From Peptides to Proteins: Exploring Modular Evolution Through the β-Trefoil Fold

by

Robert Aron Broom

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Science
in
Chemistry

Waterloo, Ontario, Canada, 2010

© Robert Aron Broom 2010

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Understanding the origin of protein folds, and the mechanism by which evolution has generated them, is a critically important step on a path towards rational protein design. Modifying existing proteins and designing our own novel folds and functions is a lofty but achievable goal, for which there are many foreseeable rewards.

It is believed that modern proteins may have arisen from a primordial set of peptide precursors, which were initially only pseudo-stable or stable only as complexes with RNA, and later were able to self-assemble into multimeric complexes that resembled modern folds. In order to experimentally examine the feasibility of this theory, an attempt was made at reconstructing the evolutionary path of a beta-trefoil. The beta-trefoil is a naturally abundant fold or superfold, possessing pseudo-threefold symmetry, and usually having a sugar-binding function. It has been proposed that such a fold could arise from the triplication of just one small peptide on the order of 40-50 amino acids in length.

The evolutionary path of a ricin, a family within the beta-trefoils known to possess a carbohydrate binding function was the chosen template for evolutionary modelling. It was desirable to have a known function associated with this design, such that it would be possible to determine if not only the fold, but also the function, could be reconstructed. A small peptide of 47 amino acids was designed and expressed. This peptide not only trimerized as expected, but possessed the carbohydrate binding function it was predicted to have. In an evolutionary model of the early protein world, the gene for this peptide would undergo duplication and later, triplication, eventually resulting in a completely symmetrical beta-trefoil, which would represent the first modern beta-trefoil fold. Such a completely symmetrical protein was also designed and expressed by triplicating the gene for the aforementioned small peptide. This hypothetical first modern beta-trefoil is: well folded, stable, soluble, and appears to adopt a beta-trefoil fold.

Together these results demonstrate that an evolutionary model of early life: that proteins first existed as self-assembling modular peptides, and subsequent to gene duplications or fusions, as what we now recognize as modern folds, is experimentally consistent and not only generates stable structures, but those with function, which of course is a prime requisite of evolution. Moreover the results show that it may be possible to use this modular nature of protein folding to design our own proteins and predict the structure of others.

Acknowledgements

I would like to acknowledge all of my labmates for answering my many questions and demonstrating the use of equipment. In particular I would like to thank Martin T.J. Smith for his help with NMR, and Ming Sze Tong and Helen Stubbs for showing me the ropes of differential scanning calorimetry and dynamic light scattering.

Additionally, I wish to thank Andrew Doxey for our many insightful an involved conversations regarding protein evolution.

I would also like to thank Cathy Van Esch for answering my innumerable questions regarding policies and forms.

Finally, I wish to thank my supervisor Dr. Elizabeth Meiering for not only being a fantastic mentor, but also for putting up with my many odd tendencies, and obliging my scientific interests.

Dedication

I would like to dedicate this thesis to my parents for their constant support throughout my life and in particular since my leaving for University.

I would also like to dedicate my work to the concepts of truth and critical thinking. That we should always question, always search for an answer that is flawless, even if rationale tells us that such a goal is impossible.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1RAB: One Repeat Assembly of Beta-elements

3RAB: Three Repeat Assembly of Beta-elements

CD: Circular Dichroism

DLS: Dynamic Light Scattering

DSC: Differential Scanning Calorimetry

EDTA: Ethylene Diamine Tetraacetic Acid

GuHCl: Guanidine Hydrochloride

^{1}H-NMR: Proton Nuclear Magnetic Resonance

HEPES: 4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid

HIV: Human Immunodeficiency Virus

IPTG: Isopropyl-β-D-thiogalactopyranoside

MW: Molecular Weight

NCBI: National Center for Biotechnology Information

NMR: Nuclear Magnetic Resonance

PCR: Polymerase Chain Reaction

PMSF: Phenylmethylsulphonyl fluoride

SEC: Size Exclusion Chromatography

SOB: Super Optimal Broth

SOC: Super Optimal Broth with Catabolite Repression

SDS-PAGE: Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis

TIM: Triose Isomerase

WD: Tryptophan-Aspartate

1 Introduction

1.1 Proteins, Prediction and Design

Proteins are ubiquitous to life, and their structures and sequences are as diverse as their functions. Proteins perform necessary functions for our survival, from enzymatic reactions (1), nutrient binding/uptake (2), signalling (2), structural roles (3), and pathogen recognition (4). In addition to their necessary roles, we often use proteins commercially to− improve the effectiveness of certain products like detergents (5), allow the cheap production of products such as lactose-free milk (6), and medicinally as seen in the use of insulin to treat diabetes (7) or collagen as a regenerative matrix (3). While proteins offer so much possibility, our ability to exploit this possibility is limited. In order to understand the working of biological systems and thereby engineer methods of controlling them (such as anti-viral drugs) we must know the structures of the proteins involved and which partners they interact with. In order to design better commercially or medicinally relevant proteins we must be able to modify both form and function rationally to fit our needs. To this end many attempts have been made at predicting the structure of a protein from its amino acid sequence alone, as recent advances have allowed us to easily obtain sequence information from a wide variety of organisms (8). Additionally, many design attempts have been made in which a certain structure is given, and an attempt is made to predict a compatible sequence for that structure (9). Although many advances have been made, the methods used are still far from perfect, often failing to predict the correct structure from a sequence (10) or supplying a sequence for a structure that fails to adopt a compact fold (11). And yet, there already exists an effective mechanism for design, in the form of evolution, which has already developed such a wide variety of forms and functions. Understanding the mechanism of this natural process would provide invaluable information in our quest to creating a world of synthetic proteins.

1.2 Early Proteins and Beta-trefoils

In order to address the question of the evolutionary mechanism behind protein fold evolution, an examination of the early protein world and how modern folds may have come into being may be a necessary starting point, after all, why not start at the beginning. It has been suggested that the early protein world may have existed in the form of peptides which were only stable when complexed with RNA (12). Subsequently they would have evolved the ability to be somewhat stable on their own and formed multimeric peptide complexes that may have resembled our modern folds (13, 14). After gene fusions or duplications, these peptide complexes would be covalently linked together into single proteins, enhancing their stability and likely function as well. In order to demonstrate a proof of principle for this model, it was thought that developing a small peptide which would be stabilized by RNA, but could also self-assemble into a modern fold would be ideal.

A modern fold, the beta-trefoil, provided an excellent starting point. The beta-trefoil fold is composed of a 6-stranded beta barrel, capped by a 6-stranded beta-hairpin triplet (15) (Figure 1.1). The entire structure may be described as having pseudo-threefold structural symmetry, and in some cases also has evidence of matching sequence symmetries. Beta-trefoils appear to always perform a binding function, whether on their own as a signalling molecule (16) or as a part of a multidomain protein as the binding domain (17, 18). Often their binding partners are carbohydrates, suggesting a possible link with early RNA binding (19). The total size of the beta-trefoil fold is approximately 120-150 amino acids, which makes each symmetrical unit approximately 40-50 amino acids in length.

The aforementioned characteristics make the beta-trefoil an ideal fold for examining the proposed model, by expressing a single 40-50 amino acid sequence representing one symmetrical unit, henceforth referred to as a trefoil element. Binding to a substrate or ligand can sometimes improve the stability and folding of a protein (20). Therefore, binding of a trefoil element to a carbohydrate could stabilize the peptide, mimicking the proposed early stabilization when complexing with RNA.

Additionally, large-scale sequence analysis experiments have demonstrated that the small stable peptides which would have existed in an early protein world would have had an average size between 40 and 60 amino acids (21), approximately the same size as the trefoil element. Moreover, because each symmetrical unit would share one third of the hydrophobic core amino acids, and because hydrophobic patch binding has been shown to be the most common method of multimerization (22), it is reasonable to suggest that each trefoil element could come together with two others in order to bury their collective hydrophobic patches, thereby creating a new hydrophobic core and a multimeric structure resembling a beta-trefoil. Subsequently, gene duplication and fusion would generate a single gene in which the peptides would now be covalently bonded as a larger protein, perhaps enhancing both stability and function. In fact, this sequential model of beta-trefoil evolution has been previously suggested, making it an even more attractive fold (23).

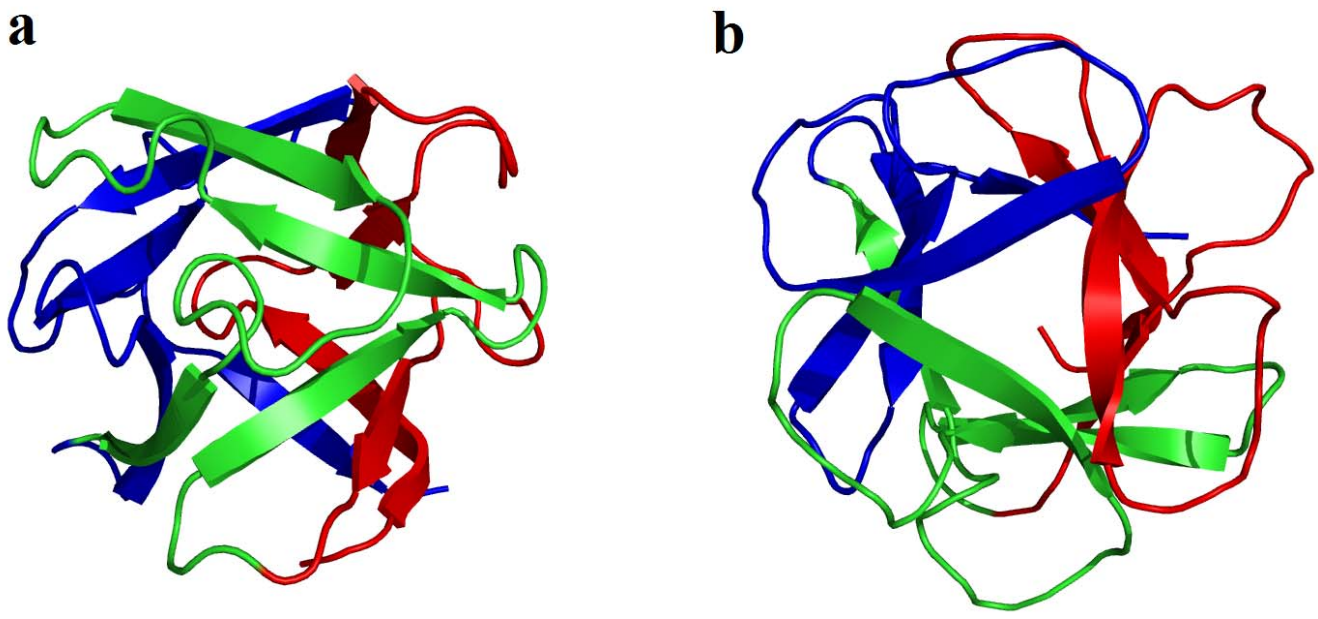

Figure 1.1: Ribbon representation of a beta-trefoil. Showing the barrel capped by hairpin triplet (a), and a view from above the hairpin triplet looking down the axis of symmetry (b). Pymol was used to generate the images using a ricin (a xylan binding domain, PDB code 1KNM) as the backbone.

1.3 Prediction and Design

Demonstrating experimentally that stable structures can be formed by the association of peptide elements and that these structures resemble modern folds would suggest that folds can be formed from these peptide precursors. A method of protein structural design and prediction could be explored whereby a library of stable peptide elements are combinatorially assembled to generate a new stable designed fold, or, the sequences of those elements are used to generate a partially folded starting point for prediction, respectively.

The idea that the structure of proteins can be not only predicted but rationally designed using small segments of known structure is not new (9, 24, 25). This concept of using small segments of known structure relies on the idea that proteins are inherently modular in nature. That is to say, that one can swap in or out certain portions, without drastically affecting the neighbouring pieces. At a large scale this is clearly true, as there are many proteins which have multiple functional domains, and each domain can remain functional on its own (26). Obviously at the polar extreme (considering one amino acid at a time) the protein can not be considered modular, as the structure of a small segment of residues depends very heavily upon its composition, and mutating a single position can have drastic effects. One can imagine that trying to predict the structure of an unknown protein using only known domains in order to obtain the general shape of the backbone would only be effective if the unknown structure was composed entirely of domains that were already known. Any reasonable deviation towards some novel domain would not be recognized, simply because the piece being used is too large, and cannot accommodate internal changes. The objective then, is to build a library of small peptides which are large enough such that their structure is not greatly affected by neighbouring peptides, and yet small enough to be combined to sample all known structures.

The aforementioned approach to design and prediction would be very similar to how we envision evolution has constructed the proteins we see today, and is therefore quite attractive as a

method that we know is capable of producing effective results. The beta-trefoil not only possesses many attributes suggesting it could have arisen via this evolutionary mechanism (symmetry, carbohydrate binding), but statistical analysis of beta-trefoil sequences and structures has also shown that some members of this fold may still be evolving through a related mechanism even today (27). Beta-trefoil evolution today is thought to occur by a mechanism through which a single repeat unit is triplicated to form a new, completely symmetrical beta-trefoil (27). Therefore, we need only find an example of a recently formed beta-trefoil, and it may provide the ideal template for experimentally demonstrating this mechanism.

2 Sequence Design

2.1 Introduction

In order to demonstrate the hypothesis that a stable peptide could self-assemble to form a modern fold, the sequence for such a peptide needed to be elucidated. Not only must this sequence on its own be capable of this self-assembly, but a multiplication of that sequence should also form a stable structure, such that an evolutionary path to modern proteins is feasible. The idea being that a small and yet stable peptide could not only exist, but perform a useful function, and thereby have an evolutionary pressure towards being maintained. Once maintained it would only be a matter of time before a duplication or triplication in the gene occurred generating a new sequence. If this new sequence is more stable than the single peptide there may be an even greater evolutionary pressure towards its maintenance, thereby allowing the triplicated sequence to predominate in nature.

Since the beta-trefoil fold is an abundant fold (23), there exist a large number of sequences that are annotated as forming beta-trefoils (more than 1000 (28)), and also a number of solved structures (more than 50 (29)). In general, strategies for determining an ancestral sequence involve multiple sequence alignments of known sequences followed by choosing residues based upon frequency within that alignment, as was successful for the design of a tryptophan-aspartate (WD) repeat (30). In the case of the beta-trefoil, while it might be possible to use all available sequence information to generate a sequence for a progenitor peptide, the considerable diversity of sequences and likely evolutionary distance of those sequences (23, 31), makes this an unattractive approach. Far more desirable would be identifying the most symmetrical beta-trefoil sequence, which may have very recently triplicated, followed by working backwards only a short evolutionary distance to determine the identity of any positions which are currently not completely symmetrical.

After obtaining a sequence which is already very symmetrical, the closest related sequences could be used in an attempt to elucidate the remaining sequence information, using multiple sequence alignments. Alternatively, or in addition to the previous method, computational structure modelling could be used to determine which residues might provide the lowest free energy at each of the given ambiguous positions (11, 32).

Once a completely symmetrical sequence is obtained which represents the likely result of a recent triplication of a peptide precursor, both the symmetrical beta-trefoil and the peptide can be expressed and characterized to determine the feasibility of each existing independently as steps in the proposed evolutionary model.

2.2 Methods and Results

As a general note to the reader, the methods for designing the sequence are not being reported separately from the results of those methods. The reason for doing so is that the results themselves are intricately entwined with the methods, and in some cases the method is itself a kind of result. Overall this is likely a result of the fact that designing a novel protein is not a well understood process and therefore lacking in well accepted methodologies.

2.2.1 Template Sequence

The initial template sequence chosen was the most symmetrical beta-trefoil sequence available in the Conserved Domain Database (28), a database of 1167 sequences. The sequence chosen was a putative glycosidase (NCBI accession #AAV45265) from *Haloarcula marismortui*, an archaeon from the dead sea, where external salt concentrations reach the saturation limit (4 M NaCl). The sequence was manually dissected into its three repeats (the reasons for manually identifying the repeats rather than using an automated algorithm are discussed in Section 2.2.2). The alignment of these repeats is

shown in Figure 2.1, where it can be seen that 26 of 47 amino acids are already conserved in all three positions. Additionally, the putative core hydrophobic residues (as determined by the sequence pattern of beta-trefoils) (AF, AM), were conserved in 5 of 6 repeat positions.

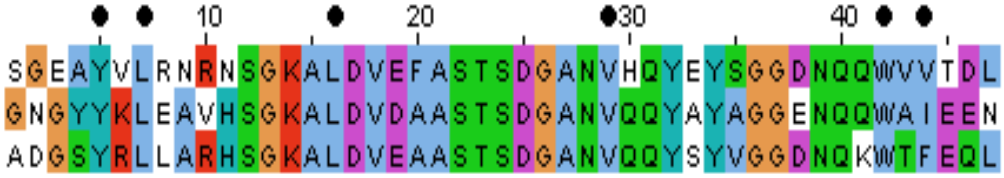

Figure 2.1: Alignment of the template sequence. Alignment of the three 47 residue repeats from NCBI accession #AAV45265, a putative glycosidase from *Haloarcula marismortui*. Black circles indicate core hydrophobic positions. All alignment images were generated using Jalview.

2.2.2 Family of Closest Relatives

BLAST (33) was used to identify the 13 closest relatives to the template sequence. All sequences were manually broken into their repeats by following the sequence patterns observed in beta-trefoils of known structure. Although there exist many algorithms for automatically identifying internal repeats within a sequence, such as RADAR (34), they often poorly define the borders of the repeats. In the case of the beta-trefoil this problem is confounded by the fact that the repeat borders occur in loops between strands, which often show considerable sequence variation. After manually separating the sequences into repeats they were aligned in an all-by-all alignment using CLASTALW (35) (Figure 2.2).

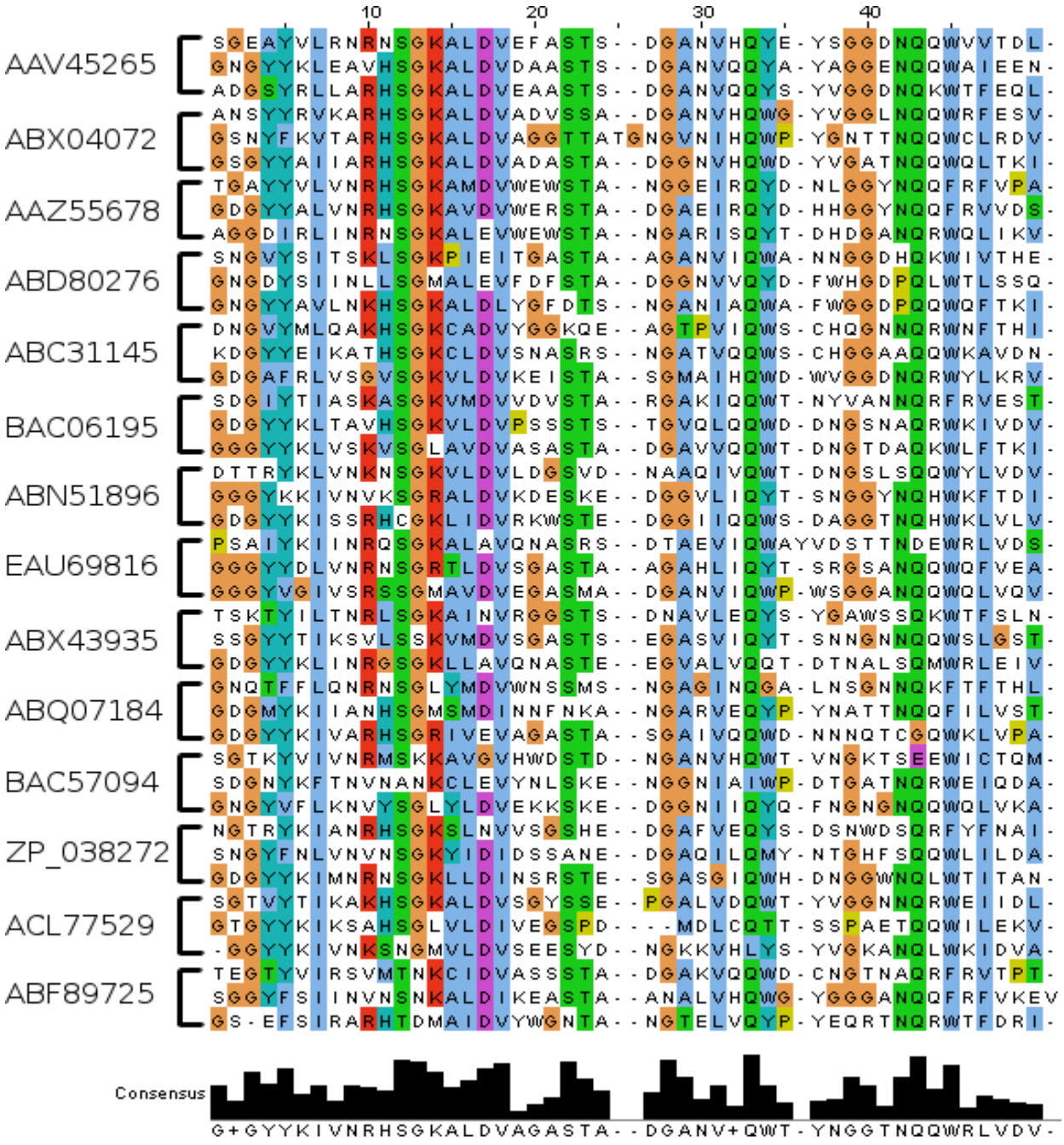

Figure 2.2: Alignment of 13 sequences most closely related to the template. Sequences are annotated with their NCBI accession numbers. The chosen design sequence is listed first, and the consensus sequence for these alignments is shown in black at the bottom.

2.2.3 Filling in the Gaps

As mentioned, we see from Figure 2.0 that 26 of the 47 positions are already completely conserved in the starting sequence. These 26 positions were automatically chosen to be in the final sequence (Figure 2.4, sequence A). Moreover, an additional 11 residues are conserved in 2 of the 3 positions. In these cases, the related sequences (Figure 2.2) can often confirm that these partially conserved residues are in fact, the likely progenitor residue at that position. For example, this is seen in Figure 2.3 where residues 9, 10 and 11 show conservation in only 2 of the 3 repeats of the template sequence, but where the closest relative shows these same residues conserved at all 3 positions. This conservation pattern not only indicates that these residues are a good choice at this position, but also that they were likely present in the common ancestral repeat for both these sequences. Following this approach many of the remaining positions were assigned a residue (Figure 2.4, sequence B).

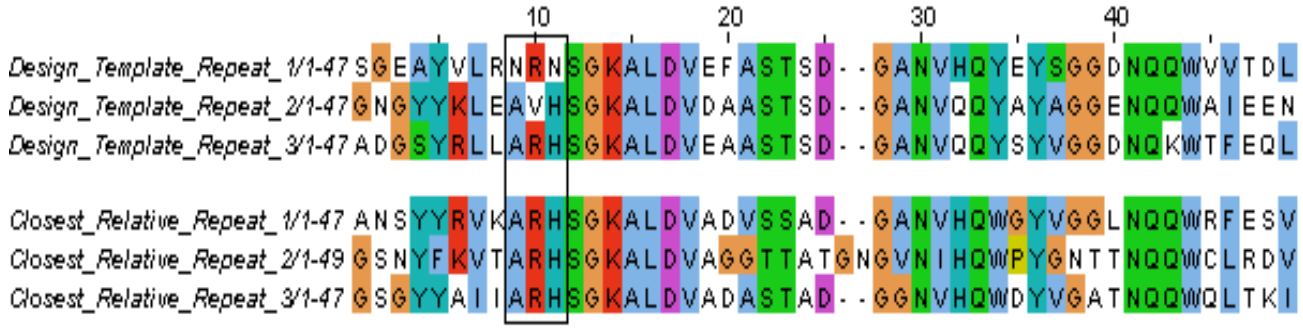

Figure 2.3: Related sequences used to identify specific positions. The template sequence alignment is shown first, with the closest relative (lowest BLAST (33) score) shown below. A black box indicates an area where the closest relative's conserved residues are particularly helpful for inferring the putative progenitor sequence.

In some cases, however, the related sequence information still leaves the residue choice ambiguous, particularly in cases where the original sequence shows no conservation at a given position

(different residues in all three repeats). In these cases an additional tool was needed to pick the best residue.

It was desirable that the chosen residue be an energetically favourable choice in order to improve the chances of generating a stable structure, but also because naturally occurring residues are often among the most energetically favourable choices unless functional constraints are involved (36). To this end, ROSETTA (11, 24) was used to determine the most energetically favourable residues at the remaining unassigned positions. Because ROSETTA requires a structural model in order to perform its algorithm, homology models were generated using MODELLER9 (37). Homology modelling requires not only the backbone of a close homologue, but also the sequence of the target. The current best estimated sequence (Figure 2.4, sequence C) was used, along with backbones from the three most closely related structures: 2IHO (a galactose binding lectin) (38), 1YBI (a hemagglutinin) (18), and 1KNM (a xylan binding domain) (17). While ROSETTA was allowed to choose from all naturally occurring amino acids (except: proline, cysteine, and methionine, which do not exist in the template sequence), it repeatedly chose amino acids which were already well represented in the closest relatives at those positions, perhaps illustrating its ability to make native-like choices.

In order to choose the identity of the remaining positions, each amino acid was given a frequency based on an equal weighting between the original sequence, the closest relatives, and the ROSETTA choices. Thus, a residue which was present in one repeat of the original sequence, represented in some of the closest relatives, and chosen by ROSETTA, would be chosen over a residue with, for example, the same properties but not present in the original sequence. This approach was chosen in order to maximize both the use of evolutionary information from the closest relatives, energetic considerations from ROSETTA, and also to choose a residue which may in fact have been present in the progenitor sequence. The final sequence generated from applying the aforementioned approach is shown in Figure 2.4, sequence D.

