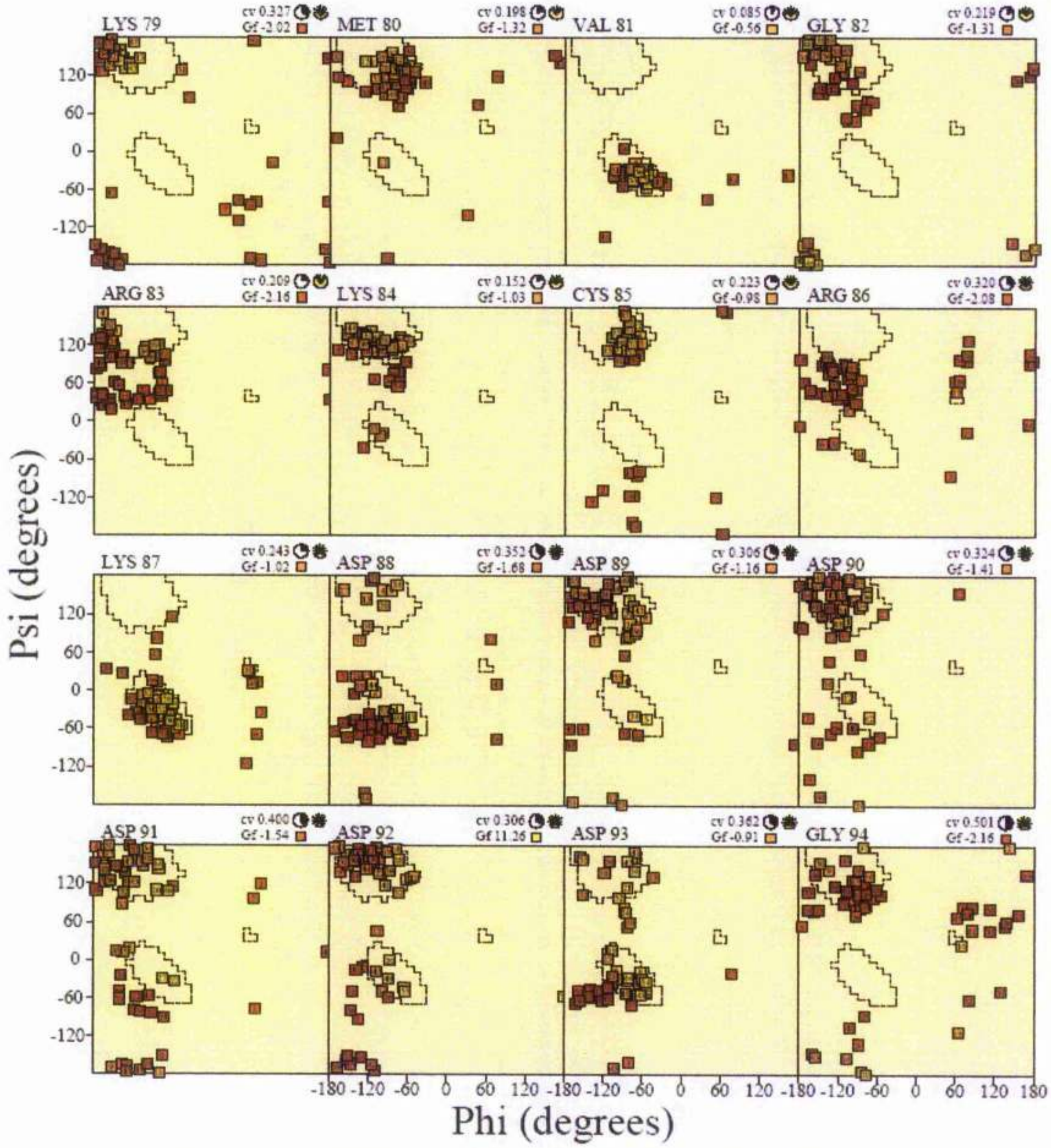
Data points coloured according to G-factor: Favourable Unfavourable



cv = Circular Variance (low values signify high clustering of the data points). ✹ Accessible ◐ Buried

Gf = Average G-factor for the residue (the higher the value the more favourable the conformations) based on analysis of high-res. Xstal structures

Data points coloured according to G-factor: Favourable Unfavourable



cv = Circular Variance (low values signify high clustering of the data points). ✸ Accessible ◉ Buried
Gf = Average G-factor for the residue (the higher the value the more favourable the conformations) based on analysis of high-res. Xstal structures
Data points coloured according to G-factor: Favourable Unfavourable

12 References

[1] Barr, R. G., Kurth, T., Stampfer, M. J., Buring, J. E., Hennekens, C. H., Gaziano, J. M. "Aspirin and Decreased Adult-Onset Asthma: Randomized Comparisons from the Physicians' Health Study" (2007) *Am. J. Respir. Crit. Care Med.* **175** 120-125

[2] Bai, R., Pettit, G. R., Hamel, E. "Receptor-mediated endocytosis of macromolecular conjugates in selective drug delivery" (1990) *Biochem Pharmacol* **39** 1941-1946

[3] Von Mehren, M., Balcerzak, S. P., Kraft, A. S., Edmonson, J. H., Okuno, S. H., Davey, M., McLaughlin, S., Beard, M. T., Rogatko, A. *Sarcoma* "Phase II Trial of Dolastatin-10, a Novel Anti-Tubulin Agent, in Metastatic Soft Tissue Sarcomas" (2004) *Sarcoma* **8** 107-111

[4] Francisco, J. A., Cerveny, C. G., Meyer, D. L., Mixan, B. J., Klussman, K., Chace, D. F., Rejniak, S. X., Gordon, K. A., DeBlanc, R., Toki, B. E., Law, C. L., Doronina, S. O., Siegall, C. B., Senter, P. D., Wahl, A. F. "cAC10-vcMMAE, an anti-CD30–monomethyl auristatin E conjugate with potent and selective antitumor activity" (203) *Blood* **102** 1458-1465

[5] Cassady, J. M., Chan, K. K., Floss, H. G., Leistner, E. "Recent Developments in the Maytansinoid Antitumor Agents" (2004) *Chem. Pharm. Bull.* **52** 1-26

[6] Fleming A. "On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of *B. influenzae*" (1929) *Br. J. Exp. Pathol.* **10** 226-236

[7] Ryadnov, M. G., Woolfson, D. N. "Engineering the morphology of a self-assembling protein fibre" (2003) *Nature Materials* **2** 329-332

[8] Savage, D., Mattson, G., Nielander, G., Morgensen, S., Conklin, E. "*Avidin-Biotin Chemistry: A Handbook*" **2nd Ed.** Pierce Chemical Co.

