

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

In search of novel folds: Protein evolution via non-homologous recombination

A dissertation presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biochemistry

at Massey University, Albany, New Zealand

Mayank Saraswat

2014

© 2014
Mayank (Mack) Saraswat
All rights reserved.

Abstract

The emergence of proteins from short peptides or subdomains, facilitated by the duplication and fusion of the minigenes encoding them, is believed to have played a role in the origin of life. In this study it was hypothesised that new domains or basic elements of protein structure, may result from non-homologous recombination of the genes coding for smaller subdomains.

The hypothesis was tested by randomly recombining two distantly related ($\beta\alpha$)₈-barrel proteins: *Escherichia coli* phosphoribosylanthranilate isomerase (PRAI), and β subunit of voltage dependent K⁺ channels (Kv β 2) from *Rattus norvegicus*. The aim was to identify new, folded structures, which may or may not be ($\beta\alpha$)₈-barrels. Incremental truncation (ITCHY), a method for fragmenting and randomly recombining genes, was used to mimic *in vivo* non-homologous recombination and to create a library of chimeric variants. Clones from the library were selected for right reading frame and solubility (foldability) of the recombined chimeras, using the pSALect selection system. Out of the six clones identified as soluble by pSALect, only one (P25K86) was found to be actually soluble. The protein, P25K86, was found to form oligomers and on treatment with a reducing agent, β -mercaptoethanol the multimeric state disappeared. The protein has three cysteines and one of the cysteines (Cys56) was found to mediate in the bond formation, thus giving a dimeric state. An engineered version of P25K86 that has the Cys56 replaced by serine was expressed as a monomer and additionally it was found to be

more stable.

As the pSALect folding selection system reported false positives, i.e. only one of the six chimeras was actually soluble, it was concluded that the *in vivo* solubility selection system was leaky. A series of experiments were conducted so as to improve pSALect that led to the creation of pFoldM – a more stringent selection system, discussed in chapter 4. Comparing the newer improved version with the old, two more interesting chimeras were discovered.

A total of 240,000 non-homologous recombination events were created *in vitro* and three soluble chimeras (evolutionary solutions) were found. Data from circular dichroism spectroscopy (CD) combined with heteronuclear single quantum coherence (HSQC) spectra suggest that the proteins, P24K89 and P25K86, are present in a molten globule state. ITCHY, as a means of mimicking the subdomain assembly model, was applied *in vitro*. The discovery of two interesting chimeras (P25K86 and P24K89) using high-throughput engineering experiments widens the possibilities of exploring the protein structure space, and perhaps offers close encounters with these *never born proteins* that may be trapped in an ensemble of fluctuating (structured and unstructured) states.

Acknowledgements

Boy it has been a philosophical journey! I have learnt a lot about myself and that is because of the people I met and the places I went. When I left the UK for New Zealand, I was passionate to embark on this journey to explore the “protein universe”. My supervisor, Dr. Wayne Patrick gave me this opportunity and taught me some really clever techniques. One in particular is my favourite and is called ITCHY. For that I thank him a lot. I also thank him because he coped with my craziness, my bipolarity and my entrepreneurial spirit.

Not long ago, Dr. Jane Allison did rescue me from sinking. I was very close to quit my PhD and move on with my ventures. She motivated me and told me that that there is light at the end of this tunnel. So for that I thank her a lot!

You can't study proteins if you don't have a crystallographer or a structural biochemist on board. My second supervisor, Dr. Gill Norris gave me the chance to work in her lab, for which I am extremely grateful. Trevor Loo has been phenomenal in showing me how things work in the lab and for that I will always be obliged.

Everything comes down to money, so I would like to thank Massey University for offering the scholarship and support to pursue this research. I would also like to thank the New Zealand Society for Biochemistry and Molecular Biology (NZSBMB) and Protein Society (USA) that provided the travel grant, which made my journey to USA possible. I am a huge fan of Marc

Ostermeier and I had the chance to meet him in the conference. Sadly I am leaving academia but I am grateful to Wayne, who arranged a lunch in some food court in Boston where I had the chance to talk to him about my research.

Several members from the lab need special mention: Daniela Quintana, Paulina Hanson-Manful, Matilda Newton (Oh I am so going to miss you pal), Matteo Ferla and Valerie Soo. These folks always make me happy.

I would also like to thank my really good friends: Ralph Grand, Lutz Gehlen, Katrin Hammerschmidt, Matt Woods, Peter Deines, Caroline Cotty, Frazer Nobel, Brig Kreig, Anthony Bloom, Stephen Carr, Jon Atherton, Leon Kapetas, Charlie Kerr, Stephen Wallace and Sharif Zamir.

Being an entrepreneur, I have also met some really encouraging people who have constantly motivated me to finish this thesis. Some of the best people I know are: Jim McNaught (my Guru), John Sullivan, Steve Corbett, John Harrison and lots of other good people (the list is way too long).

Lastly, I would like to thank my Mum and Dad. Mum for always nagging me to finish what I have started and Dad for helping me buy the toys I can't live without. And of course my two sisters, Meetu and Nikki for their endless support and love.

Table of contents

| | | |
|-----------|---|-----------|
| 1. | Introduction - The Protein Universe | 1 |
| 1.1 | Protein domains and folds defined | 2 |
| 1.2 | Protein Evolution | 3 |
| 1.2.1 | The role of subdomains in protein evolution | 5 |
| 1.2.2 | The primordial minigenes coding short peptides | 8 |
| 1.3 | Non homologous recombination and its role in generating novelty | 10 |
| 1.3.1 | Non-homologous recombination: definition and examples | 10 |
| 1.3.2 | Generating novelty <i>via</i> NHR | 12 |
| 1.4 | The subdomain assembly model: a central theme of this research project- | 16 |
| 1.5 | Aims of this study | 18 |
| 2. | Mimicking non-homologous recombination | 20 |
| 2.1 | Premise of the chapter | 21 |
| 2.2 | Introduction | 22 |
| 2.2.1 | TIM barrels – A proof-of-principle | 22 |
| 2.2.1.1 | The two distantly related barrels | 24 |
| 2.2.2 | Protein chimeragenesis: an engineering approach to mimic non - homologous recombination | 27 |
| 2.2.2.1 | Making the THIO-ITCHY library | 35 |
| 2.3 | Results | 37 |
| 2.3.1 | Library design | 37 |
| 2.3.1.1 | Making the long fusion PCR product | 37 |
| 2.3.1.2 | Making a unidirectional library | 38 |
| 2.3.2 | Crossover plots | 42 |
| 2.3.3 | The distribution of crossovers | 43 |
| 2.3.4 | Solubility validation | 44 |
| 2.3.4.1 | Proteins from cluster A | 45 |
| 2.3.4.2 | Proteins from cluster B | 49 |
| 2.4 | Discussion | 52 |
| 2.4.1 | pSAlect is not robust enough | 52 |
| 2.4.2 | The outcome of random recombination | 53 |
| 2.5 | Materials and methods | 56 |

- 2.5.1 Construction of pSAlect-trPRAI and pSAlect-Kv β 2 56
- 2.5.2 ITCHY library construction 57
- 2.5.3 Screening the clones 62
- 2.5.4 Scaling up and harvesting the library 63
- 2.5.5 Folding selection 63
- 2.5.6 Sequencing 64
- 2.5.7 Construction of the expression vector pLAB101 64
- 2.5.8 Expression and purification of chimeric proteins 66

3. Characterisation of P25K86 68

- 3.1 Premise of the chapter 69
- 3.2 Introduction 70
 - 3.2.1 Borrowing some tools from biophysics to study P25K86 70
 - 3.2.2 Circular dichroism spectroscopy 71
 - 3.2.2.1 Secondary structure analysis 72
 - 3.2.3 Nuclear magnetic resonance spectroscopy 74
 - 3.2.3.1 Chemical shift 74
 - 3.2.4 Protein NMR 75
- 3.3 Results 77
 - 3.3.1 Oligomerisation 77
 - 3.3.1.1 Size exclusion chromatography 77
 - 3.3.1.2 High performance liquid chromatography (HPLC) 80
 - 3.3.1.3 Mass spectrometry (MS) 81
 - 3.3.2 Investigating the role of cysteine in oligomerisation and a quick test for function 82
 - 3.3.3 The P25K86_CCS transition 84
 - 3.3.3.1 Labelling attempts for NMR spectroscopy 85
 - 3.3.3.1.1 Far-UV circular dichroism spectroscopy 87
 - 3.3.4 NMR spectroscopy of P25K86_CCS 92
- 3.4 Discussion 97
 - 3.4.1 Switching from P25K86 to P25K86_CCS 97
 - 3.4.2 What can we learn from this experiment? 100
- 3.5 Materials and methods 101
 - 3.5.1 Expression and purification of P25K86 & its variants 101
 - 3.5.2 Buffer for biophysical experiments 102
 - 3.5.2.1 Exchanging the buffer *via* concentrator 104
 - 3.5.3 Measuring concentration 104
 - 3.5.4 Gel filtration & HPLC 105
 - 3.5.5 Mass Spectrometry 106
 - 3.5.6 Making the P25K86 mutants 106
 - 3.5.7 Immunoblotting 108
 - 3.5.8 Labelling P25K86_CCS 109
 - 3.5.8.1 Preparing M9 minimal media base 110

| | | | |
|-----------|---------|--|------------|
| | 3.5.8.2 | Modified Media | 110 |
| | 3.5.9 | Circular dichroism spectroscopy | 111 |
| | 3.5.10 | NMR spectroscopy | 112 |
| 4. | | Improving the folding selection system | 113 |
| 4.1 | | Premise of the chapter | 114 |
| 4.2 | | Introduction | 115 |
| | 4.2.1 | Switch to pInSAlect | 115 |
| | 4.2.2 | Protein translocation via Tat and Sec pathways | 116 |
| | | 4.2.2.1 Tat <i>versus</i> Sec | 117 |
| | | 4.2.2.2 Export promiscuity | 118 |
| | 4.2.3 | Other improvements | 120 |
| | | 4.2.3.1 Improving the desalting step prior to transformation | 120 |
| | | 4.2.3.2 Switching to time-dependent ITCHY | 121 |
| | 4.2.4 | Focus of the chapter | 122 |
| 4.3 | | Results | 124 |
| | 4.3.1 | Attempts to improve the folding selection system | 124 |
| | | 4.3.1.1 Step 1 of the two-step strategy to reengineer pSAlect | 124 |
| | | 4.3.1.2 Step 2 of the two-step strategy to reengineer pSAlect | 126 |
| | 4.3.2 | Initial tests with pFoldM-KRK and pFoldM-KR | 129 |
| | 4.3.3 | A new library | 133 |
| | | 4.3.3.1 Test library results – a disappointment | 134 |
| | 4.3.4 | Time-dependent ITCHY of PRAI-Kvβ2 | 135 |
| | | 4.3.4.1 Optimized time-dependent ITCHY protocol | 136 |
| | 4.3.5 | A library that pushed things forward | 138 |
| | 4.3.6 | Comparing the three folding selection systems | 139 |
| | | 4.3.6.1 Is temperature the key to filtering out false positives? | 141 |
| | | 4.3.6.2 Crossover distribution of PRAI-Kvβ2 in the three-selection systems | 142 |
| | 4.3.7 | An overall ITCHY comparison | 145 |
| | 4.3.8 | Solubility validation | 147 |
| | | 4.3.8.1 P19K93: a quirky selection | 149 |
| | | 4.3.8.2 The protein P24K89 | 150 |
| | | 4.3.8.2.1 Purification attempts for biophysical characterization | 151 |
| | | 4.3.8.2.2 CD spectrum of P24K89 | 153 |
| | | 4.3.8.2.3 Impediments in NMR trials | 155 |
| | 4.3.9 | Support from our collaborators | 156 |
| | 4.3.10 | Comparing P24K89 and P25K86_CCS | |
| 4.4 | | Discussion | 162 |
| | 4.4.1 | An improved selection for protein folding? | 162 |
| | 4.4.2 | Comparing the three folding selection systems | 164 |
| | 4.4.3 | Comparing and contrasting the three soluble chimeras | 166 |

| | | |
|----------|--|-----|
| 4.5 | Materials and methods | 170 |
| 4.5.1 | Construction of pFoldM-KR and pFoldM-KRK | 170 |
| 4.5.1.1 | Step I of the strategy | 170 |
| 4.5.1.2 | Step II of the strategy | 172 |
| 4.5.2 | Sub-cloning P25K86 and P69K149 in both versions of pFoldM | 174 |
| 4.5.2.1 | Screening the clones | 175 |
| 4.5.3 | Sub-cloning PRAI and Kv β 2 in the vector backbone pInSAlect | 176 |
| 4.5.3.1 | Constructing the test library | 176 |
| 4.5.4 | THIO-ITCHY library | 177 |
| 4.5.5 | Screening the clones | 177 |
| 4.5.6 | Time-dependent ITCHY library | 178 |
| 4.5.6.1 | Steps of the optimized protocol | 178 |
| 4.5.7 | Column <i>versus</i> drop dialysis | 180 |
| 4.5.8 | Scaling up and harvesting the library | 181 |
| 4.5.9 | Frame selection | 181 |
| 4.5.10 | Sequencing | 182 |
| 4.5.11 | Comparing the fold selection system | 182 |
| 4.5.11.1 | Plating experiment | 183 |
| 4.5.12 | Expression and purification of chimeric proteins | 184 |
| 4.5.13 | Exchanging the buffer via concentrator and dialysis | 185 |
| 4.5.14 | Labelling P24K89 | 186 |
| 4.5.15 | Circular dichroism spectroscopy | 187 |
| 4.4.16 | Nuclear magnetic resonance spectroscopy | 187 |

5. Conclusions 161

| | | |
|-----|---|-----|
| 5.1 | The findings | 189 |
| 5.2 | Combinatorial assembly: a way to innovate | 191 |
| 5.3 | Engineering functions on artificially created folds | 193 |
| 5.4 | In the future | 194 |

Appendix I. General materials and methods 197

| | | |
|-------|------------------------------|-----|
| AI.1 | Reagents | 197 |
| AI.2 | Growth media and antibiotics | 197 |
| AI.3 | Storing strains | 198 |
| AI.4 | Bacterial strain | 198 |
| AI.5 | Plasmids | 198 |
| AI.6 | Oligonucleotides | 200 |
| AI.7 | Software | 201 |
| AI.8 | Agarose gel electrophoresis | 201 |
| AI.9 | SDS-PAGE | 201 |
| AI.10 | Electrocompetent cells | 202 |
| AI.11 | Sequencing | 203 |

Appendix II. Statement of contributions 204

Appendix III. Publication arising from this work 206

Appendix IV. Supplementary data 210

AIV.1 Protein sequence of chimeras 210

AIV.2 DNA and protein sequence of Kv β 2 and trPRAI 211

AIV.3 Multiple sequence alignment of all chimeras 213

AIV.4 Secondary structure elements in trPRAI; PDB-2KZH 214

AIV.5 Secondary structure elements in Kv β ; PDB-1EXB 215

AIV.6 Deconvolution of Kv β 2 216

AIV.7 Deconvolution of trPRAI 216

AIV.8 NADPH-binding subdomain experiment 217

AIV.9 SDS-PAGE gel of P25K86 218

References. 219

Abbreviations

| | |
|--------------------|--|
| α S-dNTPs | α -phosphorothioate (α S)-dNTPs (deoxyribonucleotide triphosphates) |
| c.f.u | colony-forming unit |
| DMSO | Dimethyl sulfoxide |
| DTT | Dithiothreitol |
| EDTA | Ethylenediaminetetraacetic acid |
| (His) ₆ | Hexa-histidine |
| IPTG | Isopropyl- β -D-1-thiogalactopyranoside |
| Kv β 2 | β subunit of voltage dependent K ⁺ channels |
| MWCO | Molecular Weight Cut Off |
| OD ₆₀₀ | Optical density at 600 nm |
| NADPH | Reduced form of β -nicotinamide adenine dinucleotide phosphate |
| NMR | Nuclear magnetic resonance |
| PCR | Polymerase Chain Reaction |
| PDB | Protein Data Bank |
| POI | Protein Of Interest |
| PRAI | Phosphoribosylanthranilate isomerase |
| SEM | Standard Error of the Mean |
| SOB | Super Optimal Broth |
| SOC | SOB with Catabolite repression |
| Tat | Twin arginine translocase |

| | |
|--------|-----------------------------------|
| Tris | Tris (hydroxymethyl) aminomethane |
| trPRAI | truncated version of PRAI |

Chapter I

Introduction

1. The Protein Universe

The term protein is derived from the Greek word *proteios*, which means primary. Proteins are the workhorse molecules, performing the essential activities of life (Apic & Russell 2010). A protein molecule has three fundamental properties that are related to each other. A protein is defined primarily by its amino acid sequence, which contains information such as the type of atoms and the way these atoms are connected via chemical bonds. The second property is the arrangement of the atoms of a protein in a three-dimensional (3D) space. The structure of a protein is defined by how every atom is placed in 3D. Lastly, a protein is defined by the biological function that it performs. Naturally occurring protein sequences, as opposed to synthetic proteins that lack function and are prone to aggregation, are assumed to adopt only one native structure and similar sequences are

thought to fold in similar structures (Stirling *et al.* 2003). Albeit these assumptions are true in most cases, there are proven exceptions (Kolodny *et al.* 2013). Another essential unit of protein structure is the structural domain.

1.1 Protein domains and folds defined

Protein domains constitute the basic elements of protein structure, which are spatially distant parts of a protein that fold independently of neighbouring sequences. They are compact polypeptide structures, organised around a hydrophobic core and associated with a specific function or activity (Russell 2007; Alva *et al.* 2010; Zhang 1997).

One or more of the following criteria can classify protein domains:

- Spatially separate regions of protein chains.
- Domains vary in size and may constitute 50 to 300 amino acids. Forty-nine per cent of all the domains range from between 51 and 150 residues (Petsko & Ringe 2004), although much smaller domains (12-16 residues) that can fold into a tertiary structure without the need of metal binding or a disulfide link have been reported. One such example is the tryptophan zipper peptide or trp-zips, which are monomeric β -hairpins stabilised by tryptophan-tryptophan cross-strand pairs (Cochran *et al.* 2001).
- Sequence and/or structural resemblance to an entire chain from another protein.
- A specific function associated with a region of protein structure.
- A substructure in another protein that meets one or more of the above requirements.
- Repeating substructures within a single chain that meet one or more of the above requirements.

Each domain has a 3D structure (fold) arising from its amino acid sequence. A fold is a feature of most protein domains where secondary structural elements in space are arranged in a certain way. Combinations of different numbers and types of domains form complex proteins with various biological functions. It is estimated that there are between 1000 and 10,000 folds, which is a small number given the diversity of sequences (Aloy & Russell 2004; Kolodny *et al.* 2013). Fundamentally the protein universe is built up of a string of amino acids (sequence space), whose order determines their arrangement in 3D (structure space) and ultimately gives rise to phenotypes (function space) in the context of a living cell or an entire organism. There are only 20 different standard amino acids and given a polypeptide chain of length 100, there are 20^{100} possible scenarios in the sequence space, which is a number much larger than the number of electrons in all the galaxies of the physical universe (Dill 1999; Kolodny *et al.* 2013). Given the limited number of folds, broadly between 1000 and 10,000 (Schaeffer & Daggett 2011) it appears that nature re-uses these folds again and again to perform totally new functions.

1.2 Protein evolution

Until today, it is still not entirely clear how the 'modern' proteins evolved. There are two models that help us understand these steps in protein evolution. The single-birth model suggests that the present-day proteins evolved from proteins that were present in the last universal common ancestor (LUCA), whereas the multiple-birth model suggests emergence of ancestral proteins at different times (Bukhari & Caetano-Anollés 2013; Choi & Kim 2006). In the multiple-birth model for protein evolution, not every present-day protein evolved from a single set of proteins in LUCA, rather there was emergence of new common ancestral proteins throughout

evolutionary time (Choi & Kim 2006). Irrespective of their evolutionary emergence, it is important to note that protein structures are more conserved than the sequences from which they are derived. Structures can withstand sequence variations provided the perturbations to their physical (thermodynamic stability) and chemical properties (biological function or loss of structural integrity due to bond cleavage) are minimal. This means that proteins with no apparent sequence identity can have the same overall folds. Examples include the polymerase fold (both DNA and RNA polymerase share this fold but have no sequence identity) and Ig fold (52 members of the immunoglobulin-fold family (Ig) are evolutionarily unrelated but share the Ig fold) (Patel & Loeb 2000; Halaby *et al.* 1999). Nature uses a limited number of stable folds relative to what is theoretically possible in protein sequence space. The folds that exist today may have arisen through divergent evolution from a limited set of ancient proteins (single-birth model) or repeated convergent evolution to stable structures (multiple-birth model) (Yang *et al.* 2009; Sadowski & Taylor 2010).

The dominance of duplication followed by divergence is evident from the proportion of duplicated domains in various organisms. Animal, fungi and bacterial genomes contain at least 93%, 85% and 50% duplicated domains respectively. In the event of duplication the original copy retains its function, whereas the new divergent copy 'explores' alternative functions (Chothia & Gough 2009; Kolodny *et al.* 2013). The imprints of intragenomic duplication can be seen in the human genome as large-scale inter- and intra-chromosomal similarities, and as small-scale DNA repeats (Lupas *et al.* 2001). Lateral transfer is a genetic event in which one organism takes up fragments of the genome or even an operon (for example, lateral transfer of the *Mycobacterium tuberculosis* Rv0986-8 virulence operon from a γ -

proteobacterium donor to a *M. tuberculosis* ancestor) of another organism, which can be traced as “imperfect duplications” of parts of the host genome, i.e. the fragments acquired by the host organism (Rosas-Magallanes *et al.* 2006). Events like duplication, gene fusion and gene sharing explain some of the evolution in the protein world, i.e. the origin of paralogous and multidomain proteins from an ancestral “vocabulary” of protein domains (Ruepp *et al.* 2000; Lupas *et al.* 2001). However, a pertinent question, namely, how the domains themselves, the building blocks of all proteins arose, still remains unanswered.

Some genetic events may cause fundamental changes in the structure of protein domains, such as circular permutation (occurs by gene duplication, fusion and partial deletion) and illegitimate recombination (that occurs between unrelated genes, potentially leading to new folds when the recombined parts prove structurally compatible) (Ponting & Russell 1995; Eisenbeis *et al.* 2012). Owing to advances in sequence and structure data comparison and increasing *in silico* sensitivity to structural data analysis, a different view on the evolution of domains is starting to emerge. It has been speculated that the contemporary domains, which are single chain, arose from ancestral counterparts that were oligomeric and formed from assembly of short polypeptides (Lupas *et al.* 2001).

1.2.1 The role of subdomains in protein evolution

Subdomains are portions of proteins that are bigger than individual secondary structure elements but smaller than domains (Ostermeier & Benkovic 2001). All known protein structures, in the existing protein families, can be assigned to one of only four major classes, which include all- α , all- β , α/β (combination of β - α - β motifs) and $\alpha+\beta$ (mixture of all- α and all- β) as

defined by structural classification of proteins (SCOP) (Lo Conte *et al.* 2000). The presence of supersecondary structural elements ($\beta\beta$ -hairpin, $\alpha\alpha$ -hairpin and $\beta\alpha\beta$ -element) in these classes and their occurrence as repeats indicate a possibility of multiple recombination events thus making them the building blocks of ancient proteins. This suggests that the existing protein families did not evolve from the LUCA but *via* the multiple birth model, where independent recombination of subdomains or supersecondary structural elements gave rise to proteins (Choi & Kim 2006).

Using *in silico* approaches there is a growing support for the idea that domains might have been formed from the assembly of subdomains. Bioinformatics methods that do not focus on similarity across entire domains, but instead on localised regions of sequences and structures, have shown that there are similar folds whose evolutionary relationships remain ambiguous and that there are different folds that contain similar localised structures. Russell (1998) performed an all against all comparison of known structures searching for common protein 3D side chain patterns within different protein folds. The search led to the identification of several similarities in structurally clustered side chains. The majority of examples are likely to be the result of convergent evolution as amino acids packed in a similar pattern occurred in different stretches along the polypeptide chain. An example is the Ser/His/Asp catalytic triad of trypsin like and subtilisin like serine proteases (Russell 1998). Lupas *et al.* have detected proteins with different folds that contained short stretches of amino acids with both similar sequences and 3D structures. Examples include the Fe-S binding site in trimethylamine dehydrogenase, 4Fe-4S ferredoxins, cytochromes c, protein phosphatases and NAD/FAD binding motifs (Lupas *et al.* 2001). The histidine kinase fold in bacteria has a topological similarity to domains in other eukaryotic kinases. These two folds, which are interrelated by circular

permutation, bind nucleotides at equivalent sites (Koretke *et al.* 2000). Structural alignment of other such different folds suggests that fragments from non-homologous proteins appear to have common origin. Furthermore, *in silico* analysis of the twenty four structures of PLP (Pyridoxal 5' phosphate) dependent enzymes, which use an identical phosphate binding cup, has shown that these structures fall into five different (non-homologous) structural families despite all of them anchoring PLP. These examples do point towards a probable pathway to domain evolution via subdomain assembly (Denesyuk *et al.* 2002).