Figure 2.4: Sequence history of the rebuilt progenitor repeat. Starting from only the conservation in the initial template sequence (A), residues were added as related sequence information (B, C), and finally homology modelling and energy minimization (D) were incorporated.

2.3 Summary of Results

The final sequence chosen as a candidate for generating a stable peptide capable of self-assembling into a beta-trefoil-like fold (Figure 2.4, sequence D) was simply repeated three times in order to generate the final sequence for the triplicated gene (Figure 2.5). The single repeat was named, one repeat assembly of beta-elements (1RAB), while the triple repeat was name three repeat assembly of beta-elements (3RAB). The theoretical molecular weights of 1RAB and 3RAB are 5.1 and 15.3 kDa respectively.

Figure 2.5: The sequence of 3RAB. The final sequence for 1RAB (Figure 2.4, sequence D) was triplicated to yield the final design sequence for a completely symmetrical beta-trefoil, 3RAB.

2.4 Discussion

2.4.1 Feasibility

It is extremely likely that the triplicated sequence (3RAB) will properly generate a beta-trefoil protein, as it possesses 73% identity to the template sequence, which presumably, is a naturally occurring beta-trefoil. Moreover, the triplicated sequence also possesses 40% identity to a beta-trefoil of known structure, 1KNM (a xylan binding domain) (17). More importantly, the majority (15 out of 18) of the core hydrophobic residues have not been changed, which is perhaps the most critical portion of the protein required for forming a compact fold (31). In the case of single peptide, homology modelling reveals that the peptide should have a large exposed hydrophobic patch. Since hydrophobic patch association is the primary mechanism for oligomerization of proteins, this will likely allow for a strong interaction between three such peptides, which could bury their hydrophobic patches together to form a hydrophobic core. Additionally, the hairpin triplet portion of the protein may be particularly prone to trimerization, as has been seen in a structurally similar designed small peptide (39) (Figure 2.6).

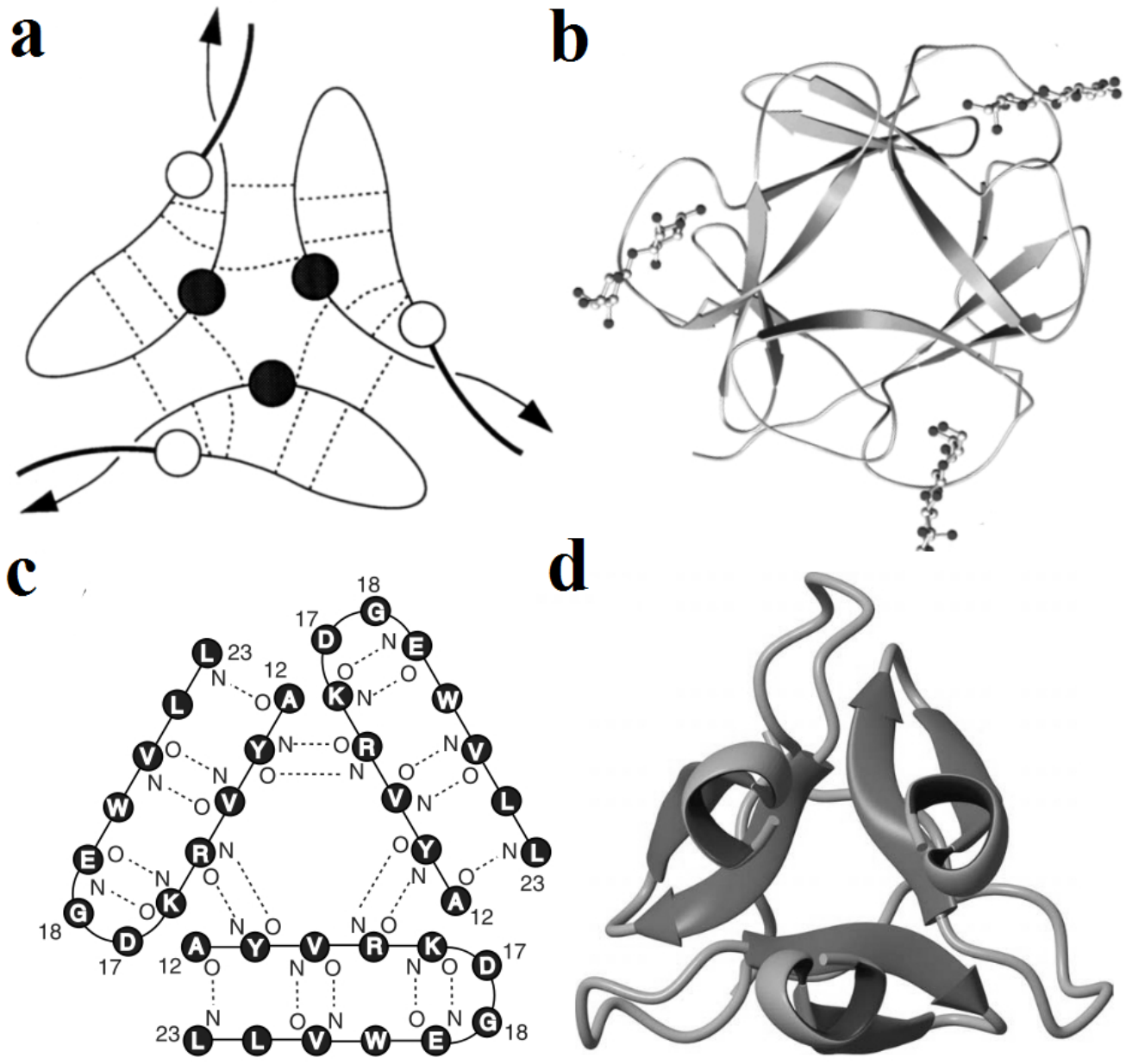

Figure 2.6: Comparison of the beta-trefoil hairpin triplet with the trimerization domain of T4 Fibritin. Beta-trefoil hairpin triplet hydrogen-bonding scheme (a) (15), and ribbon diagram (b) (17), appear similar to the trimerization domain hydrogen-bonding scheme (c) (39), and ribbon diagram (d) (39).

2.4.2 Symmetry

Thus far this discussion has not addressed the idea of symmetry and whether it is a benefit or a hindrance to protein folding, but it is particularly relevant to the design of this completely symmetrical sequence. It is clear from sequence evidence (Figure 2.1) that this fold can tolerate a very symmetrical sequence, but is the lack of a completely symmetrical sequence simply a result of genetic drift, or is total symmetry incompatible with the determinants of this fold? A reasonably unsymmetrical beta-trefoil, the acidic fibroblast growth factor, has been extensively studied concerning symmetry (40, 41, 42) and it has been found that mutations which increase the symmetry of the sequence and structure make the structure more stable. Moreover, these improvements in stability are cumulative, giving a phenomenal increase in stability of over 40 kcal/mol when at least 6 sites have been mutated to improve symmetry (many of these sites are part of the protein core). Additionally, theoretical work has shown that symmetry may improve the smoothness and funnelling of the folding energy landscape, a requirement for foldable proteins (43).

While the aforementioned examples, in combination with the inherent symmetry of the beta-trefoil fold would tend to suggest that symmetry is not only compatible, but desirable for this fold, there is some evidence to suggest otherwise. If the template sequence and 13 closest relatives are examined at the final conserved hydrophobic position (position 47 in Figure 2.2), we can see that 12 of the 14 (13 closest relatives plus template sequence) sets of repeats have a phenylalanine at that position in exactly 1 of the 3 repeats. This is a very illuminating observation, as it suggests that the amount of space that needs to be taken up in that area of the fold is large enough that 3 smaller residues, such as valine, are insufficient, but small enough that 3 larger residues such as phenylalanine would be too voluminous. If this is the case, then what residue is the best choice at this symmetry-position? A small residue would be entropically unfavourable, whereas a large residue might distort the shape of the fold in a destabilizing manner. Looking at the repeats from ACL77529 in Figure 2.2, it can be seen that at

position 47 the residues are: isoleucine, leucine, and isoleucine. This suggests that the volume occupied by leucine or isoleucine is sufficient to fill the needed space without being too large. Interestingly, as detailed above, Rosetta also chose leucine as the most energetically favourable choice at this position, and henceforth it was included in the final design.

2.4.3 The Next Step

The next step in providing a proof of principle for the evolutionary model of stable peptides assembling to form stable modern-like folds, was the expression of these newly designed sequences for both the single peptide and the triplicated protein.

3 Cloning, Expression, Purification, and Refolding

3.1 Introduction

3.1.1 Histidine Tags

Many proteins of varying origin have been successfully expressed using *Escherichia coli*. Purification of the recombinant protein can be accomplished through a variety of means, but one approach which has seen considerable success in proteomics research is the use of an affinity tag. In particular, a poly-histidine tag can facilitate purification through affinity chromatography using a nickel affinity resin (44). This approach is particularly useful not only because it can lead to both a high yield and high protein purity, but also because the poly-histidine tag does not appear to affect the native structure of the protein (44). The use of a poly-histidine tag is based upon solution pH and is a very specific technique. A histidine sidechain which is neutral (pH >7) is capable of binding to a positively charged divalent metal cation, such as nickel. If the histidine is protonated (pH <6) it is rendered incapable of binding to nickel. In the case of a poly-histidine tag, all the histidine residues (usually 6, whereas 10 have been used in this case) are maintained in a neutral state via a high pH (no less than 8). This facilitates binding to the column, whereas most proteins which may only have a few isolated histidines or none at all, pass through. Afterwards any non-specifically bound proteins or nucleic acids can be eluted by a slight decrease in pH to approximately 6.5, where the histidine residues will be partially protonated (in this case the use of many histidines in the tag allows binding to be maintained). Afterwards the pure protein may be eluted under acidic conditions. If changes in pH cannot be tolerated by the protein, the use of imidazole (which is a chemical analog to histidine) in increasing concentrations may be used to compete for nickel binding sites. It should be noted, however, that using imidazole may necessitate considerable dialysis in order to ensure complete removal. While there are

many techniques for separating protein species from one another, affinity chromatography using a poly-histidine tag is particularly effective because very few contaminating proteins would have a sufficiently long string of histidine residues, or other elements capable of binding as tightly to nickel. Moreover, the use of pH to elute adds an additional element of specificity as the particular pH at which the histidine tag will stop binding will differ slightly from different structures on histidine rich contaminants.

3.1.2 Inclusion Bodies

Recombinant proteins often express as insoluble inclusion bodies, particularly when expression levels are higher than might be expected for native proteins (45). Inclusion bodies present the benefit that they contain few other contaminating proteins, and can be easily isolated from the remainder of cell components via centrifugation. They also present a difficulty in that they must be solubilized in denaturant and later refolded, which often leads to a loss of yield in the form of misfolded products (46). Many strategies can be used to reduce the degree of misfolding, which include using various additives during refolding (46), refolding slowly or at low protein concentration (46), or attempting to resolubilize in a minimum concentration of denaturant, which may leave residual secondary structure intact (45). In many applications, such as the preparation of proteins for pharmaceutical use, inclusion body formation is a desired outcome, as the improvement in protein purity is more important than the loss in yield. Although refolding should in theory be possible for any protein with a well-defined native structure, the particular parameters must be empirically determined for each protein (47). Fortunately both 3RAB and particularly 1RAB are fairly small, and should favour refolding without misfolded species as small proteins tend to fold and unfold through 2-state transitions of native to unfolded without intermediates (48).

Overall, recombinant expression followed by affinity purification and refolding represent an

attractive means of obtaining large quantities of natively structured protein with high purity, which was the chosen method of purification in this study.

3.2 Methods

3.2.1 Agarose Electrophoresis

Agarose (10%, w/v) was added to deionized water and heated until it was a solution. This was poured into a mold, and allowed to cool for 30 minutes. The gel was immersed in TBE buffer (90 mM Tris base, 88 mM boric acid, 2 mM EDTA), and run in at a constant voltage of 90 V.

3.2.2 Sodium Dodecylsulphate Polyacrylamide Electrophoresis (SDS-PAGE)

The lower gel was made of 15% polyacrylamide (37.5:1 acrylamide:bis-acrylamide) and buffered with 400 mM Tris (pH 8.8). The upper gel was 5% polyacrylamide (37.5:1 acrylamide:bis-acrylamide) buffered with 125 mM Tris (pH 6.8). Both upper and lower gels contained sodium dodecylsulphate (0.2% w/v), and polymerization was accomplished by addition of ammonium persulphate (0.001% w/v) and tetramethylethylenediamine (0.000001% v/v). Gels were immersed in SDS-PAGE running buffer (200 mM Tris base, 1.5 M glycine, 0.1% (w/v) sodium dodecylsulphate) and run at a constant voltage of 150 V.

3.2.3 Vector Digestion

All vectors used had both NdeI and BamHI cut-sites. The vectors were double-digested using NdeI and BamHI (Fermentas) in the manner suggested by the supplier. Following digestion the mixture was run on a 10% agarose gel and the bands of interest were cut from the gel and purified using the QIAquick Gel Extraction Kit (QIAGEN), in the manner suggested by the supplier.

3.2.4 Oligo Assembly

Lyophilized oligos were initially dissolved in Tris-EDTA buffer (1 mM Tris-HCl, 1 mM ethylene diamine tetraacetic acid (EDTA), pH 8.0), to a concentration of 30 pmol/L. An equimolar volume of each oligo was added to an eppendorf polypropylene tube (200 µL) and then diluted tenfold with Tris-EDTA buffer. The solution was heated to 95 °C for 10 minutes in a thermal cycler and then allowed to cool in 5 °C increments every 3 minutes until the solution had reached room temperature.

3.2.5 Vector Assembly

Assembly of the expression vector was accomplished by adding: 1 µL of the cooled oligonucleotide mixture (see Section 3.2.2), 9 µL of double-digested pET-28a, 7 µL deionized H_2O, 2 µL ligation reaction buffer and 1 µL of T4 DNA Ligase (New England Biolabs) to a polypropylene tube which was allowed to stand at room temperature for 1 hour.

3.2.6 Electroporation

Ligated vector in Tris-EDTA buffer was added (1 uL) to electrocompetent BL21 (DE3) *Escherichia coli* cells (40 uL). The resulting mixture was added to a cooled 1 mL electroporation cuvette and subjected to a 1.80 mV current for 4 ms. Immediately following electroporation the mixture was added to 1 mL of SOB media for catabolite repression (SOC media) (20 g/L Bacto-tryptone, 5 g/L Bacto Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose) in a 1.5 mL eppendorf tube and allowed to sit at 37 °C for 1 hour. A fraction of the mixture (500 uL) was then plated on Luria Broth (10 g/L Bacto-tryptone, 5 g/L yeast extract, 10 g/L NaCl, pH 7.5) plates containing kanamycin (300 µg/L).

3.2.7 Polymerase Chain Reaction (PCR)

Primers were resuspended in Tris-EDTA pH 8.5 buffer to a concentration of 100 pmol/L. To 39.5 µL of deionized H_2O in a 200 µL PCR tube, 5 µL of Vent Polymerase Buffer (10x concentration) (New England Biolabs), 1 µL of 20 mM dNTPs, 1 µL of each primer, 1 µL of template DNA, and 0.5 µL of Vent Polymerase (New England Biolabs) were added, and mixed gently via pipetting. The solution was cycled in a Px2 Thermal Cycler (Thermo Electron Corporation) for 30 cycles, with each cycle consisting of a 30 second denaturation stage at 95 °C followed by a 30 second annealing stage at 55 °C and finally a 30 second elongation stage at 72 °C.

3.2.8 Blunt-ended Ligation

Ligation of blunt-ended repeat fragments was accomplished in the same manner as vector ligation of the sticky ended fragment (see Section 3.2.3). To 15 µL of dH_2O, 2 µL of gel purified blunt ended single repeat, 2 µL of ligase buffer and 1 µL of T4 DNA Ligase were added. The reaction solution was allowed to stand at 4 °C for 24 hours, and the desired product was gel purified.

3.2.9 Induction of Expression

Expression of 1RAB and 3RAB from BL21 (DE3) cells was accomplished through induction of the pET-28a lac operon, using 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) when cell cultures had reached an optical density of ~0.7 at 600 nm in a 1 cm cuvette. Expression of 3RAB was performed at 37 °C, whereas 1RAB needed to be grown at 25 °C in order to see stable expression. Both 1RAB and 3RAB expressed optimally after approximately 24 hours.

3.2.10 Isolating Inclusion Bodies

Cells were pelleted at 10,000 g at 4 °C for 20 minutes. The cell pellets were then resuspended in Buffer A (300 mM NaCl, 100 mM phosphate, 10 mM Tris, pH 8.0) using 5 mL of buffer per gram of cell pellet. To the mixture: phenylmethylsulphonyl fluoride (PMSF) (50 μg/mL), and lysozyme (50 μg/mL) were added. The cells were broken using three rounds of freeze-thawing with liquid nitrogen and room temperature water respectively. Subsequently, DNase (50 μg/mL) and MgCl (5 mM) were added and allowed to sit for 30 minutes while shaking on ice. The resulting mixture was spun down at 10,000 g at 4°C for 20 minutes. The pellet was then solubilized in Buffer B (300 mM NaCl, 100 mM phosphate, 10 mM Tris, 8 M Urea, pH 8.0) and spun down at 10,000 g at 4°C for 20 minutes.

3.2.11 Affinity Column Purification

Individually, 1RAB or 3RAB in Buffer B were added to an Ni-NTA affinity resin (QIAGEN), and eluted into Buffer C (300 mM NaCl, 100 mM phosphate, 8 M Urea, pH 4.5) according to the manner suggested by the supplier.

3.2.12 Refolding via Dialysis

Both 1RAB and 3RAB were refolded by 24 hour dialysis using a 3 kDa cutoff regenerated cellulose membrane (Spectrapore) at room temperature from Buffer C to Buffer D (300 mM NaCl, 100 mM phosphate, pH 6.6) at a protein concentration of 20 μM and 10 μM respectively (at higher concentrations, visible aggregates were observed). The reservoir volume was 2 L with the sample volume not exceeding 100 mL. No reservoir exchanges were performed. Following dialysis the samples were filtered through a 0.2 μm filter to remove any invisible aggregates.

In the case of the initial refolding attempts the exact concentration of protein was not well controlled although they were known, and a smaller 1 kDa cutoff membrane was used, which was

made from cellulose-ester rather than regenerated cellulose (Spectrapore). Additionally, the initial dialysis was performed using a 1 L reservoir and 1 mL samples, with 16 hours of dialysis and no reservoir exchanges.

3.2.13 Concentration

1RAB and 3RAB were concentrated using a pressurized, stirred-cell device (Amicon) with 1 kDa and 10 kDa cut-off membranes (Millipore) respectively. Both were concentrated to approximately 1 mL in a 50 mL device.

3.3 Results

3.3.1 1RAB Gene Assembly

Since the length of 1RAB is a mere 47 residues, the gene for its expression was short enough to facilitate the assembly of several small oligonucleotides into the full sequence. Six oligonucleotides were ordered from Sigma-Aldrich (full sequences in Appendix A) such that they would spontaneously overlap to form the desired sequence (Figure 3.1) with overhangs for vector ligation, based on a published example (49). The oligos were assembled by allowing them to anneal at over a gradually declining temperature gradient in order to reduce the opportunity for mismatched base pairs. Ligation of the oligos prior to vector assembly was not necessary as the strength of the interactions between the overlapping segments was sufficient to maintain the structure outlined in Figure 3.1.

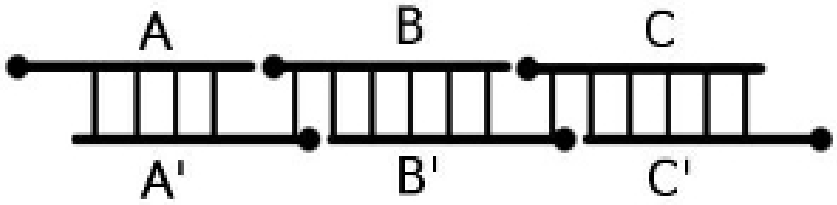

Figure 3.1: Overlap design scheme of 1RAB. Six oligonucleotides were designed to overlap such that after ligating, a single double-stranded DNA segment would result with sticky-ends for plasmid insertion.

3.3.2 1RAB Vector Assembly

The expression vector used for both 1RAB and 3RAB was a modified pET-28a plasmid (Novagen) which contained a deca-histidine (sequence in Appendix A) rather than hexa-histidine affinity tag (50). With the deca-histidine tag, the theoretical molecular weights of 1RAB and 3RAB are 7.8 and 18.0 kDa respectively. The deca-histidine tag was used rather than the standard hexa-histidine tag simply because it was readily available, although a longer histidine tag may also allow for improved purification when using an affinity resin. After digestion with NdeI and BamHI the desired vector was separated from non-digested or over-digested species via an agarose gel and then removed and purified. A band representative of the aforementioned digestion can be seen in Figure 3.1 marked with a triangular arrow. The digested vector and the assembled oligos were mixed together and ligated in order to generate an intact expression vector, which was then used for electroporation of BL21 (DE3) *Escherichia coli* cells. Electroporation was chosen over chemical transformation due to its higher transformation efficiency. BL21 (DE3) cells were chosen for several reasons: first, the lack of some native proteases in the BL21 cell line was considered advantageous as the stability of the final product was unknown, and its small size (47 residues for 1RAB) might lend it towards proteolytic cleavage. Second, as the designed proteins are not native to any known living organisms, they could

possess toxic functions or be toxic by virtue of some aggregation process, and as such the use of a very tightly controlled promoter was desirable.

Figure 3.2 shows the recircularization of cut plasmid by ligation with the assembled oligos sequence before transformation. Several different concentrations of assembled oligos were used with a constant concentration of digested vector. This was done in order to maximize the fraction of correctly ligated vector, while minimizing the degree of oligo polymerization that occurred (as seen in the most concentrated case, indicated with a square arrow in Figure 3.2). The final plasmid was also sequenced after purification from transformed cells using the University of Waterloo "in-house" sequencing facility, and gave the expected sequence (see Chapter 2: Sequence Design).

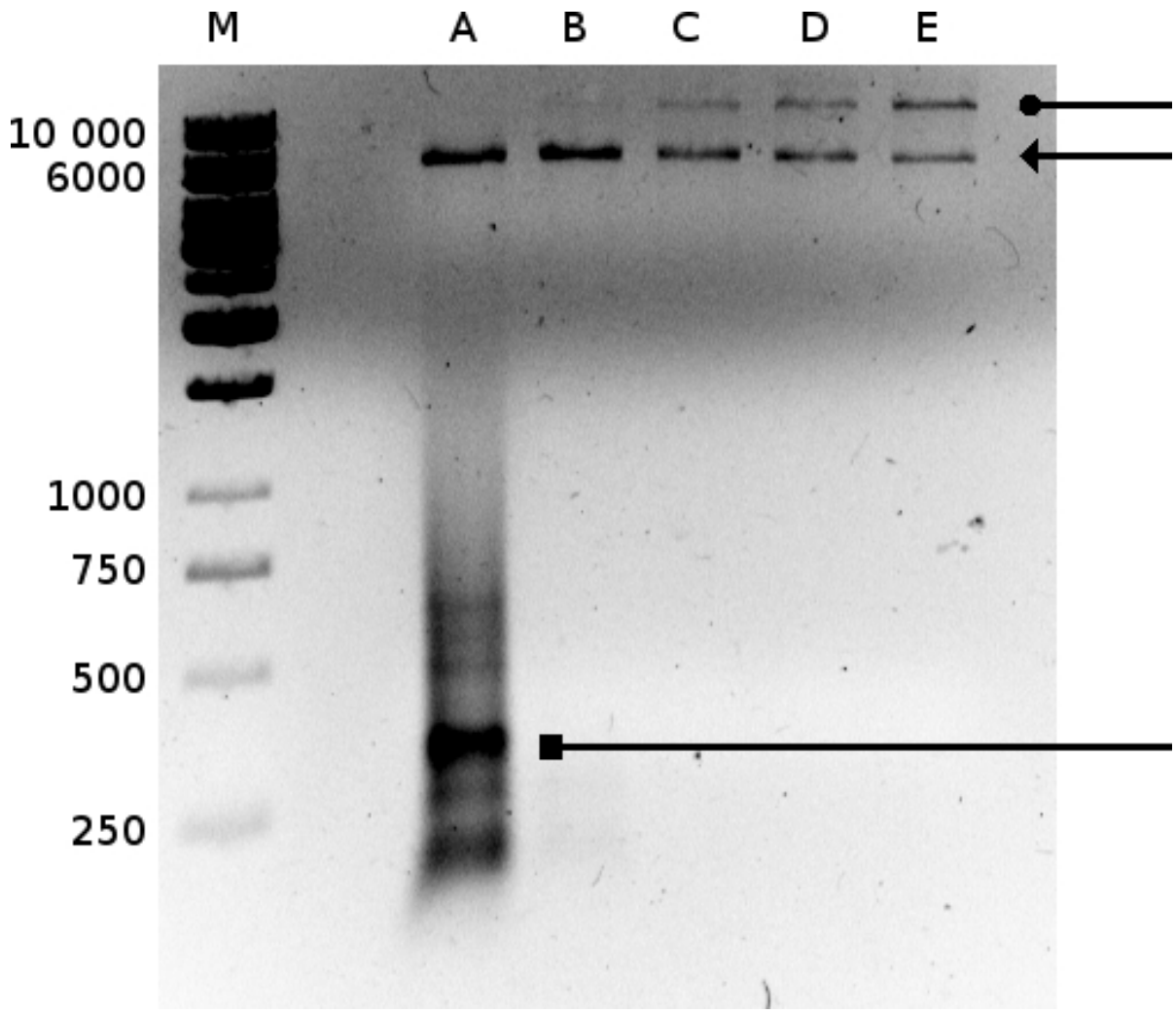

Figure 3.2: Recircularization of the expression vector via ligation of the 1RAB sequence. The molecular weight marker (lane M) is shown on the left, with sizes listed as number of base pairs. Lanes A through E represent sequential 10-fold dilutions of the 1RAB sequence (starting with the highest concentration in lane A at 30 pM) added to a constant concentration of digested plasmid. The triangular arrow indicates linear cut plasmid, and the circular arrow ligated and recircularized plasmid. The square arrow indicates the expected size of the 1RAB sequence alone. The colours in the image have been inverted for clarity.