[9] Wilchek, M., Bayer, E. A., Livnah, O. "Essentials of biorecognition: The (strept)avidin–biotin system as a model for protein–protein and protein–ligand interaction" (2006) *Immunol. Lett.* **103** 27-32

[10] Nolting, B. Andert, K. "Mechanism of protein folding" (2000) *Proteins* **41** 288-298

[11] Kolb, V. A. "Cotranslational Protein Folding" (2001) *Mol. Biol.* **35** 584-590

[12] Levinthal, C. "*Mössbauer Spectroscopy in Biological Systems, Proceedings of a Meeting Held at Allerton House, Monticello, Illinois*" DeBrunner, P. and Munck, E. (Eds) 22-24, University of Illinois Press, Illinois (1969).

[13] Zwanzig, R., Szabo, A., Bagchi, B. "Levinthal's Paradox" (1992) *Proc. Nat. Acad. Sci.* **89** 20-22

[14] Shirts, M., Pande, V. S. "Screen Savers of the World Unite!" (2000) *Science* **290** 19031904

[15] Jayachandran, G., Vishal, V., Garcia, A. E., Pande, V. S. "Local structure formation in simulations of two small proteins" (2007) *J. Struct. Biol.* **157** 491-499

[16] Yang, W. Y., Gruebele, M. "Folding at the speed limit" (2003) *Nature* **423** 193-197

[17] Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M. V., Alonso, D. O. V., Daggett, V., Fersht A. R. "The complete folding pathway of a protein from nanoseconds to microseconds" (2003) *Nature* **421** 863-867

[18] Dobson, C. M., Sali, A., Karplus, M. "Protein Folding: A Perspective from Theory and Experiment" (1998) *Angew. Chem. Int. Ed. Eng.* **37** 868-893

[19] Onuchic, J. N. Wolynes, P. G. "Theory of protein folding" (2004) *Curr. Opin. Struct. Biol.* **14** 70-75

[20] Wolynes, P. G. "Recent successes of the energy landscape theory of protein folding and function" (2005) *Quart. Rev. Biophys.* **38** 405-410

[21] Vendruscolo, M., Paci, E., Dobson, C. M., Karplus, M. "Three key residues form a critical contact network in a protein folding transition state" (2001) *Nature* **409,** 641-645

[22] Dobson, C. M. "Protein folding and misfolding" (2003) *Nature* **426** 883-890

[23] Cheng, M.-Y., Hartl, F.-U., Martin, J., Pollock, R. A., Kalousek, F., Neupert, W., Hallberg, E. M., Hallberg, R. L., Horwich, A. L. Mitochondrial heat-shock protein hsp60 is essential for assembly of proteins imported into yeast mitochondria" (1989) *Nature* **337** 620-625

[24] Hartl, F. U., Hayer-Hartl, M. "Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein" (2002) *Science* **295** 1852-1858

[25] Berg, J. M., Tymoczko, J. L., Stryer, L. "*Biochemistry*" **5th Ed.**

[26] Pauling, L., Corey, R. B.,Branson, H. R. "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain" (1951) *Proc. Natl. Acad. Sci* **37,** 205-211

[27] Eisenberg, D. "The discovery of the {alpha}-helix and {beta}-sheet, the principal structural features of proteins" (2003) *Proc. Natl. Acad. Sci.* **100** 11207-11210

[28] Kabsch, W., Sander, C. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features" (2004) *Biopolymers* **22** 2577-2637

[29] http://wiz2.pharm.wayne.edu/biochem/main.html

[30] Sipe, J. D., Cohen, A. S. "Review: History of the Amyloid Fibril" (2000) *J. Struct. Biol.* **130** 88-98

[31] International Human Genome Sequence Consortium, "Finishing the euchromatic sequence of the human genome" (2004) *Nature* **431** 931-945

[32] Ramachandran, G. N., Sasisekharan, V., Ramakrishnan, C. "Stereochemistry of polypeptide chain configurations " (1963) *J. Mol. Biol.* 7 95-99

[33] Tanford, C. "Interfacial Free Energy and the Hydrophobic Effect" (1979) *Proc. Natl. Acad. Sci.***76** 4175-4176

[34] Bessette, P. H., Aslund, F., Beckwith, J., Georgiou, G. "Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm" (1999) *Proc. Natl. Acad. Sci.* **96** 13703-13708

[35] Paoli, M., Liddinngton, R., Tame, J., Wilkinson, A., Dodson, G. "Crystal Structure of T State Haemoglobin with Oxygen Bound At All Four Haems" (1996) *J. Mol. Biol.* **256** 775-792

[36] Long, S. B., Campbell, E. B., MacKinnon, R. "Crystal Structure of a Mammalian Voltage-Dependent Shaker Family K$^+$ Channel" (2005) *Science* **309** 897-903

[37] Berman, H. M., Westbrook, J., Feng, Z., Gillard, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. *Nuc. Acid. Res.* "The Protein Data Bank" (2000) **28** 235