Proteolytic dissection was used to study the folding pathway, of *trp* repressor and horse heart cytochrome *c*. It was found that both proteins, upon mild proteolysis, were cleaved into small fragments that spontaneously associated to form subdomains *via* non-covalent interactions. These subdomains had native-like secondary and/or tertiary structural characteristics, suggesting folding could be encoded at the subdomain level (Wu *et al.* 1994). More recently, experimental support for the evolutionary pathway of a symmetric protein was demonstrated using the human fibroblast growth factor-1 (EGF-1), which has a characteristic threefold β -trefoil fold. In a top-down experiment, the β -trefoil fold was deconstructed to form monomer (monofoil) and dimer (difoil) polypeptides. A simple trefoil fold peptide motif (monofoil) was found to restore the β -trefoil fold *via* gene duplication and fusion (Lee & Blaber 2011). This experiment successfully showed that subdomain-sized polypeptides have a role as the building blocks in extant protein structure thus supporting the idea of the evolution of proteins from simple peptide motifs.

1.2.2 The primordial minigenes coding short peptides

The emergence of modern-day proteins and the diverse array of functions associated with them could perhaps be the result of combinatorial arrangements of a limited number of units of structures with partial functions. For instance, functional units such as DNA-binding motifs and metal binding motifs, reoccur in different cases to confer new functions (Dorit *et al.* 1990). Such motifs could be the transitional product of minigenes (genes-in-pieces) of primordial life that code for polypeptides of ~40-60 residues. Russell and coworkers have suggested that many current protein domains are the result of combination of smaller domain segments. Such small polypeptide segments were encoded by primitive minigenes that tended to homo-oligomerise in aqueous solution thereby forming the folded proteins of the primordial protein world. Over the evolutionary time-scale, selection pressure applied through entropic and thermodynamic factors may have encouraged the development of genes produced by fusion of these small domain segments (Lupas *et al.* 2001; Ali & Imperiali 2005).

In 1978, Blake asked the question “do genes-in-pieces imply proteins-in-pieces? He extended Doolittle’s idea that a “genes-in-pieces” structure is a primitive form present in the genomes of the common ancestors of both prokaryotes and eukaryotes. An obvious assumption was that exons correspond to integrally folded protein units – domains or subdomains. It was suggested that two conditions must be fulfilled if a new protein arises from the combination of exonic regions; firstly the new polypeptide chain must fold into a stable globular form, and secondly the newly folded chain must contain an active site with some functional property. Combinations occurring in exonic regions will be more likely to be stable, being the “sum of parts” of existing stable proteins, thereby improving the chances of fulfilling

the two conditions. Based on these ideas, it was proposed that individual protein domains might have evolved by assembly of smaller gene fragments, via exon shuffling or non-homologous recombination. This would have led to the diversification of domains and even generated novel folds (Blake 1978; Stoltzfus *et al.* 1994).

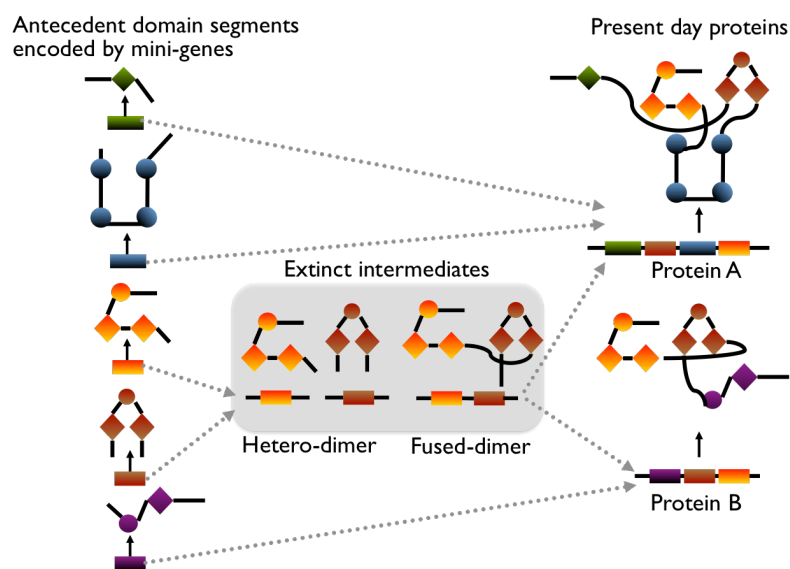


Fig. 1.1. ADS evolutionary scenario. Protein A and protein B may have arisen from a series of different Antecedent Domain Segments (ADS). α -Helices are represented by circles and β - strands by diamonds. Different coloured rectangular segments represent minigenes encoding the corresponding ADS. Adapted from Lupas *et al.* 2001

Alternatively, it has been proposed that the most ancient proteins were oligomers, comprising several antecedent domain segments (ADSs). In this model, the single chain domains of today arose from the fusion of ancient genes that encoded ADSs. These ADS encoding genes could be fused in many combinations, giving rise to modern proteins with conserved segments, which are otherwise non-homologous. Figure 1.1 shows a scenario of how two hypothetical modern proteins (A and B) may have formed from ADSs. Two minigene segments (orange and brown in Fig. 1.1) correspond to the ancient progenitors of a fused heterodimer, the common ancestor of both

proteins. This hypothesis also tends to support the idea that short peptides played a role in the origin of life (Lupas *et al.* 2001).

1.3 Non-homologous recombination and its role in generating novelty

Large-scale genomic rearrangements are involved in major evolutionary events. Studies on plants suggest that dramatic changes of genomes in response to shock may lead to the evolution of new species (McClintock 1984). This is not limited to plants. In fact, a gene transfer is believed to have resulted in the replacement of the genes of primitive eukaryotes with genes from bacteria consumed as food by the eukaryotes (Doolittle 1998). For efficient horizontal transfer between organisms, bacterial genes coding proteins necessary to produce a protein with a single metabolic function are often found in operons (Doolittle 1998). Transfer of genes can confer novel functions by allowing the recipient to exploit a new environmental niche. This mobility may facilitate speciation (Lawrence 1997). Shapiro (1997) argues that “natural genetic engineering” employs a toolbox of mechanisms for generating novelty and that only lateral and horizontal transfer events in the past, can explain certain peculiarities that exist in modern-day genomes. In addition, it is only through large-scale genomic rearrangement between species that the rapid emergence of bacterial antibiotic resistance can be explained (Shapiro 1997).

1.3.1 Non-homologous recombination: definition and examples

Non-homologous recombination (NHR) takes place in regions with no large-scale sequence similarity and it occurs more frequently than homologous recombination. Some examples of NHR include chromosome deletions, inversions and translocations (Kegel *et al.* 2006). Exon shuffling is the means by which novel combinations of exons are created *via* NHR (Long, Betrán *et al.*

2003). Exon shuffling occurs within introns, leading to new combinations of exonic regions (Long, Deutsch *et al.* 2003). The probable mechanism of NHR comprises three steps, which include insertion of introns at the borders of the protein domain, recombination in inserted introns causing tandem duplications and NHR-mediated transfer of introns to a non-homologous gene (Keren *et al.* 2010). This is known as the modularisation hypothesis. Introns basically increase the length of a chromosome, which increases the probability of NHR between the exons of different genes (Long, Betrán *et al.* 2003; Long, Deutsch *et al.* 2003). The EGF-like and C-type lectin domains are examples that demonstrate this mechanism (Keren *et al.* 2010). As well as NHR occurring in introns at the DNA level, retroposition (sequence derived from RNA integrated into DNA) is another way for NHR to occur at the RNA level (Brosius 2003; Brosius 1999; Long, Deutsch *et al.* 2003). The chimeric gene *jingwei* from *Drosophila teissieri* and *D. yakuba* is a good, recent example of exon shuffling by retroposition. This gene was created by reverse transcription of *alcohol dehydrogenase* messenger RNA and its insertion into the third intron of the *yellow emperor* (*ymp*) gene (Long & Langley 1993; Ding *et al.* 2013). The resultant chimeric protein participates in hormone/pheromone metabolism, and has a novel function compared to its parental gene *Adh* (Ding *et al.* 2013). Some other chimeric genes created by exon shuffling in humans include the tissue plasminogen activator (TPA) and low-density lipoprotein (LDL) receptor (Arguello *et al.* 2007). TPA (a serine protease that converts plasminogen to plasmin) contains domains that are very similar to urokinase, epidermal growth factor and fibronectin, while the LDL receptor has evolved by mixing and matching exons from the C9 component factor, the EGF precursor and other blood-clotting factors (Arguello *et al.* 2007).

1.3.2 Generating novelty *via* NHR

As discussed in the previous section, exon shuffling *via* NHR can yield new combinations of gene fragments or chimeric genes. This may result in different polypeptide sequences or chimeric structures with new functions (Koide 2009). An example of a young chimeric gene generated by NHR is *Hun* *Hunaphu* (*Hun*), which is recently evolved (within the last 2-3 million years). This young gene was created at some time before the speciation of *Drosophila simulans*, *D. sechellia* and *D. mauritiana* from *D. melanogaster* (Arguello *et al.* 2006; Arguello *et al.* 2007).

Using tools from computational biology, it has been suggested that domain recombination is the major driving force bringing about leaps in protein evolution (Bogarad & Deem 1999). Using Monte Carlo simulations, it was shown that point mutation by itself is largely unable to yield new protein folds. It was also shown that NHR is much more efficient at searching fitness space (Bogarad & Deem 1999). During any evolutionary process, proteins become trapped in local energy minima in the protein fold landscape. While major shifts are needed to bring them out of those regions, they can also be deleterious. The success of major evolutionary shifts depends on population size, generation time, mutation rate, population mixing and selective pressure. Mechanisms have evolved that increase the probabilities of successful exchanges (shifts) (Bogarad & Deem 1999). For example, transposons play a vital role in large-scale integration, where DNA moves from one position to a different one, independent of homology (Pennisi 1998). Long-terminal repeat (LTR) retrotransposons, long interspersed elements (LINEs) and helitrons can facilitate the evolution of new genes (Wang *et al.* 2006; Morgante *et al.* 2005). Another interesting pathway to diversity is non-homologous end joining (NHEJ), which repairs double-strand

breaks in DNA. What makes NHEJ different from homologous recombination is that it does not require a homologous template in order to ligate the broken ends of the DNA. NHEJ also plays a critical role in rearranging the breaks introduced during V(D)J (variable, diversity and joining gene segments) recombination – a mechanism by which T-cell and B-cell receptor diversity is produced. Despite more than 50% in-frame failures, V(D)J recombination is still effective (Popławski & Błasiak 2003; Valencia-Burton *et al.* 2006; Bogarad & Deem 1999). Bogarad and Deem further demonstrate by using simulations that evolution *via* non-homologous recombination of protein segments is many orders of magnitude more effective than point mutation in acquiring a significant new function. A very good example is the evolution of *Escherichia coli* from *Salmonella* that involved major shifts over an evolutionary timescale (horizontal gene transfer) (Lawrence 1997). None of the observed phenotypic differences between them could be attributed to point mutations (Lawrence 1997). Furthermore the rate of evolution caused by major shifts account for 31,000 bases/million years whereas for point mutation it is 22,000 bases/million years, which suggests that even though these major shifts or dramatic changes can be deleterious and are less likely to be tolerated than point mutations, there is an overall higher rate of evolution. Thus it seems plausible that non-homologous recombination, in addition to base substitution and homologous recombination are required to explain the observed diversity in protein structure space (Bogarad & Deem 1999).

Exon shuffling has been widely shown, as a non-homologous process for the generation of new proteins (Gilbert 1978). Examples discussed earlier (*jingwei* and *Hun*) show that new combinations of exons from different genes, subunits of structure, through NHR may give rise to new proteins with novel functions. The *Tre2* (*USP6*) oncogene (*FUT*) is yet another example of exon

shuffling, where this gene is the result of a chimeric fusion of *USP32* (*NY-REN-60*) and *TBC1D3*. In this case, an ancient highly conserved gene (*USP32*) has fused with a more recent gene (*TBC1D3*), which is a product of segmental duplication and is absent in most mammals but is amplified and dispersed through the primate lineage (Paulding *et al.* 2003). The first identified example of exon shuffling was in the low-density lipoprotein receptor gene (Südhof *et al.* 1985). Exon shuffling has also been shown to occur in many genes of plants, vertebrates and invertebrates. One of the mechanisms that drive exon shuffling is illegitimate recombination (Long 2001). Bloemendal and coworkers have shown illegitimate recombination gives rise to exon shuffling in a small heatshock protein, α A crystallin, in hamsters. They transfected a construct of the hamster α A crystallin gene into a mouse muscle cell line, which led them to identify a mutant α A crystallin gene with a large intragenic duplication. The sequence of the mutant suggested that it was created by illegitimate recombination between two CCCAT sites in two α A crystallin genes, one at intron 3 and the other in exon 2 (an intron in another isoform of α A crystallin), resulting in an internally duplicated exon structure (Van Rijk *et al.* 1999; Van Rijk *et al.* 2000). This specific sequence requirement would, however, result in a low rate of recombination, which does not explain the higher frequency of exon shuffling observed. On the other hand, the L1 element mediated 3' transduction (L1 is a retrotransposon that can reverse transcribe and get inserted in the mammalian genome) has been reported to be efficient in generating diversity *via* exon shuffling. Based on the location of L1, a whole nuclear gene or exons downstream of L1 can be taken by L1 thus recombining with the exons of the recipient locus (Moran *et al.* 1999; Long 2001).

The movement of transposable elements (TE) has also been shown to generate protein diversity. Makalowski and Boguski have examined vertebrate genomes in which they detected 200 cases where mobile elements were inserted into genes coding proteins and altered the functions of the proteins that received the TE insertions (Makalowski & Boguski 1998; Long 2001). They found that the *Alu* TE, which is abundant in primate genomes, was responsible for new peptides within new proteins. An example is the human decay-accelerating factor (DAF), which contains an *Alu* fragment. This new protein received a novel hydrophilic carboxy terminal region as a result of an *Alu* insertion, resulting in an intracellular location that is different from that of the original protein (Caras *et al.* 1987).

Lateral gene transfer is another mechanism whereby non-homologous recombination can generate diversity in the protein structure. De Koning *et al.* (2000), showed that the gene encoding N acetyl neuraminate lyase in the protozoan *Trichomonas vaginalis* shares a high protein sequence similarity (80%) with neuraminate lyase from the bacterium *Haemophilus influenza* suggesting that this is a recent transfer event. The newly transferred gene has a new leader sequence encoding 24 amino acids, which acts as a signal peptide. The *T. vaginalis* N acetyl neuraminate lyase is a secreted protein, whereas the bacterial neuraminate lyase is cytosolic, which suggest that the protein in *T. vaginalis* may have evolved a new function (Koning *et al.* 2000).

Interestingly, comparisons of oncogenes and proto oncogenes show that alterations in domain architecture through non-homologous recombination can trigger dramatic changes in phenotype, such as carcinogenesis. The examples include BCR Abl (fusion of the dimeric BCR protein to Abl kinase) and v Src (elimination of a phosphor Tyr site for intramolecular SH2

interaction), both of which result in kinase deregulation. These mechanisms in natural proteins suggest the effectiveness of non-homologous recombination in mediating large leaps in protein structure and function (Huse & Kuriyan 2002; Koide 2009).

1.4 The subdomain assembly model: a central theme of this research project

In this project, the hypothesis that protein domains may have evolved by non-homologous recombination of smaller subdomains was tested. According to this model, combinatorial assembly of smaller polypeptides could contribute a basic unit of structure or function to a newly assembled domain (Fig. 1.2).

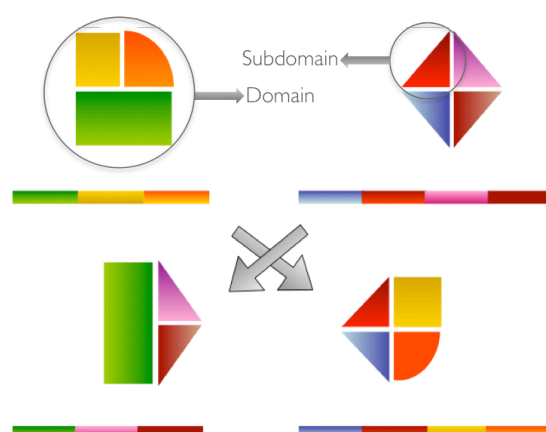


Fig. 1.2. The subdomain assembly model. Two domains (top) consist of various subdomains (coloured shapes), each of which may (right) or may not (left) be related structurally. Non-homologous recombination of their genes yields new domains with different subdomain combinations (bottom).

The idea that protein domains may have evolved by combinatorial assembly of smaller gene fragments (coding stable subdomains) is appealing because assembly of smaller subdomains by means of non-homologous recombination could contribute a basic unit of structure (such as a $\beta\beta$ hairpin) or function (such as a phosphate binding motif) to a newly

assembled domain (Frenkel & Trifonov 2008). For example, a very recent *in silico* study suggests that the outer membrane β barrels evolved by duplication of an ancestral $\beta\beta$ hairpin (Remmert *et al.* 2010).

To date, experiments to investigate the subdomain assembly model have focused on assembling subdomain sequences from structurally related proteins. This has been used to reconstruct the β propeller and $(\beta\alpha)_8$ barrel folds (Yadid & Tawfik 2007). The assembly of two related $(\beta\alpha)_4$ subdomains to build a functional $(\beta\alpha)_8$ barrel like protein is one such example (Bharat *et al.* 2008). Furthermore, the coexpression of the two $(\beta\alpha)_4$ units resulted in self association into a functional hetero dimer with activities similar to the wild type protein (Höcker *et al.* 2004). It has been suggested that the $(\beta\alpha)_4$ half barrel itself may have originated from even smaller elements such as individual $(\beta\alpha)$ units (Janecek & Baláz 1993).

Another fascinating experiment is where a gene coding a domain with a flavodoxin-like fold (CheY) was recombined with a gene coding a $(\beta\alpha)_8$ -barrel (HisF), which resulted in a "hopeful monster". The new protein had an extra β -strand in the core of the new barrel (CheYHisF) (Eisenbeis *et al.* 2012; Shanmugaratnam *et al.* 2012). The chemotaxis response regulator (CheY) is a doubly wound $(\beta\alpha)_5$ -protein, whereas HisF (imidazole glycerol phosphate synthase) is a $(\beta\alpha)_8$ -barrel. The half-barrel topology of $(\beta\alpha)_8$ -barrel resembles that of the flavodoxin-like fold. As both proteins share structural similarities, fragments from both folds were recombined to create a nine-stranded chimera, CheYHisF (Eisenbeis *et al.* 2012; Bharat *et al.* 2008). The new β -strand occupies the space between β_1 and β_2 of CheY and is formed by residues of the C-terminus of HisF and the histidine tag, which was strategically placed for protein purification (Shanmugaratnam *et al.* 2012; Eisenbeis *et al.* 2012). Interestingly, upon removing the residues responsible

for making the new strand (variant CheYHisF-sfr), the protein aggregated, suggesting that this new structural element is responsible for holding this monster together. However, upon introduction of five point mutations in the CheYHisF-sfr variant, the aggregation problem was overcome and this new variant, called CheYHisF-sfr_RM, was solubly expressed (Eisenbeis *et al.* 2012). Another interesting aspect of this study was that CheYHisF-sfr_RM resumed the $(\beta\alpha)_8$ -barrel. Furthermore, both CheYHisF (nine-stranded) and CheYHisF-sfr_RM (eight-stranded) were functional as both could bind phosphorylated compounds (Eisenbeis *et al.* 2012; Shanmugaratnam *et al.* 2012). This study highlights the possibility of hopeful monsters, generated *via* non-homologous recombination, that could potentially be new folds that have not yet been sampled by evolution, or if sampled, have been converged back to the stable $(\beta\alpha)_8$ -barrel fold (Höcker 2013). The five targeted mutations were part of a futile cycle, which changed CheYHisF to CheYHisF-sfr_RM. CheYHisF could also be a seed for new folds. Höcker (2013) further argues, "*We view the protein chimaeras that we engineered by recombination of fragments from different folds as hopeful monsters that are worthwhile exploring*".

In another example, a chimeric protein (1B11) was obtained by non-homologous recombination of two subdomain-sized segments - one from the cold shock protein CspA and the other from the S1 domain of the *Escherichia coli* 30S ribosomal subunit. These two proteins are distantly related, sharing the nucleic acid binding OB fold, yet 1B11 is a tetramer comprising four six-stranded β -barrels (De Bono *et al.* 2005).

1.5 Aims of this study

In this thesis, high-throughput engineering approaches have been used to test the subdomain assembly model. For the proof-of-concept, two distantly

related ($\beta\alpha$)₈ barrel proteins were randomly recombined in an attempt to make a library of shuffled subdomain fragments. For this, the *Escherichia coli* phosphoribosylanthranilate isomerase (PRAI), and rat KvBeta2 (Kv β 2) proteins were used. Incremental truncation (ITCHY), a method for fragmenting and randomly recombining genes, was used to mimic *in vivo* non-homologous recombination. To identify folded structures in the library of recombined polypeptides, ITCHY variants were cloned into pSALect. In this plasmid, the sequence of interest is positioned between an N-terminal Tat signal sequence and a C-terminal β -lactamase reporter (TEM-1). Cells containing the plasmid are spread on carbenicillin-containing plates. Only PRAI-Kv β 2 chimeras with in-frame crossovers yield full-length (Tat-chimera-TEM-1) fusion proteins. More importantly, Tat directs the export of folded proteins to the periplasm, and TEM-1 must be in the periplasm to confer resistance to carbenicillin. Using this system, six proteins were picked and validated for solubility. This forms the basis of Chapter 2.

Chapter 3 describes the exploration of a chimeric protein, P25K86 that was discovered in chapter 2. Biophysical characterisation of the protein was done using circular dichroism (CD) and nuclear magnetic resonance (NMR). Chapter 4 focuses on addressing the pitfalls and improving those that were observed in creating a hybrid library and the *in vivo* solubility screening system. The improvements led to the discovery of another interesting chimeric protein, P24K89 whose structural properties is also reported.

Overall, this research offers experimental support for the evolution of protein domains from assembly of partially structured subdomains. Chapter 5 discusses the whole project, and establishes a link with similar events that occur in nature. Future directions and further expansion of the strategy to search for novel folds are also discussed.

Chapter II

Mimicking non-homologous recombination

Acknowledgement: **Dr Monica Gerth** (Department of Biochemistry, University of Otago) for helping in setting up ITCHY experiments and for other valuable discussions.

2.1 Premise of the chapter

This chapter extends the idea of the subdomain assembly model by mimicking non-homologous recombination *in vitro*. To achieve this, two non-homologous proteins were chosen and both were truncated randomly and recombined to make a library of shuffled chimeric fragments. The recombined chimeras from the library were selected for correct reading frame and their potential to fold using the pSALect selection system. This proof-of-principle study was initiated based on the condition (for this study) that only the soluble chimeras might be folded and those candidates were explored further. Six clones were randomly selected and the chimeric proteins expressed and validated for solubility. One clone in particular, out of the six, was found to be highly soluble.

2.2 Introduction

2.2.1 TIM barrels – A proof-of-principle

To demonstrate the feasibility of the subdomain assembly model, i.e. the idea that protein domains may have evolved by non-homologous recombination of smaller subdomains, two distantly related triosephosphate isomerase (TIM) barrels were examined. The $(\beta\alpha)_8$ or TIM-barrel fold provides an excellent model system to address the subdomain assembly model. About 10% of all proteins with known three-dimensional structure contain a $(\beta\alpha)_8$ -barrel fold and these are versatile enzymes that act as oxidoreductases, transferases, lyases, hydrolases and isomerases. The fold of the canonical $(\beta\alpha)_8$ barrel consists of a closed eight-stranded parallel β -sheet, forming a central barrel, which is surrounded by eight α -helices. The active site residues are located on the catalytic face of the barrel, which comprises the C-terminal ends of the β -strands and the loops that link β -strands with the subsequent α -helices. In contrast, the loops that link the α -helices with the subsequent β -strands, which are located on the opposite face of the barrel, are important for stabilising the fold (Höcker *et al.* 2001).