3.3.3 3RAB Gene Assembly

All PCR reactions were performed as described for 1RAB. As the repetitive nature of the 3RAB gene presented many experimental problems, assembly of that gene using a related method to the that used for 1RAB was found to be insufficient and costly. Although the method was eventually abandoned for the aforementioned reasons, the general details of the method, and a handful of the plethora of PCR results are shown for sake of completeness and perhaps to serve as a warning against this kind of endeavour. Particularly, the author wishes to implore any reader considering the manual construction of a large or repetitive sequence to consider simply ordering the sequence complete and tested in a vector. The cost is comparable to the cost of the reagents needed to attempt it on one's own, and the savings in time is almost immeasurable.

The single gene as assembled for 1RAB was amplified using modified primers such that no additional nucleotides were added to the DNA sequence encoding the 47-residue peptide, 1RAB. These oligos would prime in place of A and C' in Figure 3.1, and are listed in Appendix A, as Ax and Cx respectively. Once amplified, the blunt ended fragments encoding the 47-residue sequence were ligated together. Figure 3.3 shows the results of this ligation, which include not only duplications of the gene, but also triplications further states of oligomerization. Since ligation of blunt ended fragment imposes no intrinsic means of stopping the reaction, the presence of many different oligomers was not unexpected. The PCR products that were of the correct size for a duplication (282 bp) were cut from an agarose gel (10%) and purified. Although the end goal was to obtain a gene for a triplicated sequence (3RAB), the duplicated sequence was needed at this stage, as will be demonstrated.

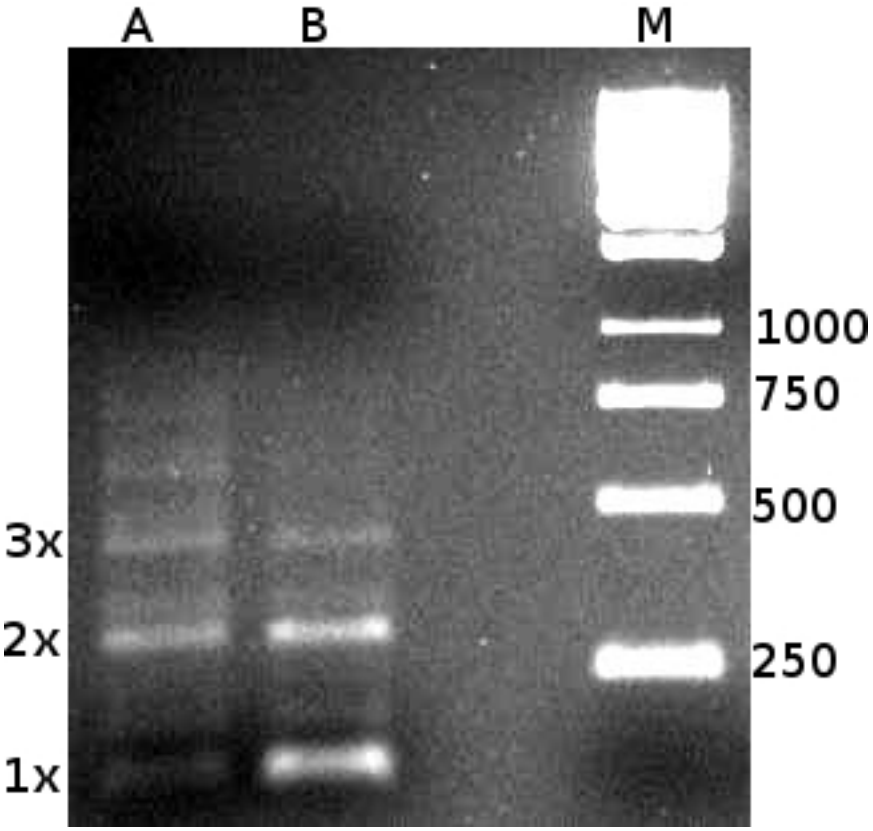

Figure 3.3: Ligation of 1RAB sequence to form duplicated and triplicated sequences. Starting from only the single 1RAB sequence (with an estimated size 141 base pairs, marked as 1x on the left), blunt ended ligation resulted in observable quantities of both duplicated (marked as 2x) and triplicated (marked as 3x) 1RAB sequences. The size marker (lane M) is in base pairs, with lanes A and B representing two identical ligation attempts.

Since the fragment used for ligation was blunt ended, there should in theory be no preference between the desired orientation of the ligated fragments, where all sequences are ordered beginning to end, and those where one of the fragments is inverted with respect to the first. The particular details of these kind of possibilities are shown in Figure 3.4.

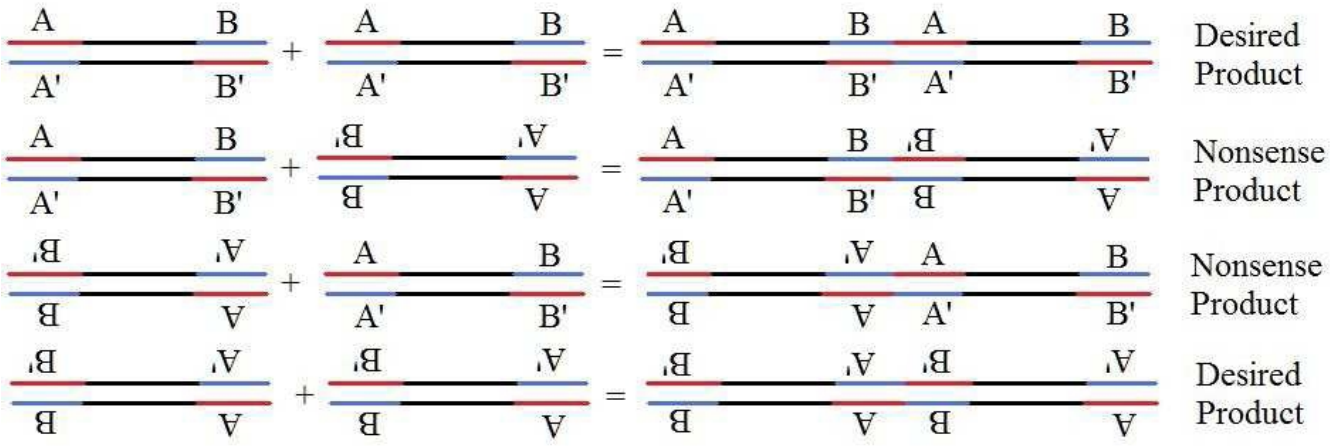

Figure 3.4: A schematic representation of the possible products from blunt-ended ligation. The single repeat fragment may align in an undesirable manner, generating inverted repeats.

Since all the possibilities listed in Figure 3.4 are the same size, they cannot be distinguished using an agarose gel or any other sizing method. Moreover, since they in fact have the same net chemical composition, it would seem very difficult to determine any method capable of separating them. Fortunately there are several inherent properties of these nonsense products which lend themselves to being isolated from the desired products.

During PCR it is of critical importance that secondary structures do not form within the primers or within the template sequence, as even moderately stable secondary structures may greatly inhibit proper annealing of the primer. That the nonsense products are inverted repeats is of particular interest, because it means that a single strand of the nonsense product can fold back on itself and form a very strong hairpin. This idea is shown in Figure 3.5, where it can be seen that the desired product is not self complementary, whereas the nonsense product is. Therefore, if PCR is performed on the duplicated fragments, only the desired product should be amplified, while the nonsense product will be unable to anneal to the primers, due to self-annealing.

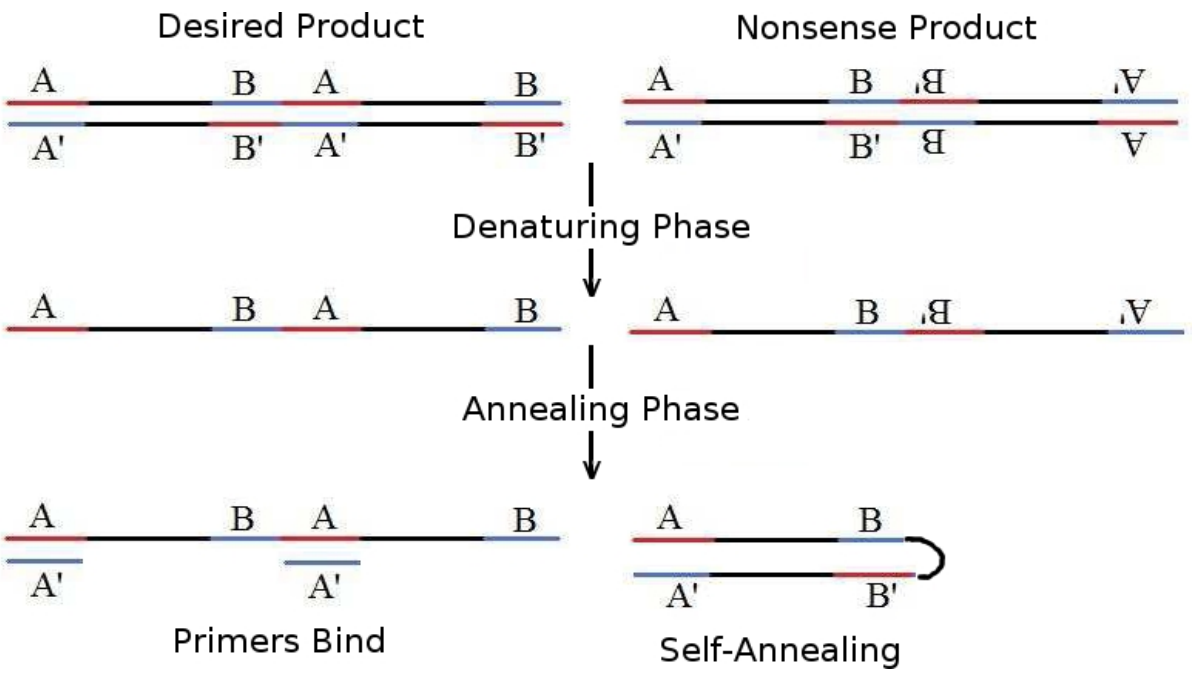

Figure 3.5: A schematic representation of PCR performed on the mixture of products seen in Figure 3.4. The desired product is unable to form secondary structures and is therefore free to bind primers and undergo amplification. Whereas the nonsense product, being an inverted repeat, naturally forms a hairpin secondary structure, blocking primer annealing and therefore preventing amplification.

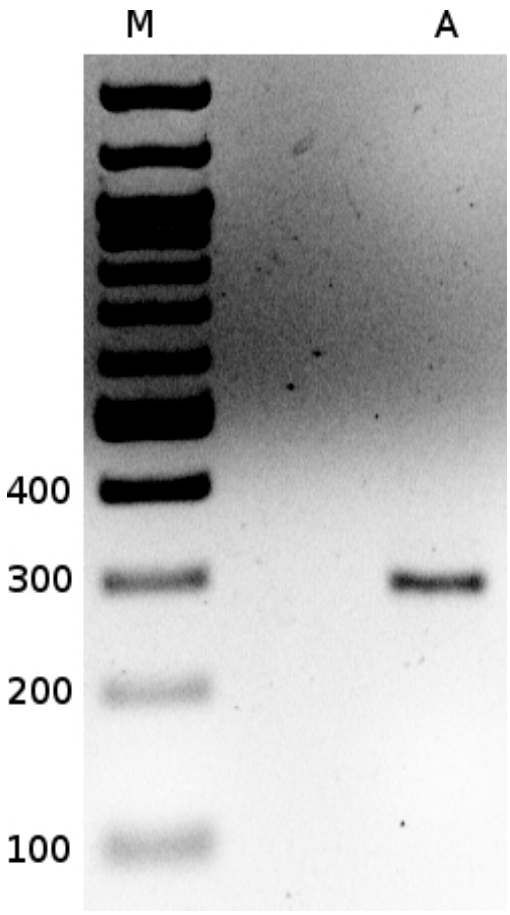

Figure 3.6: Results of digesting duplicated 1RAB sequences which were amplified as in Figure 3.5. The molecular size marker (lane M) is in base pairs, with the expected size of a duplicated sequence being 282 base pairs. As seen in lane A, the attempted digestion with PstI, no digestion occurs indicating that only desired product was present. The results for digestion with NarI are identical and are not shown.

After amplification, it was desirable to test if the aforementioned hypothesis was correct, and desired product was in fact amplified. Fortunately, the fact that the nonsense products are inverted repeats suggested that using restriction enzymes, which recognize small inverted repeat sequences would be an effective choice. Based on the sequence information, the restriction enzymes Pst1 and Nar1 were chosen in order to digest the nonsense products, but not react with the desired products. The

results of this were as expected, none of the amplified product was digested (Figure 3.6), indicting that it is majority desired product.

One can perhaps imagine at this point and from examining both Figures 3.4 and 3.5, that using the duplicated sequence rather than the triplicated one was preferable due to the considerable added complications that adding yet another blunt ended fragment would create. Of course, this still leaves the question of how to obtain the desired triplicated sequence. In order to get the triplicated sequence from the duplicated one a very simple method of converting duplicated sequences to triplicated ones was devised. The basics of this method are shown in Figure 3.7, where the duplicated sequences act as self-primers for generation of a triplicated sequence only when they mis-align during the annealing phase of PCR. Note that since no additional primer is added, this reaction is merely a conversion rather than an amplification. The results are shown in Figure 3.8, where we can see that this method is successful, and also creates some larger fragments, likely through the same mechanism.

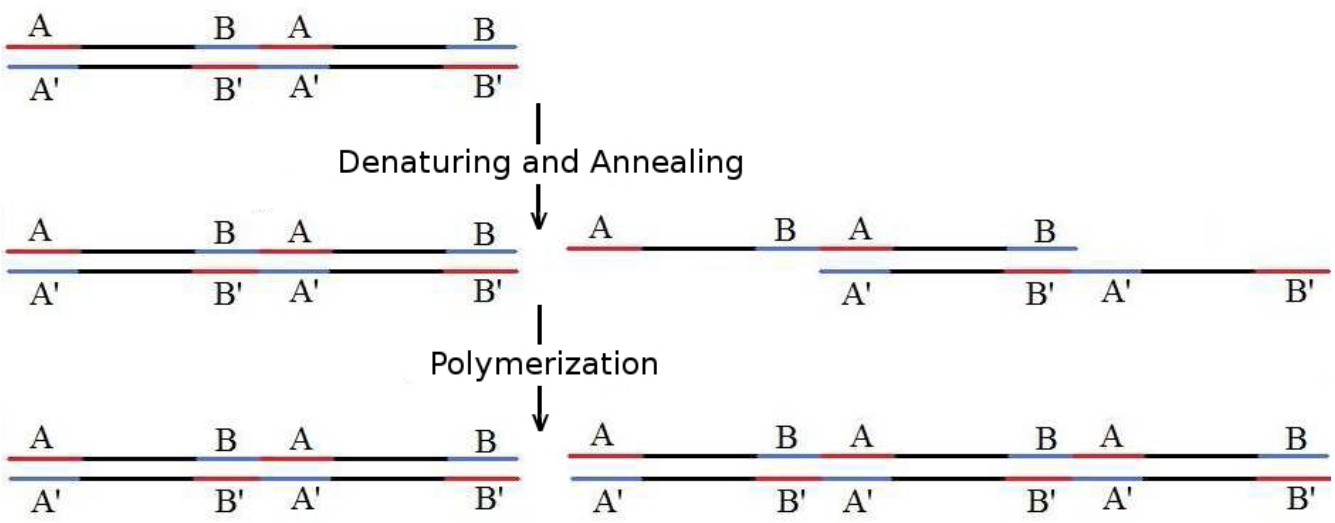

Figure 3.7: Schematic representation of a duplicated sequence being converted into a triplicated one. After denaturation and annealing without added primers, some strands have annealed in a mis-aligned manner, allowing for polymerization to form a complete triplicated strand.

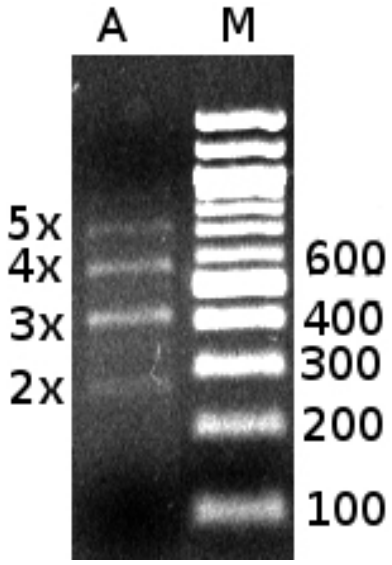

Figure 3.8: Using the duplicated 1RAB sequence to generate the triplicated 1RAB sequence. The size marker (lane M) is reported in base pairs. The duplicated sequence (marked as 2x, which is expected at 282 base pairs) is used as a template to generate larger sequences. The primary product is the intended triplicated sequence (marked as 3x, which is expected at 423 base pairs).

This method was employed, and although it was successful, the final stage of sequence preparation, the addition of sticky ends for orientation specific insertion into the ligated pET-28a vector, could not be performed successfully. Of more than 100 distinct PCR reactions involving this final process, all either failed to produce observable product, or produced a product which was too short, or too long (data not shown). Overall it was concluded that the primers were likely binding not only to the terminal complementary sequences, but also those internal to the triplicated sequences. While this likely caused no problem with the standard PCR reported earlier, the use of different primers with additional nucleotides encoding the sticky ends, would only partially anneal to the internal complementary sequences, leaving a number of trailing nucleotides, which likely interfered with the progression of the polymerase.

In the end, the full sequence was supplied in a pUC57 vector by GenScript. The vector was double digested using NdeI and BamHI and the sequence for 3RAB isolated using an agarose gel (10%). The isolated sequence with sticky-ends was then ligated into a similarly double digested pET-28a vector (as per 1RAB), and used to transform electrocompetent BL21 (DE3) cells (also as per 1RAB). The cost of simply ordering the gene was on the order of the cost for purchasing the reagents needed to attempt assembly, and was accompanied by a validated sequence and had a turn-around time of 2 weeks (but none of that time was the author's), overall a beneficial choice.

3.3.4 Expression of 1RAB and 3RAB

Expression of 1RAB and 3RAB from BL21 (DE3) cells was induced using IPTG. Expression of 1RAB was seen to be most stable at 25 °C (Figure 3.9). While expression of 3RAB was induced the manner, expression was stable at the more conventional 37 °C (Figure 3.11), perhaps indicating that 1RAB is less stable than 3RAB and is degraded by the host cell machinery under normal conditions. Both 1RAB and 3RAB expressed optimally after approximately 24 hours (Figure 3.10 and 3.11 respectively).

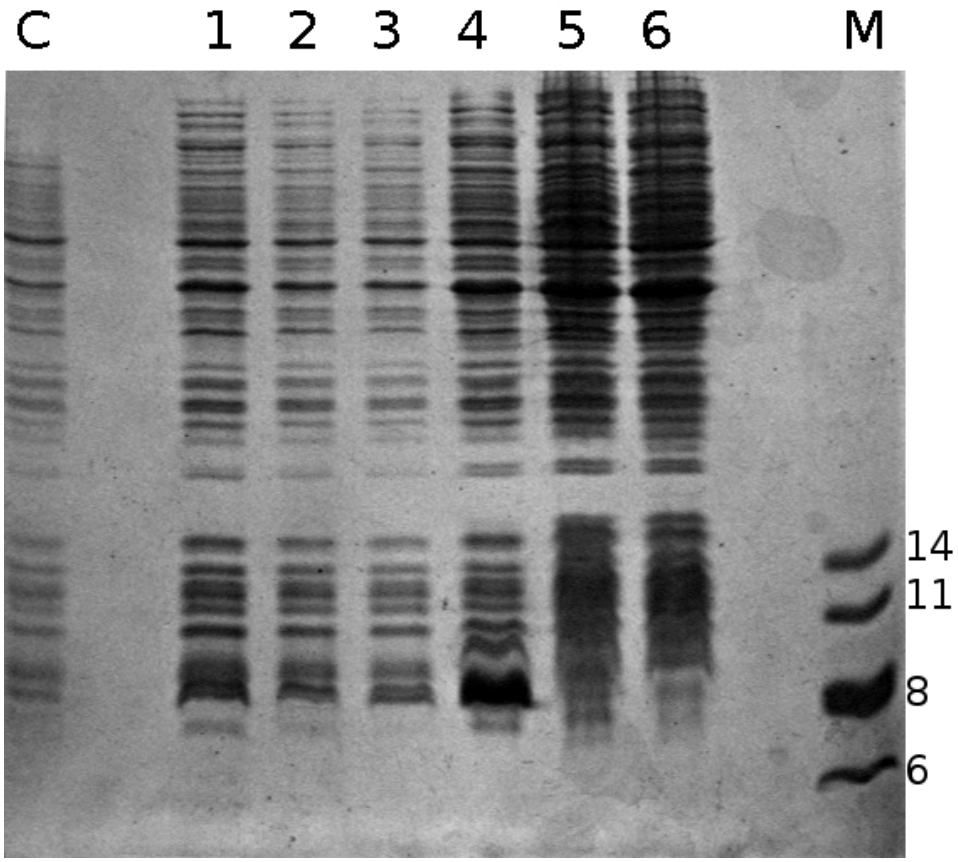

Figure 3.9: Temperature dependence of 1RAB expression. Pre-induction control (lane C), as compared against: 4 hours of expression at 25 (lane 1), 30 (lane 2), and 37 °C (lane 3), where little expression is seen, and 20 hours of expression at 25 (lane 4), 30 (lane 5) and 37 °C (lane 6), where considerable expression is seen at 25 °C. Marker (lane M) sizes are given in kDa, with expected size of tagged 1RAB being 7.8 kDa.

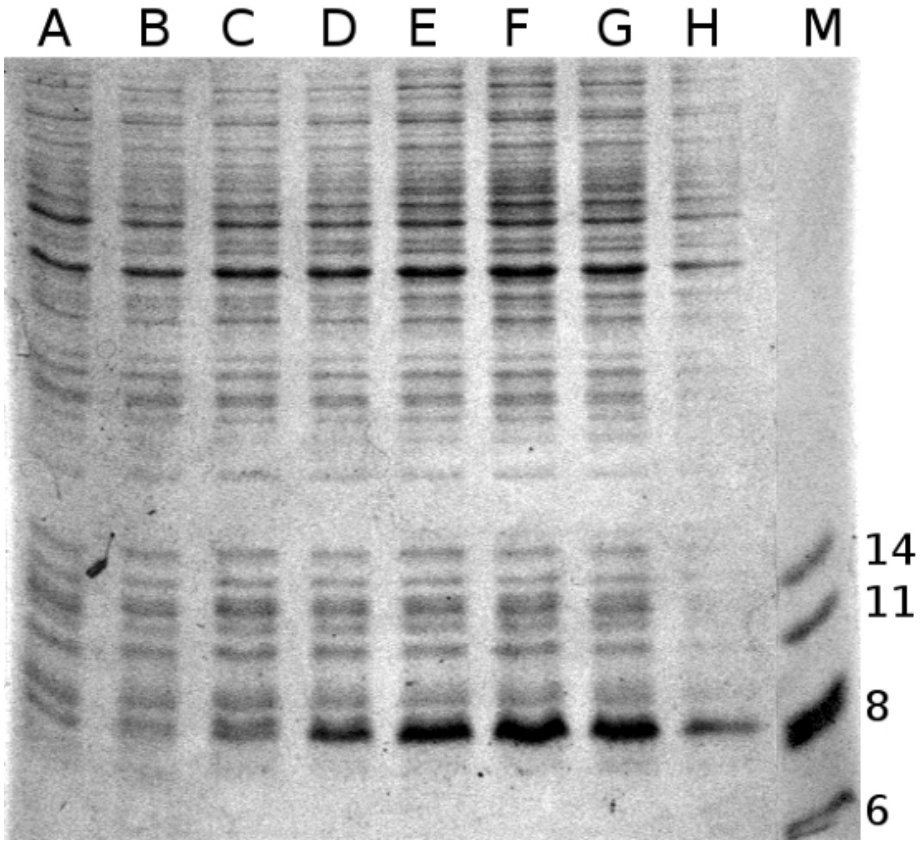

Figure 3.10: Timecourse of 1RAB expression at 25 °C. Increasing expression time at: 0, 4, 8, 12, 16, 24, 32, and 43 hours, lanes A to H respectively. Marker (lane M) sizes are given in kDa, tagged 1RAB is expected to be 7.8 kDa.