[38] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., Thornton, J. M. "CATH – a hierarchic classification of protein domain structures" (1997) *Structure* **5** 1093-1108

[39] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. "SCOP: A structural classification of proteins database for the investigation of sequences and structures" (1995) *J. Mol. Biol.* **247** 536-540

[40] Day, R., Beck, D. A. C., Armen, R. S., Daggett, V. "A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary" (2003) *Prot. Sci.* **12** 2150-2160

[41] Boeger, H., Bushnell, D. A., Davis, R., Griesenbeck, J., Lorch, Y., Strattan, J. S., Westover, K. D., Kornberg, R. D. "Structural basis of eukaryotic gene transcription" (2005) *FEBS Lett.* **579** 899-903

[42] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., Phillips, D. C. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis" (1958) *Nature* **181** 662-666

[43] Barnes, W. J. P., Oines, C., Smith, J. M. *J. Comp. Physiol, A* "Whole animal measurements of shear and adhesive forces in adult tree frogs: insights into underlying mechanisms of adhesion obtained from studying the effects of size and scale" (2006) **192** 1179-1191

[44] Costanzo, J. P., Lee Jr., R. E. "Cryoprotection by urea in a terrestrially hibernating frog" (2005) *J. Exp. Biol.* **208** 4079-4089

[45] Weldon, C., du Preez, L. H., Hyatt, A. D., Muller, R., Speare, R. "Origin of the Amphibian Chytrid Fungus" (2004) *Emerg. Infect. Dis.* **10** 2100-2105

[46] Richardson, P. M., Chapman, J. "The Xenopus Tropicalis Genome Project" (2003) *Curr. Genomics* **4** 645-652

[47] Symula, R., Schulte, R., Summers, K. *Proc. R. Soc. Lond. B* "Molecular phylogenetic evidence for a mimetic radiation in Peruvian poison frogs supports a Müllerian mimicry hypothesis" (2001) **268** 2415-2421

[48] Darst, C. R., Cummings, M. E. "Predator learning favours mimicry of a less-toxic model in poison frogs" (2006) *Nature* **440** 208-211

[49] Smith, S. Q., Jones, T. H. "Tracking the cryptic pumiliotoxins" (2004) *Proc. Nat. Acad. Sci.* **101** 7841-7842

[50] Jilek, A., Mollay, C., Tippelt, C., Grassi, J., Mignogna, G., Müllegger, J., Sander, V., Fehrer, C., Barra, D., Kreil, G. "Biosynthesis of a D-amino acid in peptide linkage by an enzyme from frog skin secretions" (2005) *Proc. Nat. Acad. Sci.* **102** 4235-4239

[51] Conlon, J. M. "The therapeutic potential of antimicrobial peptides from frog skin." (2004) *Rev. Med. Microbiol.* **15** 17-25

[52] Brand, G. D., Leite, J. R. S. A., de Sa Mandel, S. M., Mesquita, D. A., Silva, L. P., Prates, M. V., Barbosa, E. A., Vinecky, F., Martins, G. R., Galasso, J. H., Kuckelhaus, S. A. S., Sampaio, R. N. R., Furtado Jr., J. R., Andrade, A. C., Bloch Jr., C. "Novel dermaseptins from *Phyllomedusa hypochondrialis* (Amphibia)" (2006) *Biochem. Biophys. Res. Commun.* **347** 739-746

[53] VanCompernolle, S. E., Taylor, R. J, Oswald-Richter, K., Jiang, J., Youree, B. E., Bowie, J. H., Tyler, M. J., Conlon, M., Wade, D., Aiken, C., Dermody, T. S. "Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T Cells" (2005) *Journal of Virology* **79** 11598-11606

[54] Brand, G. D., Krause, F. C., Silva, L. P., Leite, J. R. S. A., Melo, J. A. T., Prates, M. V., Pesquero, J. B., Santos, E. L., Nakaie, C. R., Costa-Neto, C. M., Bloch Jr., C. "Bradykinin-related peptides from *Phyllomedusa hypochondrialis*" (2006) *Peptides* **27** 2137-2146

[55] Petranka, J. W., Thomas, D. A. G. "Explosive breeding reduces egg and tadpole cannibalism in the wood frog, *Rana sylvatica*" (1995) *Anim. Behav.* **50** 731-739

[56] Rödel, M. O., Range, F., Seppänen, J. T., Noë, R. "Caviar in the rain forest: monkeys as frog-spawn predators in Taï National Park Ivory Coast" (2002) *J. Trop. Ecol.* **18** 289-294

[57] Laurence, W. F., McDonald, K. R., Speare, R. "Epidemic Disease and the Catastrophic Decline of Australian Rain Forest Frogs" (1996) *Conserv. Biol.* **10** 406-413

[58] Ryan, M. J., Fox, J. H., Wilczynski, W., Rand, A. S. "Sexual selection for sensory exploitation in the frog *Physalaemus pustulosus*" (1990) *Nature* **343** 66-67

[59] Bosch, J., Rand, A. S., Ryan, M. J. "Acoustic Competition in *Physalaemus pustulosus*, a Differential Response to Calls of Relative Frequency" (2000) *Ethology* **106** 865-871

[60] Dobkin, D. S., Gettinger, R. D. "Thermal Aspects of Anuran Foam Nests" (1985) *J. Herpetol.* **19** 271-275

[61] Shepard, D. B., Caldwell, J. P. "From Foam to Free-living: Ecology of Larval *Leptodactylus Labyrinthicus*" (2005) *Copeia* 803-811

[62] Cooper, A., Kennedy, M. W., Fleming, R. I., Wilson, E. H., Videler, H., Wokosin, D. L., Su, T., Green, R. J., Lu, J. R. "Adsorption of Frog Foam Nest Proteins at the Air-Water Interface" (2005) *Biophys. J.* **88** 2114-2125