Sequence and structural similarity suggest that several $(\beta\alpha)_8$ -barrel enzymes from the pathways of tryptophan and histidine biosynthesis have evolved by divergence from a common ancestral enzyme. Single amino acid exchanges can confer PRAI (phosphoribosylanthranilate isomerase, the product of the *trpF* gene), activity on both HisF (imidazole glycerol phosphate synthase) and HisA (phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide isomerase ProFAR isomerase). Both HisF and HisA, which catalyse two consecutive reactions of histidine biosynthesis, are

characterised by internal two - fold symmetry. In an attempt to generate a $(\beta\alpha)_8$ barrel from two apparent $(\beta\alpha)_4$ half barrels, Hocker *et al.* duplicated then joined the C - terminal half barrel from HisF to yield a stable and monomeric HisF - C**C* barrel. In addition, the N- and C-terminal half barrels of HisA and HisF were fused to yield chimeric HisAF and HisFA proteins, which showed that $(\beta\alpha)_8$ barrels could be assembled from $(\beta\alpha)_4$ half barrels. This led the authors to speculate that “a primordial gene encoding a $(\beta\alpha)_4$ half barrel as a subunit of a homodimeric enzyme was duplicated and fused to yield a monomeric, ancestral $(\beta\alpha)_8$ barrel, from which HisA, HisF and arguably TrpF evolved by a series of further gene duplication and diversification events” (Höcker *et al.* 2004).

Recently, Richter *et al.* have demonstrated that $(\beta\alpha)_8$ barrels may have evolved from even smaller, quarter barrel subdomains. This suggests that barrels have evolved by two gene duplication and fusion events from an ancestral $(\beta\alpha)_2$ quarter barrel and through an intermediate $(\beta\alpha)_4$ half barrel. Richter *et al.* used phylogenetic approaches to reconstruct the sequence of HisF - LUCA (HisF last universal common ancestor). They showed that the quarter - barrels of HisF - LUCA are more similar to each other than the corresponding quarter barrels in extant HisF proteins. The HisF - N1 protein, purified from *Thermotoga maritima*, was found to be a quarter barrel ($(\beta\alpha)_2$) that is more stable than other quarter barrels and forms a tetramer. The fusion of two HisF - N1 quarter barrels yielded a stable half barrel HisF - N1N1 (Richter *et al.* 2010). However, all these studies are based on joining very closely related partial barrels. It remains unknown whether $(\beta\alpha)_8$ barrels can be assembled from subdomains or motifs from distantly related sequences. Therefore, it could be argued that the symmetry of the HisA and HisF barrels is the exception and not the rule, in $(\beta\alpha)_8$ barrels.

The objective of this study was to build on the aforementioned studies by using ITCHY (a method to randomly recombine non-homologous genes - see section 2.2.2) in combination with the pSAlect fold selection system (see section 2.2.2) to explore the subdomain assembly model (see Chapter 1, section 1.4) for distantly related $(\beta\alpha)_8$ barrels. This was achieved by randomly truncating two distantly related barrels and making a fusion library of chimeric variants. The library was screened for folded soluble chimeras by exploiting the Tat quality control feature in the pSAlect fold selection system (see section 2.2.2).

2.2.1.1 The two distantly related barrels

PRAI (Fig. 2.1) catalyses the Amadori rearrangement of N - (5' - phosphoribosyl) anthranilate (PRA) to 1' - (2' - carboxyphenylamino) - 1' - deoxyribulose 5' - phosphate (CdRP). This is the third step in the synthesis of tryptophan from chorismic acid. CdRP is in turn the substrate for indoleglycerol - phosphate synthase (IGPS). Although PRAI is part of a bifunctional IGPS-PRAI enzyme in *E. coli*, the two domains have been separated genetically and expressed as stable, monomeric proteins with virtually full catalytic activity (Patrick & Blackburn 2005).



Fig. 2.1. Ribbon diagram of PRAI from *E. coli*, **PDB ID: 1PII**. Molecular visualisation was done via PyMOL (DeLano 2002).

The PRAI enzymes from *E. coli* and *Saccharomyces cerevisiae* were also the targets of a number of pioneering protein engineering experiments undertaken by Luger and colleagues. The yeast enzyme was modified by circular permutation, duplication of the final two ($\beta\alpha$) units, and fragmentation into ($\beta\alpha$)₁₋₆ and ($\beta\alpha$)₇₋₈ substructures; *E. coli* PRAI was subjected to an internal duplication of the fifth ($\beta\alpha$) unit (Luger *et al.* 1989; Luger *et al.* 1990). A stable subdomain of PRAI, corresponding to ($\beta\alpha$)₁₋₅ β_6 of the ($\beta\alpha$)₈ barrel, has been described previously (Patrick & Blackburn 2005) and recently has had its structure solved by NMR spectroscopy (Fig. 2.2). The biophysical characterisation of this subdomain ('truncated PRAI', trPRAI) revealed the robustness of the ($\beta\alpha$)₈-barrel structure. One quarter of the barrel in trPRAI is missing, yet circular dichroism (CD) and tryptophan fluorescence data suggested that trPRAI (134 amino acids) retained the same proportion of α/β content (confirmed by the NMR structure) and that it was almost as thermostable as full - length PRAI (200 amino acids) (Setiyaputra *et al.* 2011).



Fig. 2.2. Ribbon diagram of trPRAI (Setiyaputra *et al.* 2011), PDB ID: 2KZH. Molecular visualisation was done via PyMOL (DeLano 2002).

In addition, trPRAI contains the two catalytic residues, Cys260 at the C - terminal end of β_1 and Asp379 at the C - terminal end of β_6 . However, it lacks the phosphate-binding motif located in the $\beta_7\alpha_7$ loop and helix 8' of

PRAI. This missing portion is shared with IGPS and also with a large number of other $(\beta\alpha)_8$ -barrel proteins (Setiyaputra *et al.* 2011; Patrick & Blackburn 2005).

Voltage-gated potassium (Kv) channels are integral membrane proteins that play a fundamental role in regulating the membrane potential and the frequency of action potential firing of excitable cells. Channels in the Kv1 family consist of hetero - oligomeric complexes comprising four pore - forming alpha (Kv α) subunits and four regulatory beta subunits (Kv β), the latter forming $(\beta\alpha)_8$ barrels. The Kv β 2 subunit structure in the complex coordinates a NADPH cofactor, predominantly through the interactions between the adenine dinucleotide moiety and α 7, α 8 and the first of the two α - helices that form a sizeable insertion into the β 7 α 7 loop of the barrel. The Kv β 2 subunit from *Rattus norvegicus* is an oxidoreductase and residues Asp85, Tyr90, Ser188 and Arg189 are considered to be important in orientating the cofactor (Gulbis *et al.* 2000). Examination of the structure of the Kv β 2 subunit showed that the C - terminal residues, which correspond to $(\beta\alpha)_{7-8}\alpha_6$ of the $(\beta\alpha)_8$ barrel in addition to the predominantly α - helical insertion into the β 7 α 7 loop, comprise the NADPH-binding subdomain (Fig. 2.3). It was postulated that this subdomain from Kv β 2 could complement the $(\beta\alpha)_1-5\beta_6$ subdomain of trPRAI.

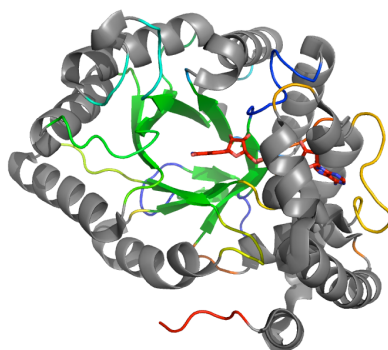


Fig. 2.3. Ribbon diagram of Kv β 2 (Gulbis *et al.* 2000), PDB ID: 1EXB. The bound NADPH cofactor is shown in ball-and-stick representation. Molecular visualisation was done using PyMOL (DeLano 2002).

In order to test the subdomain assembly model, it was thought that a $(\beta\alpha)_7$ - α_6 subdomain from a related barrel might complement trPRAI. A topologically similar yet evolutionarily distant barrel should be an ideal candidate for such complementation. A curiosity-driven experiment, using high - throughput methods from protein engineering to construct a chimera from random recombination to complement trPRAI and regenerate the folded $(\beta\alpha)_8$ barrel was designed. In this experiment, Kv β was not codon-optimised for expression in *E. coli*.

2.2.2 Protein chimeragenesis: an engineering approach to mimic non - homologous recombination

High - throughput tools to mimic non-homologous recombination and to identify novel folds were used in this study. As discussed earlier, many enzymes with little or even no sequence similarity can have similar structures, and shuffling the sequences of these proteins requires non-homologous recombination tools. In eukaryotic genomes the intron-exon makeup tends to enable non-homologous recombination by recombining genes with non-homologous introns, which maintains the functionality and accuracy of the exons or the coding region of DNA sequence (Kolkman & Stemmer 2001). In nature, this creates proteins that belong to distant families yet share conserved functional domains because of exon swaps between unrelated genes (Kolkman & Stemmer 2001; Ostermeier, Shim *et al.* 1999). Some of the methods for recombination of genes with limited or no sequence similarity that can be applied *in vitro* to mimic non-homologous recombination include exon shuffling, SHIPREC (sequence-homology independent protein recombination), DHR (degenerate homoduplex recombination), RM-PCR (random multi-recombinant PCR), YLBS (Y-ligation based shuffling) and ITCHY

(incremental truncation for the creation of hybrid enzymes). The latter method was chosen in this study and will be discussed in more detail.

Table. 2.1. Alternative approaches for non-homologous recombination. Five methods, including ITCHY, and their advantages and disadvantages are discussed.

| Method | Definition | Advantages | Disadvantages | References |
|-----------------------|--|--|--|---|
| Exon shuffling | In exon shuffling, exon fragments are amplified with chimeric primers (oligonucleotides), which splice the fragments into variable sizes. The fragments are then assembled by a primerless PCR (fragments prime against each other) to generate a shuffled library of chimeric sequences. | Maintains the functionality aspect of the exons. | Requires the knowledge of the intron-exon organisation of the gene to be shuffled. | (Kolkman & Stemmer 2001) |
| SHIPREC | In this method, two genes are joined with a linker that contains a unique restriction site. The fusion product is digested with DNase I (to form a chimeric library) followed by S1 nuclease (to produce blunt ends). The fragments are ligated (re-circularised) and re-linearised by digestion (restriction site within linker). | Crossover generally occurs at structurally related sites. | Only a single crossover per gene can be achieved. | (Sieber <i>et al.</i> 2001) |
| DHR | Using this method a polymorphic gene can be swapped randomly amongst a collection of polymorphic genes. It is quite a complicated method as it involves annealing, gap-filling and ligation steps. | Has a higher recombination rate with the added advantage of reduced recombination bias. | Requires synthesis of complementary oligonucleotides | (Coco <i>et al.</i> 2002) |
| RM-PCR | RM-PCR is a kind of exon shuffling where reassembly of fragments into full-length genes is done using overlap-extension PCR, with the primers containing the crossovers. | Incorporates variable size DNA fragments. | The chimeras can be longer or shorter than expected, with possible frame shifts. | (Tsuji <i>et al.</i> 2001) |
| YLBS | YLBS is a technique that involves block shuffling of DNA and subsequently of protein by using Y-ligation (ligation of blocks with a stem and two branches). | Can shuffle large exon regions. | There is often low product recovery and frame shifts may also occur. | (Nishigaki <i>et al.</i> 1998; Kitamura <i>et al.</i> 2002) |
| ITCHY | ITCHY is a method for creating a library of every one base truncation of dsDNA. | Does not require the structural knowledge of the genes (nor consequently the proteins) and also eliminates recombination bias. | Can only be applied to two parents and two-thirds of the library is out of frame. | (Ostermeier, Nixon, <i>et al.</i> 1999) |

DNA shuffling works efficiently between sequences with >70% sequence identity. Below this threshold, it generates libraries with crossovers that are biased towards regions of locally higher identity and reassembly of parental sequences becomes a dominant factor. The limitations of DNA shuffling and its dependence on sequence identity have led to the development of homology-independent methods to create hybrid libraries (Williams *et al.* 2004). One such tool is Incremental Truncation for the Creation of Hybrid enzYmes (ITCHY). Incremental truncation is a method for creating a library of every one base truncation of dsDNA. There are two ways to create an incremental truncation library; either by time-dependent digestion with the use of exonuclease (Exo III), or by the random incorporation of α -phosphorothioate dNTPs (α S-dNTPs). The latter method was used to create a novel fusion library of shuffled polypeptide fragments (Ostermeier, Nixon, *et al.* 1999).

ITCHY can be used to generate chimeras with a single crossover, or the protocol can be modified to allow multiple crossovers (Fig. 2.4). Single-crossover hybrids consist of the N - terminal section of one protein and the C-terminal section of another protein. With multiple crossovers, one or more internal stretches of amino acid sequence is/are replaced by the corresponding segment from another gene (Lutz *et al.* 2001). ITCHY was used to generate fusions of glycinamide ribonucleotide transformylases, which are encoded by *E. coli* (PurN) and its human counterpart (GART) genes. They share only 50% DNA sequence identity. The ITCHY library was created between the N-terminal of PurN and the C-terminal of GART (Ostermeier, Shim, *et al.* 1999). Selection of clones was made in an *E. coli* auxotroph that lacked GAR transformylase activity. In order to compare ITCHY with a DNA shuffling experiment, the N-terminal (PurN) and C-terminal (GART) fragments

were reassembled to create a shuffled hybrid library. It became evident from this experiment that clones from the ITCHY library had a wider crossover spectrum than those generated using the shuffling method. In terms of functional hybrids, the shuffled library contained more positive clones but was limited in diversity. Single crossover hybrids were found in the region where crossovers had high DNA homology. On the other hand, clones from the ITCHY library had crossovers in the non-homologous regions with more diverse fusion points. In fact, a PurN-GART hybrid from the ITCHY library had ~500-fold reduced activity compared to the wild-type PurN (Ostermeier, Shim *et al.* 1999; Michnick & Arnold 1999; Petrounia & Arnold 2000).

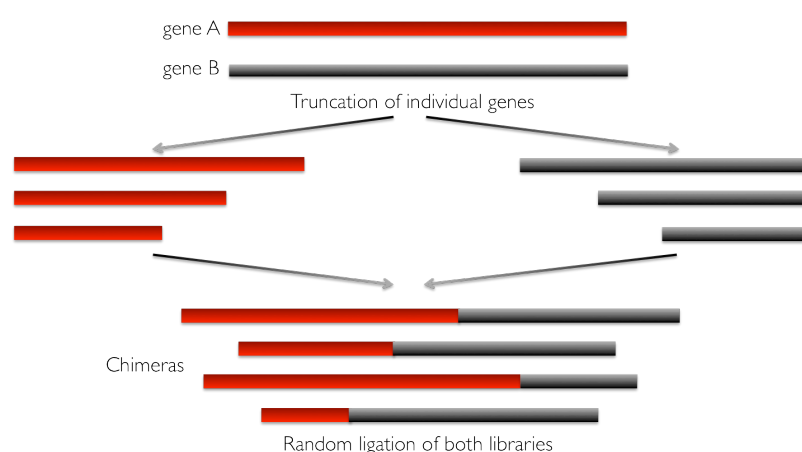


Fig. 2.4. The process of creating chimeras. Chimeras are created from two parental genes. The fusion of two incremental truncation libraries is called an ITCHY library. A single-crossover library consists of variants comprising different portions of the N-terminal region of one protein fused with the C-terminal region of the other.

Another interesting experiment involving ITCHY that generated hybrids with higher catalytic activity than either of their parental enzymes was the non-homologous recombination of human thymidine kinase 2 (hTK2) and deoxyribonucleoside kinases (dNKs) from *Drosophila melanogaster*. Chimeras from the ITCHY library were found to produce phosphorylate nucleoside

analogues with much higher activity than htk2 and dNK from *D. melanogaster* and also to have new activity towards the anti-HIV prodrug 2',3'-didehydro-3'-deoxythymidine (d4T) (Gerth & Lutz 2007). Overall, ITCHY has not been commonly used compared to DNA shuffling. It is, however, an effective method for exploring the origins of new folds.

Sampling interesting candidates from a large population of variants in a library is a crucial step in the process of making an ITCHY library. The use of *E. coli* as a genetic vehicle to carry the protein-of-interest (POI) is preferred by some biologists primarily due to the ease of genetic tinkering. The POI is expressed and ample quantities can be made for further downstream studies or applications (An *et al.* 2011). However, a common problem when expressing recombinant proteins is solubility. Often there are low levels of expression, or the formation of insoluble inclusion bodies. Solubilisation of the protein in inclusion bodies can be achieved by using strong denaturing conditions followed by *in vitro* refolding, but these techniques are often very limited in their approach to purify a soluble protein. An *in vivo* selection for folded proteins is highly desirable in both therapeutic and non-therapeutic applications (Batas *et al.* 1999). Protein solubility and foldability are two distinct attributes of a protein molecule. A protein's electrostatic charge is one of the crucial determinants for its solubility. A molecule carries a charge based on its amino acid sequence and the pH of the solvent it is dissolved in. At any given pH, charge of the functional groups of amino acid residues can vary (Trevino *et al.* 2007). For example, at neutral pH, aspartic acid and glutamic acid have a negative charge whereas lysine, arginine and histidine have positive charges. One of the factors that determines the net charge in a protein molecule is the pH of the solvent (Pace *et al.* 2009). At a specific pH, known as the isoelectric point, the net charge on a protein molecule is zero. So when the solvent is at pH 10 and 12, lysine and arginine respectively lose

their positive charge to become neutral, whereas in an acidic environment (pH 4.5) aspartic and glutamic acids lose their negative charge and become neutral (Pace *et al.* 2009; Trevino *et al.* 2007). The charge on the protein surface (negative or positive) attracts water molecules (water being dipolar) which form a layer around the protein molecule, which is responsible for the solubility of the molecule. It can also prevent like-charged protein molecules interacting with each other. On the other hand, an absence of net charge causes the protein molecules to interact with each other rather than with water (Trevino *et al.* 2007), making the protein molecule more prone to precipitation or aggregation. This is a very simplistic model used to help understand solubility (Sheinerman *et al.* 2000). Other factors such as hydration layers, charge distribution and polarisability can also influence the solubility of a protein molecule (Bostrom *et al.* 2005). Folding, on the other hand, is due to molecular interactions, which are responsible for the thermodynamic stability of the protein molecule, and include Van der Waals interactions, hydrophobic interactions, hydrogen bonding, charge-charge interactions and the formation of disulfide bonds (Onuchic & Wolynes 2004). The thermodynamic state of the molecule is a crucial factor in folding, because without the structure in its lowest energy conformation it will continue to move towards the most stable (lowest energy) state (Onuchic & Wolynes 2004).

Genetic selection systems that screen for solubility have been developed over the years. In some of these systems the POI is fused at the C-terminal to a reporter protein, which acts as a proxy for solubility, i.e. a misfolded POI may cause the reporter protein to misfold. This has its limitations as some reporter proteins stay active despite the POI being insoluble. Efforts have been made to bypass these proxy solubility screens and to engineer a

selection system utilising the cell's quality control mechanisms to export only correctly folded proteins across biological membranes (Choi & Lee 2004; Mansell *et al.* 2010).

Recently, Fisher and co-workers utilised the twin arginine translocase (Tat) pathway and its "folding quality control" feature to discover and engineer proteins that are both correctly folded and have increased solubility. In this *in vivo* screen, a target protein is fused with the Tat signal peptide and a reporter protein such as β -lactamase, which confers resistance to the antibiotic ampicillin/ β -lactams (Fig. 2.5). If the target protein or POI is not in-frame and folded, then this tripartite fusion is not compatible with Tat export, i.e. β -lactamase is not exported into the periplasm which renders the cells susceptible to the action of the antibiotic. Using this fold selection system alongside directed evolution experiments, the *in vivo* solubility of the aggregation-prone Alzheimer's A β 42 peptide and poorly folded single-chain Fv antibody fragment was improved (Fisher *et al.* 2006).

For this study, the pSAlect plasmid system was used to identify folded structures in an ITCHY library (Lutz *et al.* 2002). Cells containing the plasmid, with an ITCHY derived chimeric gene inserted, were grown on ampicillin-containing plates. Proteins that are in-frame and folded, i.e. competent for Tat export, will be transported to the periplasm (Fig. 2.5, right), and therefore only these clones will survive when plated on ampicillin.

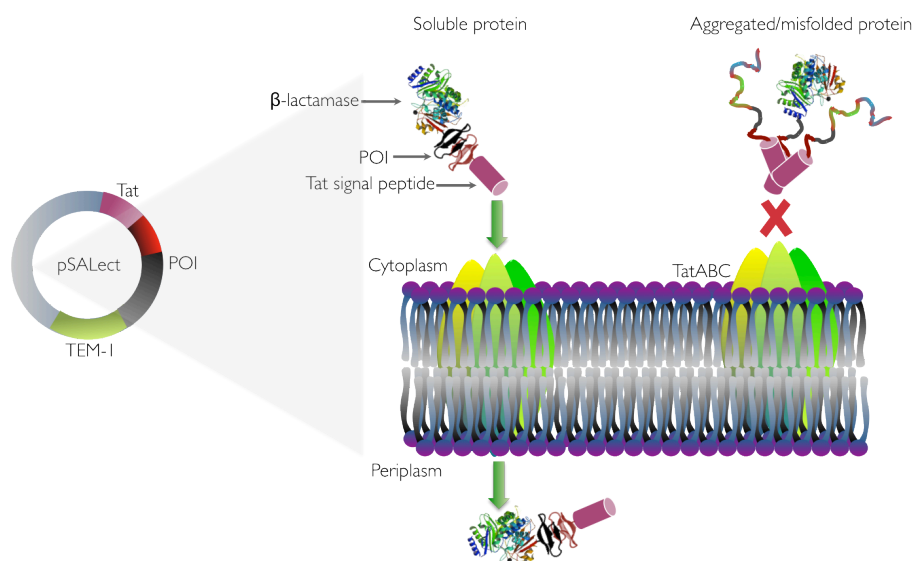


Fig. 2.5. The twin arginine translocation (Tat) pathway in *E. coli*. (Left) The plasmid pSAlect contains the POI (which could be ITCHY chimeras – represented by red and grey) cloned between the Tat signal peptide and the TEM-I β-lactamase. (Right) The quality control feature of Tat selects for folded proteins. The Tat signal peptide is fused to the N-terminus of the POI with the TEM-I β-lactamase at the C-terminus. POIs that are competent for Tat export (correctly folded) co-localise to the periplasm and render cells resistant to ampicillin. Proteins incapable of Tat export due to incorrect folding render cells sensitive to ampicillin.

For this study, the pSAlect plasmid system was used to identify folded structures in an ITCHY library (Lutz *et al.* 2002). Cells containing the plasmid, with an ITCHY derived chimeric gene inserted, were grown on ampicillin-containing plates. Proteins that are in-frame and folded, i.e. competent for Tat export, will be transported to the periplasm (Fig. 2.5, right), and therefore only these clones will survive when plated on ampicillin. This study was based on two prerequisites: 1) only proteins that appear in the soluble fraction of the cell lysate will be explored further (this was also based on the assumption that soluble proteins will be folded); and 2) that the pSAlect folding system is able to screen for folded proteins *in vitro*.

2.2.2.1 Making the THIO-ITCHY library

To explore the role of subdomain shuffling in fold evolution, THIO - ITCHY was used to randomly recombine Kv β 2 and trPRAI. This method involves the incorporation of α -phosphorothioate dNTPs, which are nucleotide analogues, during PCR amplification (Fig. 2.6) (Lutz, Ostermeier & Benkovic 2001).