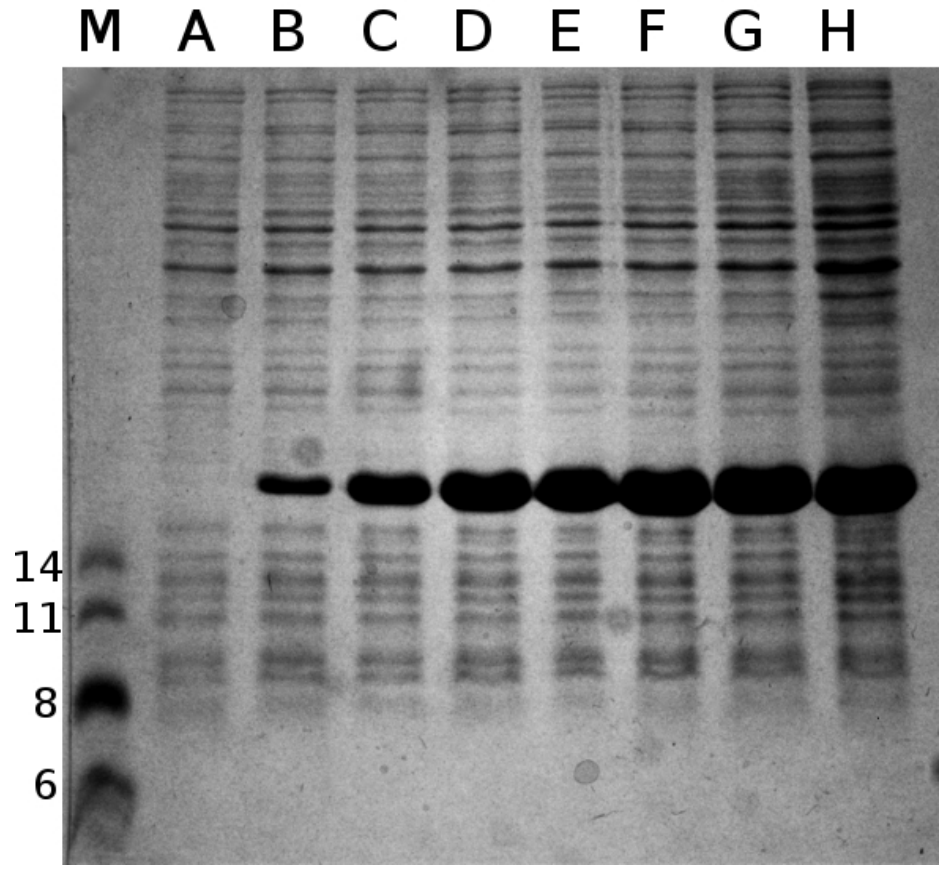

Figure 3.11: Timecourse of 3RAB expression at 37 °C. Increasing expression time at: 0, 4, 8, 12, 16, 24, 32, and 43 hours, lanes A to H respectively. Marker (lane M) sizes are given in kDa, tagged 3RAB is expected to be 18 kDa.

3.3.5 Purification of 1RAB and 3RAB

After successfully expressing both 1RAB and 3RAB, the task of isolating those proteins from the native host proteins began with determining if 1RAB and 3RAB were soluble in the cytoplasm, or localized as inclusion bodies. Inclusion body formation was briefly discussed (see Section 3.1) and is a likely outcome of the expression of foreign proteins, or in fact, of any expression that may greatly exceed the natural levels of expression for a given protein in a given host. Figure 3.12 shows the results of separating the soluble from insoluble fractions, where it can clearly be seen that 1RAB and

3RAB both exist predominantly in the insoluble fraction, suggesting that they exist as inclusion bodies (as there are no known examples of isolated beta-trefoils forming membrane bound complexes).

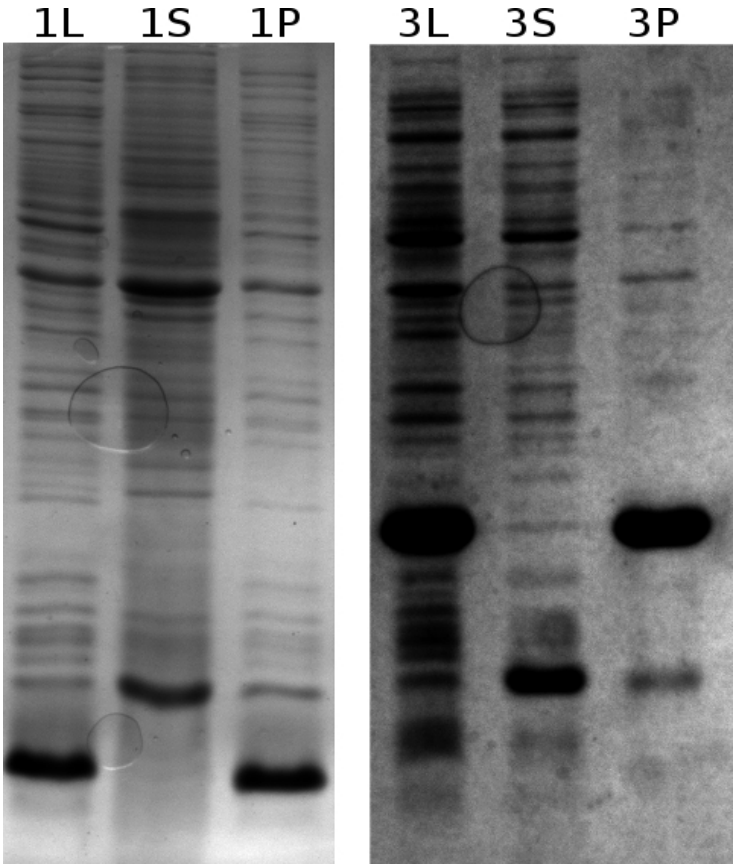

Figure 3.12: Inclusion bodies of 1RAB and 3RAB solubilized in urea. Over-expressed protein is seen in the initial lysate for 1RAB (lane 1L), and 3RAB (lane 3L). Following lysis, the soluble supernatant for 1RAB (lane 1S) and 3RAB **(**lane 3S**)** show no over-expressed protein, although a lower running band appears amplified in the case of 3RAB. The insoluble pellet contains all of the over-expressed protein for both 1RAB (lane 1P) and 3RAB (lane 3P).

Although the existence of 1RAB and 3RAB as inclusion bodies meant that extra steps were required in order to properly refold these proteins, it is quite likely that the final degree of purity is

much higher, as can be seen from Figure 3.12, where the insoluble fraction (pellet) already contains mostly the desired protein (particularly in the case of 3RAB).

After isolation of the inclusion bodies the protein must be completely purified from host contaminants, and additionally, solubilized in order to facilitate refolding. Since the pellets shown in Figure 3.12 are already solubilized in urea, that was a logical form in which to apply them to an affinity column. Also, the buffer used to solubilize the pellets had intentionally been set to a pH of 8, allowing the deca-histidine tag to bind to the affinity resin which contained nickel. The elutions from the column were monitored via absorbance at 280 nm, since the proteins of interest contain both tryptophan and tyrosine. During the initial washing at pH 8, a considerable quantity of absorbing material elutes (data not shown), but during the subsequent wash steps at pH 6.5 very little additional material is detected (data not shown). The results of the final elution of bound proteins at a pH of 4.5 can be seen in Figure 3.13. In both cases the proteins appear to be quite pure (note that while some contaminating bands can be seen for 3RAB, whereas none are seen for 1RAB, the density of the 3RAB band is in considerable excess to those contaminants. Additionally the yield for both proteins was within normal ranges, ~50 mg/L of initial culture for 1RAB and ~100 mg/L of initial culture for 3RAB.

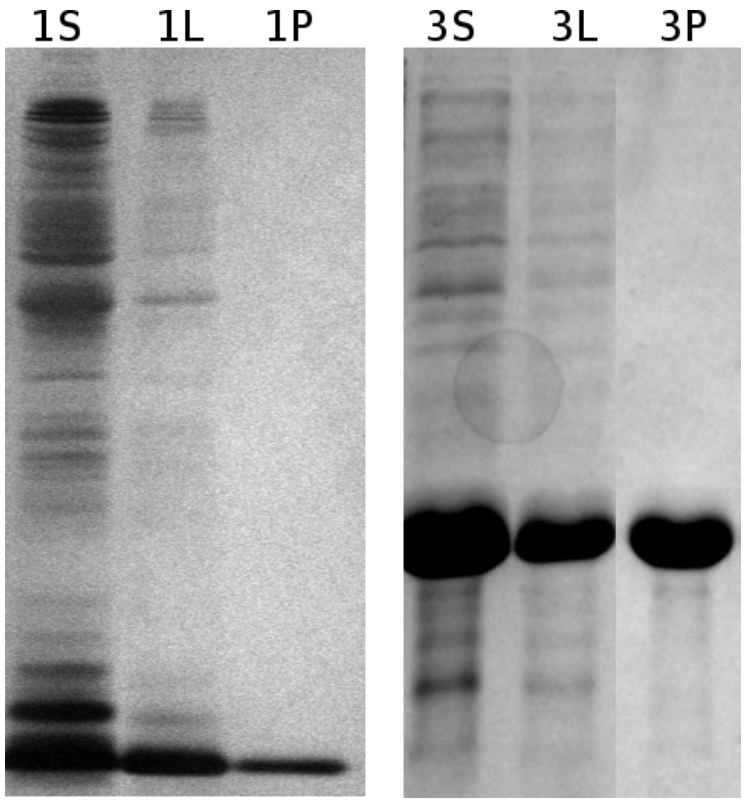

Figure 3.13: Column purification of 1RAB and 3RAB. The initial cell lysate, 1S and 3S for 1RAB and 3RAB respectively, shows over-expression of the proteins. The insoluble fractions, 1L and 3L show that much of the protein has been maintained, and the final purified fractions 1P and 3P show excellent purification. Note that in the case of 3RAB, although some faint lower running bands can be seen, the 3RAB band is extremely intense by comparison.

3.3.6 Refolding of 1RAB and 3RAB

Since protein net charge can often strongly influence aggregation, initial refolding attempts were performed at a variety of pHs in order to determine the optimum pH for successful refolding, along with some examination of salt concentration and choice of buffer. The results can be seen in Table 3.1.

Table 3.1: 1RAB and 3RAB refolding conditions				
pH	**[NaCl]**	**Buffer**	**% Yield 1RAB**	**% Yield 3RAB**
6	300 mM	100 mM phosphate	56	74
6.5	300 mM	100 mM phosphate	99	100
7	300 mM	100 mM phosphate	74	78
7.5	300 mM	100 mM phosphate	90	96
8	300 mM	100 mM phosphate	78	97
6.5	3000 mM	100 mM phosphate	78	N/A
6.5	0 mM	100 mM phosphate	34	28
6.5	0 mM	50 mM Tris	31	6
6.5	0 mM	50 mM HEPES	22	5
6.5	0 mM	50 mM MES	25	8
Approximate yields after refolding using a 1 kDa cutoff membrane, filled to ~1 mL in a 1 L reservoir.				

In Table 3.1 it can be seen that the ideal pH for refolding of both 1RAB and 3RAB is 6.5. It also appears as though phosphate is the best buffer for refolding, although the data may be misleading, as it is clear that some salt is required in order to obtain a good yield, and buffers such as Tris or HEPES may have lower ionic strengths then phosphate at the same concentration. Moreover, buffers such as Tris do not buffer well at a pH of 6.5, and may have shifted slightly towards more acidic conditions, as unfolded protein solution was at a pH of 4.5. Based on Table 3.1, it was decided that refolding should take place using a buffer containing 300 mM NaCl and 100 mM phosphate buffer near pH 6.5. One may note the reduction in yield near a pH of 7 for both 1RAB and 3RAB, which may be a result of pI, as the predicted pI for 1RAB and 3RAB is 7.4 and 6.8 respectively.

After determining good refolding conditions using small samples and a small cutoff membrane (which might allow more urea to persist than would be desired), the refolding was attempted using much larger volumes (up to 100 mL) and a larger membrane (3 kDa cutoff). Absorbance measurements at 280 nm revealed that after refolding and filtration the apparent yield of refolded

species for 1RAB and 3RAB was ~75% and ~100% respectively. The reduced yield of 1RAB in comparison to 3RAB may be due to an increased propensity to form insoluble aggregates, or, may be due to a loss of the smaller peptide through the dialysis membrane, as the preliminary refolding results (Table 3.1) used a more expensive 1 kDa cutoff membrane, instead of the 3 kDa cutoff membrane used for the later, large scale preparations. Additionally, the poorer results with 1RAB using the 3 kDa cutoff membrane may be a result of a different membrane composition, as the 3 kDa membrane was made from cellulose, and the 1 kDa membrane from cellulose-ester. After refolding, both 1RAB and 3RAB were concentrated to 100 μM and 695 μM respectively.

The molar absorption coefficient at 280 nm used to determine the concentration of both 1RAB and 3RAB was determined by a simple theoretical calculation in which each tryptophan and tyrosine residue is presumed to absorb as it would in an unstructured peptide (51). The result was 11200 $Lmol^{-1}cm^{-1}$ for 1RAB and 33600 $Lmol^{-1}cm^{-1}$ for 3RAB.

3.4 Summary of Results

Both 1RAB and 3RAB express well in BL21 (DE3) cells as insoluble inclusion bodies, although 1RAB requires growth at 25 °C rather than 37 °C. Both can be solubilized and purified using denaturant and a nickel affinity resin, and subsequently refolded via dialysis. At present, 3RAB can be concentrated to at least 695 μM (12.5 mg/mL), whereas 1RAB does not appear soluble in Buffer D beyond 100 μM (0.78 mg/mL). The concentration ranges reached for 3RAB make certain characterization experiments possible, including proton nuclear magnetic resonance (^{1}H-NMR), dynamic light scattering (DLS), and crystallization attempts. 1RAB is not suitably concentrated for ^{1}H-NMR or crystallization, yet can be analyzed through fluorescence, circular dichroism (CD) and differential scanning calorimetry (DSC).

3.5 Discussion

The fact that both 1RAB and 3RAB were able to express to a reasonable degree is in of itself exciting and indicative of at least some stability, as unstructured protein tends to be degraded. A similar attempt to generate a repetitive protein (although not a globular one) showed that stable expression was only seen for larger covalent assemblies of the peptide (at least a four-fold repeat) (30). This might indicate that the sequence chosen for 1RAB is itself fairly stable, or that it rapidly associates to form a homotrimer, thereby avoiding host proteases.

The use of a deca-histidine tag appears to have been very effective, as a high degree of purity is seen from SDS-PAGE and also good yield. The tag itself is cleavable via thrombin, which could prove invaluable particularly for the single repeat (1RAB). Although the histidine tag does not generally appear to affect structure (see Section 3.1), it can impede crystallization, which may be necessary in order to determine the structure of both 1RAB and 3RAB. Moreover, the relatively large size of the histidine tag (24 residues) in comparison to 1RAB (47 residues) suggests that in this case the tag may in fact affect structure, and possibly stability also.

It is likely that different solution conditions are required in order to properly concentrate 1RAB, such as a different pH or buffer, different salt concentrations, or additives such as glycerol or sucrose. That both 1RAB and 3RAB are suspected to bind carbohydrates from their sequence information, suggests that adding carbohydrates to the buffer may be a particularly effective tactic for obtaining a more concentrated sample, where spurious association of 1RAB is prohibited.

Even without a concentration of 1RAB comparable to that of 3RAB, many characterization techniques can be applied to both, in order to begin determining if the overall goal – the creation of a stable homotrimer capable of binding sugar, and triplicated monomer – has been successful.

4 Characterization

4.1 Introduction

Since the designed proteins: 1RAB and 3RAB are supposed to represent a beta-trefoil fold, they should display the properties of the beta-trefoil fold: a well packed core, cooperative folding, soluble, and should also have the structure of a beta-trefoil.

There exist many experimental techniques for determining a protein's capacity to meet the above requirements, and in many cases more than one technique can be applied to determine a particular property. Unfortunately, because 1RAB could not be concentrated to sufficient levels, some experiments were only performed on 3RAB. Also, direct structural determination through crystallography or multidimensional nuclear magnetic resonance (NMR) could not be performed in the case of either 1RAB or 3RAB due to concentration and time constraints respectively.

In general the experiments that were focused on were those that gave a reasonable quantity of information very quickly and without destroying or using up large quantities of protein. Many techniques such as NMR, circular dichroism (CD) and dynamic light scattering (DLS) do not affect the sample, and are therefore quite attractive.

In order to understand the significance of some results, fluorescence in particular, it is important to have a model for the positioning of certain key residues such as the tryptophans. Figure 4.1 shows a homology model of 3RAB built using the closest matching structure, 2IHO (a galactose binding lectin) as a backbone. It is clear that the tryptophans are expected to be within the hydrophobic core of the protein. Moreover, if 1RAB has a similar shape, where the core hydrophobic residues from each repeat form a hydrophobic patch with which to trimerize, it is expected that those tryptophans should move from being solvent exposed when a monomer, to a hydrophobic environment after trimerization.

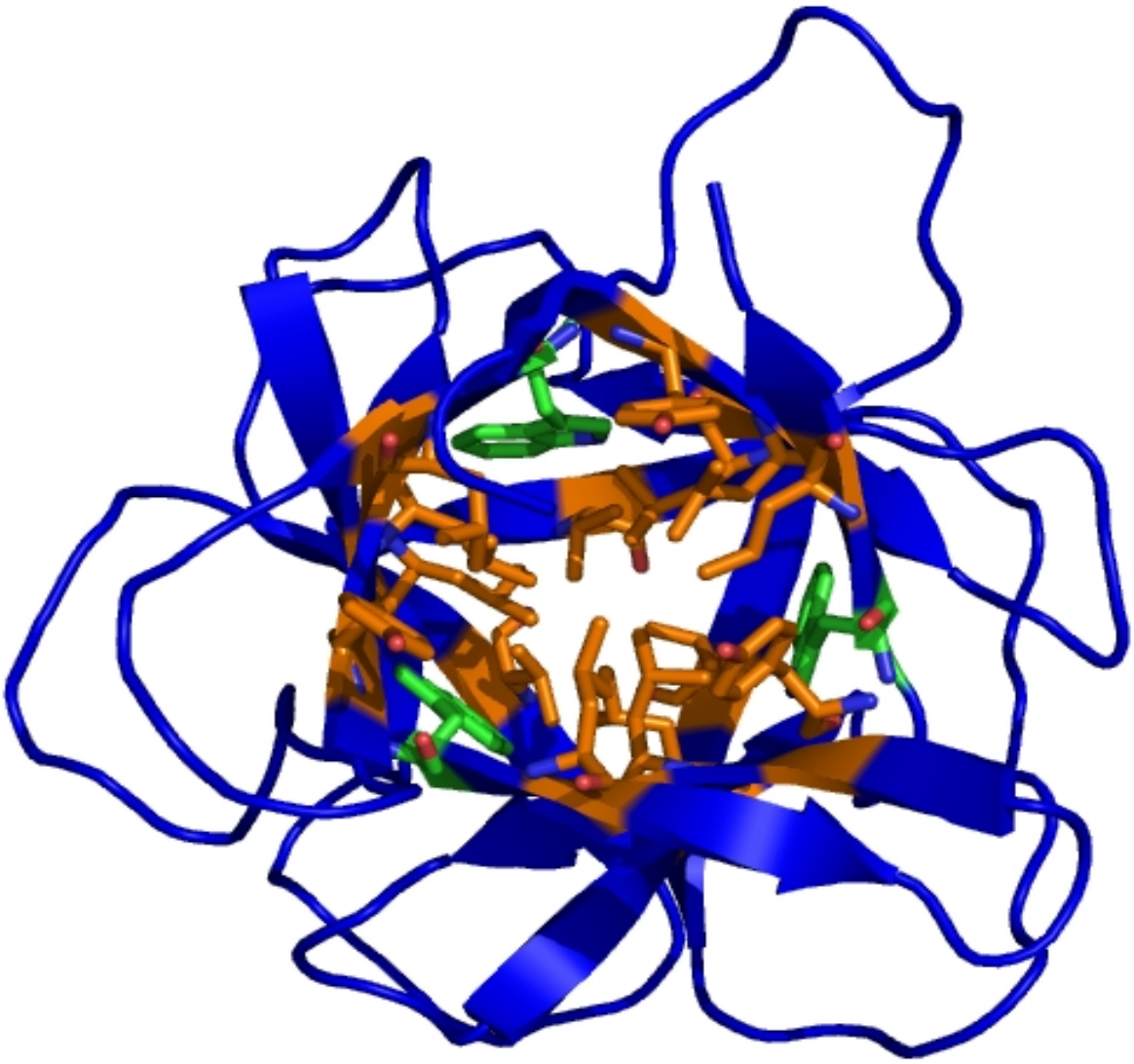

Figure 4.1: Homology model of 3RAB. A homology model of 3RAB built using the backbone of a galactose binding lectin (PDB code, 2IHO), one of the three closest matching beta-trefoil structures found using BLAST (33). The structure is shown as a ribbon representation, with 15 of the 18 core hydrophobic residues (5 of 6 from each repeat unit) shown in orange, with the remainder being tryptophan residues critical for structure assessment, shown in green. The image was made using Pymol.

4.2 Methods

4.2.1 Circular Dichroism (CD)

CD was performed using a JASCO-715 with 0.1 cm path length, 0.1 nm step size and 50 nm/min scan rate, using 0.12 mg/mL (3RAB) and 0.093 mg/mL (1RAB) in buffer Z (300 mM KF, 100 mM phosphate, pH 6.6). KF was necessary over NaCl because chloride ions absorb UV light below 200 nm.

4.2.2 Proton Nuclear Magnetic Resonance (^{1}H-NMR)

^{1}H-NMR was performed on a Bruker 600 MHz using a TXI probe and water presaturation at a concentration of 12.5 mg/mL 3RAB at 25 °C in buffer Z.

4.2.3 Dynamic Light Scattering (DLS)

DLS and SLS measurements were performed using a Malvern, Zetasizer Nano, with 45 uL cuvette, and 633 nm laser at room temperature. In the case of 1RAB a concentration of 0.78 mg/mL was used in buffer Z, and in the case of 3RAB a concentration range from 1.5 to 12.5 mg/mL were used also in buffer Z, which facilitated not only a direct determination of hydrodynamic radius, but also the inference of molecular weight and second virial coefficient through a Debye plot (52).

4.2.4 Differential Scanning Calorimetry (DSC)

Differential scanning calorimetry (DSC) measurements were performed on a MicroCal VP-DSC, using a 0.51 mL cell, and a 0.5 °C/min scan rate, and a concentration of 0.43 mg/mL 3RAB in buffer Y (300 mM NaCl, 100 mM phosphate, pH 6.6).

4.2.5 Fluorescence Spectroscopy

Fluorescence spectra measurements for 1RAB and 3RAB were performed on a Fluorolog (ISA), using a 1 cm path length cuvette, with an excitation wavelength of 280 nm at room temperature with 1 mm slit widths. 1RAB and 3RAB were in Buffer Y plus urea. Renaturation and denaturation curves for 1RAB were performed using a 0.5 cm path length 45 uL cuvette, monitored at 360 nm. In the case of 3RAB refolding curves, fluorescence was monitored at 313 nm, and measured were performed as for 3RAB spectra, but using a Photon Technology International QuantaMaster 4.

4.2.6 Size Exclusion Chromatography (SEC)

SEC was performed using a Superose 12 10/300 GL (Amersham Bioscience) at a flow rate of 0.5 mL/min buffer Y (300 mM NaCl, 100 mM phosphate, pH 6.6), and an injection volume of 100 uL of either 1RAB of 3RAB in buffer Y. In the case of analysing sugar binding using SEC, buffer Y was supplemented with D-galactose.

4.3 Results

Although the primary objective of this research was to establish if 1RAB could form a homotrimer with a beta-trefoil fold, 3RAB was much easier to work with, and as the results for 3RAB have a more easily understood interpretation, those results have been presented first (contrary to what one might intuitively expect).

4.3.1 Fluorescence of 3RAB

Fluorescence spectra for 3RAB (0.020 mg/mL) reveal a marked blue shift from 360 nm when unfolded (8 M Urea) to a double peak at 313 and 323 nm when folded (300 mM Urea) (Figure 4.2). A peak at 360 nm is characteristic of unfolded proteins, whereas a considerable blue shift is seen only for

the tryptophan sidechain in a hydrophobic environment, indicating that the symmetrical tryptophans become buried within the hydrophobic core (53). It may also be possible that the peak at 313 nm is from tyrosine (79), as there are a large number of tyrosine residues in 3RAB (12 in total). This possibility was mostly ruled out by examining the shape of the spectrum using an excitation wavelength of both 280 and 295 nm. In the case of 280 nm (Figure 4.2), the fluorescence can often come from both tryptophan and tyrosine, but at 295 nm (data not shown), the fluorescence should be primarily due to tryptophan (79). In this case no difference in the shape of the spectrum was seen between the aforementioned excitation wavelengths (data not shown), which indicates that tyrosine is not responsible for either of the peaks seen in the spectrum. A renaturation curve at 0.12 mg/mL also shows a cooperative transition (Figure 4.3), but a curve was not observed for denaturation samples, as it appeared to take multiple weeks for any appreciable degree of unfolding to occur even in 8 M urea or 6 M GuHCl. As a result, denaturation curve fluorescence for 3RAB appears the same as folded protein (data not shown). Without both a denaturation and renaturation plot, it is difficult to determine if the samples are at equilibrium, and therefore, impossible to determine the thermodynamic stability by fitting these measurements. This results from the nature of thermodynamic analysis in that it requires a reversible equilibrium in order to determine correct parameters.