[63] Fleming, R. I., Cooper, A., Kennedy, M. W. *Unpublished data*

[64] Morris, M. *The Daily Telegraph* 24th October 2006

[65] McMahon, S. A., Walsh, M. A., Ching, R. T. Y., Carter, L. G., Dorward, M., Johnson, K. A., Liu, H., Oke, M., Bloch Jr., C., Kennedy, M. W., Latiff, A. A., Cooper, A., Taylor, G. L., White, M. F., Naismith, J. H. "Crystallization of Ranasmurfin, a blue-coloured protein from *Polypedates leucomystax*" (2006) *Acta Cryst.* **62** 1124-1126

[66] Parry, S., Bhandal, J., Haslam, S., Dell, A. *Unpublished data*

[67] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, Z., Miller, W., Lipman, D. J. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs" (1997) *Nuc. Acids. Res.* **25** 3389-3402

[68] Kelley, L. A., MacCallum, R. M., Sternberg, M. J. "Enhanced Genome Annotation Using Structural Profiles in the Program 3D-PSSM" (2000) *J. Mol. Biol.* **299** 499-520

[69] Lehman, W. D., Bohne, A., Lieth, C. W. "The information encrypted in accurate peptide masses - improved protein identification and assistance in glycopeptide identification and characterization" (2000) *J. Mass. Spectrom.* **35** 1335-1341

[70] Kyte, J., Doolittle, R. F. "A simple method for displaying the hydropathic character of a protein" (1982) *J. Mol. Biol.* **157** 105-132

[71] Wierenga, P. A>, Meinders, M. B. J., Egmond, M. R., Voragen, A. G. J., de Jongh, H. H. J. "Quantitative Description of the Relation between Protein Net Charge and Protein Adsorption to Air-Water Interfaces" (2005) *J. Phys. Chem. B* **109** 16946

[72] Jaspe, J., Hagen, S. J. "Do Protein Molecules Unfold in a Simple Shear Flow?" (2006) *Biophys. J.* **91** 3415-3424

[73] Johansson, J., Curstedt, T. "Molecular structures and interactions of pulmonary surfactant components" (1997) *Eur. J. Biochem.* **244** 675-693

[74] Horne, D. S. "Casein structure, self-assembly and gelation" (2002) Curr. Opin. Colloid Interface Sci. 7 456-461

[75] Wösten, H. A. B., van Wetter, M. A., Lugones, L. G., van der Mei, H. C., Busscher, H. J., Wessels, J. G. "How a fungus escapes the water to grow into the air" (1999) H. *Curr. Biol.* **9** 85-88

[76] Wösten, H. A. "HYDROPHOBINS: Multipurpose Proteins" (2001) *Annu. Rev. Microbiol.* **55** 625-646

[77] Hakanpää, J., Paananen, A., Askolin, S., Nakari-Setälä, T., Parkkinen, T., Penttilas, M., Linder, M., Rouvinen, J. "Atomic Resolution Structure of the HFBII Hydrophobin, a Self-assembling Amphiphile" (2004) *J. Biol. Chem.* **279** 534-539

[78] Szilvay, G. R., Paananen, A., Laurikainen, K., Vuorimaa, E., Lemmetyinen, H., Peltonen, J., Linder, M. B. "Self-Assembled Hydrophobin Protein Films at the Air-Water Interface: Structural Analysis and Molecular Engineering" (2007) *Biochem.* **46** 2345-2354

[79] Bianchet, M. A., Odom, E. W., Vasta, G. R., Amzel, L. M. "A novel fucose recognition fold involved in innate immunity" (2002) *Nat. Struct. Mol. Biol.* **9** 628-634

[80] Odom, E. W., Vasta, G. R. "Characterization of a Binary Tandem Domain F-type Lectin from Striped Bass (*Morone saxatilis*)" (2006) *J. Biol. Chem.* **281** 1698-1713

[81] Cammarata, M., Benenati, G., Odom, E. W., Salerno, G., Vizzini, A., Vasta, G. R., Parrinello, N. "Isolation and characterization of a fish F-type lectin from gilt head bream (*Sparus aurata*) serum" (2006) *Biochim. Biophys. Acta* **1770** 150-155

[82] Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., Higgins, D. G. "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools" (1997) *Nucleic Acids Res.* **25** 4876-4882

[83] Zhu, Y., Ng, P. M., Wang, L., Ho, B., Ding, J. L. "Diversity in lectins enables immune recognition and differentiation of wide spectrum of pathogens" (2006) *Int. Immunol.* **18** 1671-1680

[84] Galliano, M., Minchiotti, L., Campagnoli, M., Sala, A., Visai, L., Amoresano, A., Pucci, P., Casbarra, A., Cauci, M., Perduca, M., Monaco, H. L. "Structural and biochemical characterization of a new type of lectin isolated from carp eggs" (2003) *Biochem. J.* **376** 433-440

[85] Dodd, R. B., Drickamer, K. "Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity" (2001) *Glycobiol.* **11** 71R-79R

[86] Xie, J., Wen, J. J., Chen, B., Gui, J. F. "Differential gene expression in fully-grown oocytes between gynogenetic and gonochoristic crucian carps" (2001) *Gene* **271** 109-116

[87] Funayama, N., Nakatsukasa, M., Kuraku, S., Takechi, K., Dohi, M., Iwabe, N., Miyata, T., Agata, K. "Isolation of Ef silicatein and Ef lectin as molecular markers for sclerocytes and cells involved in innate immunity in the freshwater sponge *Ephydatia fluviatilis*." (2005) *Zool. Sci.* **22** 1113-1122

[88] Bradford, M. M. "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding" (1976) *Anal. Biochem.* **72** 248-254

[89] Laemmli, U. K. "Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4" (1970) *Nature* **227** 680-685

[90] Cleland, W. W. "Dithiothreitol, a New Protective Reagent for SH Groups" (1964) *Biochem* **3** 480-482