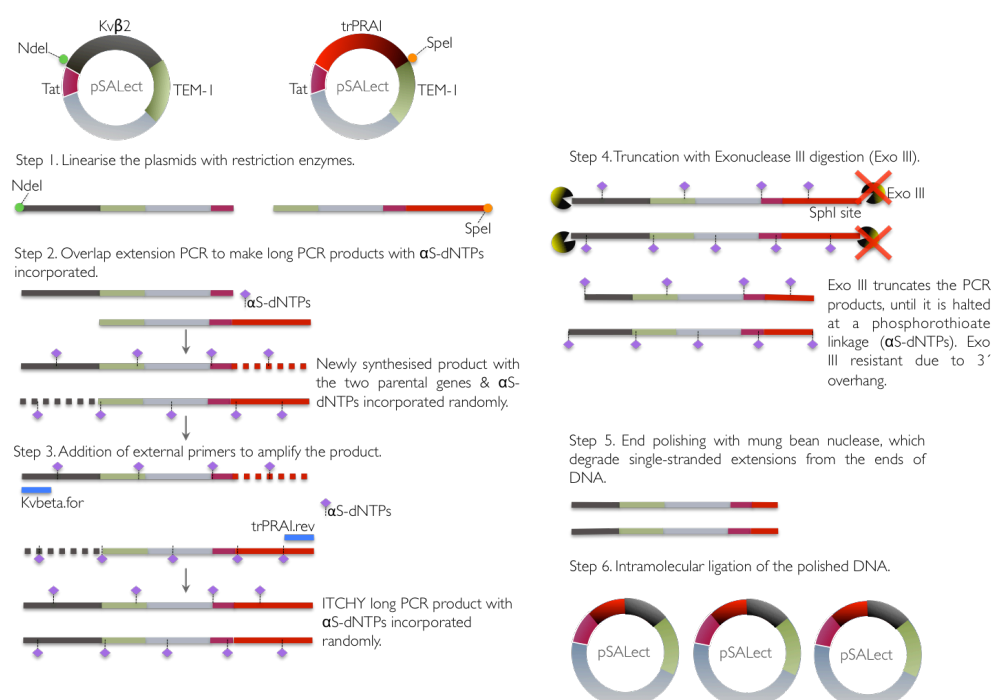


Fig. 2.6. The procedure used to generate Kv β 2 and trPRAI chimeras. (1) The pSALect plasmids with the two parental genes (Kv β 2 and trPRAI) are linearised with different enzymes (NdeI and SpeI). (2) and (3) The genes are recombined by overlap extension PCR (owing to the identical plasmid backbone) in the presence of α S-dNTPs. External primers, forward and reverse (Kv β .for and trPRAI.rev), are added to make more copies. (4) Exonuclease truncates the PCR products, until it is halted at a phosphorothioate linkage (from incorporation of α S-dNTP). Truncation via Exo III at 3' can be prevented using the SphI site. For more details refer to section 2.3.1.2. (5) and (6) End polishing and ligation yields a library of randomly recombined variants.

These nucleotide analogues protect the DNA from exonuclease III (Exo III) digestion, which leads to variation in truncation lengths. By adjusting the concentration ratio between dNTPs and α -phosphorothioate (α S)-dNTPs in the PCR reaction, the frequency of incorporation of phosphorothioate internucleotide linkages can be controlled. After PCR, in the presence of α S-dNTPs, the product is treated with Exo III, which hydrolyses the unmodified phosphodiester linkages but stops upon encountering a phosphorothioate linkage. Due to the random distribution of phosphothioates in the DNA, a truncation library upon digestion with Exo III is generated (Lutz, Ostermeier & Benkovic 2001).

2.3 Results

2.3.1 Library design

A stable subdomain of PRAI (trPRAI), corresponding to $(\beta\alpha)_1 - \beta_6$ of the $(\beta\alpha)_8$ barrel, was randomly recombined with the Kv β 2 subunit using ITCHY. To complement trPRAI with portions of Kv β 2, a single crossover and unidirectional library was generated. Unidirectional in this library means that the 3' (C-terminus) end of the long PCR product was protected from exonuclease digestion and that only the 5' (N-terminus) end (Kv β 2) was truncated. The first step in making an ITCHY library is to generate the long PCR product, which required many optimisations.

2.3.1.1 Making the long-fusion PCR product

The single-crossover trPRAI - Kv β 2 library was generated using the genes for trPRAI (402 bp) and full length Kv β 2 (996 bp). First, both genes were excised from their parent plasmids and sub-cloned into the plasmid pSALect, which contains the restriction sites *Nde*I and *Spe*I. Both genes were amplified using oligonucleotides with an *Nde*I site at the 5' end (N-terminus) and *Spe*I at the 3' end (C-terminus). Subcloning of the PCR products yielded pSALect - trPRAI and pSALect - Kv β 2 (see section 2.5.1). The plasmid pSALect - trPRAI was linearised by digestion with *Spe*I and pSALect - Kv β 2 with *Nde*I. The linearised plasmids were mixed and recombined in an overlap extension PCR by virtue of their identical backbone sequences. After several attempts to optimise the reaction, a protocol to amplify the PCR product was developed, the key parameter being ramp rate, the time it takes for the thermocycler to reach a certain temperature in the PCR reaction. The use of a shorter ramp rate

(0.4°C/s) in making a long PCR product ensured that spurious annealing was minimised. If the ramp speed of the thermocycler is low, it provides more time for non-specific binding (Kurata *et al.* 2004). Also, to avoid template depurination, the extension was done at 68°C instead of 72°C. Taq DNA polymerase was used for PCR amplification (Cheng *et al.* 1994). Thermocycling conditions were: 95°C for 2 min, three cycles of 94°C for 10 s and 68°C for 1 min, ramped at 0.4°C/s; the primers (Kvβ₂.for and TrPRAI_Nsil_rev) were added 5 s before the end of the third cycle, followed by 30 cycles of 94°C for 10 s, 58°C for 20 s, 68°C for 250 s and one cycle of 68°C for 5 min. The size of the long PCR product was 4.6 kb (pSALect backbone: 3192 bp; Kvβ₂: 996 bp; trPRAI: 402 bp) and this product was used for making the trPRAI - Kvβ₂ library. Two ratios of αS - dNTPs (1/8th and 1/10th of the total dNTP concentration) were used to monitor the frequency of incorporation with no - αS-dNTP and all - αS - dNTPs reactions as controls (see section 2.5.2).

2.3.1.2 Making a unidirectional library

To complement trPRAI with portions of Kvβ₂, the primary objective was to protect the 3' end of trPRAI from exonuclease digestion and achieve a unidirectional library in the following steps of the protocol (Fig. 2.7). Exo III is not active on single - stranded DNA, and thus 3' (C-terminus)-protruding termini are resistant to digestion. The degree of resistance depends on the length of the extension, with extensions of four bases or longer being essentially resistant to cleavage. The restriction enzyme *SphI* cleaves to leave a 3' CATG extension, which will be resistant to exonuclease treatment and is present just inside the *SpeI* site in the trPRAI gene fragment. Therefore, the PCR products from the overlap extension step were digested overnight with

the restriction endonuclease *SphI* (see section 2.5.2, step 3). This was followed by Exo III treatment to truncate the DNA from one direction, i.e. from the 5' end of the fusion product.

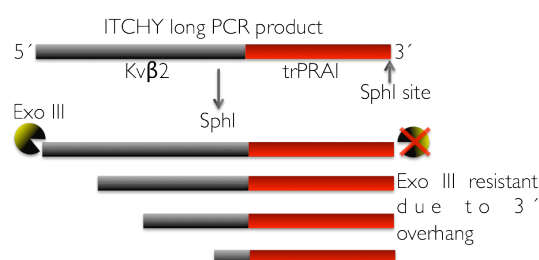


Fig. 2.7. Making the unidirectional library. In order to make a unidirectional library, the ITCHY long PCR product was digested with *SphI*, which cleaves to leave a 3' CATG extension or an overhang. Since Exo III is not active on single-stranded DNA, this overhang makes the 3' end or *trpF* coding trPRAI resistant to any truncation via Exo III.

After the Exo III step, the DNA was treated with mung bean nuclease to remove nucleotides from single-stranded DNA molecules, which might be overhanging after treating with Exo III. End polishing with mung bean nuclease was done to prepare blunt-ended DNA. After this step, the DNA was ready for size-selection, which was achieved by loading the DNA onto agarose gel and separating it by electrophoresis. A portion of the agarose gel was excised (in this case between 3500 and 4000 bp as anything outside this range will not have intact trPRAI or will have full length trPRAI+Kvβ2) and the DNA in the desired range was purified. This constituted the library DNA, which was then used to set-up intramolecular ligations, i.e. reclosure of the plasmid. The re-circularised plasmid DNA was then ready to transform *E. coli* DH5α-E cells (see section 2.5.2, step 6).

Electrocompetent *E. coli* DH5α-E cells were transformed by the re-circularised plasmids, and in total 11 transformations were pooled together

and plated on LB agar containing chloramphenicol. Clones from the resulting library were screened for inserts by means of colony PCR (see section 2.5.3).

Dilution plates provided an estimate that the library consisted of 52,000 variants (see section 2.5.4). Assuming that the 3' end of trPRAI was protected from exonuclease digestion, in making a unidirectional library, then the theoretical diversity of crossovers was equal to the length of the Kv β 2 gene: 996 bp. That is, 996 variants with every single Kv β 2 base truncated and fused with intact trPRAI. Therefore, the library would have over-sampled the total number of possible variants by 50-fold, ensuring that all possible trPRAI-Kv β 2 hybrids were represented. As crossovers can occur at any base, two thirds of all ITCHY library members will contain frameshifts, which result in premature termination or non-functional, frameshifted progeny. This meant that the library contained ~17,000 variants (one third of 52,000) that were in-frame. The likely selection for potentially folded chimeras, using the Tat system, was carried out on carbenicillin (a more stable derivative of ampicillin) plates, and approximately five to six times the library size was plated in order to ensure that every possible clone was represented (see section 2.5.5). In one such selection attempt, ~270,000 cells were plated and only 5100 clones passed the selection process, giving a survival percentage of 1.9%. In another attempt, only 1.3% survived. Together, these experiments suggested that there appears to be selection for both reading frame and perhaps folding, as desired.

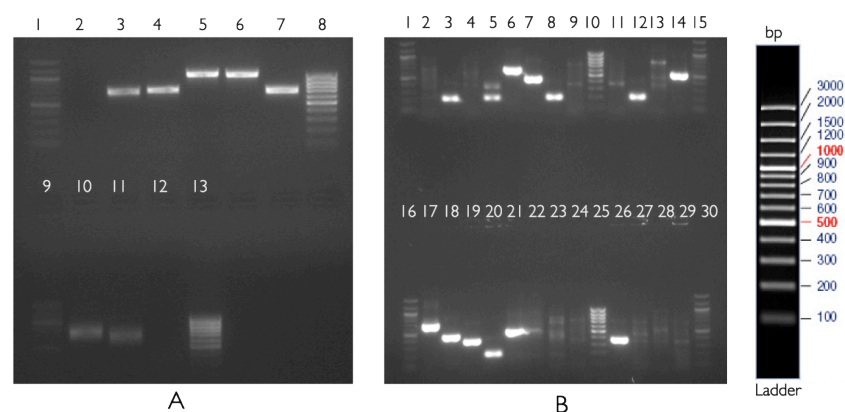


Fig. 2.8. An agarose gel showing products of the PCR screen from the pre-selection (A) and post-selection (B) libraries. Chimeric inserts (trPRAI-Kv β 2) of different sizes were analysed by sequencing with specific primers and the fusion points were plotted on a crossover plot (Fig. 2.9). (A) – lane 1, 8, 9 and 13: ladder; 2-7 and 10-12: inserts of various sizes ranging between 400 and 1000 bp. (B) – lane 1, 10, 15, 16, 25: ladder; All other lanes: inserts of variable sizes.

Clones were randomly picked from the pre-selection and post-selection libraries, and screened using primers that bound at the 5' and 3' ends of the chimeric variants (Fig. 2.8). The PCR screen of 30 clones from the pre-selection library and 28 clones from the post-selection library revealed inserts of different sizes. The location of the crossover between the parental genes in each variant was investigated by sequencing the inserts (see section 2.5.6) from the randomly selected colonies. The trPRAI and Kv β 2 fragment sizes and points where the crossover between the two parental genes occurred were plotted (Fig. 2.9).

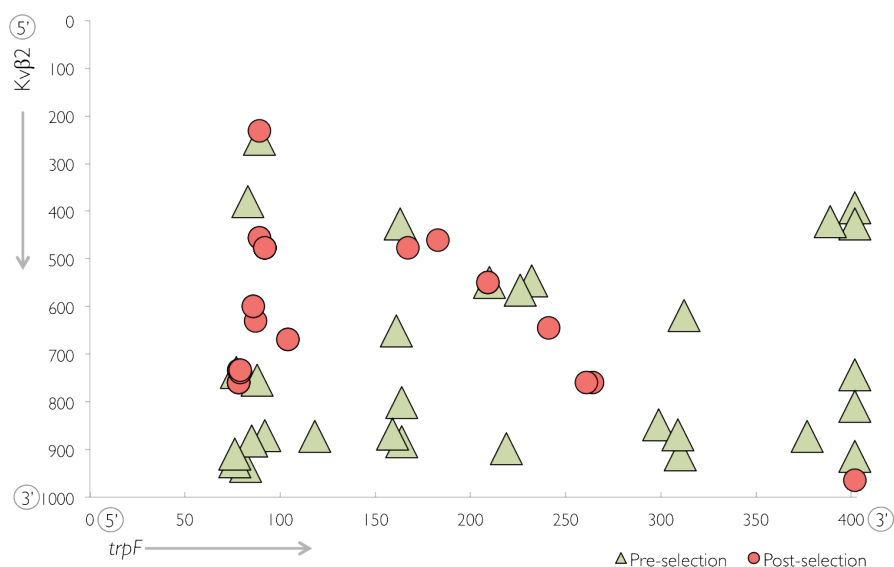


Fig. 2.9. Crossover plot. Distribution of 30 pre-selection library members (green triangles) and 28 post-selection members (red dots) in sequence space. On the X-axis is part of the *trpF* gene that codes for trPRAI and on the Y-axis is Kvβ2.

2.3.2 Crossover plots

Crossover plots (Fig. 2.9) are a way of representing, in a two-dimensional space, fusion points between two parental genes in an ITCHY library. On the x-axis is one parental gene, e.g. 'A', and on the y-axis is the other parental gene, e.g. 'B'. The coordinates (x, y) represent (last base of gene A, first base of gene B), i.e. if a crossover has the coordinate (97, 466), it means that the hybrid gene comprises the first 97 bp of gene A, fused to the fragment of gene B that runs from its 466th base to its 3' end.

2.3.3 The distribution of crossovers

Sequencing 30 clones from the pre - selection library revealed that only 17% of the sequences had trPRAI intact and fused to randomly sized fragments of Kv β 2, while the remaining clones showed truncations of both trPRAI and Kv β 2 (Fig. 2.9). This indicated that the attempt to protect the 3' end of trPRAI from exonuclease digestion by treating the PCR product with *Sph*I was only partially effective. Clones from the pre-selection library, which had trPRAI intact, were fused to variably sized fragments of Kv β 2, ranging from 80 to 600 bp. If these fusion chimeras were in-frame, they would have resulted in proteins constituting between 160 and 330 residues. However, only one of the 28 post-selection clones contained full-length trPRAI, and this was fused to a short portion of Kv β 2 (33 bp). While the sample size was small, this may suggest that folded chimeras can only result if trPRAI is fused to a small portion of Kv β 2, or that trPRAI must be truncated further in order to accommodate larger pieces of Kv β 2.

While the experiment was designed to protect trPRAI from exonuclease digestion, its undesired truncation did reveal some interesting patterns. In particular, 20 of the 28 post-selection clones contained a narrowly defined fragment of trPRAI (the first 75–102 bp), fused to 237–765 bp from the 3' end of the Kv β 2 gene. The high frequency with which truncation of trPRAI occurred between amino acids 25 and 34 indicates that this fragment may form a folded subdomain that can stabilise a variety of fusion partners (i.e. different sized fragments of Kv β 2). This cluster represents a β - α - β motif in PRAI (Fig. 2.10A). Although the remaining sequences cover a wide range of crossovers, they tend to be in between the range 100 - 260 bp of trPRAI,

which corresponds to slightly more than a quarter barrel fused with variable segments of the Kv β 2 gene.

2.3.4 Solubility validation

As the ultimate goal of the study was to search for folded structures, a solubility validation experiment was designed. Most of the clones in the post-selection library were either in the cluster that contained 75-102 bp or 160-260 bp of trPRAI. Three clones were picked from each of these two clusters. The cluster A clones were the chimeras P25K86, P28K132 and P29K122, where P stands for PRAI and K stands for Kv β 2. The numerals next to P and K indicate the number of residues contributed to the chimera from each protein. For example, P25K86 comprises 25 residues from the N-terminal of PRAI, fused to 86 residues from the C-terminal of Kv β 2. The three chimeras from cluster B were P55K173, P69K149 and P88K79. Cluster B clones contain 55-88 residues from PRAI. The N-termini of clusters A and B contain the β - α - β motif and $(\beta\alpha)_3\beta\Delta\alpha$ from PRAI respectively (Fig. 2.10).

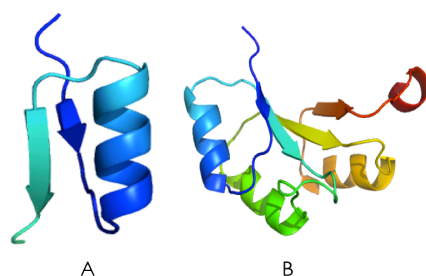


Fig. 2.10. Cartoon representation of the N-termini of chimeras from cluster A (left) and B (right) that contain a β - α - β (30 residue) motif and $(\beta\alpha)_3\beta\Delta\alpha$ (89 residue) from PRAI respectively. Molecular visualisation was done via PyMOL (DeLano 2002).

All six clones were sub-cloned in the expression vector pLAB101 and tested for solubility (see section 2.5.7). In order to test the candidates from both clusters, the proteins were over-expressed using IPTG induction, and fractions were run on SDS-PAGE gels.

2.3.4.1 Proteins from cluster A

1. Protein: P25K86

Status: **Soluble** (Fig. 2.11)

This chimera has a predicted molecular weight of 13,128 Da and contains 19% of trPRAI and 26% of the C-terminus of the Kv β 2. A 50 mL culture of *E. coli* DH5 α -E cells, harbouring pLAB101-P25K86, was over-expressed and the protein was His₍₆₎-tag purified using Talon™ resin with an IPTG-induced T7 expression system (see section 2.5.8). The expression conditions were the same for every chimera.

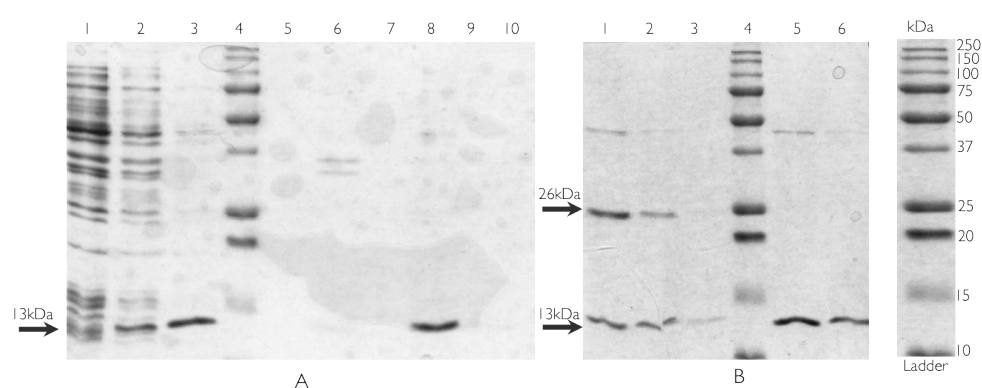


Fig. 2.11. (A) SDS-PAGE showing expression of P25K86. Lane 1: pre-induction; lane 2: induction of P25K86 after 2 h; lane 3: soluble lysate; lane 4: protein precision ladder (size shown on the right-hand side of the figure); lane 5: blank lane; lane 6: insoluble pellet; lane 7: blank lane; lane 8: unbound fraction (fraction that did not bind the resin); lane 9: resin wash; lane 10: column wash. (B) The purified protein (13 kDa) and a possible dimer (26 kDa) are indicated by arrows. Elution washes (E) of P25K86, with and without β -mercaptoethanol (BME). Lane 1: E1; lane 2: E2; lane 3: E3; lane 4: protein precision ladder; lane 5: E1+BME; lane 6: E2+BME. See section 2.5.8 for more details.

It can be seen from the SDS-PAGE (Fig. 2.11A, lane 3) that the protein appeared in the soluble fraction and no trace of the protein was found in the insoluble fraction. After IMAC chromatography, some protein remained in the unbound fraction. This was remedied by using batch methods in which the (His)₆-tagged protein in the soluble lysate was exposed to the resin for longer. The amount of protein purified was therefore increased by using batch methods with overnight exposure to the resin to bind the protein to the ion exchange resin, or by increasing the resin volume. SDS polyacrylamide gels of the eluted fractions also showed that P25K86 might be dimerising, as a band can be seen corresponding to ~26 kDa (Fig. 2.11B). On examining the sequence of P25K86, it was found that there are three cysteines (Cys7, 46 and 56), giving rise to the possibility of an intermolecular disulfide linkage (explained in more detail in Chapter 3, section 3.3.2). In order to confirm the presence of an oligomeric state in P25K86, the first two-elution fractions were mixed with β -mercaptoethanol (BME), a strong reducing agent, to reduce disulfide bonds, to a final concentration of 1.7 M, 10 μ L, of which was loaded on the gel. After treatment with BME, the 26 kDa band disappeared, indicating that indeed P25K86 could dimerise by the formation of an intermolecular disulfide bond.

2. Protein: P28K132

Status: **Partially soluble** (Fig. 2.12)

This chimera has a predicted molecular weight of 18,385 Da and contains 21% of trPRAI and 39% of the C-terminus of the Kv β 2.

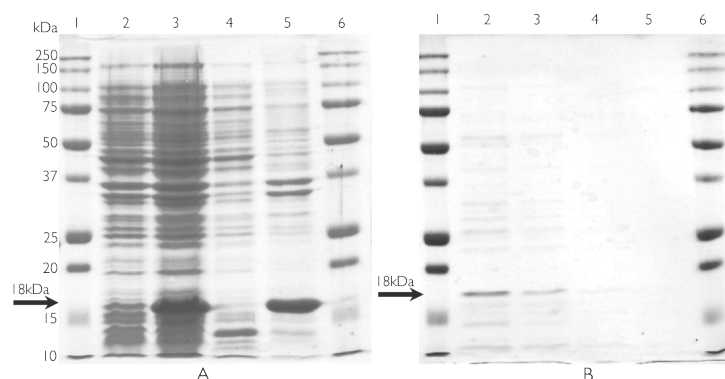


Fig. 2.12. Expression of P28K132. **(A)** The protein is indicated by an arrow. Lane 1: protein precision ladder; lane 2: pre-induction; lane 3: induction after 2 h; lane 4: soluble lysate; lane 5: insoluble pellet; lane 6: protein precision ladder. **(B)** Elution washes (E). Lane 1: protein precision ladder; lane 2: E1; lane 3: E2; lane 4: E3; lane 5: E4; lane 6: protein precision ladder. See section 2.5.8 for more details.

The protein was found to be partially soluble (Fig. 2.12) under the expression conditions described in section 2.5.8. As there was some soluble protein in the elution fractions (Fig. 2.12B), it suggested that yields of the protein might be able to be further optimised. However, given that the objective of the study was to search for folded and soluble proteins by using the Tat filtering machinery, it was decided not to optimise partially soluble or insoluble proteins but to select only those proteins that were well soluble. Therefore, this chimera was not studied further.

3. Protein: P29K122

Status: **Insoluble** (Fig. 2.13)

This fusion chimera has a molecular weight of 17,367 Da and contains ~22% of trPRAI gene and 37% of the C-terminus of the Kv β 2. Given its poor yields and appearance in the insoluble fraction (Fig. 2.13A, lane 4), this chimera was not studied further.