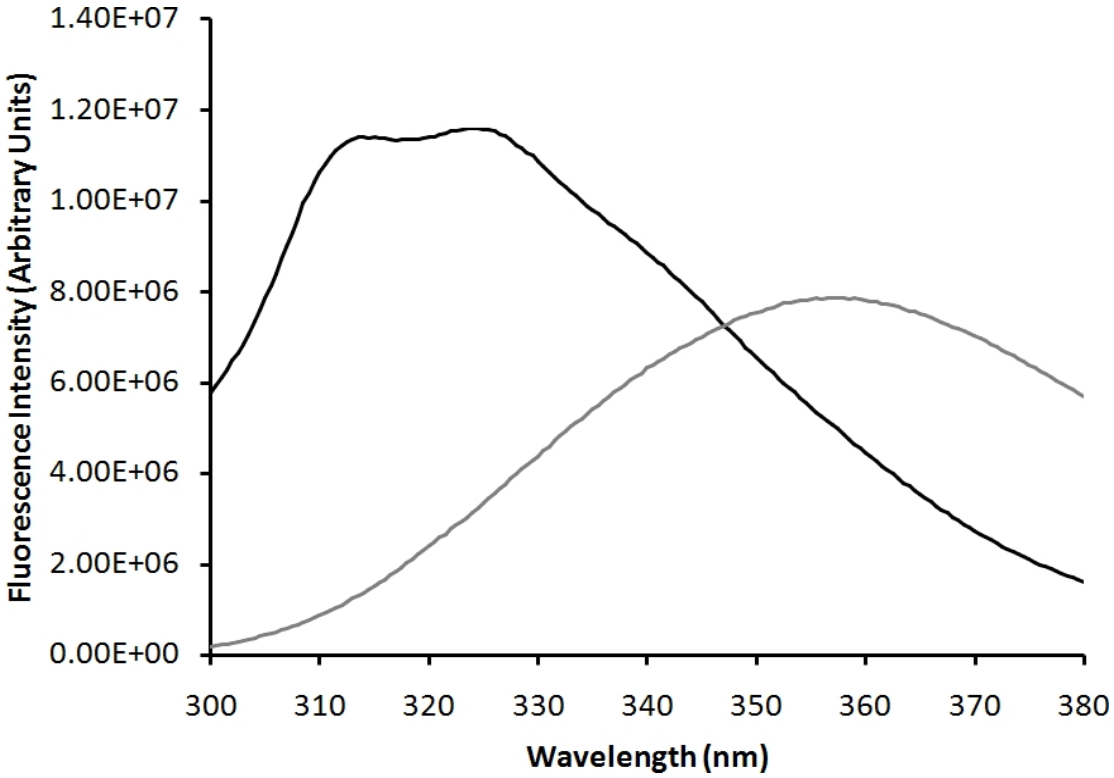

Figure 4.2: Fluorescence spectra of folded and unfolded 3RAB. The folded spectra (black) shows a peak at 313 nm and 323 nm, whereas the unfolded spectra (grey line) shows a broad peak at 360 nm. The excitation wavelength was 280 nm.

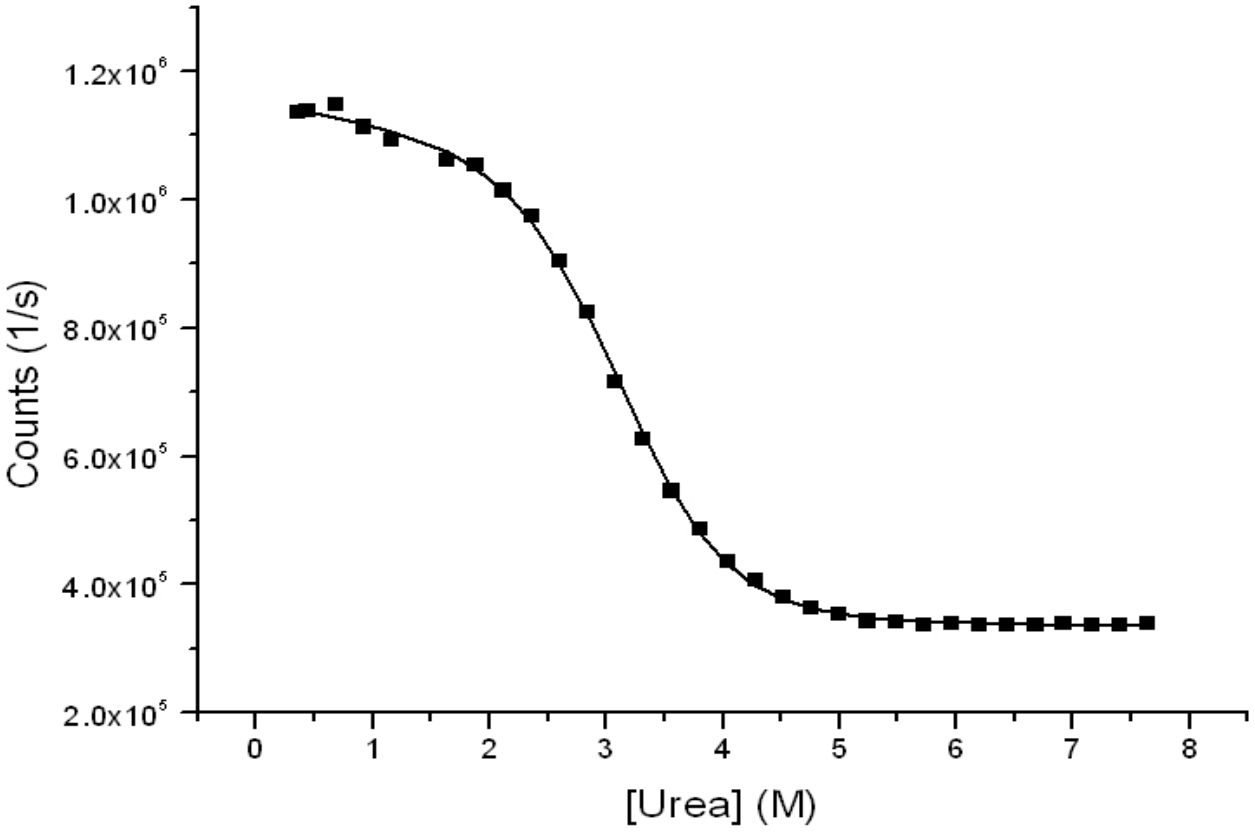

Figure 4.3: Refolding fluorescence curve of 3RAB. Monitored at 313 nm with 280 nm excitation. Although the samples may not be at equilibrium, precluding the possibility for thermodynamic analysis, it is possible to conclude that 3RAB refolds cooperatively, given the sharp transition centred at 3 M urea.

4.3.2 Differential Scanning Calorimetry of 3RAB

The DSC plot for 3RAB (Figure 4.4) shows that it has a very high apparent melting point of 94 °C, and the single sharp peak indicates that it unfolds cooperatively, which is in agreement with the aforementioned fluorescence measurements.

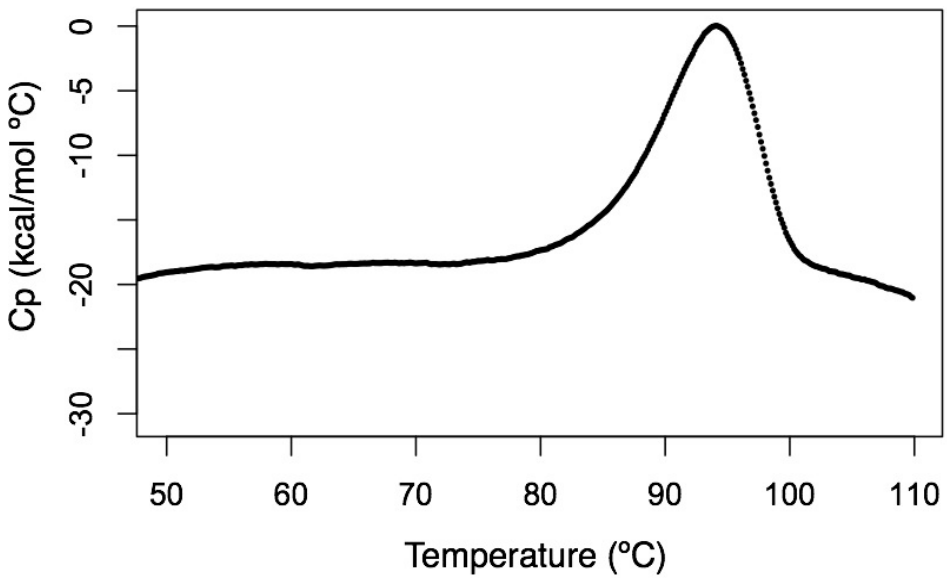

Figure 4.4: Differential scanning calorimetry of 3RAB. The single smooth exothermic peak indicates a cooperative unfolding transition, with a melting point of 94 °C, exceedingly thermo-stable.

4.3.3 Circular Dichroism of 3RAB

The CD spectrum of 3RAB (Figure 4.5) was submitted to DICHROWEB (54), and analyzed using several algorithms which predict secondary structural content (CDSSTR (55), CONTIN (56), SELCON3 (57), using the SP175 dataset (58)). The prediction results (Table 4.1) indicate that 3RAB forms a mainly beta structure with approximately 12 strands in total, which is the number of strands seen in beta-trefoils. Also, the SELCON3 algorithm predicted the closest matching structure in the dataset to be that of Kunitz soybean trypsin inhibitor (59) (PDB code, 1AVW) a beta-trefoil. The CD spectrum of Kunitz soybean trypsin inhibitor (60) is shown in Figure 4.6, which shows the same shape as that for 3RAB, although the magnitude of signal is different. A difference in magnitude appears to be a standard observation for beta-trefoil proteins, as the magnitude of the mean residue ellipticity ([θ]) near 200 nm can range from -5000 (60) to -500,000 (61). Together these results indicate that 3RAB likely adopts a beta-trefoil fold.

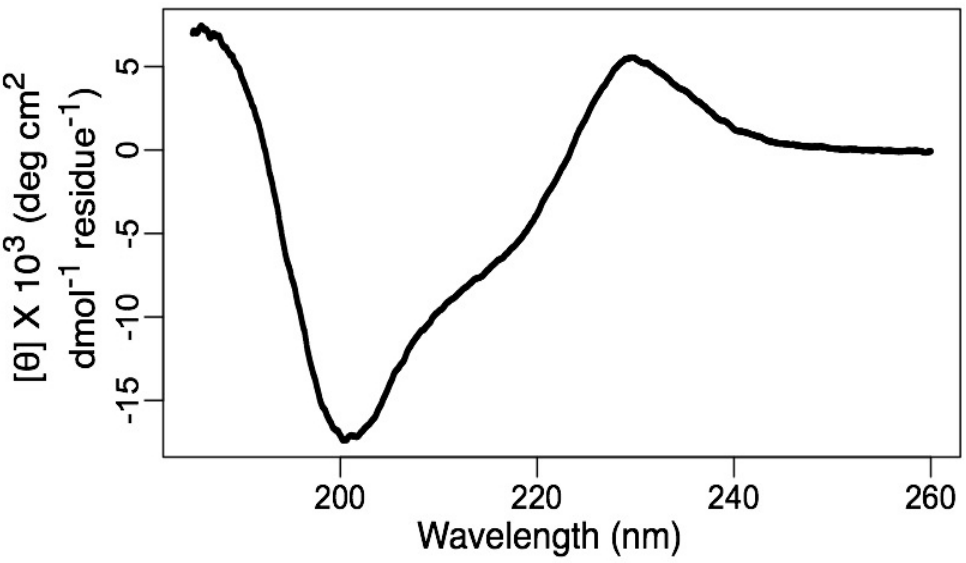

Figure 4.5: Circular dichroism spectrum of folded 3RAB. Characteristic beta-trefoil characteristics are seen, such as a maximum at ~230 nm and a minimum between 200-205 nm.

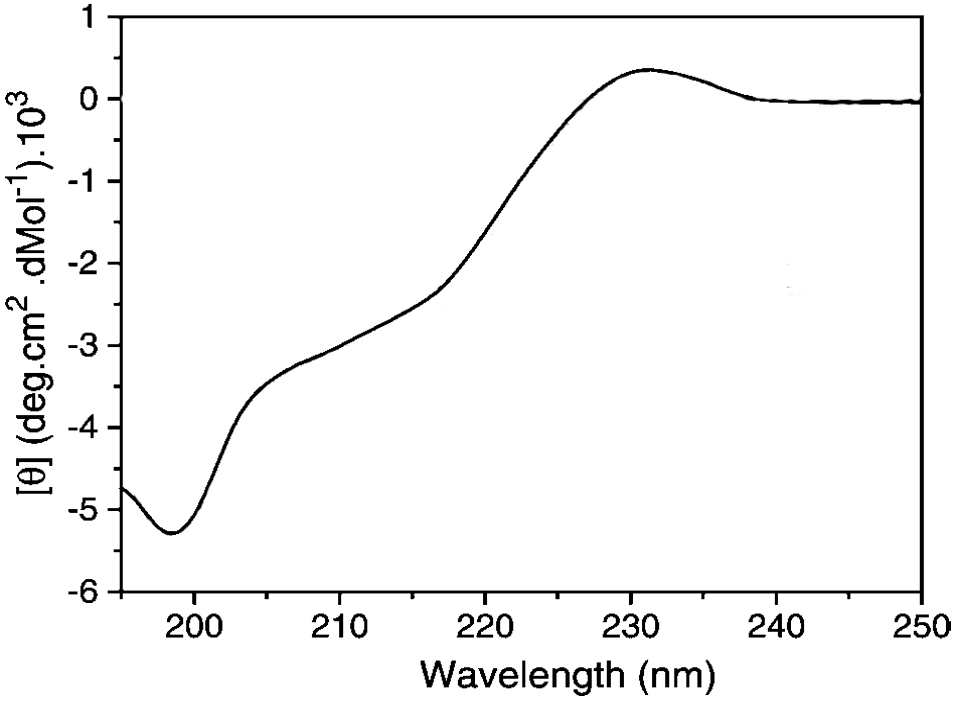

Figure 4.6: CD spectrum of STI is similar to 3RAB. The same features that are seen in 3RAB, such as a maximum at ~230 nm, a shoulder at ~210 nm, and a minimum at ~200 nm are seen in the soybean trypsin inhibitor, a beta-trefoil protein.

Table 4.1: Frequency of secondary structure in 3RAB from prediction algorithms.								
Algorithm	**Ordered Helix**	**Disordered Helix**	**Ordered Strand**	**Disordered Strand**	**Turn**	**Unordered**	**Strands per protein**	**Average Strand Length**
CDSSTR	-0.01	0.02	0.32	0.15	0.11	0.39	12.5	6.3
CONTIN	0.00	0.05	0.30	0.15	0.09	0.40	12.4	6.0
SELCON3	-0.03	0.04	0.31	0.15	0.10	0.41	12.2	6.2

4.3.4 Proton Nuclear Magnetic Resonance of 3RAB

^{1}H-NMR spectroscopy of 3RAB (Figure 4.7) shows characteristic beta-features such as amide resonances >9.5 ppm and alpha resonances >5 ppm. It also demonstrates that 3RAB possesses a well packed hydrophobic core (methyl resonances <1 ppm). These key chemical shifts together with the well dispersed spectrum, demonstrate that 3RAB has a well defined globular structure, characteristic of a beta-trefoil.

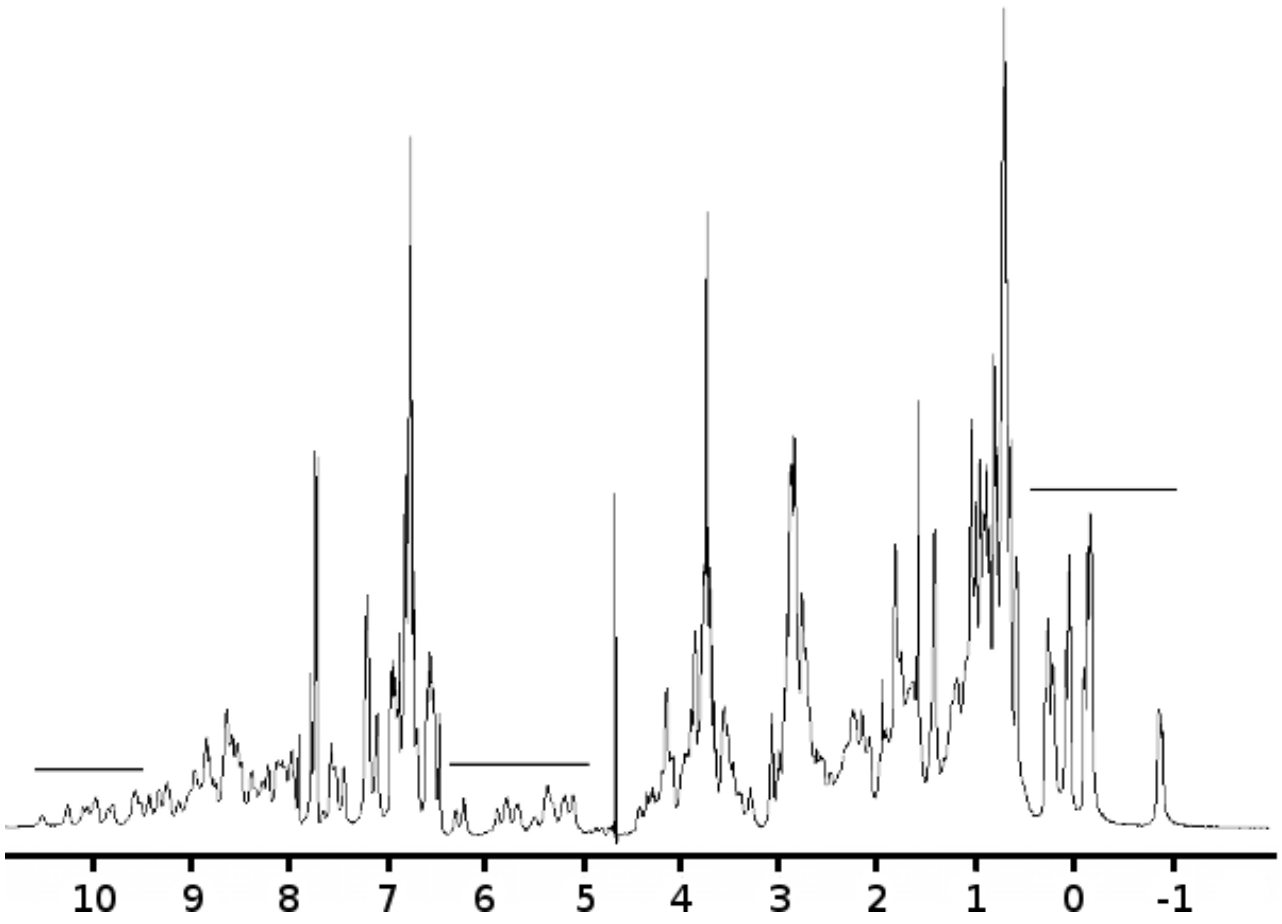

Figure 4.7: ^{1}H-NMR of 3RAB. The spectrum is well dispersed and shows characteristic features of beta structure such as amide resonances >9.5 ppm, and alpha resonances >5 ppm. Methyl resonances are seen <1 ppm, indicating a packed hydrophobic core. Areas of interest are highlighted with a bar.

4.3.5 Dynamic Light Scattering of 3RAB

DLS measurements gave a predicted diameter of 3 nm, in good agreement with the theoretical diameter of 3 nm from homology modelling. Additionally, a Debye plot (Figure 4.8), shows that the molecular weight is estimated as 18.4 kDa, in good agreement with the theoretical weight of the tagged monomer at 18.0 kDa. Debye analysis also gave a second virial coefficient of 6.72×10^{-5} mLmolg^{-1} for 3RAB indicating that it prefers interactions with the solvent over self interactions (62), and is therefore quite soluble. Moreover, DLS measurements were monodisperse, even at concentrations as high as

12.5 mg/mL (data not shown). The maintained monodispersity, estimated molecular weight and positive second virial coefficient all indicate that 3RAB forms a monomer in solution.

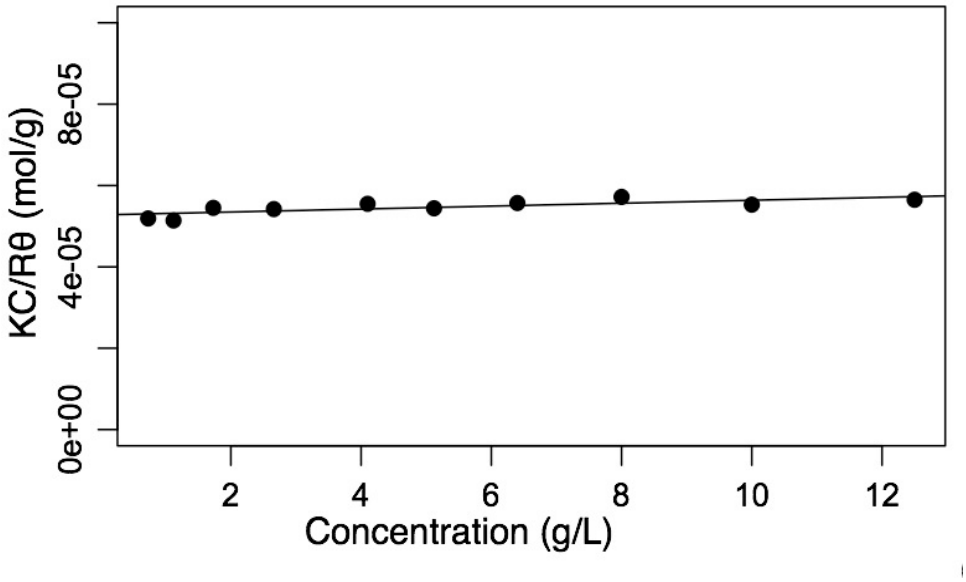

Figure 4.8: Debye plot of 3RAB. The data is well fit linearly over a large concentration range, indicating a good estimate of molecular weight (from intercept), and second virial coefficient (from slope). In this case the molecular weight is estimated as 18.4 kDa (18.0 kDa theoretical using the sequence), and a positive second virial coefficient indicates a preference for solvent association rather than self-association.

4.3.6 Size Exclusion Chromatography of 3RAB

SEC results were confounded by a problem whereby 3RAB had favourable interactions with the resin, slowing its progress and making it appear to be 8.3 kDa (Figure 4.9) (note that in size exclusion chromatography smaller molecules elute later than larger ones). Since all data to this point (theoretical estimates, SDS-PAGE, DLS) indicated a minimum monomer size of 18.0 kDa and that 3RAB was completely intact during the SEC run (SDS-PAGE data not shown), it was concluded that indeed some

form of column interaction had occurred. The possible transient binding to the column is not entirely unexpected, as the resin used is made from covalently linked galactose units, and the ricin family to which this sequence would belong, are well known to bind carbohydrates (19). In particular the closest known structure which was used for homology modelling (1KNM, a xylan binding domain) has been shown to bind galactose (17).

In order to validate the theory that binding to the resin was causing the anomalous elution times, a series of experiments using increasingly high concentrations of galactose were performed. Standards and 3RAB were rerun at each concentration of galactose, and the final results (Figure 4.10) show a hyperbolic curve in which the apparent weight of 3RAB increases with increasing galactose concentration, consistent with the proposition that 3RAB binds to galactose units in the resin, slowing its elution time. Since the results demonstrate a hyperbolic shape, it is possible to make a very rough estimate of the asymptote of the curve, which represents the true molecular weight of 3RAB, when the concentration of galactose in the running buffer out-competes all resin interactions. Fitting to a hyperbolic function ($y = [y_{max}*x]/[x\_at\_y_{max}/2 + x]$) gave an asymptotic value of 19 kDa, in good agreement with the theoretical monomer weight of 18 kDa, demonstrating not only that 3RAB is a monomer in solution, but also that it binds to galactose. Since the number of binding sites on the resin is unknown, it is impossible to make predictions about the binding affinity to galactose, and therefore impossible to estimate a K_d.

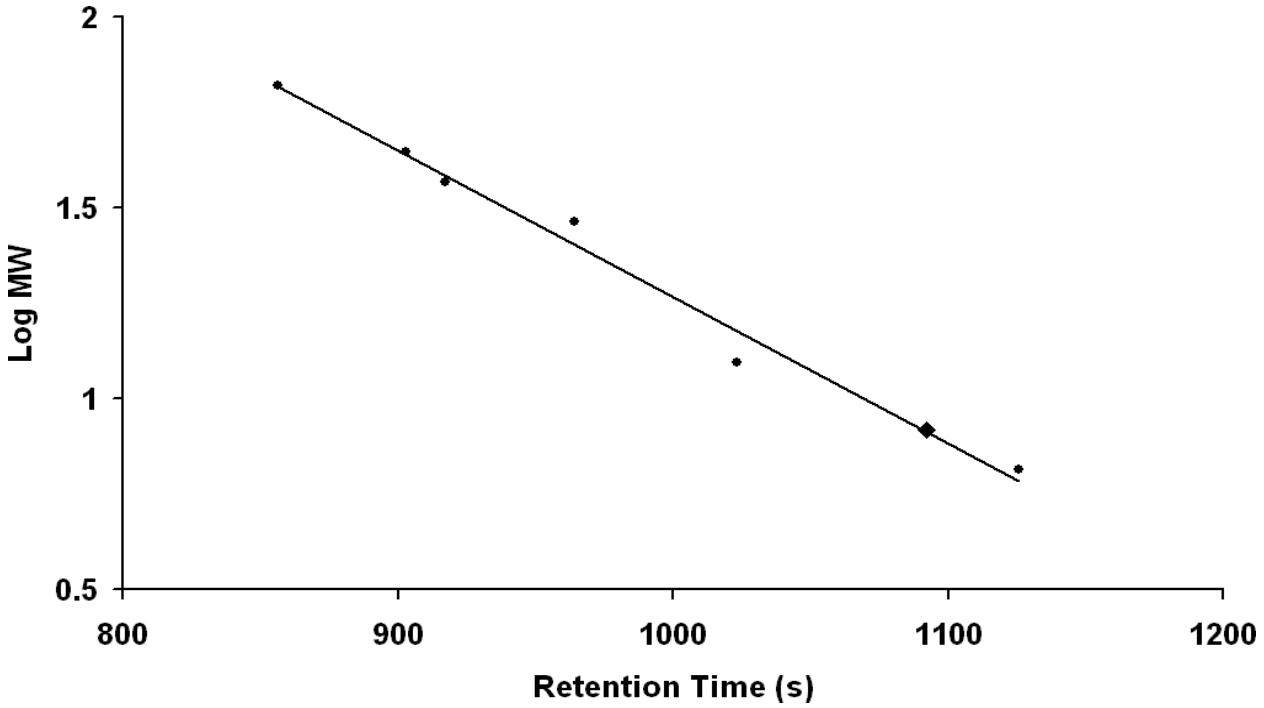

Figure 4.9: Size exclusion standards and 3RAB. Standards: Bovine Serum Albumin (66.3 kDa), Ovalbumin (44.3 kDa), Beta-lactoglobulin (36.8 kDa), Carbonic Anhydrase (29.0 kDa), Cytochrome C (12.4 kDa), and Aprotinin (6.5 kDa), are shown with small circles, and a linear best fit as a solid black line. 3RAB is shown as a large black diamond, resulting in an estimated size of 8.3 kDa.