[91] Edman, P. "Mechanism of the Phenyl Isothiocyanate Degradation of Peptides" (1956) *Nature* **177** 667-668

[92] Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., Khorana, H. G. "Studies on polynucleotides: XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerases" (1971) *J. Mol. Biol.* **56** 341-361

[93] Liepnieks, J. J., Light, A. "The preparation and properties of bovine enterokinase" (1979) *J. Biol. Chem.* **254** 1677-1683

[94] Chang, J. Y. "Thrombin specificity. Requirement for apolar amino acids adjacent to the thrombin cleavage site of polypeptide substrate" (1985) *Eur. J. Biochem.* **151** 217-224

[95] Bjellqvist, B., Hughes, G., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J., Frutiger, S., Hochstrasser, D. "The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences" (1993) *Electrophoresis* **14** 1023-1031

[96] Builder, S. E. (Ed), "*Hydrophobic Interaction Chromatography, Principles and Methods*" **Edition AB** AmershamPharmacia Biotech

[97] Kaltschmidt, E., Wittmann, H.G. "Ribosomal proteins. VII : Two-dimensional polyacrylamide gel electrophoresis for fingerprinting of ribosomal proteins" (1970) *Anal. Biochem* **36** 401-412

[98] O'Farrell, P.H. "High resolution two-dimensional electrophoresis of proteins" (1975) *J. Biol. Chem.* **250** 4007-4021

[99] Neuhoff, V., Arold, N.,Taube, D., Ehrhardt, W. "Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250" (1988) *Electrophoresis* **9** 255-262

[100] Wilkins, M.R., Lindskog, I., Gasteiger, E., Bairoch, A., Sanchez, J.-C., Hochstrasser, D.F., Appel, R.D. "Detailed peptide characterization using PEPTIDEMASS - a World-Wide-Web-accessible tool" (1997) *Electrophoresis* **18** 403-408

[101] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J., Frutiger, S., Hochstrasser, D.F. "The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences" (1993) *Electrophoresis* **14** 1023-1031

[102] Purcell, E. M., Torrey, H. C., Pound. R. V. "Resonance Absorption by Nuclear Magnetic Moments in a Solid" (1946) *Phys. Rev.* **69** 37-38

[103] Bloch, F. "Nuclear Induction" (1946) *Phys. Rev.* **70** 460-474

[104] Bloch, F., Hansen, W. W., Packard, M. "The Nuclear Induction Experiment" (1946) *Phys. Rev.* **70** 474-485

[105] Cavanagh, J., Fairbrother, W. J., Palmer III, A. G., Skelton, N. J. "Protein NMR Spectroscopy: Principles and Practice" **1st Ed.**, 1996

[106] Roberts, G. C. K. (Ed.), "NMR of Macromolecules: A Practical Approach" 1993

[107] Wüthrich, K., Wider, G., Wagner, G., Braun, W. "Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance" *J. Mol. Biol.* (1982) **155** 311-319

[108] Pervushin, K., Riek, R., Wider, G., Wüthrich, K. "Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution" (1997) *Proc. Natl. Acad. Sci.* **94** 12366-12371

[109] Fiaux, J., Bertelsen, E. B., Horwich, A. L., Wüthrich, K. "NMR analysis of a 900K GroEL GroES complex" (2002) *Nature* **418** 207-211

[110] Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Ono, A. M., Güntert, P. "Optimal isotope labelling for NMR protein structure determinations" (2006) *Nature* **440** 52-57

[111] Sambrook, J., Fritsch, E.F., Maniatis, T. "*Molecular Cloning*" **2nd Ed.**, Vol. 3

[112] Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., Laue, E. D. "The CCPN data model for NMR spectroscopy: Development of a software pipeline" (2005) *Proteins* **59** 687-696

[113] Piotto, M., Saudek, V., Sklenář, V. "Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions" (1992) *J. Biomol. NMR* **2** 661-665

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

[115] Kay, L. E., Xu, G., Y., Singer, A. U., Muhandriam, D. R., Forman-Kay, J. D. "A Gradient-Enhanced HCCH-TOCSY Experiment for Recording Side-Chain ^{1}H and ^{13}C Correlations in H2O Samples of Proteins" (1993) *J. Magn. Reson. B* **101** 333-337

[116] Tjandra, N., Bax, A. "Measurement of Dipolar Contributions to $^1J_{CH}$ Splittings from Magnetic-Field Dependence of *J* Modulation in Two-Dimensional NMR Spectra" (1997) *J. Magn. Reson.* **124** 512-515

[117] Marion, D., Kay, L. E., Sparks, S. W., Torchia, D. A., Bax, A. "Three-Dimensional Heteronuclear NMR of ^{15}N-Labelled Proteins" (1989) *J. Am. Chem. Soc.* **111** 1515-1517

[118] Zuiderweg, E. R. P., McIntosh, L. P., Dahlquist, F. W., Fesik, S. W. "Three-dimensional -^{13}C-resolved proton NOE spectroscopy of uniformly ^{13}C-labelled proteins for the NMR assignment and structure determination of larger molecules" (1990) *J. Magn. Reson.* **86** 210-216

[119] Nilges, M. "Calculation of Protein Structures with Ambiguous Distance Restraints. Automated Assignment of Ambiguous NOE Crosspeaks and Disulphide Connectivities" (1995) *J. Mol. Biol.* **245** 645-660

[120] Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L. "Crystallography and NMR System (CNS): A new software system for macromolecular structure determination" (1998) *Acta Crystallographica D* **54** 905-921

[121] Linge J.P., Williams M.A., Spronk C.A.E.M., Bonvin A.M.J.J., Nilges M. "Refinement of protein structures in explicit solvent" (1995) *Proteins* **50** 496-506

[122] Nilges M., Macias M.J., ODonoghue S.I., Oschkinat H. "Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin" (1997) *J Mol Biol.* **269** 408-422