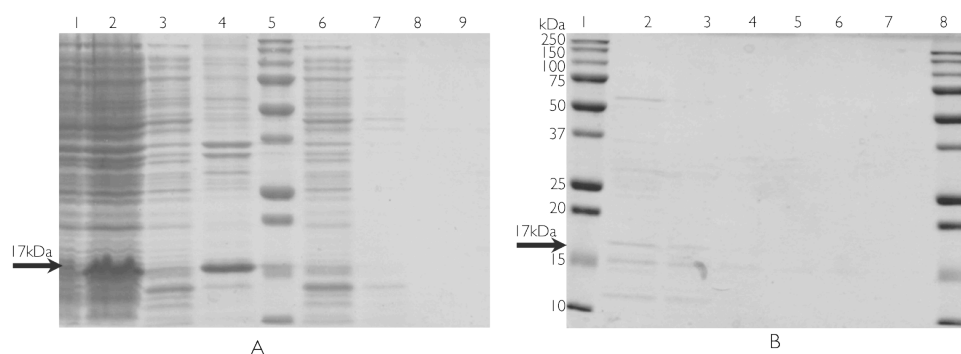


Fig. 2.13. Protein expression of P29K122. **(A)** The protein is indicated by an arrow. Lane 1: pre-induction; lane 2: induction after 2 h; lane 3: soluble lysate; lane 4: insoluble pellet; lane 5: protein precision ladder; lane 6: unbound fraction; lane 7: resin wash; lane 8: column wash-1; lane 9: column wash-2. **(B)** Elution washes (E). Lane 1: protein precision ladder, lane 2: E1; lane 3: E2; lane 4: E3; lane 5: E4; lane 6: E5; lane 7: E6, lane 8: protein precision ladder. See section 2.5.8 for more details.

2.3.4.2 Proteins from cluster B

1. Protein: P55K173

Status: **Insoluble** (Fig. 2.14)

This fusion chimera has a molecular weight of 26,174 Da and contains 41% of trPRAI and 52% of the C-terminus of the Kv β 2. The protein was insoluble and had very low yields. As P55K173 came in the insoluble fraction (see Fig. 2.14A, lane 5), no further analysis was carried out on this protein.

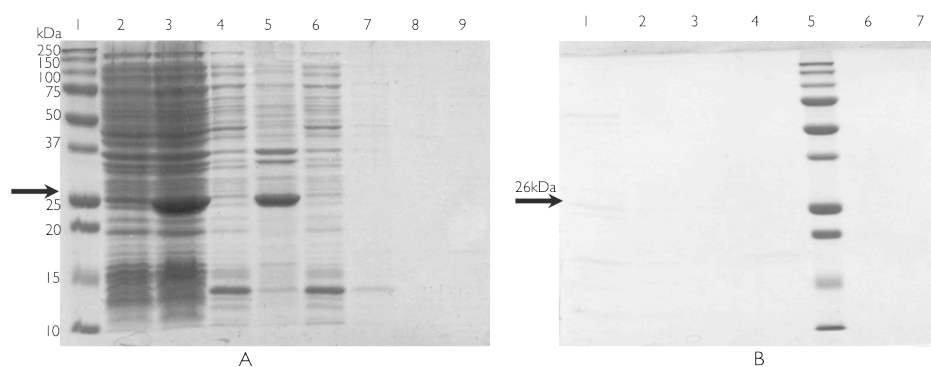


Fig. 2.14. Protein expression of P55K173. **(A)** The protein is indicated by an arrow. Lane 1: protein precision ladder; lane 2: pre-induction; lane 3: induction after 2 h; lane 4: soluble lysate; lane 5: insoluble pellet; lane 6: unbound fraction; lane 7: resin wash; lane 8: column wash-1; lane 9: column wash-2. **(B)** Elution washes (E). Lane 1: E1; lane 2: E2; lane 3: E3; lane 4: E4; lane 5: protein precision ladder; lane 6: E5; lane 7: E6. See section 2.5.8 for more details.

2. Protein: P69K149

Status: **Insoluble** (Fig. 2.15)

This fusion chimera has a molecular weight of 24,950 Da and contains 51% of trPRAI and 45% of the C-terminus of Kv β 2. No further analysis was done on this protein, as most of it was present in the insoluble fraction (see Fig. 2.15A, lane 4).

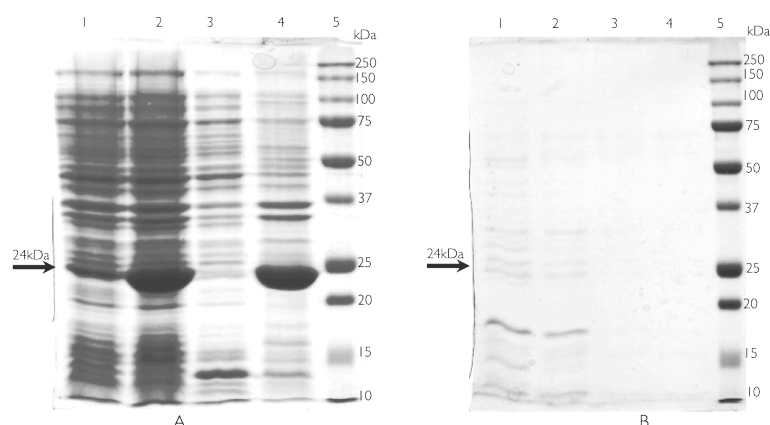


Fig. 2.15. Protein expression of P69K149. **(A)** The protein is indicated by an arrow. Lane 1: pre-induction; lane 2: induction after 2 h; lane 3: soluble lysate; lane 4: insoluble pellet; lane 5: protein precision ladder. **(B)** Elution washes (E). Lane 1: E1; lane 2: E2; lane 3: E3; lane 4: E4; lane 5: protein precision ladder. See section 2.5.8 for more details.

3. Protein: P88K79

Status: **Partially soluble** (Fig. 2.16)

This fusion chimera has a molecular weight of 19,023 Da and contains 66% of trPRAI and 24% of the C-terminus of Kv β 2. The protein was found to be partially soluble (Fig. 2.16A, lane 3) and it also appeared in the elution fractions (Fig. 2.16B, faint band).

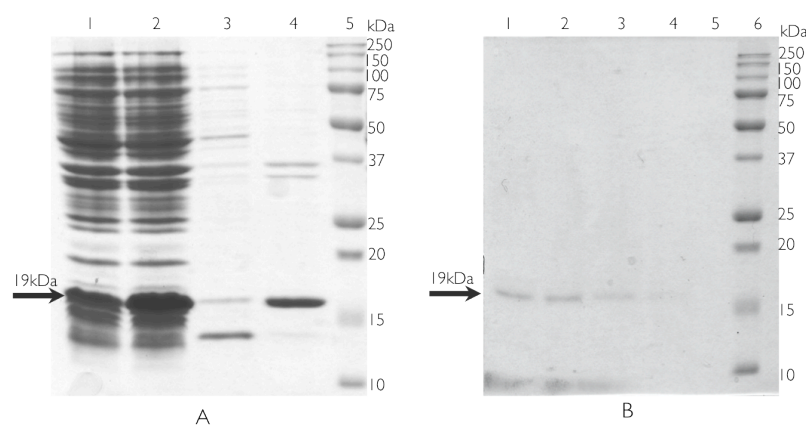


Fig. 2.16. Protein expression of P88K79. **(A)** The protein is indicated by an arrow. Lane 1: pre-induction; lane 2: induction after 2 h; lane 3: soluble lysate; lane 4: insoluble pellet; lane 5: protein precision ladder. **(B)** Elution washes (E). Lane 1: E1; lane 2: E2; lane 3: E3; lane 4: E4; lane 5: blank lane; lane 6: protein precision ladder. See section 2.5.8 for more details.

From the solubility validation experiment, it was found that from cluster A, protein P25K86 was soluble and protein P28K132 was partially soluble, whereas from cluster B, only P88K79 was found to be partially soluble. However, both P28K132 and P88K79 started to precipitate when the proteins were concentrated and the buffer was exchanged for further biophysical experiments. It is to be noted that at this stage of the study, selection of chimeras for further experiments was based solely on whether the protein produced by the heterologous host was soluble or insoluble. It was simply a qualitative experiment, designed to narrow down the selection of soluble proteins for further biophysical experiments, and protein concentrations were not measured. It was concluded, therefore, that only one of the six proteins (P25K86) created using the pSALect selection system was a true positive. The large proportion of false positives suggested that perhaps the Tat fold selection system has its limitations or that it may have been bypassing the proteins through an alternative Sec pathway, which can translocate unfolded proteins. This idea is explored further in Chapter 4.

2.4 Discussion

2.4.1 pSALect is not robust enough

The search for folded proteins from the trPRAI-Kv β 2 ITCHY library revealed candidates that were mostly insoluble. The selection for potentially folded and in-frame chimeras in the trPRAI-Kv β 2 ITCHY library was made using the pSALect system. This system utilises the Tat export machinery to translocate only fully folded polypeptides. Testing this system to see whether it was suitable for this study was one of the aims. If the screening capability (i.e. to select folded proteins) of this system held true, then it has the potential to be applied on a whole-genome-wide scale to search for novel folded proteins. This idea is discussed in greater detail in Chapter 5.

The low survival rate of 1.3% observed in selection experiments suggested that pSALect was effectively identifying rare and folded chimeras. However, only one (P25K86) of the six clones that were validated for solubility was found to be a true positive. It thus appeared that the selection system was leaky, i.e. reporting false positives as true positives, and could be improved. The translocation of many proteins from the cytoplasm to the periplasm is *via* the general secretory (Sec) pathway, which exports a newly synthesised polypeptide in a largely unfolded state. What makes the Tat pathway different from Sec is that Tat only exports pre-folded proteins across the inner membrane. The targeting of a protein to a given pathway is done by signature signal peptides that differ in the hydrophobicity of what is known as the “H-region” of the peptides. Generally, the Sec-specific peptides are more hydrophobic than their Tat counterparts (Bagos *et al.* 2010). However, recently it has been found that the majority of the signal peptides are not

completely specific, and that they can route their substrates by either of the two pathways (DeLisa et al. 2003). It has been shown that increasing the positive charge in the N-terminus of the mature protein blocks export by Sec, without hindering Tat export (Tullman-Ercek *et al.* 2007). As Sec can translocate proteins in an unfolded state, it implies that the β -lactamase domains of the false positives are folding in the periplasm (to give carbenicillin resistance). This hypothesis will be revisited in Chapter 4.

It is plausible, therefore, that the false positives observed here arose because some chimeras were routed via the Sec pathway instead of Tat. This would compromise the folding quality control feature of the pSAlect system. Another possibility was the pore size of the Tat translocase machinery. However, this is unlikely, because the upper size limit for Tat-mediated translocation has been estimated between 160 and 180 kDa and all the chimeras that were tested for solubility were well below this threshold (Strauch & Georgiou 2007). The shape of the protein (cargo) may also affect the translocation, i.e. only tightly folded shapes are translocated as opposed to long strings (Palmer & Berks 2003).

2.4.2 The outcome of random recombination

The attempt to make a library by randomly recombining two distantly related TIM barrels revealed clustering of chimeric candidates in sequence space. It was found that most of the selected clones contained either 75-102 bp ('cluster A') or 160-260 bp ('cluster B') of *trpF*. In particular, the clones in cluster A are packed closely in terms of the dominant N-terminal half of these chimeras. Nearly 70% of the 28 clones sequenced from the post-selection library were present in this cluster. These represent clones with a

β - α - β motif. The dominance of this motif in cluster A suggests a “triggering” mechanism, whereby partially structured units nucleate folding in proteins. The theory proposed by Fersht (1995) suggests the formation of aggregates of secondary structure motifs or “foldons”, which are simultaneously stabilised by tertiary interactions that strengthen as the structure extends. This facilitates the folding of the protein polymer (Fersht 1995; Salem *et al.* 1999). A possible explanation then could be that the β - α - β motif from trPRAI plays an important role in triggering folding in these chimeric variants.

In addition, this subdomain might be crucial in giving some shape or a partial structure to these chimeras to get them to translocate *via* the Tat pathway. However, it may also just be a “Tat-friendly” polypeptide. There appears to be a size threshold in accommodating fragments from both genes in order to be selected as a folded chimera *via* the pSALect selection. Sequences from the library indicated that as the size of fragments from trPRAI increased, fragment size from Kv β 2 decreased, and vice versa. This might suggest some favourable configurations of subdomains in play for rapid and convenient folding, along with topological constraints to accommodate fragments in the pSALect fold selection system. What is remarkable in this study is the clustering of folded recombinants, despite a decent spectrum of crossovers in the library, thus hinting at hotspots in folding space. The only soluble protein amongst the six that were tested for solubility was P25K86. The presence of an oligomeric state could be a result of an SS-bond or disulfide link. This is quite obvious as there are three cysteines (Cys7, 46 and 56) present in P25K86 and any one of them could be facilitating the linkage. Despite the small sample size for solubility screens, one soluble chimera was discovered in this proof-of-concept study. Although the production of more soluble proteins would have been desirable, this experiment highlighted

some key points, including: limitations in the Tat based fold selection system; probable bias in selection (clustering of chimeras, post-selection); and that ITCHY can be applied to randomly recombine non-homologous genes.

In this chapter, a method that mimics a non-homologous recombination event was employed to search for novel folds, and six candidates were tested for solubility. Of these, only P25K86 is soluble both in its oligomeric and monomeric state. The next chapter describes a full biophysical characterisation of this protein.

2.5 Materials and methods

All reagents were purchased from Sigma-Aldrich unless stated otherwise. Common molecular biology materials, techniques and primer sequences are described in Appendix I. Specific materials and methods used in this chapter are described in the following subsections.

Note: The *trpF* gene codes for PRAI and a portion of this gene (402 bp) that codes trPRAI was used in this experiment. For simplicity, the notation trPRAI is used to represent the portion of the *trpF* gene when plasmids were made.

2.5.1 Construction of pSALect-trPRAI and pSALect-Kv β 2

In order to construct trPRAI-Kv β 2 ITCHY library, the parent genes were subcloned into the pSALect plasmid (Lutz *et al.* 2002). *E. coli* strains harbouring pDEP-Kv β 2 and pDEP-trPRAI were cultured overnight in LB containing carbencillin (Carb; 100 μ g/mL). *E. coli* harbouring pSALect-PRAI were cultured in LB supplemented with chloramphenicol (Cam; 34 μ g/mL). Plasmid DNA was prepared from each overnight culture using the QIAGEN Plasmid Mini Kit. DNA was eluted from each Qiagen column in 40 μ L of elution buffer (10 mM Tris-Cl, pH 8.5). The total yield of each plasmid ranged from 3 μ g to 4 μ g. Each plasmid (3 μ g) was digested with 20 U of *Nde*I and *Spe*I (NEB), 1 \times BSA, 1 \times NEBuffer 4 in a total reaction volume of 40 μ L. After incubation at 37°C for 6 h, the restriction enzymes were inactivated at 80°C for 20 min. Following this, the samples were loaded on a 1% agarose gel and electrophoresed at 70-90 v for 30-40 min in 1 \times TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 8.0). The bands corresponding to the two inserts (trPRAI, 402 bp; and Kv β 2, 996 bp) and the pSALect vector backbone

(3192 bp) were excised and the DNA recovered with the Qiagen QIAQuick Gel Extraction Kit. DNA was eluted from each column in 30 μ L elution buffer. Ligation reactions to make pSALect-Kv β 2 and pSALect-trPRAI were performed using a three-fold molar excess of insert DNA (48 ng of Kv β 2 and 19 ng of trPRAI) over vector DNA (52 ng), 1 \times Quick Ligation Buffer (NEB) and 2000 unit of the Quick Ligase (NEB) to a final volume of 21 μ L. After 15 min of incubation at 25°C, the reactions were purified using the QIAQuick PCR Purification Kit (Qiagen). DNA was recovered from each column with 30 μ L elution buffer. Aliquots (1.5 μ L) of the purified ligation products were used to transform electrocompetent *E. coli* DH5 α -E (Invitrogen) by electroporation, using a Bio-Rad Gene Pulser II (see Appendix I for electrocompetent cells). The parameters used were 2.5 kV, 25 μ F, 200 Ω and cuvettes with 0.2 cm gaps (Bio-Rad). The transformed cells were spread on LB-chloramphenicol (34 μ g/mL) agar plates and the number of colonies formed was counted after incubation (37°C, 16 h). The colonies were screened for the presence of the desired inserts by colony PCR with one primer specific to the pSALect backbone (pSALect.for; Appendix I), and another specific to the 3' ends of the inserts. The reverse primers trPRAI_Nsil_rev and Kv β .rev were used for trPRAI and Kv β 2, respectively. The verified clones were stored at -80°C for future use (see Appendix I).

2.5.2 ITCHY library construction

The *E. coli* strains harbouring pSALect-trPRAI and pSALect-Kv β 2 were each used to inoculate 5 mL of LB-chloramphenicol medium. After overnight incubation at 37°C, the plasmids were extracted using the Qiagen Plasmid Mini Kit (Qiagen). Purified DNA was eluted from each column with 50 μ L elution buffer.

STEP 1:

Plasmids pSAlect-trPRAI and pSAlect-Kv β 2 were linearised with *SpeI* and *NdeI* respectively (Table 1).

Table 1: Linearisation

| Reagent | pSAlect-Kv β 2 | pSAlect-trPRAI |
|-----------------------------|--------------------------|-------------------------|
| NEB buffer 4 | 4 μ L (1 \times) | 4 μ L (1 \times) |
| Plasmid | 34 μ L (1.7 μ g) | 30 μ L (3 μ g) |
| <i>NdeI</i> (20 U/ μ L) | 2 μ L (40 U) | -- |
| <i>SpeI</i> (10 U/ μ L) | -- | 2 μ L (20 U) |
| 10 \times BSA | -- | 4 μ L (1 \times) |
| Total volume | 40 μ L | 40 μ L |

This was followed by incubation at 37°C for 4 h, then the reaction products were loaded on a 0.8% DNA agarose gel. After electrophoresis, each product band (corresponding in size to the linearised plasmid) was excised and recovered using the QIAQuick Gel Extraction Kit (Qiagen) and eluted in 30 μ L of elution buffer (EB).

STEP 2:

An overlap extension PCR was set up to recombine the two-linearised plasmids (Fig. 2.6) under robust conditions (Taq polymerase - i-Taq from iNtRON). It is important to note is that the primers were added after the first three cycles of the PCR (indicated with green in Tables 2 and 3).

Table 2: Reaction mixture for overlap extension PCR.

| Reagent | Volume |
|---------------------------------------|----------------------------|
| H ₂ O | 34 μ L |
| 10 \times Taq buffer | 5 μ L (1 \times) |
| 2.5 mM dNTPs | 6 μ L (0.3 mM) |
| Template (kv β -20 ng/ μ L) | 1 μ L (20 ng) |
| Template (trPRAI-20 ng/ μ L) | 1 μ L (20 ng) |
| Taq polymerase (5 U/ μ L) | 1 μ L (5U) |
| DMSO | 1.5 μ L |
| Kv β .for (100 μ M) | 0.25 μ L (0.5 μ M) |
| trPRAI_Nsil_rev (100 μ M) | 0.25 μ L (0.5 μ M) |
| Total volume | 50 μ L |

Table 3: PCR cycling parameters.

| | | |
|----|--------|----------|
| 1 | 95°C | 2:00 min |
| 2 | 94°C | 0:10 s |
| 3 | 68°C | 1:00 min |
| | Ramp @ | 0.4°C/s |
| 4 | Goto 2 | 3X |
| 5 | 94°C | 0:10 s |
| 6 | 58°C | 0:20 s |
| 7 | 68°C | 4:10 min |
| 8 | Goto 5 | 29X |
| 9 | 68°C | 5:00 min |
| 10 | 4°C | Hold |

The conditions used in the protocol yielded long PCR products of the correct size (4.6 kb). The next step in ITCHY library construction was to repeat the long PCR protocol in the presence of phosphorothioate dNTPs (α S-dNTPs), as described by Lutz *et al.* (2001). Four ITCHY PCRs were set up, as listed in Table 4.

Table 4: Reaction mixture for long PCR in the presence of α S-dNTPs.

| Reagent | No α S-dNTPs | 1/8 α S-dNTPs | 1/10 α S-dNTPs | All α S-dNTPs |
|-------------------------------|----------------------------|-----------------------------|----------------------------|----------------------------|
| H ₂ O | 34 μ L | 30.1 μ L | 30.8 μ L | 2.5 μ L |
| 2.5 mM dNTPs | 6 μ L (300 μ M) | 5.2 μ L (262.5 μ M) | 5.4 μ L (270 μ M) | 0 |
| 10 \times Taq buffer | 5 μ L (1 \times) | 5 μ L (1 \times) | 5 μ L (1 \times) | 5 μ L (1 \times) |
| Linearised pSALect-Kv β | 1 μ L (20ng) | 1 μ L (20ng) | 1 μ L (20 ng/ μ L) | 1 μ L (20 ng/ μ L) |
| Linearised pSALect-trPRAI | 1 μ L (20ng) | 1 μ L (20ng) | 1 μ L (20 ng/ μ L) | 1 μ L (20 ng/ μ L) |
| DMSO | 1.5 μ L | 1.5 μ L | 1.5 μ L | 1.5 μ L |
| 0.4mM α S-dNTPs | 0 μ L | 4.7 μ L (37.5 μ M) | 3.8 μ L (30 μ M) | 37.5 μ L (300 μ M) |
| Intron Taq (5 U/ μ L) | 1 μ L (5U) | 1 μ L (5U) | 1 μ L (5U) | 1 μ L (5U) |
| Kv β .for | 0.25 μ L (0.5 μ M) | 0.25 μ L (0.5 μ M) | 0.25 μ L (0.5 μ M) | 0.25 μ L (0.5 μ M) |
| trPRAI_Nsil_rev | 0.25 μ L (0.5 μ M) | 0.25 μ L (0.5 μ M) | 0.25 μ L (0.5 μ M) | 0.25 μ L (0.5 μ M) |
| Total volume | 50 μ L | 50 μ L | 50 μ L | 50 μ L |

The PCR products were purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from each column in 35 μ L EB buffer. The concentration of each purified product was estimated spectrophotometrically (ACTGene).

STEP 3:

In order to protect trPRAI from exonuclease digestion (to make a unidirectional library in the ITCHY protocol), the PCR products were digested (Table 5) with *SphI* (NEB). The reactions were incubated at 37°C for 16 h. After this, 20 units of *DpnI* (NEB) were added to each reaction and the incubation

was continued for further 1 h, before the enzymes were inactivated by heating to 80°C for 20 min.

Table 5: Treatment with *SphI*.

| Reagent | No αS-dNTPs | 1/8 αS-dNTPs | 1/10 αS-dNTPs |
|-----------------------|------------------|------------------|------------------|
| NEB buffer 4 (10x) | 3 µL (1x) | 3 µL (1x) | 3 µL (1x) |
| <i>SphI</i> (10 U/µL) | 1.5 µL (15U) | 1.5 µL (15U) | 1.5 µL (15U) |
| DNA template | 25.5 µL (7.4 µg) | 25.5 µL (5.2 µg) | 25.5 µL (6.6 µg) |
| Total volume | 30 µL | 30 µL | 30 µL |

The PCR products, followed by *SphI* digestion, were purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from each column in 40 µL EB buffer.

STEP 4:

Next, the DNA was treated with Exo III (NEB). For the exonuclease step, ~120 U of enzyme was used per microgram of DNA.

Table 6: Treatment with Exo III.

| Reagent | No αS-dNTPs | 1/8 αS-dNTPs | 1/10 αS-dNTPs |
|--------------------|----------------|----------------|----------------|
| NEB buffer 1 (10x) | 4 µL (1x) | 4 µL (1x) | 4 µL (1x) |
| Exo III (100 U/µL) | 4 µL | 4 µL | 4 µL |
| DNA template | 32 µL (4.2 µg) | 32 µL (2.9 µg) | 32 µL (3.5 µg) |
| Total volume | 40 µL | 40 µL | 40 µL |

The reactions were digested with Exo III and the reaction mixture was incubated at 37°C for 30 min. It is important to thoroughly mix the reaction components upon addition of Exo III. The PCR products were purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from each column in 40 µL EB buffer.

STEP 5:

The next step was mung bean nuclease treatment. Into the reaction, ~6 units of mung bean nuclease per microgram of DNA were added. After briefly

centrifuging, the samples were incubated at 30°C for 30 min. The PCR products were purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from each column in 40 µL EB buffer.

Table 7: Treatment with Mung bean nuclease.

| Reagent | 1/8 αS-dNTPs | 1/10 αS-dNTPs |
|---------------------------------|----------------|----------------|
| Mung bean nuclease Buffer (10x) | 4.2 µL (1x) | 4.2 µL (1x) |
| Mung bean nuclease (10 U/µL) | 0.8 µL (8 U) | 0.8 µL (8 U) |
| DNA template | 37 µL (1.3 µg) | 37 µL (1.2 µg) |
| Total | 42 µL | 42 µL |

Note: *The No αS-dNTPs sample was discarded as the DNA was completely digested, as expected.*

STEP 6:

To polish the ends of the randomly truncated DNA molecules prior to their intramolecular ligation, the two samples (1/8 αS-dNTPS and 1/10 αS-dNTPS) were pooled together and treated with T4 DNA polymerase.