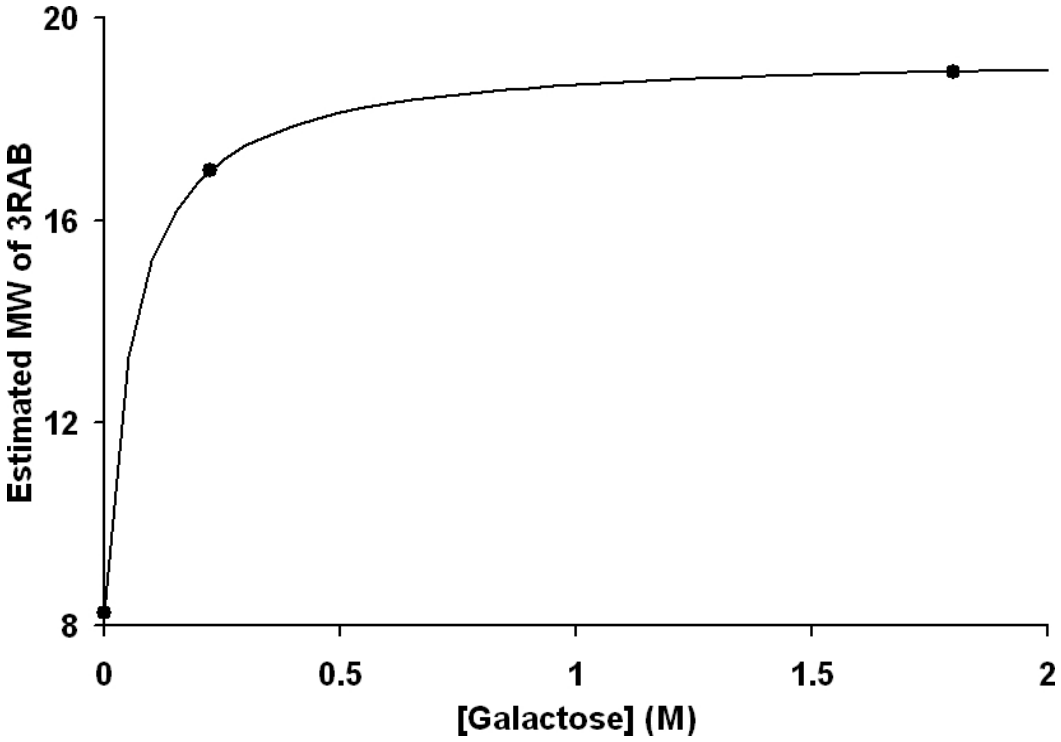

Figure 4.10: Effect of galactose on apparent 3RAB size. Increasing concentrations of D-galactose used as running buffer in the size exclusion column, resulted in increasingly large estimates of the molecular weight. A fit to a hyperbolic function gave an estimate for true weight as 19 kDa.

4.3.7 Fluorescence of 1RAB

Fluorescence spectra for 1RAB (0.018 mg/mL) show at a blue shift of at most 3 nm from a high denaturant (8 M urea) peak at 360 nm to a low denaturant (<100 mM urea) peak at 357 nm (Figure 4.11). A lack of blue shift would indicate that there is not an appreciable difference in the solvent exposure or chemical environment of the tryptophan upon moving from high denaturant to low denaturant (53). Since fluorescence under these conditions does not appear to be a good indicator of 1RAB structure, it is not surprising that renaturation and denaturation curves of 1RAB (0.016 mg/mL) monitored at 360 nm were linear (Figure 4.12). Incubation times for the denaturation and renaturation plots were several weeks, whereas spectra remain consistent beyond 1 month on incubation, making it

extremely unlikely that the lack of an observable difference is merely due to a slow equilibrium. The lack of an obvious transition in the renaturation or denaturation curves would normally indicate that the protein does not fold/unfold in a cooperative manner. The lack of a blue shift in the spectra with and without denaturant, however, suggests that while this may be true, it is also possible that tryptophan fluorescence under these conditions simply fails to give any information.

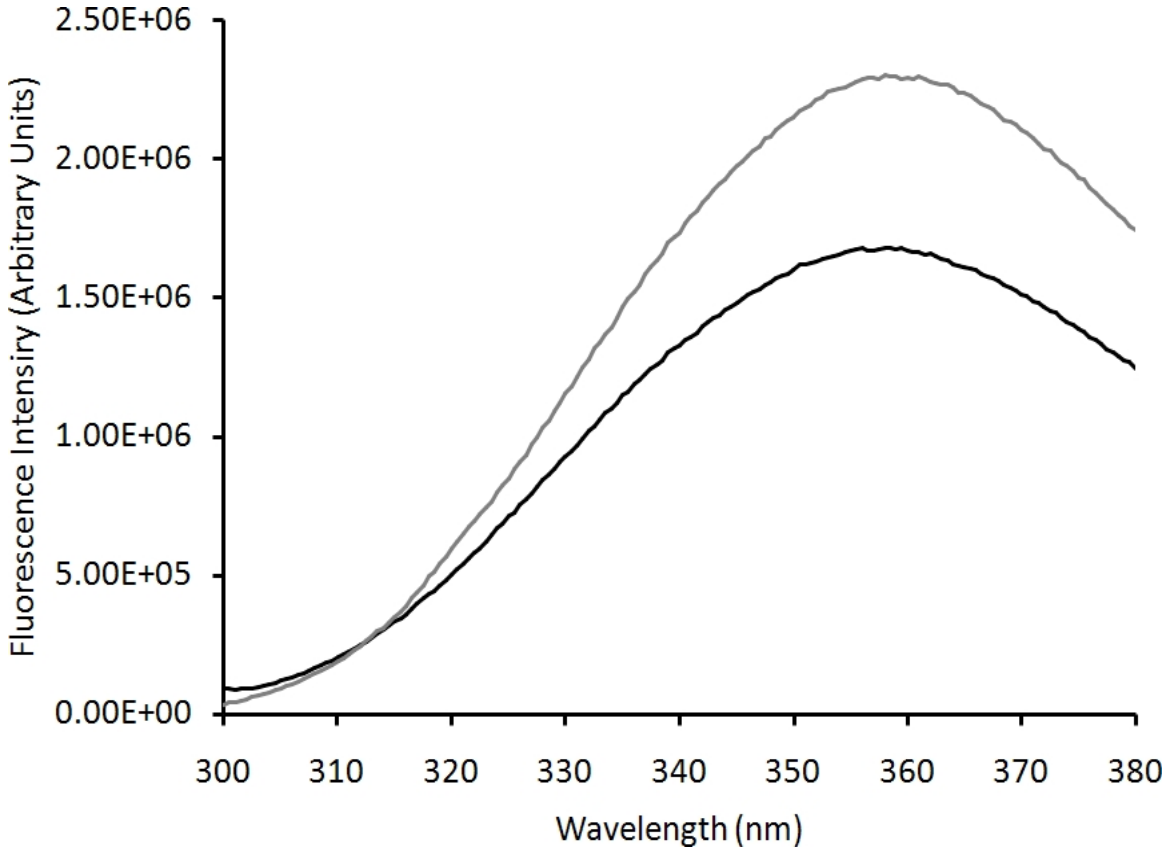

Figure 4.11: Fluorescence spectra of 1RAB. The peak for 1RAB both in denaturant (grey line) and without denaturant (black line) appear nearly the same, exhibiting at most a 3 nm blue shift upon putative folding of 1RAB. The excitation wavelength was 280 nm.

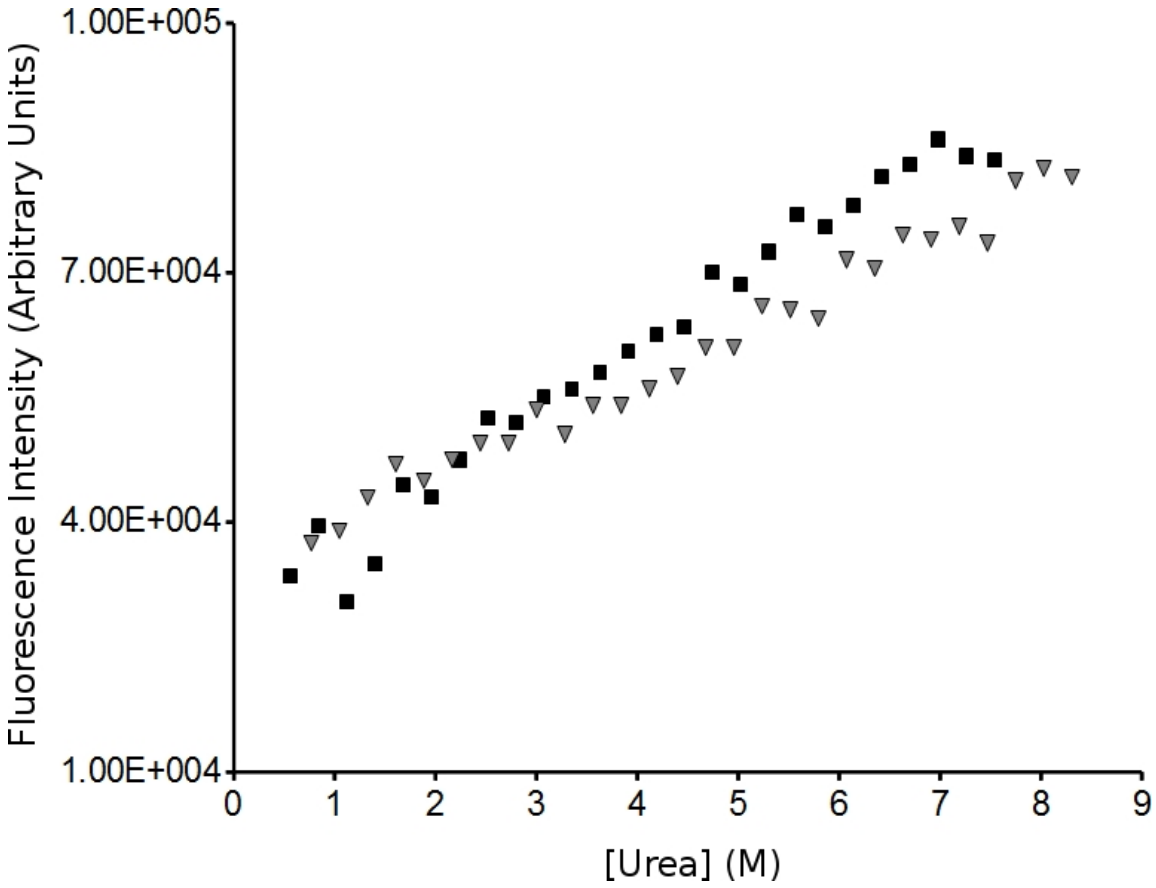

Figure 4.12: Fluorescence equilibrium renaturation and denaturation curves for 1RAB. Denaturation (black squares) and renaturation (grey triangles) for 1RAB after 4 days of equilibration are shown. There is no evidence from this data of a cooperative transition under different denaturant concentrations. The excitation wavelength was 280 nm.

There does appear to be a systematic deviation of the curves in Figure 4.12, with the renaturation values being consistently lower than those for denaturation. This likely results from the fact that the same molar extinction coefficient was used for both refolded and unfolded (in 8 M urea) stock solutions, whereas the difference in solvent conditions can give up to a 10-15% variation in the real molar extinction coefficient between having urea and not (51).

Interestingly, when high concentrations of salt (3 M NaCl) were added, 1RAB did begin to

show a blue-shift from 360 nm with high salt and denaturant (3 M NaCl, 8M Urea) to 338 nm with only high salt (3 M NaCl) (Figure 4.13). Although this blue shift is not as marked as was seen with 3RAB, it is still well within the normal range of blue-shifts for folding of compact globular folds (53).

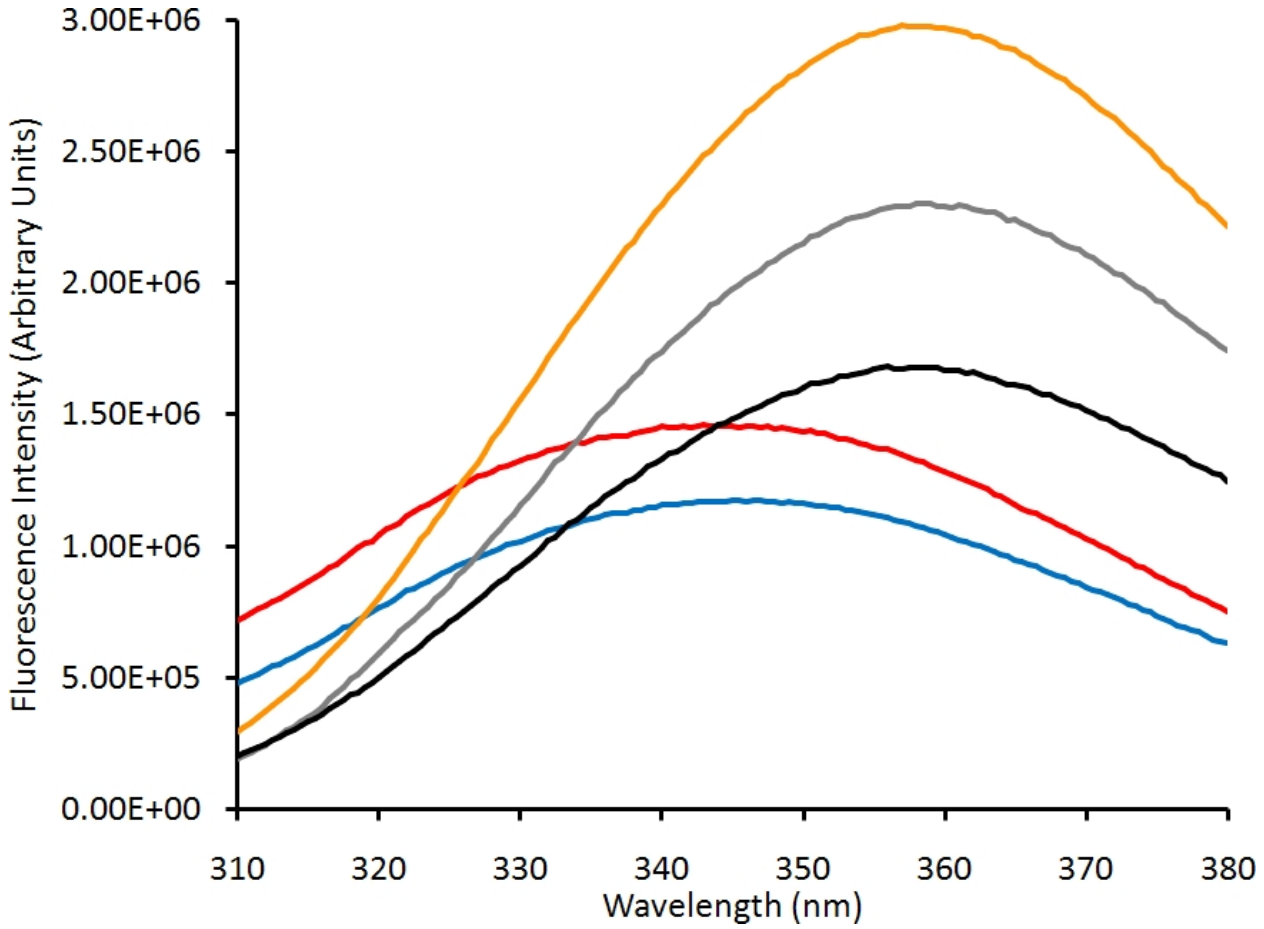

Figure 4.13: Fluorescence spectra of 1RAB with and without high salt concentrations. 1RAB with a moderate (300 mM NaCl) salt concentration and no denaturant (black line) does not differ greatly from that with denaturant (8 M urea) (grey line). By contrast, 1RAB with high (3 M NaCl) and no denaturant (red line), has a marked blue shift and decrease in fluorescence intensity when compared to the same conditions with denaturant (8 M urea) (orange line). Additionally, a high concentration of stabilizing agent, sodium sulphate (1 M) shows a similar but less pronounced shift without denaturant (blue line). The excitation wavelength was 280 nm.

The single tryptophan in 1RAB is predicted from homology modelling to be in the putative trimerization interface. The data collected thus far for 1RAB would indicate that with only a moderate concentration of salt (300 mM) 1RAB either does not trimerize, or does so without sufficiently burying the putative trimerization interface, but that burial of this interface does indeed occur with high salt (3 M NaCl). Within this context it is interesting to note that the template sequence for this design (see Chapter 2: Sequence Design) is from an organism that dwells in the dead sea, an environment where the concentration of sodium chloride is extremely high (~4M NaCl) (63). In fact, it has already been shown that several enzymes from this organism are only properly folded and active when the concentration of salt is in excess of 2M sodium/potassium chloride (63, 64). Although uncommon, it may be that some proteins have evolved to handle high salt to an extent that they in fact rely upon it, and 1RAB may have inherited this trait by virtue of its template.

4.3.8 Circular Dichroism of 1RAB

The CD spectrum of 1RAB (Figure 4.14) was analysed in the same manner as 3RAB. The prediction results (Table 4.2) indicate that it has considerable beta-characteristics, with little alpha-helix. Additionally, SELCON3 suggests the most closely related structure in the dataset is Elastase (PDB code, 1TRN), whose fold displays a 6-stranded barrel with many extended loops, and is part of the trypsin-like serine protease fold family. These results indicate that 1RAB likely has some beta-structure, but less than 3RAB. It may be the case that 1RAB is only partially folded under the conditions studied, or perhaps it adopts a different structure than 3RAB (such as a beta-propeller, or simply a beta-barrel). As Elastase's principal secondary structural motif is a 6-stranded barrel (like a beta-trefoil) but lacking the well-defined beta-hairpin triplet, it is possible that 1RAB may adopt a quasi-beta-trefoil structure.

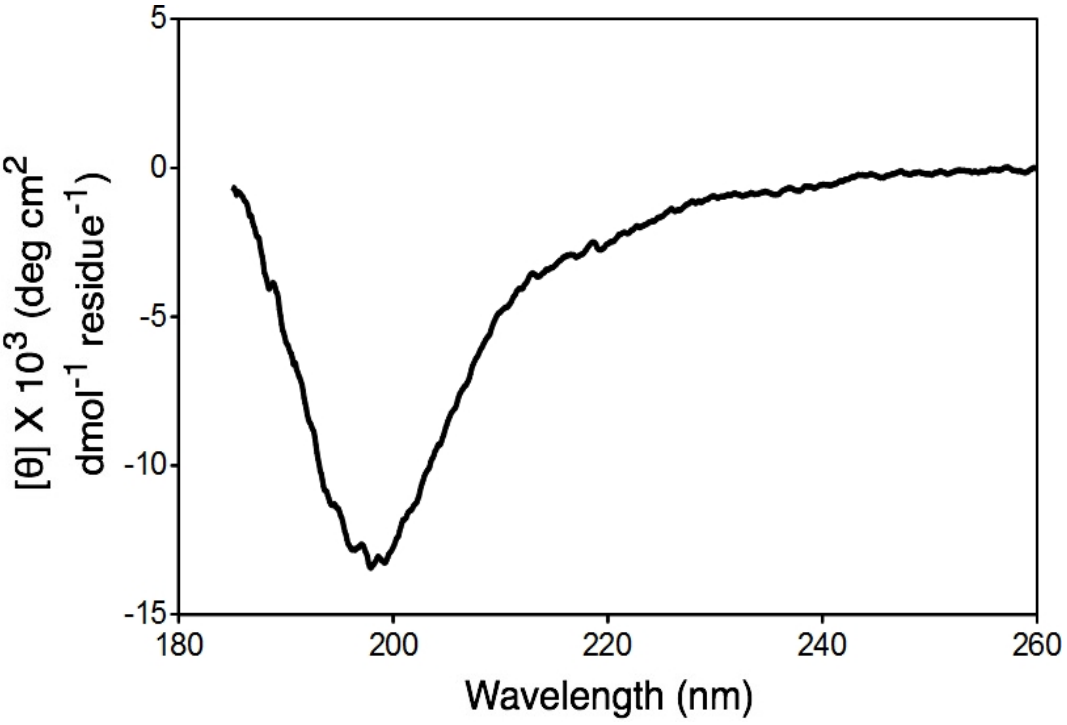

Figure 4.14: Circular dichroism spectrum of folded 1RAB. The spectrum shows some similarity to 3RAB and other beta-trefoils (minima near 200 nm), but appears most similar to Elastase, a trypsin-like serine protease which displays a 6-stranded beta-barrel with many long loops.

Table 4.2: Frequency of secondary structure in 1RAB from prediction algorithms.

Algorithm	Ordered Helix	Disordered Helix	Ordered Strand	Disordered Strand	Turn	Unordered	Strands per protein	Average Strand Length
CDSSTR	-0.01	0.05	0.23	0.13	0.14	0.45	4.58	5.5
CONTIN	0.01	0.07	0.2	0.13	0.15	0.44	4.73	5.1
SELCON3	0	0.09	0.15	0.12	0.17	0.41	4.44	4.3

4.3.9 Dynamic Light Scattering of 1RAB Compared with 3RAB

DLS measurements to predict the hydrodynamic diameter of 1RAB gave a result of 3.0 nm, which would indicate a trimer, as the theoretical diameter from homology modelling of 3RAB is ~3 nm, and the hydrodynamic diameter of 3RAB using DLS was also found to be 2.7 nm (Figure 4.15). It is interesting to note that 1RAB is appears slightly larger than 3RAB by DLS. Assuming this is not simply error due to a single measurement, there are several reasonable explanations. It could simply be that the added size of having three histidine tags present in the 1RAB trimer, as opposed to a single tag in 3RAB, results in this additional apparent diameter. It may also be that the 1RAB trimer is less compact.

Although the DLS data for 1RAB are useful for examining the hydrodynamic diameter, it is not possible to make strong conclusions about the molecular weight of the protein without performing a Debye plot, which would require a much higher protein concentration. Although it is possible that 1RAB may not have a completely compact fold (as suggested by CD measurements), it is unlikely that even a completely unfolded monomer of 1RAB would have the same diameter as 3RAB. The most parsimonious explanation for both 1RAB and 3RAB having the same diameter is that 1RAB forms a trimer in solution.

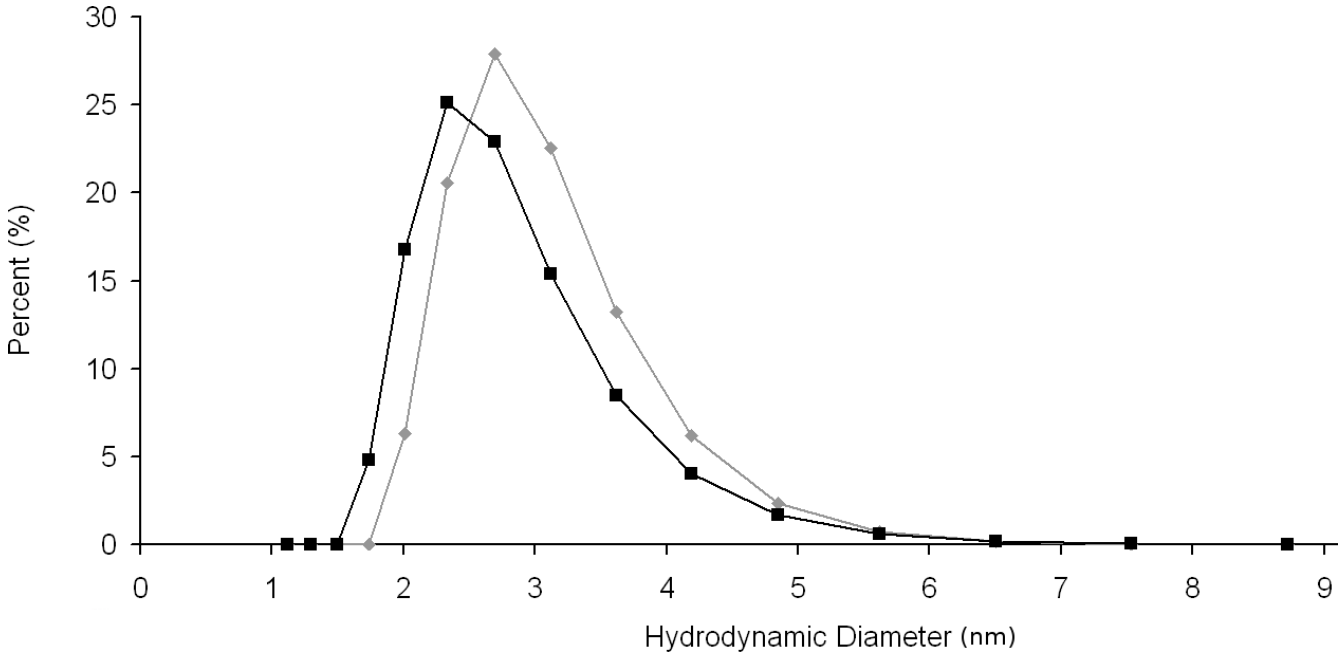

Figure 4.15: DLS measurements of 1RAB and 3RAB. Measurements of the hydrodynamic diameter of both 1RAB (grey) and 3RAB (black) show that they are both in close agreement with the theoretical diameter from homology modelling of ~3 nm. 1RAB appears slightly larger than 3RAB, likely owing to its additional theoretical molecular weight due to additional histidine tags.