[123] Sprangers, R., Kay, L. E. "Quantitative dynamics and binding studies of the 20S proteasome by NMR" (2007) *Nature* **445** 618-622

[124] Kay, L. E., Torchia, D. A., Bax, A. "Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease" (1989) *Biochem.* **28** 8972-8979

[125] Lipari, G., Szabo, A. "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity" (1982) *J. Amer. Chem. Soc.* **104** 4546-4559

[126] Lipari, G., Szabo, A. "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results" (1982) *J. Amer. Chem. Soc.* **104** 4559-4570

[127] Grzeisiek, S., Bax, A. "The importance of not saturating H_2O in protein NMR. Application to sensitivity enhancement and NOE measurements" (1993) *J. Am. Chem. Soc.* **115** 12593-12594

[128] Palmer III, A. G., Rance, M., Wright, P. E. "Intramolecular motions of a zinc finger DNA-binding domain from Xfin characterised by protein-detected natural abundance ^{13}C heteronuclear NMR spectroscopy" (1991) *J. Am. Chem. Soc.* **113** 4371-4380

[129] Mandel, M. M., Akke, M., Palmer III, A. G. "Backbone Dynamics of *Escherichia coli* Ribonuclease HI: Correlations with Structure and Function in an Active Enzyme" (1995) *J. Mol. Biol.* **246** 144-163

[130] Tjandra, N., Feller, S. E., Pastor, R. W., Bax, A. "Rotational diffusion anisotropy of human ubiquitin ^{15}N relaxation" (1995) *J. Am. Chem. Soc.* **117** 12562-12566

[131] Nabuurs, S. B., Spronk, C. A. E. M., Vuister, G. W., Vriend, G. "Traditional Biomolecular Structure Determination by NMR Spectroscopy Allows for Major Errors" (2005) *PLoS Comp. Biol.* **2**

[132] Hooft, R. W. W., Vriend, G., Sander, C., Abola, E.E. "Errors in protein structures" (1996) *Nature* **381** 272

[133] Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., Thornton, J. M. "AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR" (2004) *J. Bio. NMR.* **8** 477-486

[134] Vriend, G. *J. Mol. Graph.* "WHAT IF: a molecular modeling and drug design program" (1990) **8** 52-6, 29

[135] Nabuurs, S. B., Nederveen, A. J., Vranken, W., Doreleijers, J. F., Bonvin, A. M., Vuister, G. W., Vriend, G., Spronk., A. E. M. "DRESS: a database of REfined solution NMR structures" (2004) *Proteins* **55** 483-486

[136] Hartshorn, M. J., Caves, L. S. D. *Unpublished program*

[137] Theobold, D. L., Wuttke, D. S. "Empirical Bayes hierarchical models for regularising maximum likelihood estimation in the matrix Gaussian Procrustes problem" (2006) *Proc. Nat. Acad. Sci.* **103** 18521-18527

[138] Theobald, D. L., Wuttke, D. S. "THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures" (2006) *Bioinformatics* **22** 2171-2172

[139] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. "SCOP: A structural classification of proteins database for the investigation of sequences and structures" (1995) *J. Mol. Biol.* **247** 536-540

[140] Schubert, M., Labudde, D., Oschkinat, H., Schmieder, P. "A software tool for the prediction of Xaa-Pro peptide bond conformation in proteins based on ^{13}C chemical shift statistics" (2002) *J. Biomol. NMR* **24** 149-154

[141] Baker, N. A., Sept, D., Joseph. S., Holst, M. J., McCammon, J. A. "Electrostatics of nanosystems: Application to microtubules and the ribosome" (2001) *Proc. Natl. Acad. Sci* **98** 10037-10041

[142] http://cti.itc.virginia.edu/~cmg/Demo/wheel/wheelApp.html

[143] Aagaard, A., Listwan, P., Cowieson, N., Huber, T., Ravasi, T., Wells, C.A., Flanagan, J.U., Kellie, S., Hume, D.A., Kobe, B., Martin, J.L. "An Inflammatory Role for the Mammalian Carboxypeptidase Inhibitor Latexin: Relationship to Cystatins and the Tumor Suppressor TIG1" (2005) *Structure* **13** 309-317

[144] Spadaccini, R., Crescenzi, O., Tancredi, T., De Casamassimi, N., Saviano, G., Scognamiglio, R., Di Donato, A., Temussi, P. A. "Solution structure of a sweet protein: NMR study of MNEI, a single chain Monellin" (2001) *J. Mol. Biol.* **305** 505-514

[145] Brown, W. M., Dziegielewska, K. M. "Friends and relations of the cystatin superfamily--new members and their evolution" (1997) *Prot. Sci.* **6** 5-12

[146] Otto, H., Schirmeister, T. "Cysteine Proteases and Their Inhibitors" (1997) *Chem. Rev.* **97** 133-172

[147] Grzonka, Z., Jankowska, E., Kasprzykowski, F., Kasprzykowska, R., Łankiewicz, L., Wiczk, W., Wieczerzak, E., Ciarkowski, J., Drabik, P., Janowski, R., Kozak, M., Jaskólski, M., Grubb, A. "Structural studies of cysteine proteases and their inhibitors" (2001) *Acta Biochim. Pol.* **48** 1-20

[148] Oliveira, A. S., Xavier-Filho, J. Sales, M. P. "Cysteine proteinases and cystatins" (2003) *Braz. Arch. Biol. Technol.* **46** 91-104

[149] Abrahamson, M., Alvarez-Fernandez, M., Nathanson, C. M. "Cystatins" (2003) *Biochem. Soc. Symp.* **70** 179-199

[150] Rzychon, M., Chmiel, D., Stec-Niemczyk, J. "Modes of inhibition of cysteine proteases" (2004) *Acta Biochim. Pol.* **51** 861-873