Table 8: Treatment with T4 DNA polymerase.

| Reagent | Volume |
|---------------------------|---------------|
| NEB buffer 2 (10x) | 10 µL (1x) |
| dNTP (2.5mM) | 4 µL (100 µM) |
| T4 DNA polymerase (3U/µL) | 0.5 µL (1.5U) |
| DNA | 80 µL (800ng) |
| Water | 5.5 µL |
| Total | 100µL |

The reaction was incubated at 12°C for 20 min. After 20 min, the reaction was stopped by adding 1.9 µL of water and 2.1 µL of 0.5 M EDTA, to a final concentration of 10 mM EDTA, and incubated at 75°C for 20 min.

Next, the reaction mixture was loaded on 1% agarose gel and electrophoresed at 50 v. DNA in the size range 3500-4000 bp was excised from the gel. It was recovered from the gel using the QIAQuick Gel Extraction

Kit (Qiagen) and the DNA was eluted from the column with 50 μ L of EB. The total yield of randomly truncated DNA was 150 ng.

Two ligation reactions were set up, with each containing 20 μ L of DNA (60 ng), 3 μ L of 10x T4 ligase buffer (NEB), 2 μ L of T4 DNA ligase (10 Weiss unit of T4 DNA ligase; Fermentas) and 5 μ L of water. The ligation reactions were incubated at 16°C for 16 h and then purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted in 30 μ L EB (the ligation reaction, prior to purification was pooled together).

For making a test library, 5 μ L of the purified, self-circularised DNA was used to transform a 50 μ L aliquot of electrocompetent *E. coli* DH5 α -E cells. The cells were recovered in 500 μ L SOC and incubated at 37°C for 1 h. Three different volumes (10, 50 and 100 μ L) were spread on LB agar containing chloramphenicol (LB-Cam34 μ g/mL) plates and incubated overnight at 37°C.

2.5.3 Screening the clones

Randomly chosen clones from the test library and the scaled-up library were screened for the presence of hybrid trPRAI-Kv β 2 inserts by means of colony PCR. A single colony was picked, resuspended in 10 μ L of water, and heated at 95°C for 5 min. A 1 μ L aliquot of the lysed cells was used as the template for colony PCR, in a reaction with 1x GoTaq buffer (Promega), 0.25 mM of each dNTP, 5 μ M of each primer (PRAI.for and Kv β .rev) and 2.5 units of Taq polymerase (iNtRON), in a total volume of 20 μ L. The thermocycling conditions were: 94°C for 2 min; 30 cycles of 94°C for 10 s, 58°C for 20 s, 72°C for 70 s; and one final cycle of 72°C for 5 min. The PCR products were loaded on a 1% agarose gel and electrophoresed at 90 v for 30-40 min in 1×

TAE buffer. After this the PCR products with the inserts (chimeras of variable sizes) were sent for sequencing (see section 2.5.6).

2.5.4 Scaling up and harvesting the library

To construct the big library, electrocompetent *E. coli* DH5 α -E cells were transformed by the re-circularised plasmids (1 μ L) and in total 11 transformations were pooled together and plated on LB agar containing chloramphenicol (LB-Cam34 μ g/mL). Pre-selection was done on LB agar containing chloramphenicol (Cam-34 μ g/mL), while fold selection was done on carbenicillin (Carb-100 μ g/mL) containing plates. The colonies from big plates were picked and screened for inserts as discussed in section 2.5.3. To make -80°C freezer stocks of the entire pre-selection library for future use, the clones were scraped off with a glass spreader using 20 mL of LB chloramphenicol liquid media and harvested. After harvesting, the cells were pelleted by centrifuging at 3000 g, at 4°C for 15 min and the supernatant removed. The pelleted cells were then resuspended in a smaller volume of LB chloramphenicol liquid media and aliquoted for freezer stocks. While making freezer stocks of the pre-selection library, a 1000-fold dilution of the resuspended library was performed and its optical density at 600 nm was measured to record the number of cells present in the library.

2.5.5 Folding selection

For the selection of folded chimeras, an aliquot of the frozen pre-selection library was used to inoculate 40 mL of LB-Cam (34 μ g/mL) liquid media and grown at 37°C until it reached an optical density (OD) of 0.4. The optical density was measured using the BioPhotometer (Eppendorf). Cells that

represented 6 times the pre-selection library coverage were then plated on LB agar containing carbenicillin (100 µg/mL). The plates were then incubated at 28°C for 24 h. Random clones from the selection plate were picked for sequencing in order to determine the spectrum of crossover locations in the trPRAI-Kvβ2 library.

2.5.6 Sequencing

Randomly chosen clones from the pre-selection library, and those that survived the folding selection, were sequenced by the Massey Genome Service, Massey University, Palmerston North. The primer used for sequencing was PRAI.for (see Appendix I).

2.5.7 Construction of the expression vector pLAB101

The expression vector pLAB101 was constructed by modifying pMS401 (Patrick & Blackburn, 2005) to remove an *Nde*I restriction site from the backbone, and then to introduce *Nde*I and *Spe*I sites into the multiple cloning cassette. Plasmids were extracted from overnight cultures of *E. coli* cells harbouring pMS401, grown in LB containing ampicillin (100 µg/mL). To eliminate the *Nde*I site from pMS401, an inverse PCR was set up with 1x Phusion HF buffer, 0.2 mM dNTPs, 5 µM primer PMS_Nde_EL.for, 5 µM PMS_Nde_EL.rev, 10 ng of template pMS401, and 1 unit of Phusion polymerase in a final volume of 50 µL. The primers used were phosphorylated at their 5' ends. The PCR cycling conditions were: 98°C for 30 s; 30 cycles of 98°C for 10 s, 68°C for 20 s, 72°C for 70 s; and then one final cycle of 72°C for 5 min. The PCR product was 4944 bp. It was purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from the column with 40 µL EB. The purified product was then treated with *Dpn*I (NEB) to

digest any of the pMS401 PCR template. DNA (1 µg) and 20 units of *DpnI* were incubated in 1x NEB Buffer 4 (final volume of 20 µL) at 37°C for 2 h, followed by heat inactivation at 80°C for 20 min.

The digested sample was PCR purified and then a ligation reaction was set up to self-circularise the DNA. Approximately 60 ng of DNA, 1x T4 DNA ligase buffer (Fermentas), and 10 Weiss unit of T4 DNA ligase (Fermentas) were mixed in a final volume of 20 µL and incubated at 22°C for 1 h. The ligase was inactivated by heating at 65°C for 10 min, then DNA was purified using the QIAQuick PCR Purification Kit and eluted from the column with 20 µL EB.

A 5 µL aliquot of the DNA was used to transform 50 µL electrocompetent *E. coli* DH5α-E. The cells were allowed to recover from electroporation by adding 500 µL SOC and incubating at 37°C for 1 h. The cells were then spread on agar plates with LB and carbenicillin (100 µg/mL). After overnight incubation at 37°C, colonies were picked and used to inoculate LB medium containing carb-100 (5 mL). The plasmids were extracted and screened by restriction mapping to check that the *NdeI* site in the backbone had been eliminated. The new plasmid was named pMS501 (Appendix I).

After verifying the intermediate template pMS501, a PCR reaction was set up to introduce the new *NdeI* and *SpeI* restriction sites, as well as sequence encoding a His₆-tag, downstream of the *SpeI* site. The inverse PCR was set up with 1x Phusion HF buffer, 0.2 mM dNTPs, 5 µM pMS501_Spe.for, 5 µM pMS501_Nde.rev, 10 ng of template pMS501, 1 unit of Phusion polymerase, in a final volume of 50 µL. The PCR cycling conditions were: 98°C for 30 s; 30 cycles of 98°C for 10 s, 54°C for 20 s, 72°C for 70 s; and then one final cycle of 72°C for 5 min. All downstream steps for clean-up, ligation and screening

were identical to those described above for constructing pMS501. The resulting plasmid was named pLAB101 (Appendix I).

2.5.8 Expression and purification of chimeric proteins

As the ITCHY inserts were in pSAlect, which are flanked between the restriction sites *NdeI* and *SpeI*, the ITCHY clones were digested with the restriction enzymes *NdeI* and *SpeI* and then sub-cloned into the expression plasmid pLAB101 to incorporate a C - terminal (His)₆ - tag for purification via metal affinity chromatography. The plasmid pLAB101 (3 µL) harbouring the variants (all inserts were validated for solubility, e.g. P25K86) was used to transform 50 µL aliquots of *E. coli* DH5α-E (Invitrogen) by electroporation using a Bio-Rad Gene Pulser II (see Appendix I for competent cells protocol). The cells were allowed to recover from electroporation by adding 500 µL SOC and incubating at 37°C for 1 h. The transformed cells were spread on LB-carbenicillin (100 µg/mL) agar plates and the number of colonies formed was counted after incubation (37°C, 16 h). The colonies were screened for the presence of the desired inserts by colony PCR with one primer specific to the pLAB101 backbone (301_seq.for; Appendix I), and another specific to the 3' ends of the inserts (Kvβ.rev; Appendix I). The verified clones were stored at -80°C for future use.

A 50 mL culture of *E. coli* DH5α-E cells, harbouring pLAB101-(inserts) in LB - Carb-100, was inoculated with 4 mL of an overnight starter culture incubated at 37°C. The OD₆₀₀ was monitored, until it reached 0.6 and then IPTG was added to a final concentration of 0.5 mM. Following the addition of IPTG, the cells were incubated at 28°C for 4 h. Four hours following induction, the cells were centrifuged at 4000 g for 15 min and the cell pellets were stored at -

80°C. After thawing, a cell pellet was resuspended in 10 ml of column buffer. The column buffer used comprised 40 mM Tris-HCl (pH 8), 300 mM NaCl, 1 mM imidazole, 10% (v/v) glycerol and 1 mM β - mercaptoethanol. Lysozyme to a final concentration of 0.2 mg/mL and 100 μ L of protease inhibitor cocktail (Sigma) were also added. The resuspended cells were then sonicated with amplitude of 50, pulse-on time of 10 s (15 cycles), pulse-off time of 12 s (MISONIX-4000). Following sonication, the cells were centrifuged at 20,000 g at 4°C for 40 min. The soluble lysate was filtered through a 0.2 μ m syringe filter and was then allowed to bind, by means of rocking for 2 h at 4°C, with Talon™ resin (Clontech). Prior to this step, the resin (0.5 mL bed volume) had been equilibrated with the column buffer by pelleting a 1 mL aliquot of Talon™ resin (800 g at 4°C for 2 min) and washing it with 2 x 8 mL of column buffer. The resin, after rocking for 2 h, was pelleted at 1000 g for 2 min and washed with 5 mL of column buffer. Both the unbound fraction and resin wash were stored for gel analysis. The resin was then resuspended in 1 mL of column buffer and transferred to a Bio-Rad gravity flow column. The column was washed with 2 x 5 mL of column buffer containing 10 mM imidazole and finally, the (His)₆-tagged protein was eluted by increasing the imidazole to a concentration of 150 mM in 8 x 0.5 mL fractions. Protein fractions were run on 15% SDS-PAGE gels (Appendix I).

Chapter III

Characterisation of P25K86

Acknowledgements: **Trevor Loo** (from Gill Norris's laboratory, Institute of Fundamental Sciences, Massey University, Palmerston North) performed the mass spectrometry analyses and assisted in all biophysical experiments and **Dr Alexander Goroncy** (Chemistry & Biophysics group, Institute of Fundamental Sciences, Massey University, Palmerston North) performed NMR as described in section 3.3.4.

3.1 Premise of the chapter

This chapter explores the physical attributes of P25K86, which was discovered in Chapter 2. The protein was found to form oligomers and on treatment with a reducing agent, β -mercaptoethanol the multimeric state disappeared. The protein has three cysteines and one of the cysteine (Cys56) was found to mediate in the bond formation, thus giving a dimeric state. While preparing P25K86 for biophysical experiments it was found that it precipitated on when concentrated, and with age. However, an engineered version of P25K86 in which Cys56 was replaced by serine remained monomeric and was moderately stable. Hence, this engineered version, P25K86_CCS, was further explored. It was found using circular dichroism (CD) spectroscopy that it contained secondary structure. Also, the nuclear magnetic resonance (NMR) spectroscopy data suggested that it was partially structured and that it was in a molten globule state.

3.2 Introduction

3.2.1 Borrowing some tools from biophysics to study P25K86

In Chapter 2, two topologically similar yet functionally unrelated proteins were randomly recombined to mimic non-homologous recombination *via* ITCHY. One was a subdomain from an *E. coli* enzyme, trPRAI (PDB: 2KZH) and the other was a rat protein, Kv β 2 (PDB: 1EXB). From a library of 5.2×10^4 variants, of which 28 clones were tested, one clone, P25K86, which contains 25 residues from trPRAI and 86 from Kv β 2, was the most well-behaved and soluble chimeric candidate (refer to Chapter 2, section 2.3.4.1). This chapter focuses on the biophysical and structural characterisation of P25K86. To this end, size-exclusion chromatography and mass spectrometry were used to study the oligomeric state of P25K86.

Size exclusion chromatography (SEC) is an analytical technique that allows separation of proteins based on their size. As there may be many proteins in a sample, with wide differences in molecular weights, this property can be used to purify and separate proteins (Hong *et al.* 2012). Mass spectrometry (MS) is used to determine the elemental composition of a sample, such as a protein. In MS the chemical compounds are ionised to create gaseous charged molecules whose mass-to-charge ratio is measured and used to calculate the monoisotopic mass (molecular weight) of the compound (Bantscheff *et al.* 2007). Two techniques were used in this study in order obtain structural information about the chimeric protein was circular dichroism (CD) and nuclear magnetic resonance spectroscopies (NMR) and these will be discussed in more detail in the following section.

3.2.2 Circular dichroism spectroscopy

CD spectroscopy is a fast way to determine whether a protein is folded, by characterising its secondary structure. Proteins with α -helices have a different spectrum compared to ones with β -sheets, or even a mix of both. In fact, random coil can be predicted using this powerful technique (Ranjbar & Gill 2009). An optically active (structurally asymmetric) molecule absorbs right- and left-handed circularly polarised light to different levels. CD spectroscopy measures the difference in the absorption of left- and right-handed polarised light. A molecule with no regular structure will have a zero CD intensity, while for a molecule with ordered structures a spectrum with both positive and negative signals will be observed (Greenfield 1996). CD spectroscopy is used to characterise secondary (far-UV) or tertiary structure (near-UV) of a protein molecule (Kelly *et al.* 2005). The far-UV spectral region is between 190 and 250 nm and in this range the chromophore is the peptide bond. If the protein molecule is folded, the signal for α -helix, β -sheet and random coil structure has a distinctive CD spectrum. As the signals are an average of the complete protein molecule, one can only determine the amount of α -helix or β -sheets present and thus exactly which residue is involved in giving the helical or sheet signal is beyond the scope of this technique (Kelly & Price 2000; Kelly *et al.* 2005). The near-UV spectral region is between 250 and 350 nm, and in this range the chromophores are the aromatic amino acids and disulfide bonds. Aromatic amino acids have characteristic signatures (near-UV signals) and are used to predict whether a protein is folded into a well-defined structure or not. For example, phenylalanine, tyrosine and tryptophan give a signal between 250-270, 270-290 and 280-300 nm respectively (Greenfield 2006; Kelly *et al.* 2005). Broad, weak signals over the entire near-UV spectrum are representative of

disulfide bonds. An unfolded or molten globule structure gives no signal (nearly zero) in the near-UV region (Greenfield 2006; Kelly *et al.* 2005).

3.2.2.1 Secondary structure analysis

For the secondary structural determination of the chimeric proteins studied in this thesis, DichroWeb, a web-based application where the CD spectrum is compared to a known CD spectrum of a protein, was performed. The analysis or the deconvolution of the CD spectrum using DichroWeb requires a reference set of known secondary structures, grouped into various optimised regions, and a statistical method (analysis programme) to fit the experimental spectra to the reference set (Whitmore & Wallace 2004). Some of the statistical methods used in this study to extrapolate the secondary structure of the chimeric proteins were CONTIN (Provencher & Glöckner 1981), CDSSTR (Compton & Johnson 1986), SELCON3 (Sreerama & Woody 1993) and K2D (Andrade *et al.* 1993). The output of each method (deconvolution algorithm) gives the content of secondary structural elements. A high level of detail is provided in the results, but a key parameter, NRMSD or normalised root mean square deviation, provides goodness-of-fit between the experimental and calculated spectra and is used to assess the quality of the results. A low value of NRMSD indicates that the analysis (statistical method) used generated a good result. A high NRMSD (>0.1) suggests that the calculated secondary structure of a given protein is unlikely to correspond with that of the actual one (Whitmore & Wallace 2004). In this study, the results from four algorithms for any given chimeric protein were compared to estimate the secondary structure. These four methods, i.e. CONTIN, CDSSTR, SELCON3 and K2D, each have their strengths and weaknesses but an overall estimation of the amount of secondary structure can be obtained. The CONTIN algorithm uses a variation of the

least squares method known as ridge regression. In this method the effect of each reference spectrum on the analysed spectrum is kept low lest there is good agreement between the theoretical best-fit curve and raw data. It gives a better estimation of the β -turns in the analysed proteins (Provencher & Glöckner 1981; Greenfield 1996). The CDSSTR method uses a minimum number of reference proteins for the analysis of an unknown protein. The reference proteins are picked randomly for analysing a given CD spectrum. It uses the singular value decomposition (SVD) algorithm, in which each basis spectra, having its unique shape, are correlated to known secondary structures for the analysis of the unknown protein. The solutions generated are subjected to a selection rule, which is based on helical content. The final solution is the average of all acceptable solutions. This method gives a good approximation of the helical content and β -sheets but a poor estimation of turns (Sreerama & Woody 2000; Johnson 1999). The self-consistent method, or SELCON, is an improvement of the variable selection method and enhances the speed and accuracy of the analysis. This is made possible because the reference proteins are arranged in increasing order of root-mean-square difference from the CD spectrum of the protein being analysed, with the least related spectrum being deleted systematically. This method also places an initially predicted structure of the protein to be analysed into the reference data set, which is then deconvoluted using the SVD algorithm. It gives a very good estimation of α -helices β -sheets and β -turns of globular proteins (Sreerama & Woody 1993; Greenfield 1996). The K2D algorithm takes the neural network approach to predict the secondary structure of the protein to be analysed. A neural network program sees patterns and correlations in the data, and in the case of CD the input patterns are the CD spectra and the output is the fractional weights of secondary structures. The K2D database consists of weights and a recall

program to estimate α -helices and β -sheets and it gives the best estimates of β -sheets (Andrade *et al.* 1993; Greenfield 1996).

3.2.3 Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance (NMR) is an analytical technique in which nuclei, when exposed to an electromagnetic pulse, absorb and then release energy. Subatomic particles spin along their axes and pairing of spins in atoms leads to an overall spin of zero. However, atoms with an odd number of protons and/or neutrons can have a net overall spin and these atoms have multiple orientations. These orientations have the same energy in the absence of an external magnetic field but in its presence the energy level splits. More nuclei are present in the lower energy level and these nuclei can be excited into the higher energy level by electromagnetic radiation. The nuclei relax to the lower energy state after absorption. The frequency of radiation (transition frequency) relates to the difference in energy between the two nuclei states (Edwards & Reid 2001; Foster *et al.* 2007).

3.2.3.1 Chemical shift

Electrons form a shield surrounding the nucleus and any external magnetic field will induce the molecular electrons to produce local currents. These currents produce an alternative field that opposes the external magnetic field. The result is the reduction in the total effective magnetic field that acts on the nuclear magnetic moment (which arises from the spin of protons and neutrons) based on the strength of a locally induced magnetic field (Tolman *et al.* 1995; Židek *et al.* 2001). This effect is known as shielding or the chemical shift. The chemical shift can be upfield or downfield (Edwards & Reid 2001; Tolman *et al.* 1995; Ludwig & Viant 2010). In an upfield or diamagnetic shift,

the field produced by electrons opposes the applied magnetic field and therefore in order to reach the transition frequency the applied field strength must increase. Downfield or paramagnetic shift is the opposite of upfield, and is where the applied field strength must decrease in order to reach the transition frequency (Edwards & Reid 2001; Bain 2003; Foster *et al.* 2007).

3.2.4 Protein NMR

Currently, the upper weight limit for NMR based structure determination is ~30 kDa. Above this molecular weight, X-ray crystallography is the only method for high-resolution structure determination (Billeter *et al.* 2008). Given the relatively small size of P25K86 (13 kDa), it was decided to label the protein and collect the ^{15}N heteronuclear single quantum coherence (HSQC) spectra, which provides an early indication as to whether a protein has any tertiary structure or not (Bieri *et al.* 2011).

The two most widely used techniques in protein NMR are 1D ^1H -NMR spectrum, where one sees signals for each of the hydrogen atoms (protons), and 2D ^{15}N -HSQC (heteronuclear single-quantum coherence) spectrum, where a signal for each N-H bond, or basically an amino acid residue, can be seen (Wüthrich 2001; Kwan *et al.* 2011). A limitation with using 1D ^1H -NMR is that there is a lot of overlap between signals, which makes the data interpretation difficult, for this reason 2D NMR is more widely used (Ludwig & Viant 2010). In 2D NMR, the hydrogen nuclei are excited and the energy is transferred (transfer of magnetisation) to either ^{15}N or ^{13}C (Edwards & Reid 2001; Wüthrich 2001). The chemical shift is evolved on the nitrogen or carbon, followed by the transfer of energy back to the hydrogen, which is then detected. Using this technique, a single peak for each amide proton (N-

H correlations) can be seen, with the exception of proline as it lacks an amide proton (Ludwig & Viant 2010; Foster *et al.* 2007). For this study, an HSQC experiment was conducted using the natural abundance of ^1H - ^{13}C and ^1H - ^{15}N as the isotopic labeling failed due to the lack of expression of the recombinant protein in minimal media. Another technique, ^1H - ^1H TOCSY (TOtal Correlated Spectroscopy), was also performed on the protein samples. This technique splits proton signals into groups and the spectrum contains all cross peaks because of protons having the same spin-spin coupling. Different amino acids have protons that belong to different spin systems, a feature which is exploited to resolve amino acids and the spin systems they belong to (Edwards & Reid 2001; Xu *et al.* 2007; Tolman *et al.* 1995).

Positions of cross peaks in the TOCSY are characteristic for a set of amino acids and thus some suggestions can be made as to how much structure the protein P25K86 has. There are certain rules (expected patterns) in regards to the position of amino acids in the spectra and these were used to assign some regions of P25K86_CCS (an engineered version of P25K86 that has Cys56 replaced with serine).

As this protein is too small to be a monomeric TIM barrel, a baffling question was – “is this a new fold or merely a reinvented old one?” Answering this question will offer insights into whether new folds can be created via non-homologous recombination using techniques like ITCHY. Some structural information was gained in this attempt to explore P25K86. The current data suggests that the protein is partially folded and exists in a molten globule state.

3.3 Results

3.3.1 Oligomerisation

In Chapter 2, a 50 mL culture of the clone containing the expression plasmid, pLAB101-P25K86, was over-expressed and (His)₆-tag purified, using the Talon resin, in *E. coli* with an IPTG-induced T7 expression system (refer to Chapter 2, section 2.3.4.1). An SDS-PAGE gel of the eluted fractions indicated that the protein might be forming a dimer (Fig 2.11B). Given that the gel was run under denaturing conditions, it appeared that the three-cysteine residues, Cys7, Cys46 and Cys56, might be involved in disulfide bond formation to stabilise the conformation of P25K86. To investigate this further, the culture volume was scaled up to 500 mL and protein was eluted in 16 fractions of 0.5 mL each (refer to section 3.5.1). Twelve of the purest fractions were pooled together and buffer was exchanged while concentrating the protein (refer to section 3.5.2 and Fig. AIV.2, Appendix IV).

3.3.1.1 Size exclusion chromatography

The oligomeric states of P25K86 at room temperature were determined by size-exclusion chromatography with a Superdex 75 10/300 GL column (refer to section 3.5.4). Following injection of 0.2 mL of the protein sample, a large peak (Fig. 3.1) suggested that the protein had aggregated and therefore was eluting in the void volume. Increasing the salt concentration (from 50 mM KCl to 200 mM KCl) and adding reducing agent (10 mM DTT) was found to prevent aggregation, and a peak was detected with an adjacent shoulder (between 10 and 15 mL), indicative of two oligomeric states (Fig. 3.2).