4.3.10 Size Exclusion Chromatography of 1RAB

SEC results for 1RAB appear to be confounded by the same galactose binding problem as was seen for 3RAB (see Figure 4.10 and section 4.3.6). Although differing concentrations of galactose were not used to estimate true molecular weight (as with 3RAB), it is evident (Figure 4.16) that 1RAB undergoes a cooperative transition from a high concentration apparent size of 10 kDa to a low concentration apparent size of 3.5 kDa. The aforementioned figure (Figure 4.16) clearly indicates a cooperative transition from trimer to monomer. As SDS-PAGE analysis confirms the monomer is ~8 kDa as expected, and since no degradation of 1RAB was detected, it must be concluded that the monomeric apparent size of 3.5 kDa from Figure 4.16 results from 1RAB interacting with the column

in the same manner as 3RAB, slowing its progress and making it appear smaller than it truly is. Of course, the scarcity of data points also suggests that strong conclusions should be withheld until further evidence is obtained.

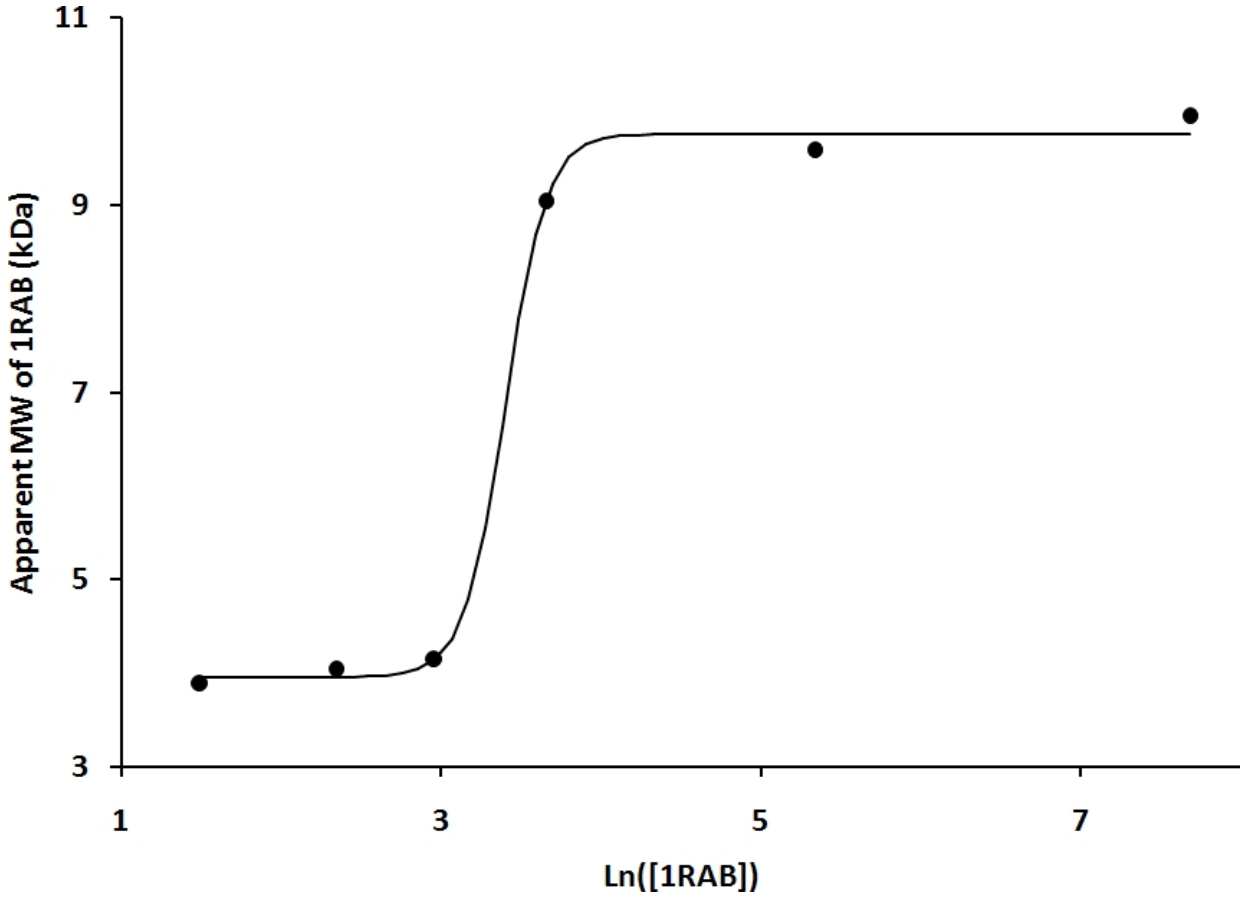

Figure 4.16: Transition of 1RAB monitored by SEC. Measuring the apparent MW of 1RAB at varying concentrations yields a plot illustrating a transition from a high MW (~10 kDa) species at high concentration to a low MW species (~3.5 kDa) at lower concentrations, indicating a trimer at higher concentrations.

4.4 Summary of Results

The results demonstrate quite clearly that 3RAB is a thermally stable (from DSC), soluble monomer (from DLS), with a compact fold (from proton NMR and fluorescence) which it reaches in a cooperative manner (from fluorescence). It also binds galactose (from SEC), and has therefore maintained the functional characteristics of the ricin family upon which it was based. Circular dichroism and NMR in conjunction with the close homology of 3RAB to beta-trefoils of known structure strongly suggest that 3RAB adopts a beta-trefoil fold.

In the case of 1RAB, both DLS and SEC data indicate that it forms a trimer in solution. Additionally, if it indeed does form a trimer, then the SEC data also demonstrates that it binds galactose (which is the primary component of the resin used for SEC in this case), indicating that it still has the function of a complete beta-trefoil, even as a small peptide or homotrimer. Although 1RAB does appear to form a sugar-binding trimer as intended, it is not evident that it forms a beta-trefoil structure. It lacks the CD spectral characteristics of the beta-trefoils that 3RAB shows (a minimum near 200 nm and a maximum near 230 nm), and additionally, it shows almost no blue shift in the fluorescence emission of its lone tryptophan, indicating that the tryptophan remains solvent exposed even upon trimerization. Although it does show a marked blue-shift in high salt which indicates hydrophobic burial, the lack of other structural analysis in high salt (such as CD) leave the structure of this high salt form unclear.

4.5 Discussion

Of the two designed proteins, 1RAB and 3RAB, 3RAB is most similar to a naturally occurring protein (its template sequence, see Chapter 2: Sequence Design), and it was therefore considered most likely to form a well behaved protein. The remarkable thermal stability of 3RAB (94 °C melting point) was not expected, as the template sequence is from an extreme halophile (*Haloarcula marismortui*) not

a thermophile. Moreover, a tryptophan fluorescence maximum of 313 nm is extremely low (53) indicating a very hydrophobic environment. Additionally, although proper equilibrium renaturation and denaturation curves were not obtained, the fact that 3RAB appears to remain folded even after weeks of incubation in 8M urea or 6M guanidine hydrochloride suggests that its chemical stability may rival its thermal stability. The positive second virial coefficient and monomeric nature of 3RAB even at 12.5 mg/mL shows that it is not only very soluble, but prefers to remain as a monomer, which was not only the desired goal, but also facilitates many experiments (such as NMR or crystallization). Moreover, although no particular data is shown to this effect, the author notes that 3RAB samples kept at room temperature for a period of months do not show any change in the fluorescence or CD spectrum, and do not show any visual signs of aggregation, suggesting that 3RAB is extremely well behaved (although the author also notes that one should not keep samples at room temperature for a period of months unless testing shelf-life is the objective, and the author also assures the reader that these samples were not used for any of the reported data).

The similarity of the CD spectra for 3RAB and known beta-trefoils along with the expected burial of its three tryptophan residues leads to the tentative conclusion that 3RAB adopts a beta-trefoil fold, although a direct structural solution (ideally through crystallization) would be needed in order to make a complete conclusion.

In the case of 1RAB the results are perhaps not as clear. Although CD analysis indicates that it has a predominantly beta-structure, it not only has less beta-structure than 3RAB, but the shape of the spectrum is lacking the characteristic peak near 230 nm that other beta-trefoils and 3RAB demonstrate. The source of the 230 nm positive peak is likely the aromatic sidechains (tryptophan and tyrosine). Which are known to contribute strongly in this region when their positions are constrained within a folded protein (65, 66). The lack of a positive 230 nm peak for 1RAB suggests that its tryptophan residues are not in a folded region of the protein, an observation which is corroborated by the

fluorescence spectra for 1RAB which indicate that tryptophan is not buried in a hydrophobic environment. It may therefore be concluded that the tryptophan residue in 1RAB is indeed solvent exposed and mobile. Both DLS and SEC data, however, indicate that 1RAB forms a trimer in solution, and so the difference in fluorescence and CD spectra seems likely to result from the 1RAB homotrimer having a different solution structure than 3RAB. That 1RAB begins to adopt a fluorescence spectra similar to 3RAB when in high salt is particularly interesting, and indicates that the stabilizing effect of the salt may cause the 1RAB homotrimer to adopt a structure more similar to 3RAB. Unfortunately, high concentrations of chloride ions interfere considerably with CD measurements in the ultra-violet region, and as a result no CD measurements on high salt 1RAB samples were obtained. It is worth considering as mentioned in Section 4.2, that the template sequence for 1RAB and 3RAB comes from an organism whose natural proteins are known to be either non-functional, or unfolded below 2 M potassium or sodium chloride (63, 64). It may simply be the case that both 1RAB and 3RAB prefer high concentrations of salt, but the added stability of covalently linking three 1RAB peptides to form 3RAB allows it to maintain its structure even in the absence of high salt, whereas 1RAB cannot properly adopt the beta-trefoil fold without high salt concentrations. Moreover, the closest matching structure by CD for 1RAB was Elastase, whose fold is characterized by a 6 stranded barrel like a beta-trefoil, but with a different pitch to the strands, and long loops rather than a hairpin triplet. This suggests that 1RAB may adopt a structure with some of the key characteristics of a beta-trefoil, but requires high salt in order to completely adopt the intended structure.

That 1RAB may, while 3RAB certainly does, bind galactose is a critical observation, as it indicates that a putative evolutionary path has perhaps been recapitulated. In particular 1RAB would be expected to not only have some stability, but also some functionality, in order to allow its continued evolutionary selection.

5 Conclusions and Future Work

5.1 Conclusions

5.1.1 The Ancient Peptide World

In order to examine the theory that modern proteins have evolved from a primordial set of small peptides, the beta-trefoil superfold has been used as a model in which a single peptide could originally have formed a functional trimer, and later, via gene multiplication, could have formed a completely symmetrical monomeric beta-trefoil. Examining the characteristics of this single peptide (1RAB) and the larger triplicated sequence (3RAB) not only addresses the general theory that modern proteins may have evolved from small stable peptides or secondary structural elements, but it also addresses the specific theory of the origin of the beta-trefoil fold. In a simple view of this ancient evolutionary path, the small peptide (1RAB) would be functional, and likely assemble to form a homotrimer with greater stability than the monomer and multivalent binding. Although this homotrimer would be stable and functional, triplicating the gene would serve several advantages: first, one can imagine that by covalently linking the three 1RAB peptides (3RAB, if the linkages are amide), stability could be improved as the structure would be restricted. Second, many modern beta-trefoils do not exist independently, but as a domain in a larger protein, usually as the carbohydrate binding domain for a carbohydrate-related enzyme (67), and generating a single gene for the beta-trefoil enables these kind of multidomain proteins to form, whereas the homotrimer would not be capable of such action. A schematic overview of the aforementioned process can be seen in Figure 5.1. Overall one can envision an ancient peptide world where stable elements modularly assemble into larger complexes, sometimes simply amplifying an existing function and perhaps occasionally creating new functions (catalytic sites at the interface of two modules). The lack of stability for these complexes, combined with the chance

for spurious associations, would eventually lead to single large proteins being favoured by evolution.

The results obtained thus far suggest that the aforementioned theory is reasonable, as the putative original small peptide, 1RAB, and the completely symmetrical beta-trefoil, 3RAB, have been successfully expressed in *Escherichia coli*, purified, and are stable enough to be partially characterized. That both proteins are capable of being expressed without degradation by host proteases already suggests that this may recapitulate a true evolutionary path.

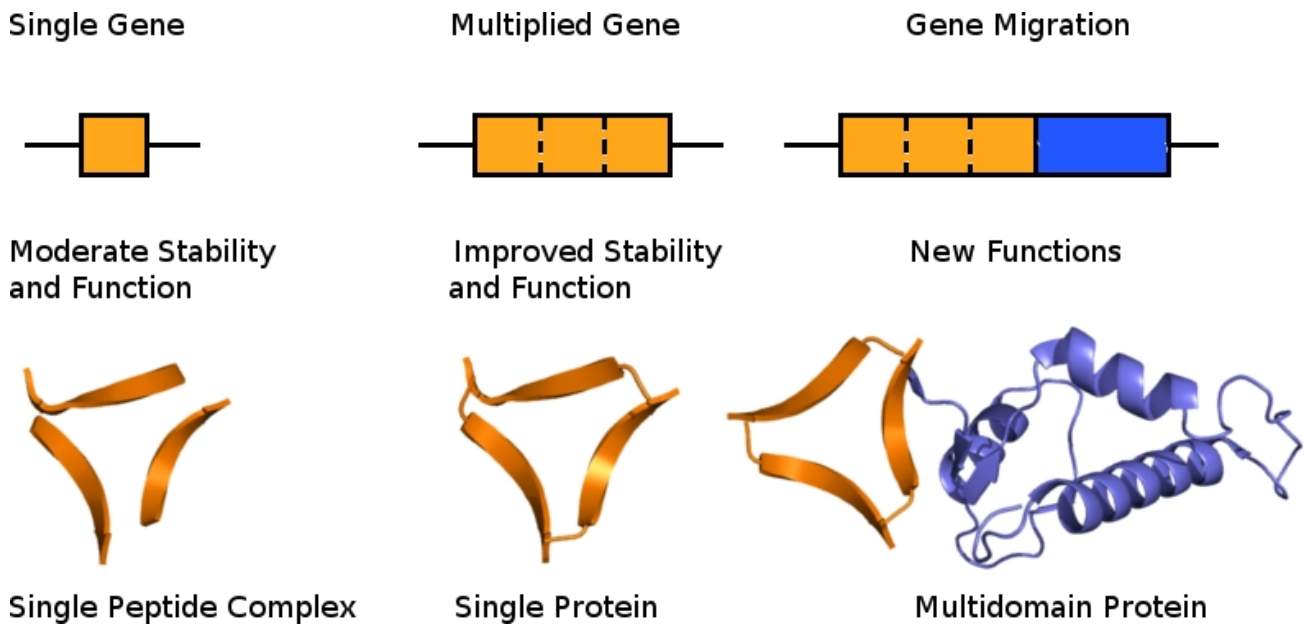

Figure 5.1: A schematic overview of modular protein evolution. Initially a small pseudo-stable peptide (such as 1RAB) is functional as a multimeric complex. After gene duplications, a larger monomeric protein (such as 3RAB) is more stable. Eventually the entire gene could migrate next to that of another protein, forming a multidomain protein with additional functions. The protein images were made using Pymol and a multidomain protein containing a beta-trefoil (PDB code, 2IHO, a galactose binding lectin).

5.1.2 Conclusions Concerning 1RAB and 3RAB

Biophysical characterization of 3RAB revealed that it not only forms a well behaved and very stable monomer, but also that it is capable of binding several carbohydrates (glucose and lactose data not shown), in particular galactose. As sugar binding is the most common function of the beta-trefoil superfold, and the only known function of the ricin family upon which the design was based, this suggests that 3RAB represents a putative ancestral structure which would have been stable and functional and therefore maintained by evolutionary pressures. Although a crystal structure has not yet been determined for 3RAB, its CD spectrum is highly similar to that of other beta-trefoils, and prediction algorithms using CD data classify it as a beta-trefoil. Additionally, both fluorescence and ^{1}H-NMR clearly indicate that it possesses a hydrophobic core with a buried tryptophan, consistent with the beta-trefoil structure.

While there are no conclusive data concerning the stability of 1RAB, it does form a trimer as expected, and that trimer is, like the subsequent beta-trefoil (3RAB), capable of binding galactose. That it forms a sugar binding trimer is very promising, as this would indicate it could have performed a useful function in an ancestral environment, and been evolutionarily maintained long enough to undergo the random gene multiplication necessary to make the more structured and stable 3RAB. Although some structural data (such as ^{1}H-NMR) could not be performed due to concentration limitations, fluorescence measurements indicate that the tryptophan remains exposed to solvent unless high concentrations of salt are added. This may indicate that the trimer formed is rather molten-globular, or that it forms a structure without a well packed hydrophobic core, unless high salt concentrations are added. Structural prediction algorithms using CD data do not classify 1RAB as a beta-trefoil when performed using low salt. Moreover these algorithms predict a reduced fraction of beta-structure and increased fraction of unstructured regions as compared to 3RAB, furthering this idea of a molten globular form or perhaps just a related structure.

It it is interesting to consider the possibility that the 1RAB homotrimer may form a different, but related structure. If this is the case, the most likely structure is one similar to Elastase, which has the most similar CD spectrum to 1RAB. Elastase is a member of the superfamily or superfold, trypsin-like serine proteases and is annotated on SCOP (29) as being defined by a six stranded beta-barrel, which is the same as the beta-trefoil fold, although no further characteristics are noted, whereas the beta-trefoil also has a beta-hairpin triplet. Although it may indeed be the case that 1RAB simply adopts a slightly different fold, perhaps with a well defined beta-barrel but less well defined structures above the barrel, there also the possibility that it does in fact have a beta-trefoil structure, but is predominantly in a molten globular state. It is known that breaking critical covalent bonds in a protein's structure can still leave the protein with a native-like fold, but in a molten globular form (68). In fact, it has been shown for a homotetramer that the isolated monomer is a molten globule, but still retains not only significant secondary structure, but more importantly, function (69). Hence, it is not difficult to imagine that 1RAB may adopt a molten globule-like form, with much of its secondary structure intact (as demonstrated by the CD analysis) and also still retaining its carbohydrate binding functions (as demonstrated by SEC using a galactose-based resin). Although this form of 1RAB does appear to be a trimer based on SEC and DLS measurements, it may be that the interface is not sufficiently stable to allow for a reduction in the molten globule structure. The result would be that the tryptophans would not be packed into a tight hydrophobic core and would be moderately solvent exposed. This idea is further supported by the fact that 1RAB appears to become more 3RAB-like when higher concentrations of salt are used (although only fluorescence data currently exist to suggest this). Many common salts (such as the NaCl used for 1RAB) are known to increase protein stability while also reducing solubility (Hofmeister series) (70). Additionally it has been found that increasing concentrations of KCl or NaCl lead to increases in the binding affinity of a halophilic protein complex (71). Therefore it may be that increasing concentrations of salt force a tighter fold and/or tighter

protein-protein interactions, thereby reducing the fraction of molten globular structure for 1RAB.

Overall the data indicate that while 1RAB maintains the function and perhaps some of the characteristics of 3RAB, it may be less stable and/or less tightly structured, demonstrating an evolutionary benefit to the gene duplications that would eventually lead to generation of 3RAB. As mentioned previously, the effective addition of covalent linkages between the 1RAB homotrimer would likely give greater stability and the possibility for multivalent binding, not to mention the opportunity for the larger sequence to migrate to other regions of the chromosome and form multidomain proteins. A schematic outline of this process is shown in Figure 5.1.

5.1.3 From Peptide to Protein: Rational Design

In addition to simply demonstrating the plausibility of an evolutionary model of beta-trefoil formation via repetition of a smaller ancestral peptide, this work demonstrates the possibility that small pseudo-structured peptide elements could be used design new sequences capable of folding into well behaved and stable proteins. Although the aforementioned design process has only been demonstrated for a fold which is structurally symmetrical and therefore very amenable to being constructed by adding together small symmetrical units, it is possible that identifying other peptide precursors from other symmetrical and non-symmetrical folds could allow for the construction of a complete peptide library. From this peptide library new structures could be designed by joining together elements which have some residual structure on their own, and possess a hydrophobic patch capable of interaction with the hydrophobic patches of the neighbouring elements, an approach which has been shown previously with much larger elements (72). It is interesting to postulate that this modular form of protein design may in fact be the method that nature has employed to give the wide variety of protein forms and function seen today. An overview of the value of continuing to assemble larger and larger structures from small modules is shown in Figure 5.1. The presumption is that once a sufficiently large library of

small peptides was created, the need to further development of these ancestral species was reduced, and future proteins were simply formed in a combinatorial manner.

5.2 Immediate Future Work

5.2.1 Removal of the Histidine Tag

Although it is usually the case that the presence of a histidine tag does not affect the structure of a protein (44), this is not always true (73). In particular, it may be that the considerable size of the histidine tag in comparison to 1RAB (~33% of the size) will mean that the tag contributes significantly to the structure of 1RAB, or at least to its stability. For this reason it is particularly important that the histidine tag be cleaved off 1RAB using the thrombin cleavage site present. Although it should also at some point be removed from 3RAB, this may not be as pressing, as 3RAB not only demonstrates that it has the correct fold, but its considerable stability suggests that it is not encountering folding problems. In general it has been noted that histidine content near the N- or C-terminus is a strong indicator of disorder (74), suggesting that the his-tag is usually disordered, which may mean that it contributes little to structure, but may destabilize a smaller protein like 1RAB.

5.2.2 Structures of 1RAB and 3RAB

Obtaining crystal and/or NMR structures for both 1RAB and 3RAB would be exceedingly valuable, as it would conclusively show what the actual structures are. Although there could be some concern that the crystal structure of 1RAB may not accurately reflect its solution structure, particularly if it is indeed a molten globule in solution, as crystallization may, by virtue of reducing the degrees of freedom of the system, force it to form a compact fold. One primary concern for obtaining crystal structures is the histidine tag. In the case of 3RAB it may be stable and soluble enough that crystallization is possible even without removing the tag, which would be ideal, as 3RAB may be

somewhat large for structural determination via NMR, not to mention that its perfect symmetry may make residue assignments extremely challenging. Since the histidine tag makes up approximately one third of the total size of 1RAB, there is not only concern that it may affect the structure and biophysical characteristics, but certainly that it may impede crystallization. Even with the tag removed through thrombin cleavage, the molten globular nature of 1RAB may mean that its structure is not rigid enough to facilitate crystallization, and NMR would be required. In either case, 1RAB has not yet been concentrated sufficiently for either NMR or crystallization, and determining more favourable solution conditions for 1RAB is a necessary task.

5.2.3 Stability of 1RAB and 3RAB

In additional to knowing the structure of both 1RAB and 3RAB, performing quantitative measurements of stability for both would be useful in understanding the evolutionary pressures towards gene fusions or duplications. As mentioned before and outlined in Figure 5.1, amongst the possible benefits of gene fusions is an increase in stability. That 3RAB appears to form a fold with a more compact hydrophobic core (from fluorescence data) already suggests that it may be more stable, but accurate denaturation and renaturation curves, plus accurate analysis of DSC data for both would be very valuable, and perhaps illuminating.

5.2.4 Differential Scanning Calorimetry in Denaturant or Acid

Although DSC was performed for 3RAB, thermodynamic properties of interest such as ΔG could not be obtained because thermal unfolding, while cooperative, was not reversible as cooling and reheating resulted in aggregate formation (data not shown). It may be possible to rescue 3RAB from forming aggregates during DSC by helping to solubilize it using perhaps denaturant, or acidic conditions as has been previously performed (75). Although this will lower the ΔG, giving only a ΔG

in that particular concentration of acid, it is possible to use varying pHs and extrapolate to the ΔG under standard solution conditions.

5.2.5 Other Folds

Although the research presented herein is specific to the beta-trefoil fold, the general mechanism of modular fold formation is thought by the author to apply to any modern fold, and in particular the mechanism of duplications to create symmetrical globular folds most likely applies to other globular symmetric folds. As a result an obvious avenue for future research would be applying this kind of design and characterization approach to folds beyond the beta-trefoil, perhaps starting with a well studied, highly abundant, and symmetrical fold such as the TIM-barrel (72).