[151] Martin, J. R., Craven, C. J., Jerala, R., Kroon-Zitko, L., Zerovnik, E., Turk, V., Waltho, J. P. "The Three-dimensional Solution Structure of Human Stefin A" (1995) *J. Mol. Biol.* **246** 331-343

[152] Tate, S., Ushioda, T., Utsunomiya-Tate, N., Shibuya, K., Ohyama, Y., Nakano, Y., Kaji, H., Inagaki, F., Samejima, T., Kainosho, M. "Solution structure of a human cystatin A variant, cystatin A2-98 M65L by NMR spectroscopy. A possible role of the interactions between the N- and C-termini to maintain the inhibitory active form of cystatin A" (1995) *Biochem.* **34** 14637-14648

[153] Stubbs, M. T., Laber, B., Bode, W., Huber, R., Jerala, R., Lenarcic, B., Turk, V. "The refined 2.4 Å X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction" (1990) *EMBO J.* **9** 1939-1947

[154] Shimba, N., Kariya, E., Tate, S., Kaji, H., Kainosho, M. "Structural comparison between wild-type and P25S human cystatin A by NMR spectroscopy. Does this mutation affect the α-helix conformation?" (2000) *J.Struct. Funct. Genom.* **1** 26-42

[155] Jenko, S., Dolenc, I., Guncar, G., Dobersek, A., Podobnik, M., Turk, D. "Crystal Structure of Stefin A in Complex with Cathepsin H: N-terminal Residues of Inhibitors can Adapt to the Active Sites of Endo- and Exopeptidases" (2003) *J.Mol.Biol.* **326** 875-885

[156] Staniforth, R.A., Giannini, S., Higgins, L.D., Conroy, M.J., Hounslow, A.M., Jerala, R., Craven, C.J., Waltho, J.P. "Three-dimensional domain swapping in the folded and molten-globule states of cystatins, an amyloid-forming structural superfamily" (2001) *EMBO J.* **20** 4774-4781

[157] Žerovnik, E., Škerget, K., Tušek-Žnidarič, M., Loeschner, C., Brazier, M. W., Brown, D. R. "High affinity copper binding by stefin B (cystatin B) and its role in the inhibition of amyloid fibrillation" (2006) *FEBS J.* **273** 4250-4263

[158] Brode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A., Brzin, J., Kos, J., Turk, V. "The 2.0 A X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases" (1988) *EMBO J.* 7 2593-2599

[159] Dieckmann, T., Mitschang, L., Hofmann, M., Kos, J., Turk, V., Auerswald, E. A., Jaenicke, R., Oschkinat, H. "The Structures of Native Phosphorylated Chicken Cystatin and of a Recombinant Unphosphorylated Variant in Solution" (1993) *J. Mol. Biol.* **234** 1048-1059

[160] Alvarez-Fernandez, M., Liang, Y.H., Abrahamson, M., Su, X.D. "Crystal Structure of Human Cystatin D, a Cysteine Peptidase Inhibitor with Restricted Inhibition Profile" (2005) *J.Biol.Chem.* **280** 18221-18228

[161] Schuettelkopf, A. W., Hamilton, G., Watts, C., Van Aalten, D. M. F. "Structural Basis of Reduction-dependent Activation of Human Cystatin F" (2006) *J.Biol.Chem.* **281** 16570-16575

[162] Nagata, K., Kudo, N., Abe, K., Arai, S., Tanokura, M. "Three-Dimensional Solution Structure of Oryzacystatin-I, a Cysteine Proteinase Inhibitor of the Rice, *Oryza sativa* L. japonica" (2000) *Biochemistry* **39** 14753-14760

[163] Evans, H. J., Barrett, A. J. "A cystatin-like cysteine proteinase inhibitor from venom of the African puff adder (Bitis arietans)." (1987) *Biochem. J.* **246** 795-797

[164] Rawlings, N. D., Morton, F. R., Barrett, A. J. "MEROPS: the peptidase database" (2006) *Nuc. Acid. Res.* **34** D270-D272

[165] Sanchez, J. F., Hoh, F., Strub, M. P., Aumelas, A., Dumas, C. "Structure of the Cathelicidin Motif of Protegrin-3 Precursor: Structural Insights into the Activation Mechanism of an Antimicrobial Protein" (2002) *Structure* **10** 1363-1370

[166] Yang, Y., Sanchez, J. F, Strub, M. P., Brutscher, B., Aumelas, A. "NMR Structure of the Cathelin-like Domain of the Protegrin-3 Precursor" (2003) *Biochem.* **42** 4669-4680

[167] Aagaard, A., Listwan, P., Cowieson, N., Huber, T., Ravasi, T., Wells, C.A., Flanagan, J.U., Kellie, S., Hume, D.A., Kobe, B., Martin, J.L. "An Inflammatory Role for the Mammalian Carboxypeptidase Inhibitor Latexin: Relationship to Cystatins and the Tumor Suppressor TIG1" (2005)*Structure* **13** 309-317

[168] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. "The Protein Data Bank" (2000) *Nuc. Acid. Res.* **28** 235-242

[169] Gill, S. C., Von Hippel, P. H. "Calculation of protein extinction coefficients from amino acid sequence data" (1989) *Anal. Biochem.* **182** 319-326

[170] Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. *The Proteomics Protocols Handbook* 571

[171] Whitmore, L., Wallace, B.A. "DICHROWEB, an online server for protein secondary structure analyses from circular discroism spectroscopic data" (2004) *Nucleic Acids Research* **32** W668-W673

[172] Lobley, A., Whitmore, L., Wallace, B.A. "Calculation of protein extinction coefficients from amino acid sequence data" (2002) *Bioinformatics* **18** 211-212

[173] Van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., Van Grondelle, R., Groen, F. C. A. "Estimation of protein secondary structure and error analysis from circular dichroism spectra" (1990) *Anal. Biochem.* **191** 110-118