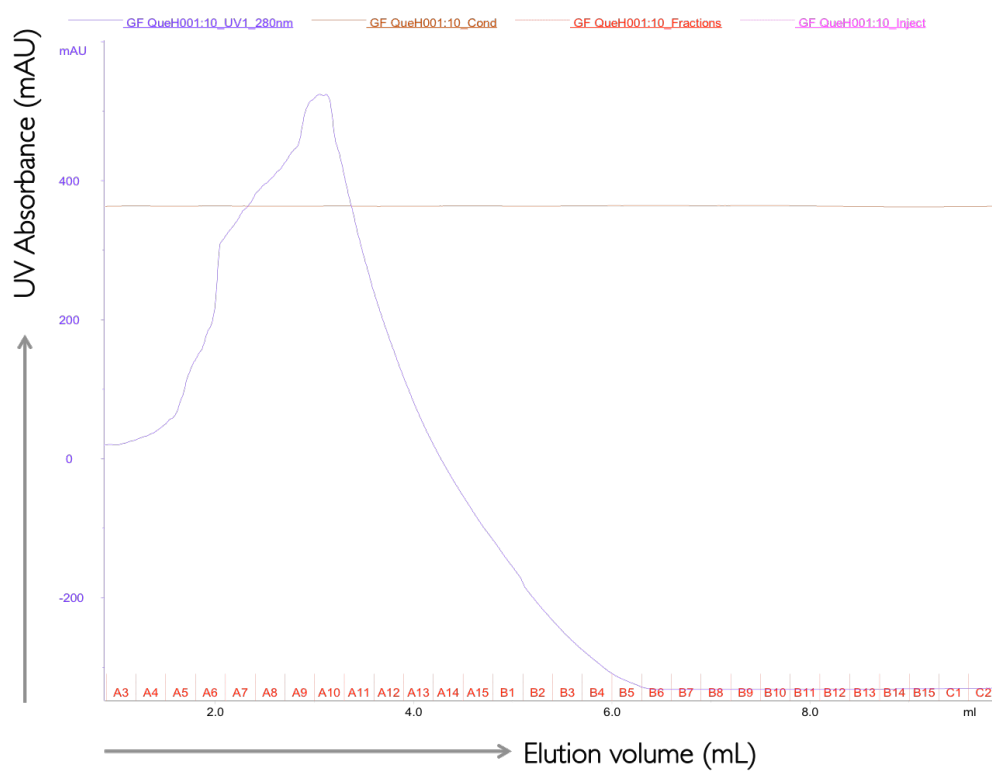


Fig. 3.1. Size-exclusion chromatography of P25K86. Initial run of P25K86, came in the void volume. On the x-axis are the fractions (elution volume) and on the y-axis is absorbance (mAU) at 280 nm.

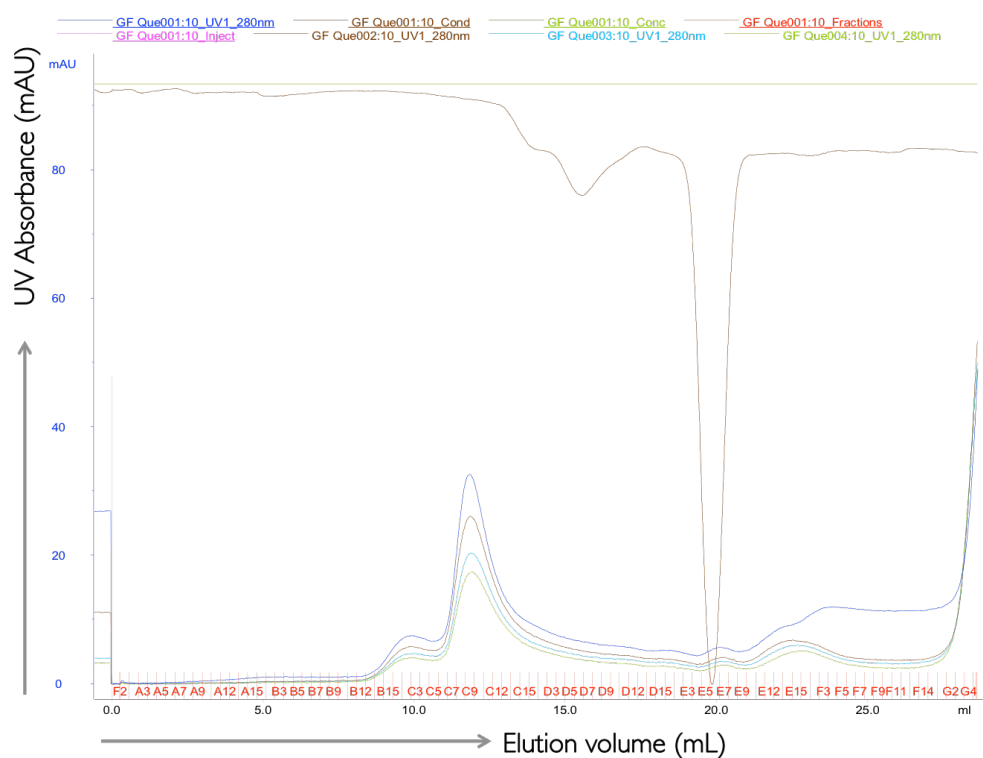


Fig. 3.2. Size exclusion chromatography of P25K86, upon addition of DTT (0.01 M) and increasing the KCl concentration to 0.2 M. On the x-axis are the fractions (elution volume) and on the y-axis is absorbance (mAU) at 280 nm. The blue and brown trace lines indicate absorbance of the protein sample and conductivity of the buffer at 280 nm respectively. *Note: Other colour trace lines can be ignored, as they pertain to background checks.*

3.3.1.2 High performance liquid chromatography (HPLC)

In a separate experiment, the protein was further purified and desalted by reverse phase-HPLC (refer to section 3.5.4). Protein fractions appearing at the retention times 20.5, 20.8 (this was later submitted for mass spectrometry) and 21.5 min (indicated by peaks 4, 5 and 6 in Fig. 3.3, top panel) were collected and concentrated to ~ 30 μ L by vacuum evaporation for accurate mass determination by electrospray ionisation mass spectrometry (ESI-MS).

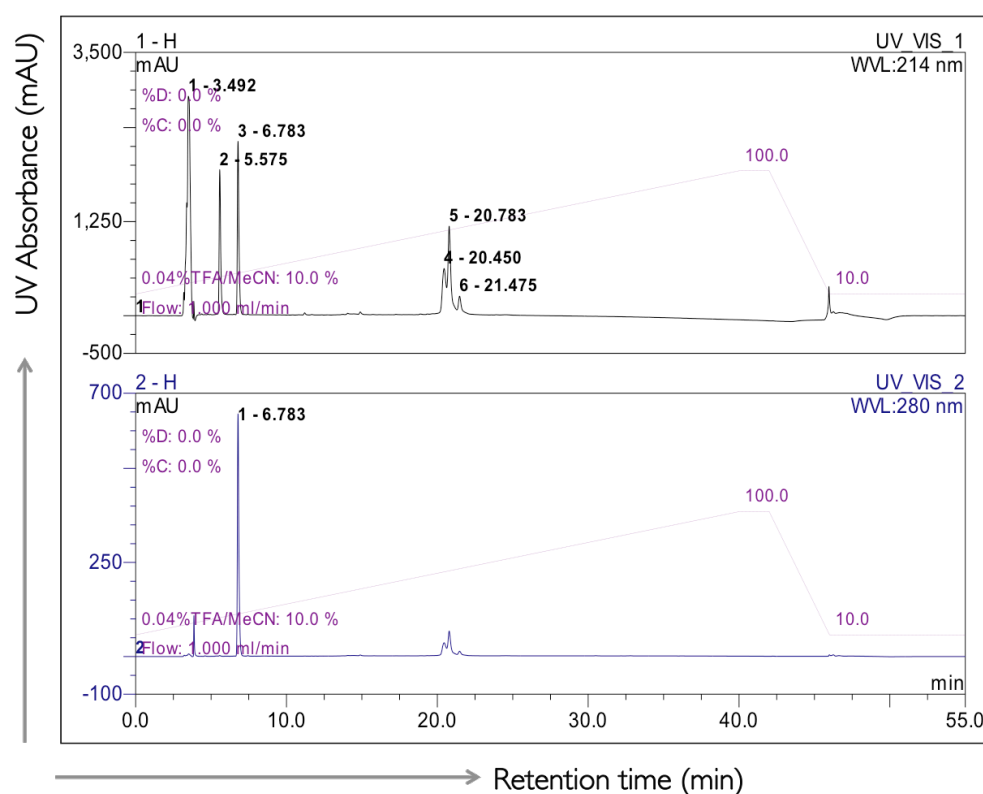


Fig. 3.3. Reverse phase-HPLC of P25K86. Peaks of the sample can be seen at retention time 20.5 min in the bottom panel (280 nm). A stronger signal response of the same peaks at 214 nm in the top panel can be seen. On the x-axis is retention time and on the y-axis is absorbance (mAU) at 214 nm (top panel) and 280 nm (bottom panel).

3.3.1.3 Mass spectrometry (MS)

In addition, the protein was further purified and desalted for MS. MS data confirmed that the protein is present in several multimeric states. MS was performed on the unreduced P25K86 protein sample. Figure 3.4 shows the deconvoluted spectrum of P25K86. The data suggested close matches for the monomer ($12981.69 + \text{water as it is ionised, so } 12999.69$) and dimer ($2 \times 12999.69 - \text{S-S, so } 25999.38 - 2 = 25997.38$). The last peak in Fig. 3.4 was speculated to be a tetramer ($4 \times 12999.69 - 2 \times \text{S-S, so } 51998.76 - 4 = 51994.76$) but was later found to be a persistent contaminant and was disproven by a western blot (Fig. 3.7). The values displayed are not an exact match as the software only gives the value of the highest peak in the isotope distribution and deconvolution can be difficult if the sample was not clean (refer to section 3.5.5).

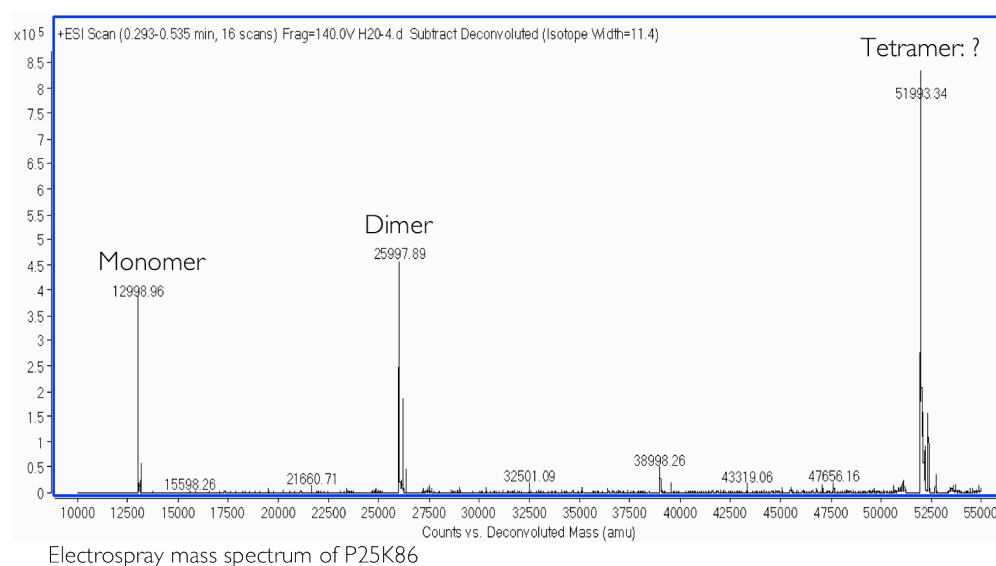


Fig. 3.4 Electrospray ionisation mass spectrometry of P25K86. The three peaks indicating monomer, dimer and possibly a tetramer can be seen. On the x-axis is the deconvoluted mass and on the y-axis are counts.

3.3.2 Investigating the role of cysteine in oligomerisation and a quick test for function

When treated with β -mercaptoethanol (BME), P25K86 is reduced to its monomeric form (refer to Chapter 2, section 2.3.4.1), therefore a strategy to test which cysteine residue(s) could be responsible was designed. To minimise disruptions to the integrity of the structure and maintain the character of the cysteine residues, serine was chosen to replace cysteine in a series of point mutants. Alanine was also considered, but its smaller size can cause a small cavity in the interior of the protein, thus tampering with the integrity. Three point mutants were made, replacing each of P25K86's cysteine residues with serine; i.e. Cys7 \rightarrow Ser (SCC), Cys46 \rightarrow Ser (CSC) and Cys56 \rightarrow Ser (CCS). Although there are nine possible ways (Fig. 3.5) to form a disulfide-linked dimer, in the first instance three genes with single point mutations were synthesised, subcloned into the expression vector pLAB101, and tested for solubility. (refer to section 3.5.6).

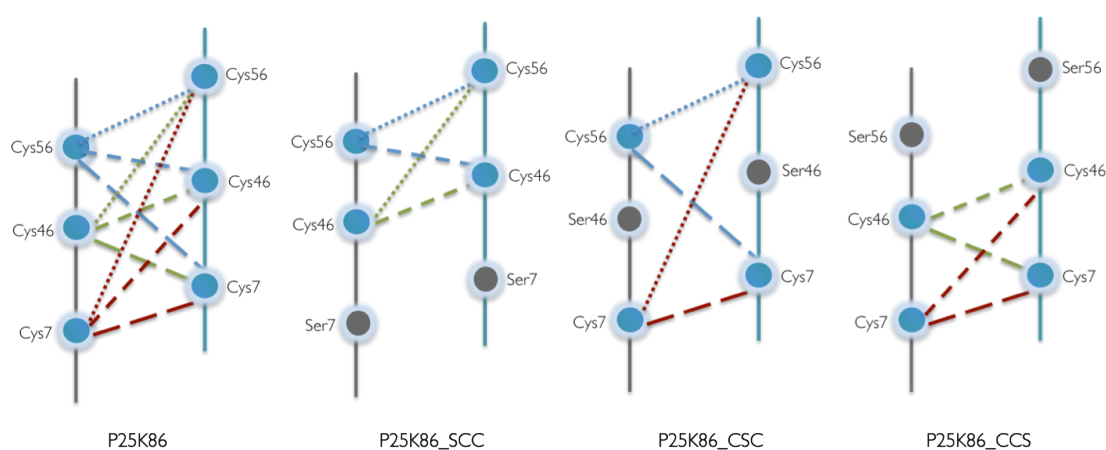


Fig. 3.5. Possible routes to disulfide bond formation in P25K86. Single point mutants were synthesised, expressed and purified to see which one of the three cysteine mutants could be contributing to the dimeric state of P25K86.

After transforming the DH5 α -E cells by plasmids (pLAB101-P25K86-SCC/CSC/CCS), the clones were screened (Fig. 3.6) for inserts with primers, 301_seq.for and Kv β .rev (Appendix I).

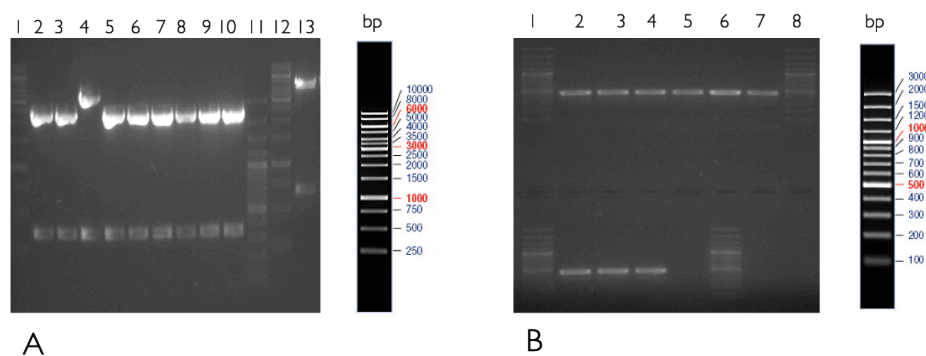


Fig. 3.6 (A) Restriction digest of plasmids (pIDTSMART) containing the SCC, CSC and CCS inserts and pLAB101-P55K173. Lane 1: 1 Kb Fermentas ladder, lane 2-4: CSC, lane 5-7: SCC, lane 8-10: CCS, lane 11: 100 bp Fermentas ladder, lane 12: 1 Kb Fermentas ladder, lane 13: vector: pLAB101. (B) PCR screen of P25K86 variants in pLAB101. Top – lane 1: 100 bp ladder, lane 2-4: CCS, lane 5-7: SCC, lane 8: 100 bp ladder; Bottom – lane 1: 100 bp ladder, lane 2-4: CSC. Product size is 450 bp.

The three protein variants were purified using Talon resin (refer to section 3.5.1). All three mutants were run on an SDS-PAGE gel, with and without BME (Fig. 3.7A). Clearly, it can be seen that the mutant P25K86_{CCS} was no longer able to form a dimer. This was further confirmed by using His-specific antibody (Fig. 3.7B). Also, as no signal was detected for a tetramer, this state was ruled out. This result demonstrated that the tetramer size band was a persistent contaminant. From the western blot analysis of the P25K86 variants, it was concluded that Cys56 in the protein P25K86 was engaging in forming a dimer. Referring back Fig. 3.5, it can be concluded that it is a Cys56-Cys56 disulfide, rather than any of the “mixed” options. Furthermore, in order to test whether the NADPH-binding subdomain of Kv β 2 in the clone P25K86 was functional, and was binding NADPH, the protein was scanned from 240 to 400 nm. It was thought that if the chimera binds NADPH, there

should be a peak at 340 nm (NADPH absorbs light at 340 nm) in addition to the peak at 280 nm. The scan suggested that NADPH was not bound, implying the domain was not appropriately folded (refer to section AIV.8, Appendix IV).

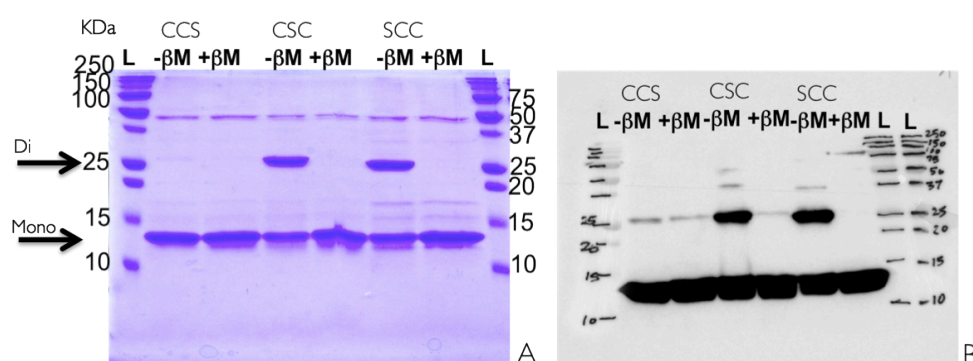


Fig. 3.7. (A) SDS-PAGE of P25K86 variants (CCS,CSC,SCC) in the absence (-βM) and presence (+βM) of β-mercaptoethanol. (B) Western blot analysis of P25K86 variants under similar conditions. The samples were His₆-tagged and blotted with His₆-specific antibody. Precision plus protein ladder (BIO-RAD) was used as protein standard.

3.3.3 The P25K86_{CCS} transition

The initial attempts to purify a clean protein sample that could eventually be used for biophysical experiments proved challenging. For instance, on many occasions the protein was precipitating while concentrating and exchanging the buffer. Even during size exclusion chromatography the salt concentration had to be increased as the protein aggregated and eluted in the void volume, and only upon addition of the reducing agent and by increasing salt concentration was a peak detected (refer to Fig. 3.2). However, for any further biophysical experiment to be possible, it was necessary to reduce the salt concentration, and thus all optimisation attempts seemed to fail. Interestingly, P25K86's oligomeric state was disintegrated to its monomeric

form in the mutant P25K86_CCS. Additionally, it was found to be soluble, yet it did not precipitate as rapidly as P25K86. Therefore, the study was continued with P25K86_CCS.

3.3.3.1 Labelling attempts for NMR

Given the relatively small size of P25K86_CCS (13 kDa), labelling the protein for NMR was initiated. A 1 L culture of P25K86_CCS in M9 minimal media was grown and ^{15}N was incorporated into the biomolecules by adding $[\text{}^{15}\text{N}]\text{H}_4\text{SO}_4$ to the media (refer to section 3.5.8). The cells were induced at 25°C with IPTG and samples were taken every 2 h for monitoring the protein expression (Fig. 3.8).

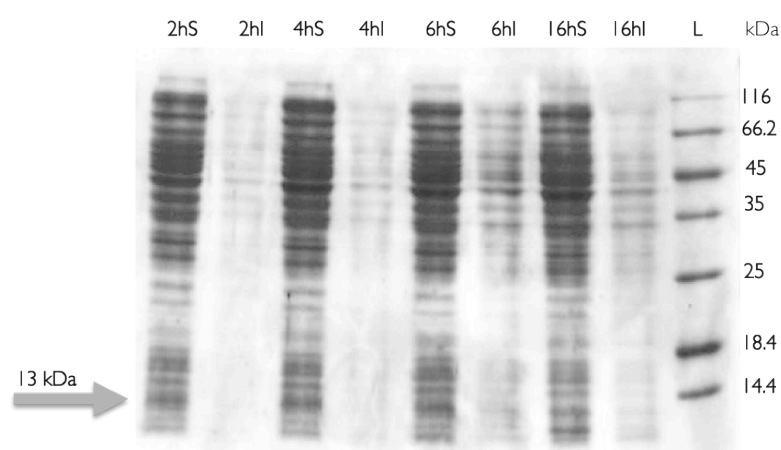


Fig. 3.8. SDS-PAGE expression profile of P25K86_CCS mutant in minimal media. The protein samples were diluted 25-fold and then 10 μL were loaded on eight lanes from left to right. The alphanumeric labels at the top represent time points i.e. **2 h** for 2 hours and soluble fractions with "S" and insoluble with "I". PageBlue protein ladder (Thermo-Scientific) represented as "L" was used as protein standard.

The expression profile (Fig. 3.8) of P25K86_CCS in minimal media indicated poor yields. The minimal media used in this experiment contained 0.3% (final volume) glucose and it was speculated that doubling the glucose in the media might increase the yields. In addition, a few trace elements (zinc and

cobalt) and boric acid were introduced, along with 0.6% (final volume) glucose in a new media (refer to section 3.5.8.2). It was found that the yields at 6 h after induction with IPTG, in the new conditions, were relatively similar to overnight expression (16 h) of the cells (Fig. 3.9A). Moreover, the cells were going into stationary phase at 6 h after induction (Fig. 3.9B). A 1 L culture of P25K86_CCS in M9 minimal media and the new labelling mix was induced for 6 h and labelled protein was purified using Talon resin.

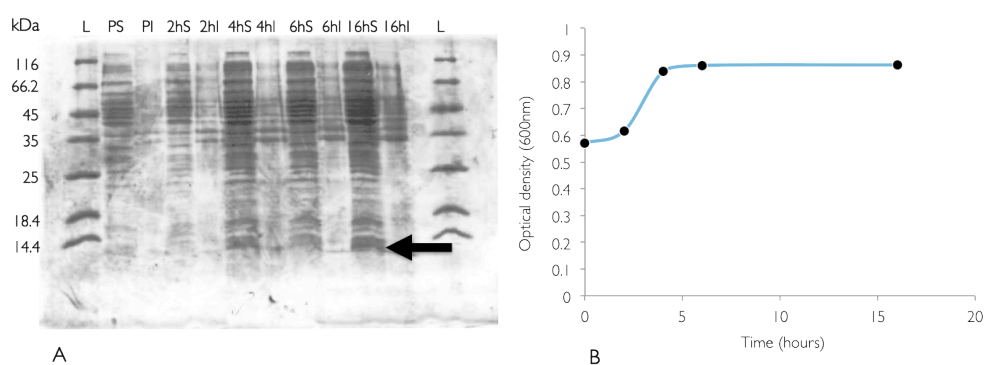


Fig. 3.9. (A) SDS-PAGE expression profile of P25K86_CCS mutant in new media with more trace elements. The protein samples were diluted 2-fold and then 10 μ L were loaded on 10 lanes from left to right. The alphanumeric labels at the top represent time points i.e. **2h** for 2 hours, soluble fractions with “S”, insoluble with “I” and **P** stands for pre-induction. PageBlue protein ladder (Thermo-Scientific) represented as “L” was used as protein standard. (B) Growth of cells in minimal media. Time zero is pre-induction and then samples were drawn out and measured 2, 4, 6 and 16 h after induction. The cells were found to be going into stationary phase 6 h after the induction with IPTG

Using the above conditions, a 1 L culture of the clone harbouring pLAB101-P25K86_CCS was used to label the protein (refer to section 3.5.8).