5.2.6 Symmetry in Reverse

Previous experiments using the beta-trefoil have demonstrated that increasing symmetry increases stability (40, 41, 42). Taking a non-symmetrical sequence and increasing the symmetry does seem a good approach to examining the relationship between symmetry and stability, but it is not on its own conclusive. At a given position that is not completely symmetrical there are at least two different residues to choose when making those positions identical, and it could perhaps be the case that the residue chosen was simply better in that context, irrespective of symmetry.

The completely symmetrical scaffold that 3RAB represents presents an alternative option for examining symmetry, namely, one can go backwards from a symmetrical starting point with the hypothesis that any mutation at a given position should decrease stability. It is the author's opinion that this will in fact not be true, and that one might find a single residue change that increases stability. Of course this could just mean that the residue is better, and so the remaining symmetrical positions would also have to be changed in order to ensure that this did not restore stability.

5.3 Long Term Future Directions and Implications

5.3.1 Computational Structure Prediction and Design

Although there are some obvious experiments and extensions to this work, the most interesting applications may involve computational or bioinformatics approaches. If the kind of result found here is general, that modern stable proteins have arisen from a combination of smaller pseudo-stable elements, then it may be possible to either predict the structure of existing proteins, or design new proteins through a combinatorial approach using small peptide modules.

In general one of the many techniques for protein prediction is to use a library of small structured fragments of known sequence to assemble a putative initial structure, followed by rounds of choosing differently structured fragments in an attempt to minimize the energy of the putative structure. This approach has seen much success in the popular program ROSETTA (24). In the case of ROSETTA fragments of 3 and 9 amino acids in length are used, but it has been suggested that using longer fragments can reduce the conformational space that needs to be searched, and as this is the primary limiting factor in computational prediction (9, 25), these longer fragments may be a much needed improvement. One existing algorithm uses fragments whose length corresponds to a single secondary structure (~9-15 amino acids) (25). If modular peptides the size of 1RAB (20-60 amino acids) are in fact the building blocks of modern proteins (21), then using these much larger fragments could drastically improve the extent to which a prediction algorithm could explore the conformational space of a protein in a reasonable period of time.

A similar argument can be made for protein design, although the visual concept of making a new protein by simply combining numerous small modules is perhaps simple and obvious enough that this idea does not warrant further arguments to demonstrate its value. Therefore, a large scale analysis of the entire protein structural database may be warranted, whereby we attempt to develop a library of peptide modules, sufficient for describing the structures of all known proteins.

An overview of the concept of protein prediction and design through small modular elements is shown in Figure 5.2.

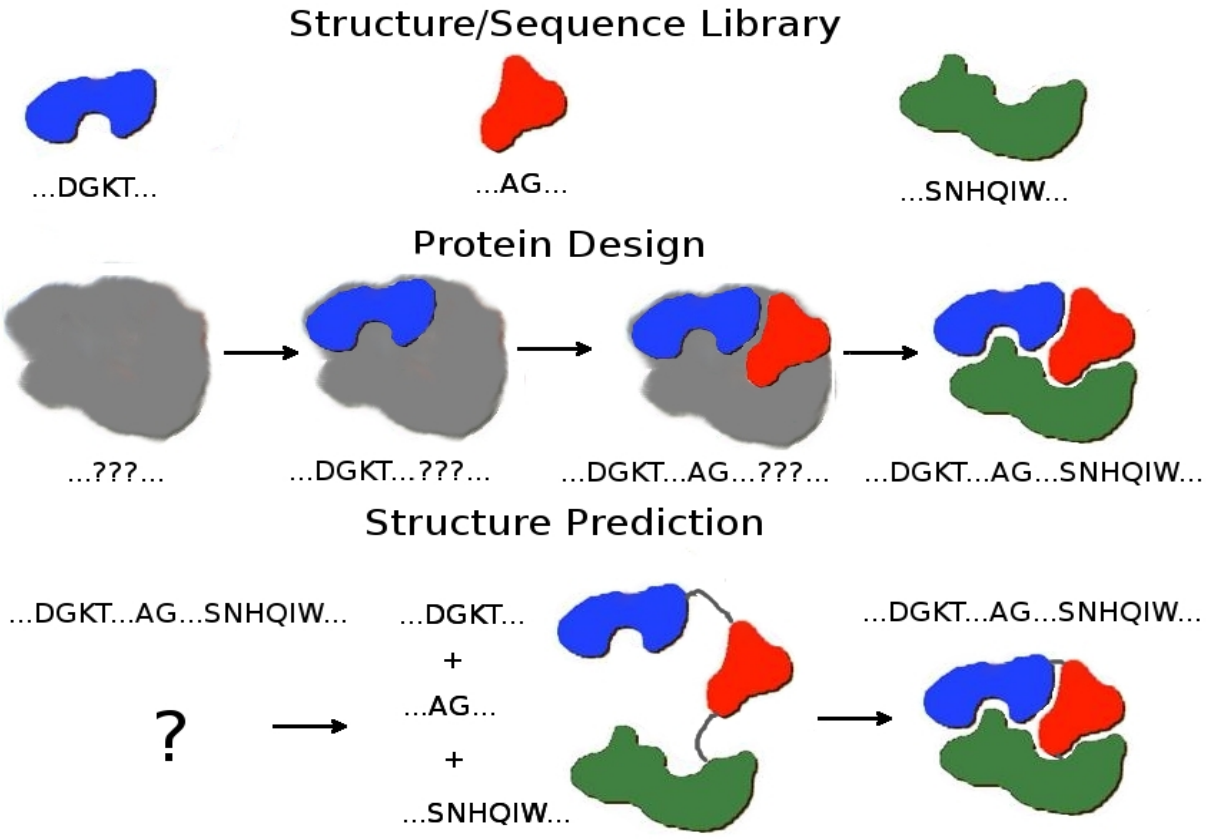

Figure 5.2: Modular Prediction and Design. First a library of super-secondary structural elements and their respective sequence profiles is constructed. Protein design is accomplished by combining elements that can fit the target structure. The sequence profiles then give the sequence of the novel protein. Structure prediction involves arranging sequence profiles from the library to match the target sequence. The structural elements associated with those profiles are then substituted in and orientations optimized to predict the structure.

5.3.2 Pharmaceutical Beta-trefoils: Exploiting Glycoproteins

In addition to protein prediction/design, there may exist some particularly valuable uses for beta-trefoils in general, and a knowledge of their structural determinants may be very useful in designing novel beta-trefoils or rationally modifying existing ones.

Beta-trefoils often appear to possess a strong binding function, which appears to favour carbohydrate binding (15). Considerable focus medicinally has been on the use custom antibodies for infectious disease prevention. Interestingly, many infectious diseases, such as HIV, have key viral proteins heavily glycosylated with the host's native carbohydrates, which can make recognition by host antibodies difficult (76). Although an attractive avenue being studied is rational design of novel antibodies, it might be more prudent to use the beta-trefoil scaffold as a starting point for designing glycoprotein binding therapeutics (77). An example is the anti-HIV protein actinohivin from an actinomycete, *Longispora albida*, which binds to the heavily glycosylated gp120 viral protein (78). While this new anti-HIV protein is very promising, it may be inherently flawed, in that it is from a microorganism, and therefore, will eventually be recognized by the immune system as foreign. As such, the very thing this protein is trying to help, will eventually become very adept at preventing its action. A solution to this problem would be to use a beta-trefoil of human origin, or modify such a protein, in order to fulfill the same role.

Perhaps it is the case that beta-trefoils will become a new class of designer drug, maybe even useful in combination with antibodies at binding to the glycoproteins of infectious particles.

5.3.3 The Scaffold

A final valuable contribution of this work is that there now exists a completely symmetrical beta-trefoil. This scaffold may perhaps be very valuable for a number of applications. Primarily one can easily envision that further studies on symmetry, and its effects upon protein folding will benefit

greatly by having a completely symmetrical starting point, from which asymmetrical mutations can be made. This is critically important because, up until now, examinations of symmetry have necessarily involved increasing the symmetry of incompletely symmetrical proteins. The second value in having this completely symmetrical beta-trefoil may be that it is quite useful for determining the critical elements of the beta-trefoil fold itself. Making a three-fold symmetrical mutation to a symmetrical fold may be a better solution to probing the fold's structural determinants than making mutations within a non-symmetrical structure. Within the context of beta-trefoils perhaps acting as drugs (see Section 5.3.2), a better understanding of the fold's determinants may be very valuable for rationally designing new beta-trefoils or altering existing ones.

References

1: Denton, RM., Randle, PJ., Bridges, BJ., Cooper, RH., Kerbey, AL., Pask, HT., Severson, DL., Stansbie, D., Whitehouse, S. (1975). Regulation of mammalian pyruvate dehydrogenase. Molecular and Cellular Biochemistry **9**:1,27-53

2: Stephens, JM., Pilch, PF. (1995). The metabolic-regulation and vesicular transport of GLUT4, the major insulin-responsive glucose-transporter. Endocrine Rev. **16**:4,529-546

3: Lee, CH., Singla, A., Lee, Y. (2001). Biomedical applications of collagen. International J. Pharmaceutics **221**:1-22

4: Skerra, A. (2000). Engineered protein scaffolds for molecular recognition. J. Molecular Recognition **13**:167-187

5: von der Osten, C., Branner, S., Hastrup, S., Hedegaard, L., Rasmussen MD., Bisgård-Frantzen, H., Carlsen, S., Mikkelsen, JM. (1993). Protein engineering of subtilisins to improve stability in detergent formulations. J. Biotechnology **28**:55-68

6: Szczodrak, J. (2000). Hydrolysis of lactose in whey permeate by immobilized β-galactosidase from *Kluyveromyces fragilis*. J. Molecular Catalysis B: Enzymatic **10**:631-637

7: Kennedy, FP. (1991). Recent developments in insulin delivery techniques – current status and future potential. Drugs **42**:2,213-227

8: Burley, SK., Almo, SC., Bonanno, JB., Capel, M., Chance, MR., Gaasterland, T., Lin, D., Sali, A., Studier, FW., Swaminathan, S. (1999). Structural genomics: beyond the Human Genome Project. Nature Genetics **23**:151-157

9: Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., Baker, D. (2005). Progress in modeling of protein structures and interactions. Science **310**:638-642

10: Dill, KA., Ozkan, SB., Shell, MS., Weiki, TR. (2008). The protein folding problem. Annual Rev. Biophysics **37**:289-316

11: Dantas, G., Kuhlman, B., Callender, D., Wong, M., Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J. Molecular Biology **332**:449-460

12: Talini, G., Gallori, E., Maurel, M. (2009). Natural and unnatural ribozymes: Back to the primordial RNA world. Research in Microbiology **160**: 457-465

13: Lupas, AN., Ponting, CP., Russel, RB. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J. Structural Biology **134**:191-203

14: Wu, LC., Grandor, R., Carey, J. (1994). Autonomous subdomains in protein folding. Protein Science **3**:369-371

15: Murzin, AG., Lesk, AM., Chothia, C. (1992). β-trefoil fold patterns of structure and sequence in the kunitz inhibitors interleukins-1-beta and 1-alpha and fibroblast growth-factors. J. Molecular Biology **223**:531-543

16: Bernett, MJ., Somasundaram, T., Blaber, M. (2004). An atomic resolution structure for human fibroblast growth factor 1. Proteins **57**: 626-634

17: Notenboom, V., Boraston, AB., Williams, SJ., Kilburn, DG., Rose, DR. (2002). High-resolution crystal structures of the lectin-like xylan binding domain from *Streptomyces lividans* xylanase 10A with bound substrates reveal a novel mode of xylan binding.
Biochemistry **41**:4246-4254

18: Arndt, JW., Gu, J., Jaroszewski, L., Schwarzenbacher, R., Hanson, MA., Lebeda, FJ., Stevens, RC. (2005). The Structure of the Neurotoxin-associated Protein HA33/A from *Clostridium botulinum* Suggests a Reoccurring beta-Trefoil Fold in the Progenitor Toxin Complex.
J.Molecular Biology **346**:1083-1093

19: Rutenber, E., Ready, M., Robertus, J. (1987). Structure and evolution of the ricin B chain. Nature **326**:624-626

20: Loo, TW., Bartlett, C., Clarke, DM. (2005). Rescue of ΔF508 and other misprocessed CFTR mutants by a novel quinazoline compound. Molecular Pharmaceutics **2**:5,407-413

21: Tateno, Y., Ikeo, K., Imanishi, T., Watanabe, H., Endo, T., Yamaguchi, Y., Suzuki, Y., Takahashi, K., Tsunoyama, K., Kawai, M., Kawanishi, Y., Naitou, K., Gojobori, T. (1997). Evolutionary motif and its biological and structural significance. J. Molecular Evolution **44**:S38-S43

22: Lijnzaad, P., Argos, P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins **28**:3,333-343

23: Ponting, CP., Russel, RB. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins. J. Molecular Biology **302**:1041-1047

24: Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, DE., Meiler, J., Misura, KMS., Baker, D. (2005). Free modeling with Rosetta in CASP6. Proteins: Structure, Function, and Bioinformatics Supplement **7**:128-134

25: Kaizhi, Y., Dill, KA. (2000). Constraint-based assembly of tertiary protein structures from secondary structure elements. Protein Science **9**:1935-1946.

26: Han, JH., Batey, S., Nickson, AA., Teichmann, SA., Clarke, J. (2007). The folding and evolution of multidomain proteins. Nature Reviews Molecular Cell Biology **8**:319-330.

27: Doxey, AC., Broom, A., Meiering, E., McConkey, B. (2010). Fold evolution repeats itself: Repeat-mediated generation of the beta-trefoil fold. Submitted to PNAS

28: Marchler-Bauer, A., Anderson, JB., Cherukuri, PF., DeWweese-Scott, C., Geer, LY., Gwadz, M., He, SQ., Hurwitz, DI., Jackson, JD., Ke, ZX., Lanczycki, CJ., Liebert, CA., Liu, CL., Lu, F., Marchler, GH., Mullokandov, M., Shoemaker, BA., Simonyan, V., Song, JS., Thiessen, PA., Yamashita, RA., Yin, JJ., Zhang, DC., Bryant, SH. (2005). CDD: a conserved domain database for protein classification. Nucleic Acids Research **33**:D192-D196

29: Hubbard, TJP., Murzin, AG., Brenner, SE., Chothia, C. (1997). SCOP: a Structural Classification of Proteins database. Nucleic Acids Research **25**:236-239.

30: Nikkah, M., Jawad-Alami, Z., Demydchuk, M., Ribbons, D., Paoli, M. (2006). Engineering of β-propeller protein scaffolds by multiple gene duplication and fusion of idealized WD repeat. Biomolecular Engineering **23**:185-194

31: Mukhopadhyay, D. (2000). The molecular evolutionary history of a winged bean α-chymotrypsin inhibitor and modelling of its mutations through structural analyses. J. Molecular Evolution **50**:214-223

32: Chivian, D., Baker, D. (2006). Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Research **34**:e112

33: Altschul, SF., Madden, TL., Schäffer, AA., Zhang, J., Zhang, F., Miller, W., Lipman, DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**:3389-3402

34: Heger, A., Holm, L. (2000). Rapid Automatic Detection and Alignment of Repeats in protein sequences. Proteins: Structure, Function, and Genetics **41**:224-237

35: Thompson, JD., Higgins, DG., Gibson, TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**:4673-80

36: Meiering, EM., Serrano, L., Fersht, AR. (1992). Effect of active site residues in barnase on activity and stability. J. Molecular Biology **225**:585-589

37: Fiser, A., Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. Methods in Enzymology **374**:461-491

38: Grahn, E., Askarieh, G., Holmner, A., Tateno, H., Winter, H.C., Goldstein, I.J., Krengel, U. (2007). Crystal structure of the *Marasmius oreades* mushroom lectin in complex with a xenotransplantation epitope. J. Molecular Biology **369**:710-721

39: Meier, S., Güthe, S., Kiefhaber, T., Grzesiek, S. (2004). Foldon, the natural trimerization domain of T4 Fibritin, dissociates into a monomeric A-state form containing a stable β-hairpin: Atomic details of trimer dissociation and local β-hairpin stability from residual dipolar couplings. J. Molecular Biology **344**:1051-1069

40: Brych, SR., Kim, J., Logan, TM., Blaber, M. (2003). Accommodation of a highly symmetric core within a symmetric protein superfold. Protein Science **12**:2704-2718

41: Brych, SR., Dubey, VK., Bienkiewicz, E., Lee, J., Logan, TM., Blaber, M. (2004). Symmetric Primary and Tertiary Structure Mutations within a Symmetric Superfold: A Solution, not a Constraint, to Achieve a Foldable Polypeptide. J. Molecular Biology **344**:769-780

42: Dubey, VK., Lee, J., Blaber, M. (2005). Redesigning symmetry-related "mini-core" regions of FGF-1 to increase primary structure symmetry: thermodynamic and functional consequences of structural symmetry. Protein Science **14**:2315-2323.

43: Wolynes, PG. (1996). Symmetry and the energy landscapes of biomolecules. PNAS **93**:14249-14255

44: Carson, M., Johnson, DH., McDonald, H., Brouillette, C., DeLucas, LJ. (2007). His-tag impact on structure. Acta Cryst. **D63**:295-301

45: Ventura, S., Villaverde, A. (2006). Protein quality in bacterial inclusion bodies. TRENDS in Biotechnology **24**:4,179-185

46: Misawa, S., Kumagai, I. (1999). Refolding of therapeutic proteins produced in *Escherichia coli* as inclusion bodies. Biopolymers **51**:297-307

47: Tsumoto, K., Ejima, D., Kumagai, I., Arakawa, T. (2003). Practical considerations in refolding proteins from inclusion bodies. Protein Expression and Purification **28**:1-8.

48: Rudolf, R., Siebendritt, R., Kiefhaber, T. (1992). Reversible unfolding and refolding behavior of a monomeric aldolase from *Staphylococcus aureus*. Protein Science **1**:654-666.

49: Hoover, DM., Lubkowski, J. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Research **30**:e43

50: Spratt, DE. (2008). Calmodulin binding and activation of mammalian nitric oxide synthases. University of Waterloo

51: Pace, CN., Vajdos, F., Fee, L., Grimsley, G., Gray, T. (1995) How to measure and predict the molar absorption coefficient of a protein. Protein Science **4**:2411-2423

52: Zimm, BH. (1948). The Scattering of Light and the Radial Distribution Function of High Polymer Solutions. J. Chem Phys. **16**:1093-1099

53: Vivian, JT., Callis, PR. (2001). Mechanisms of tryptophan fluorescence shifts in proteins. Biophysical Journal **80**:2093-2109

54: Whitmore, L., Wallace, BA. (2008). Protein Secondary Structure Analyses from Circular Dichroism Spectroscopy: Methods and Reference Databases. Biopolymers **89**:392-400.

55: Johnson, WC. (1999). Analyzing protein circular dichroism spectra for accurate secondary structures. Proteins: Structure, Function and Genetics **35**:307-312

56: Provencher, SW., Glöckner, J. (1981). Estimation of protein secondary structure from circular dichroism. Biochemistry **20**:33-37

57: Sreerama, N., Woody, RW. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. Anal. Biochem. **209**:32-44.

58: Lees, JG., Miles, AJ., Wien, F., Wallace, BA. (2006). A reference database for circular dichroism spectroscopy covering fold and secondary structure space. Bioinformatics **22**:16,1955-62

59: Song, HK., Suh, SW. (1998). Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator. J.Molecular Biology **275**:347-363

60: Araujo, APU., Hansen, D., Vieira, DF., de Oliveira, C., Santana, LA., Beltramini, LM., Sampaio, CAM., Sampaio, MU., Oliva, MLV. (2005). Kunitz-type *Bauhinia bauhinioides* inhibitors devoid of disulfide bridges: isolation of the cDNAs, heterologous expression and structural studies. Biological Chemistry **386**:561-568

61: Tateno, H., Goldstein, IJ. (2004). Partial identification of carbohydrate-binding sites of a Gal alpha 1,3Gal beta 1,4GlcNAc-specific lectin from the mushroom Marasmius oreades by site-directed mutagenesis. Archives of Biochemistry and Physics **427**:101-109

62: Hass, C., Drenth, J., Wilson, WW. (1999). Relation between the solubility of proteins in aqueous solutions and the second virial coefficient of the solution. J. Physical Chemistry B **103**:2808-2811

63: Bonnete, F., Madern, D., Zaccai, G. (1994). Stability against denaturation mechanisms in halophilic malate-dehydrogenase adapt to solvent conditions. J. Molecular Biology **244**:436-447

64: Muller-Santos, M., de Souza, EM., Pedrosa, FD., Mitchell, DA., Longhi, S., Carriere, F., Canaan, S., Krieger, N. (2009). First evidence for the salt-dependent folding and activity of an esterase from the halophilic archaea *Haloarcula marismortui*. Biochemica et Biophysica acta-molecular and cell biology of lipids **1791**:719-729

65: Haas, W., MacColl, R., Banas, JA. (1998). Circular dichroism analysis of the glucan binding domain of *Streptococcus mutans* glucan binding protein-A. Biochemica et Biophysica Acta – Protein Structure and Molecular Enzymology **1384**:112-120.

66: Woody, RW. (1978). Aromatic side-chain contributions to the far ultraviolet circular dichroism of peptides and proteins. Biopolymers **17**:1451-1467.

67: Boraston, AB., Bolam, DN., Gilbert, HJ., Davies, GJ. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochemical Journal **382**:769-781.

68: Cai, S., Singh, BR. (2001). Role of the disulfide cleavage induced molten globule state of type A Bolulinum neurotoxin in its endopeptidase activity. Biochemistry **40**:15327-15333.

69: Mitra, N., Sinha, S., Kini, M., Surolia, A. (2005). Analysis of the peanut agglutinin molten globule-like intermediate by limited proteolysis. Biochemica et Biophysica Acta – General Subjects **1725**: 283-289.

70: Cacace, MG., Landau, M., Ramsden, JJ. (1997). The Hofmeister series: salt and solvent effects on interfacial phenomena. Quarterly Reviews of Biophysics **30**:241-277.

71: Bonnete, F., Ebel, C., Zaccai, G., Eisenberg, H. (1993). Biophysical study of halophilic malate dehydrogenase in soltuion: revised subunit structure and solvent interactions of native and recombinant enzyme. J. Chemical Society Faraday Trans. **89**:2659-2666.

72: Höcker, B., Claren, J., Sterner, R. (2004). Mimicking enzyme evolution by generating new $(\beta\alpha)_8$-barrels from $(\beta\alpha)_4$-half-barrels. PNAS **101**:47,16448-16453.

73: Klose, J., Wendt, N., Kubald, S., Krause, E., Fechner, K., Beyermann, M., Bienert, M., Rudolph, R., Rothemund, S. (2004). Hexa-histidine tag position influences disulfide structure but not binding behaviour of *in vitro* folded N-terminal domain of rat corticotropin-releasing factor receptor type 2a. Protein Science **13**:2470-2475

74: Li, X., Romero, P., Rani, M., Dunker, AK., Obradovic, Z. (1999). Predicting protein disorder for N-, C- and internal regions. Genome Informatics. Workshop on Genome Informatics **10**:30-40

75: Jackson, SE., Fersht, AR. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry **30**:10428-10435

76: Scanlan, NC., Offer, J., Zitzmann, N., Dwek, RA. (2007). Exploiting the defensive sugars of HIV-1 for drug and vaccine design. Nature **446**:1038-1045.

77: Balzarini, J. (2006). Inhibition of HIV entry by carbohydrate-binding proteins. Antiviral Research **71**:237-247.

78: Chiba, H., Inokoshi, J., Nakashima, H., Omura, S., Tanaka, H. (2004). Actinohivin, a novel anti-human immunodeficiency virus protein from an actinomycete, inhibits viral entry to cells by binding high-mannose type sugar chains of gp120. Biochemical and Biophysical Research Communications **316**:203-210.

79: Kharitonenkov, IG., Siniakov, MS., Suvorova, ZK. (1980). Influenza virus haemagglutinin: estimation of tryptophan and tyrosine content and localization of tryptophan residues. J. General Virology **50**:419-422

Appendix A

Oligos

Oligonucleotides are written with 5' and 3' ends denoted, written 5' to 3'. A brief description of that oligonucleotide's function is given above each sequence.

Oligo A - First 29 bases
5'tatgggcgatggttattacaaactggttgcacg'3

Oligo B - Second 56 bases
5'ccattctggcaaagcgctggacgtggaaaacgccagcacctccgatggtgcgaacg'3

Oligo C - Third 56 bases
5'ttatccagtattcttacagcggcggtgacaaccagcagtggcgtctggtggatctgtaatag'3

Oligo A' - First 56 bases Complement
5'tccacgtccagcgctttgccagaatggcgtgcaaccagtttgtaataaccatcgccca'3

Oligo B' - Second 56 bases Complement
5'caccgccgctgtaagaatactggataacgttcgcaccatcggaggtgctggcgttt'3

Oligo C' - Third 29 bases Complement
5'gatcctattacagatccaccagacgccactgctggttgt'3

Oligo Ax - Forward Blunt Amplification Primer
Sigma = 5'ggcgatggttattacaaactgg'3

Oligo Cx - Reverse Blunt Amplification Primer
5'cagatccaccagacgccac'3

Oligo A-Primer - Forward Insertion and STOP codon Primer
5'ggactacatatgggcgatggttattacaaactgg'3

Oligo C-Primer - Reverse Insertion and STOP codon Primer
5'gttcaggatccctattacagatccaccagacgccact'3

Deca-Histidine Tag

The full sequence of the deca-histidine tag is shown with thrombin cleavage site underlined.

GSSHH HHHHH HHHSS GLVPR GSHM