[174] Sreerema, N., Venyaminov, S. Y., Woody, R.W. "Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy" (1999) *Protein Sci.* **8** 370-380

[175] Stryer, L. "The interaction of a naphthalene dye with apomyoglobin and apohemoglobin. A fluorescent probe of non-polar binding sites."(1965) *J. Mol. Biol.* **13** 482-495

[176] Cardamone, M., Puri, N. K. "Spectrofluorimetric assessment of the surface hydrophobicity of proteins." (1992) *Biochem. J.* **282** 589-593

[177] Schonbrunn, E., Eschenburg, S., Luger, K., Kabsch, W., Amrhein, N. "Structural basis for the interaction of the fluorescence probe 8-anilino-1-naphthalene sulfonate (ANS) with the antibiotic target MurA" (2000) *Proc. Natl. Acad. Sci.* **97** 6345-6349

[178] Cooper, A. *"Biophysical Chemistry"* Royal Society of Chemistry 2004

[179] Zerovnik, E., Cimerman, N., Kos, J., Turk, V., Lohner, K. "Thermal denaturation of huamn cystatin C and two of its variants; Comparison to chicken cystatin." (1997) *Biol. Chem.* **378** 1199-1203

[180] Lide, D. R. (Ed.), *"Handbook of Chemistry and Physics"* **86th Ed.**

[181] Edwards, S. A., Williams, D. R. M. "Surface tension of electrolyte solutions: Comparing the effects of ionic dispersion forces and salvation" (2006) *Europhys. Lett.* 74 854-860

[182] Roberts, G. (ed.), "*Langmuir-Blodgett Films*" 1990

[183] Franklin, B. ""Of the Stilling of Waves by Means of Oil." Extracted from sundry letters between Benjamin Franklin, William Brownrigg and the Reverend Mr. Farish. Extract of a Letter from Doctor Brownrigg to Dr. Franklin dated Jan 27, 1773 and a Letter to Doctor Brownrigg from Doctor Franklin dated Nov 7, 1773" (1774) *Philos. Trans. R. Soc. London* **64** 445-460

[184] Beeley, J. G., Eason, R., Snow, D. H. Isolation and characterization of latherin, a surface-active protein from horse sweat" (1986) *Biochem. J.* **235** 645-650

[185] Tillotson, R. D., Wösten, H. A. B., Richter, M., Willey, J. M. "A surface active protein involved in aerial hyphae formation in the filamentous fungus Schizophillum communerestores the capacity of a bald mutant of the filamentous bacterium Streptomyces coelicolor" (1998) *Mol. Microbiol.* **30** 595-602

[186] Mysels, K. J. "Surafce Tension of Solutions of Pure Sodium Dodecyl Sulfate" (1986) *Langmuir* **2** 423-428

[187] Dettre, R. H., Johnson, R. E. "Surface properties of polymers: I. The surface tensions of some molten polyethylenes" (1966) *Colloid Interface Sci.* **21** 367-377

[188] Adamson, A. W., "*Physical Chemistry of Surfaces*" **5th Ed.** 142

[189] Niño, M. R. R., Sánchez, C. C., Patino, J. M. R. "Interfacial characteristics of β-casein spread films at the air–water interface" (1999) *Colloids Surfaces B* **12** 161-173

[190] Innocente, N., Blecker, C., Deroanne, C., Paquot, M. "Langmuir Film Balance Study of the Surface Properties of a Soluble Fraction of Milk Fat-Globule Membrane" (1997) *J. Agric. Food Chem.* **45** 1559-1563

[191] Graham, D. E., Phillips, M. C. "Proteins at liquid interfaces: II. Adsorption isotherms" (1979) *J. Colloid Interface Sci* **70** 415-426

[192] Denk, W., Strickler, J. H., Webb, W. W. "Two-photon laser scanning fluorescence microscopy" (1990) *Science* **248** 73-76

[193] Gadegaard, N. (Atomic force microscopy in biology: technology and techniques" (2006) *Biotech. Histochem.* **81** 87-97

[194] Malik, I. J., Pirooz, S., Shive, L. W., Davenport, A. J., Vitus, C. M. "Surface Roughness of Silicon Wafers on Different Lateral Length Scales" (1993) *J. Electrochem. Soc.* **140** L75-L77

[195] Newlands, G. F. J., Skuce, P. J., Knox, D. P., Smith, W. D. "Cloning and expression of cystatin a potent cysteine protease inhibitor from the gut of Haemonchus contortus" (2002) *Parasitology* **122** 371-378

[196] Morita, T., Kato, H., Iwanaga, S., Takada, K., Kimura, T. "New Fluorogenic Substrates for α-Thrombin, Factor Xa, Kallikreins, and Urokinase" (1977) *J. Biochem* **82** 1495-1498

[197] Schroder, E., Phillips, C., Garman, E., Harlos, K., Crawford, C. "X-ray crystallographic structure of a papain-leupeptin complex" (1993) *FEBS. Lett.* **315** 38-42

[198] Hunt, L.T., Dayhoff, M. O. "A surprising new protein superfamily containing ovalbumin, antithrombin-III, and alpha$_1$-proteinase inhibitor" (1980) *Biochem Biophys Res Commun.* **95** 864-871

[199] Gettins, P. G. W. "Serpin, Structure, Mechanism, and Function" (2002) *Chem. Rev.* **102** 4751-4804

[200] Kudryashova, E. V., Visser, A. J. W. G., De Jongh, H. H. J. "Reversible self-association of ovalbumin at air–water interfaces and the consequences for the exerted surface pressure" (2005) *Prot. Sci.* **14** 483-493

[201] Biemann K., Cone C., Webster B. R., Arsenault G.P. "Determination of the Amino Acid Sequence in Oligopeptides by Computer Interpretation of Their High-Resolution Mass Spectra" (1966) *J. Am. Chem. Soc.* **88** 5598-5606