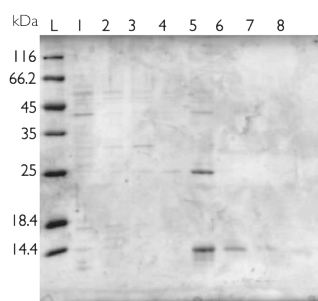


Fig. 3.10. SDS-PAGE gel of elution fractions 1 to 8 of P25K86_CCS mutant in minimal media. The protein was eluted with increasing concentrations of imidazole: Lanes 1 & 2 (10 mM), 3 & 4 (20 mM), 5 to 7 (150 mM) and the final 8th wash with 0.5 M. Fractions 6, 7 and 8 were pooled and concentrated.

A band corresponding to 25 kDa can be seen in lane 5, Fig. 3.10. This could possibly be a contaminant or a dimer formed with other cysteines. Whatever the case may be, pure fractions from lanes 6, 7 and 8 were pooled and the protein was concentrated (refer to section 3.5.2.1). The yield of the protein, purified from a 1 L culture, was ~2 mg (1.5 mL at 1.3 mg/mL). The concentration was measured using the Bradford assay (refer to section 3.5.3).

Before using the labeled protein for NMR experiments, its secondary structure was tested using CD. Also, in order to compare the CD spectrum of P25K86_CCS with its parent proteins, the control protein Kv β 2 (39 kDa) was expressed and purified under similar conditions (Fig. 3.11).

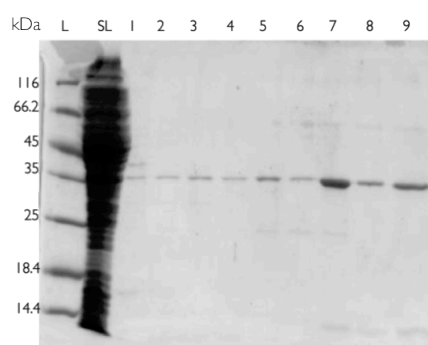


Fig. 3.11. SDS-PAGE gel of Kv β 2 elution fractions 1 to 9. "SL" represents the soluble lysate. Elution fraction 9 was chosen for further biophysical experiments. PageBlue protein ladder (Thermo-Scientific) represented as "L" was used as protein standard.

3.3.3.1.1 Far-UV circular dichroism

Far-UV circular dichroism (CD) spectroscopy was used to monitor the secondary structure of the mutant protein, P25K86_CCS. The protein P25K86_CCS was diluted (5-fold) with water to a final concentration of 0.25 mg/mL. The final concentration of the buffer used was 2 mM potassium

phosphate, 10 mM KCl, pH 7.2. CD spectra were measured at 25°C on a Chirascan spectrophotometer (Applied Photo Physics, Surrey, UK). The CD spectrum of Kv β 2 was collected under similar conditions. The far-UV CD spectrum of P25K86_CCS displays secondary structure and is not superimposable on either trPRAI or Kv β 2 (Figs 3.12 & 3.13). It appears that P25K86_CCS has a higher percentage of α -helices as compared to β -sheets. There are two negative bands at 208 nm and 222 nm, and a positive band at 190 nm. Upon deconvolution using the CDSSTR algorithm (Sreerama & Woody 2000; Compton & Johnson 1986) in the DichroWeb application (Whitmore & Wallace 2004), it can be inferred that the chimeric protein P25K86_CCS has 55% α -helix, 17% β -sheet, 11% turns and 17% of it remains disordered (Table 3.1).

| Method | SELCON3 | CONTIN | CDSSTR | K2D |
|-------------|---------|--------|--------------|------|
| NRMSD | 0.37 | 0.13 | 0.003 | 0.31 |
| Helix | 0.38 | 0.44 | 0.55 | 0.69 |
| Strand | 0.11 | 0.08 | 0.17 | 0.03 |
| Turns | 0.23 | 0.21 | 0.11 | — |
| Disordered | 0.28 | 0.28 | 0.17 | — |
| Random Coil | — | — | — | 0.27 |

Table. 3.1. Deconvolution of P25K86_CCS, using the DichroWeb application (Whitmore & Wallace 2004). Four algorithms (methods) were used to calculate the amount of secondary structure present, all of which show a similar trend. It is evident from the data that this chimeric protein contains a higher percentage of α -helices over β -sheets no matter what method is used to deconvolute.

The CDSSTR algorithm gave the smallest NRMSD value (refer to section 3.2.1.1.1) compared to SELCON3, CONTIN and K2D. However, despite the large range of NRMSD values, all four algorithms gave a good level of agreement in predicting the secondary structure elements present in P25K86_CCS. All four algorithms report a high percentage of α -helices over

β -sheets. In fact, on averaging SELCON3, CONTIN and K2D, the amount of α -helix comes to 51%, which is close to the CDSSTR value of 55%. The neural network based algorithm, K2D, estimated the highest value for α -helices (69%) and the lowest level of β -sheets (3%). The CDSSTR algorithm is known to give good estimates of α -helix and β -sheets but is poor when estimating turns (Greenfield 2006). On the other hand, the SELCON3 algorithm is known to give good approximations of all secondary structure elements, but for globular proteins (Greenfield 2006), of which P25K86_CCS is certainly not. The estimations using the CONTIN algorithm are the next lowest in the acceptable NRMSD threshold, with 44% α -helices, 8% β -sheets, 21% turns and 28% of it being disordered. The K2D algorithm also estimates the disordered proportion of the protein to be 27%, which is close to the estimated values using SELCON3 and CONTIN. Overall, it seems that the protein has some secondary structure, but that nearly 30% of the protein remains unstructured. Unfortunately, despite the presence of a few aromatic acids, the near-UV could not be measured as the protein precipitated in the cuvette.

In addition, the spectrum of P25K86_CCS does not resemble that of either trPRAI or Kv β 2, whose spectra are representative of the typical mixed α/β motifs. On a per-residue basis, P25K86_CCS has a similar amount of α -helices and β -sheets as does Kv β 2 (for deconvolutions of Kv β 2 and trPRAI see section AIV.6 and AIV.7, Appendix IV), whereas there is no similarity with trPRAI. Changes in the secondary structure of P25K86_CCS were determined by increasing the temperature (Fig. 3.12D). It can be seen that the protein loses its secondary structure at elevated temperatures (70°C).

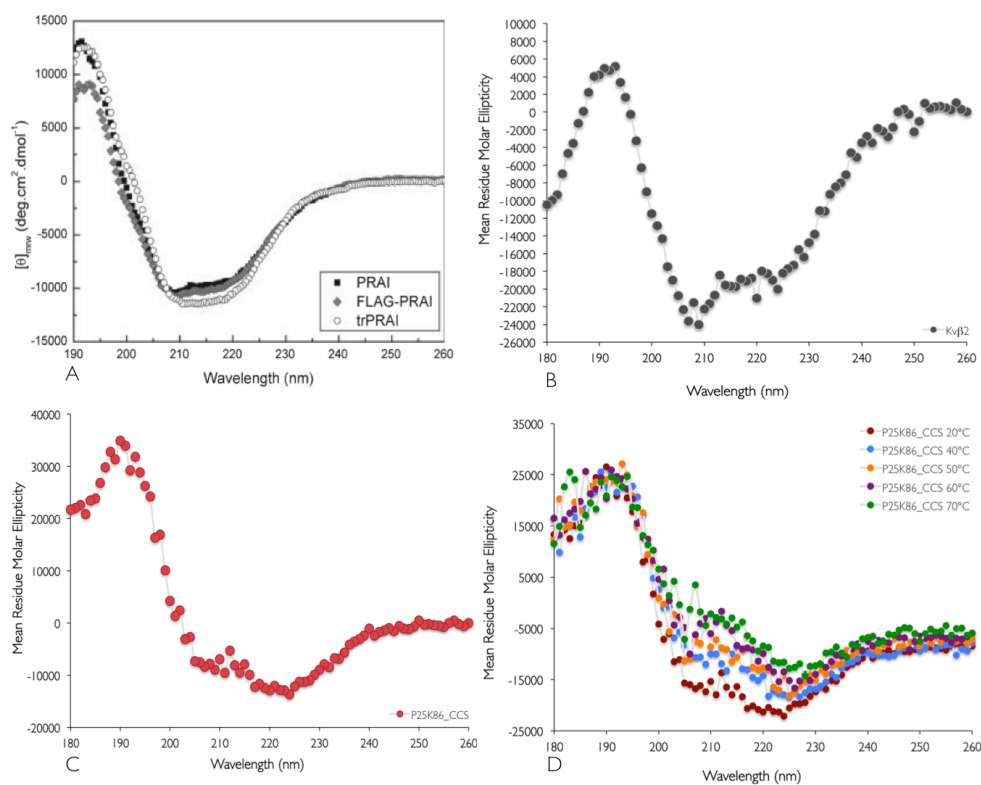


Fig. 3.12. Far-UV CD spectrum of (A) trPRAI/PRAI (adapted from Patrick & Blackburn 2005), (B) Kvβ2 at 20°C, (C) P25K86_CCS at 20°C, and (D) thermal melt of P25K86_CCS. The temperature range for thermodenaturation was: 20°C (red), 40°C (blue), 50°C (orange), 60°C (purple) and 70°C (green). On the x-axis is wavelength in nanometers (nm) and on the y-axis is mean residue molar ellipticity.

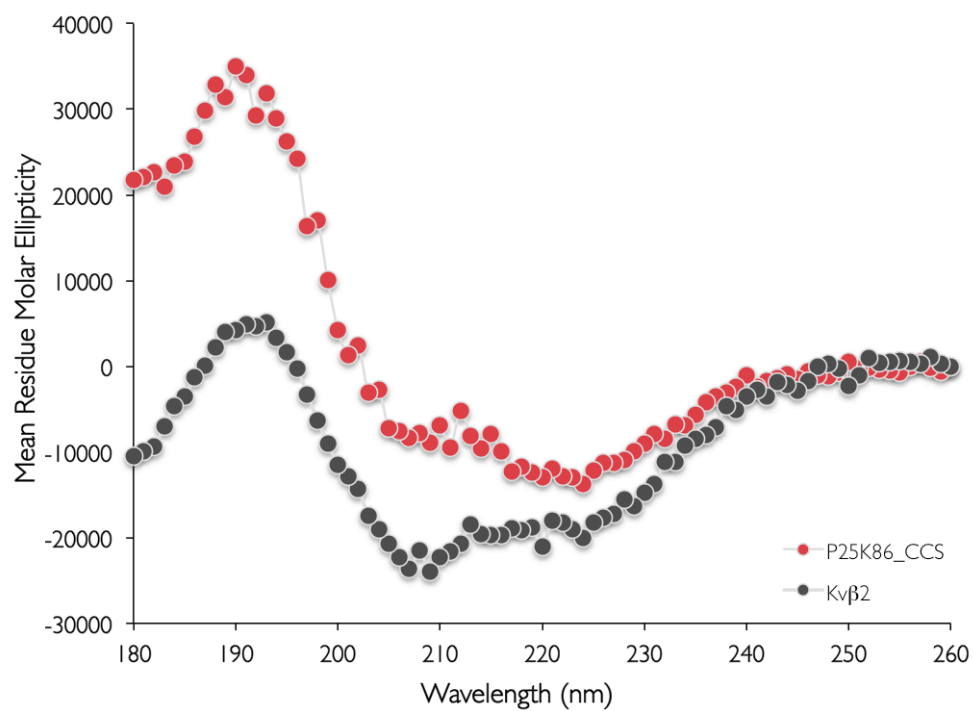


Fig. 3.13. Comparing the far-UV CD spectrum of Kvβ2 (grey) with P25K86_CCS (red) at 20°C. On the x-axis is wavelength in nanometres (nm) and on the y-axis is mean residue molar ellipticity.

3.3.4 NMR spectroscopy of P25K86_CCS

At the time of writing, Dr Alexander Goroncy had collected the NMR data of this chimeric protein. Data analysis from ^{15}N -HSQC, ^{13}C -HSQC and ^1H - ^1H TOCSY spectra suggests that only ~30 amino acids and the backbone region (NH- αH) of P25K86_CCS are visible.

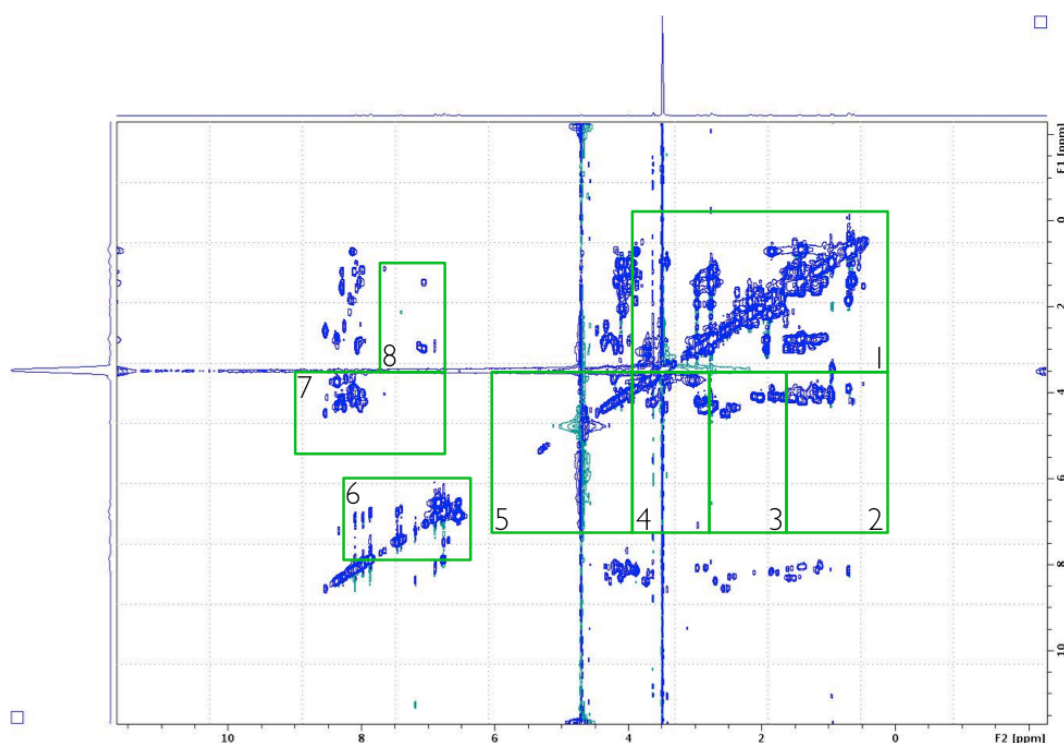


Fig. 3.14 ^1H - ^1H TOCSY of P25K86_CCS. On the x-axis is the proton NMR region of the peptide bond (NH) in parts per million (ppm) and on the y-axis is the chemical shift of the protons (αH) in the range 0.5-8.5 ppm. A few regions are annotated in green boxes and are explained in section 3.3.4.

The annotation in Fig. 3.14 is based on the expected patterns of COSY (correlated spectroscopy), but for many experiments the COSY and TOCSY spin systems look identical (Zerbe & Jurt 2013). Both provide a diagonally symmetric two-dimensional plot, but a COSY spectrum has off-diagonal

correlations between coupled spins whereas in TOCSY all ^1H in a network of coupled spins will be correlated (Zerbe & Jurt 2013). Eight regions were annotated in ^1H - ^1H TOCSY of P25K86_CCS:

1. All non-labile, non-aromatic sidechain protons.
2. αH - βCH_3 of Ala and βH - γCH_3 of Thr.
3. αH - βH of Val, Ile, Leu, Glu, Gln, Met, Pro, Arg and Lys.
4. αH - βH of Cys, Asp, Asn, Tyr, His and Trp.
5. αH - αH of Gly, αH - βH of Thr, δH - δH of Pro, αH - βH and βH - βH of Ser.
6. Aromatic ring protons, including 2H-4H of His, as well as sidechain protons from Asn and Gln.
7. Backbone NH- αH .
8. δCH_2 - ϵNH of Arg.

The fact that the backbone NH- αH (7) or the fingerprint region is present in the ^1H - ^1H TOCSY of P25K86_CCS suggests that some of the residues might be visible.

An attempt was also made to assign certain peaks (Fig. 3.15A) by using the statistics calculated for selected chemical shifts from atoms in the 20 common amino acids. The Biological Magnetic Resonance Data Bank or BMRB (<http://www.bmrb.wisc.edu/refinfo/statsel.htm>) has 6,202,473 possible chemical shifts and 4,549,638 were used to generate a table of average shift, which was used as a reference to assign the peaks for Asn, Gln, Gly, Ser, Thr, Trp, His and Ala in the ^{15}N -HSQC of P25K86_CCS.

Fig. 3.15 (A) ^{15}N -HSQC of P25K86_CCS. On the x-axis is ^1H in parts per million (ppm) and on the y-axis is ^{15}N . Assigned peaks (probable) of eight amino acids are highlighted in green circles. (B) ^{15}N -HSQC of trPRAI (adapted from Setiyaputra *et al.* 2011). Comparing A with B, it can be seen that P25K86 is only partially structured and most of it remains disordered. The peaks in trPRAI are sharp and well dispersed whereas in P25K86 they are broad and less dispersed. This is an example of a protein in a molten globule state, as discussed in section 3.4.1.

Eight residues have been assigned peaks in the ^{15}N -HSQC spectrum of P25K86_CCS, which might be present in regions of the spectrum based on data from average chemical shifts and are circled in green in Fig. 3.15A. The peaks for Asn/Gln, Gly, Ser/Thr and Ala are (7.5/6.7, 112/111), (8.2, 109.6), (8.3/8.2, 116.2/115.3) and (8.2, 123.5) respectively. The histidine-epsilon may be visible at peak (7.9, 119.6), whereas tryptophan-epsilon may be visible at (7.3, 129.2). In addition, the ^{13}C -HSQC of P25K86_CCS (Fig. 3.16) indicates the presence of a few peaks in the aromatic and aliphatic region of the spectrum. The C-epsilon of His at 137.67 and C-delta of Trp at 131 may be visible. There is also a possibility of the C-epsilon of Tyr being visible at 117.9.

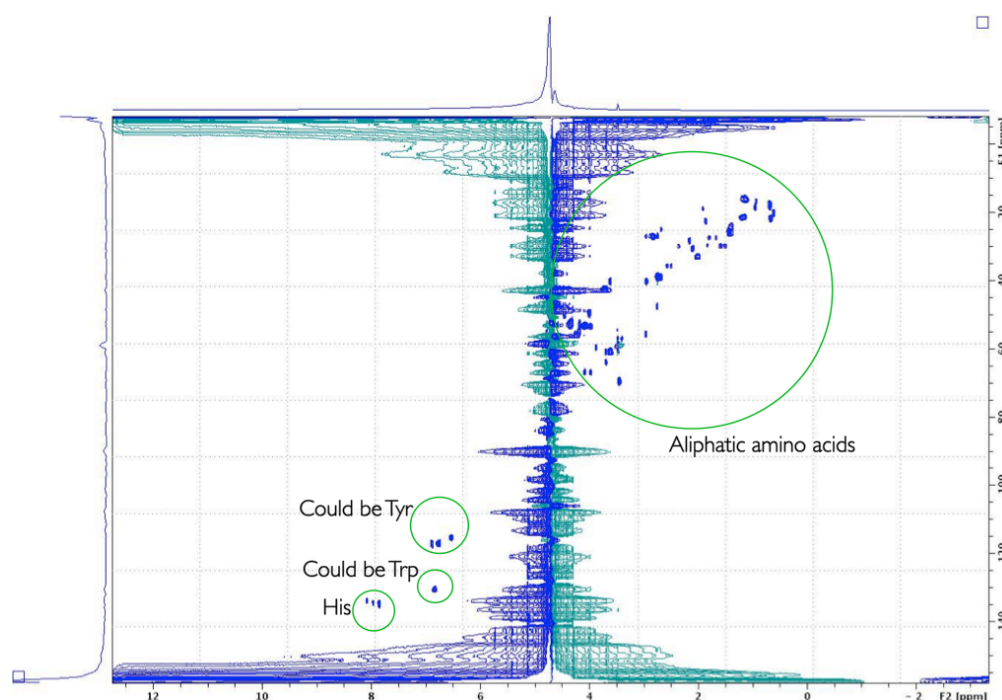


Fig. 3.16. ^{13}C -HSQC of P25K86_CCS. On the x-axis is ^1H in parts per million (ppm) and on the y-axis is ^{13}C . The aliphatic region and three aromatic amino acid residues that might be visible are highlighted in green circles.

Based on this analysis, the underlined regions of P25K86_CCS that might be visible are shown in Fig. 3.17.

Fig. 3.17. The protein sequence of P25K86_CCS. The sequence in red represents PRAI, while the sequence in grey represents Kv β 2. The C-terminal region with the His tag associated with protein is highlighted in blue. The two cysteines (Cys7 and Cys46) and serine (Ser56) are highlighted in green. Amino acid residues that may be visible are underlined. In total there are eight residues that might be visible in the NMR spectroscopy of P25K86_CCS. Exactly what specific stretch of amino acids might be contributing to the α -helical signature in the CD spectrum by analysing this NMR spectrum is hard to predict.

```

001 - MGENKVCGLT
011 - RGODAKAAYD
021 - AGAIYGGRRQ
031 - QAKLKELOAI
041 - AERLGCTLPO
051 - LAIAWSLRNE
061 - GVSSVLLGAS
071 - NAEQLMENIG
081 - AIQVLPKLSS
091 - SIVHEIDSIL
101 - GNKPYSKKDY
111 - RSTSGHHHHH
121 - H

```

Despite the small amount of structural information gathered in the study, it seems likely that the protein is only partially structured and most of it remains disordered.

3.4 Discussion

3.4.1 Switching from P25K86 to P25K86_CCS

In this proof-of-principle experiment, random recombination of trPRAI and Kv β 2 revealed a tantalising protein, P25K86. Initial attempts at its biophysical characterisation revealed that it forms a dimer. Size exclusion chromatography also indicated that it might be aggregating, as it was not interacting with the column and appeared in the void volume. Only upon addition of DTT and by increasing the salt concentration was a peak detected. Furthermore, while exchanging the buffer, it started to precipitate and stick to the membrane of the concentrator. Upon scanning the chimeric protein from 240 nm to 340 nm, it became clear that this protein does not bind NADPH. This suggested that it lacked function. It was clear that cysteine was involved in the formation of the oligomers as they diminished upon the addition of reducing agent. Therefore removal of the odd cysteine by site directed mutagenesis in order to lessen the likelihood of the formation of intermolecular disulfide bonds led to the production of the mutant P25K86_CCS.

By replacing Cys56 with serine, it was found that this cysteine was mediating in S-S bond formation in P25K86. This monomeric version of P25K86 (P25K86_CCS) was used for further biophysical experiments. This monomeric variant had a CD spectrum with two troughs between 200 and 230 nm and a peak at 190 nm, and further deconvolution suggested that it has a higher percentage of α -helices (55%) over β -sheets (17%) with nearly 20% of the protein being disordered. An interesting observation is that there is a predominant α -helical insertion at the C-terminus (β 7 α 7 loop) of Kv β 2 (Gulbis

et al. 2000) and this helical stretch (helix 15, 16, 17, 18 and 20) together with the helix 1 of trPRAI are contributing to the α -helix signal in CD. What is more interesting is that the deconvolution of P25K86_CCS *via* CDSSTR predicts 55% helices i.e. 61 residues should be contributing to helices, and by mapping the two parents for their α -helices contribution in P25K86_CCS it was found that 52 residues from the C-terminal end of Kv β 2 (AIV.5, Appendix IV) and 10 residues from the N-terminal of trPRAI (AIV.4, Appendix IV) do indeed form helices. Furthermore, the deconvolution results of Kv β 2 (57% α -helices and 20% β -sheets) suggest that the predicted values are close to the proportions found in the X-ray structure. On a per residue basis, the protein has 54% α -helices and 16.2% β -sheets. One hundred and seventy-five residues contribute to α -helices and 53 residues contribute to β -sheets in a chain of 326 residues. This validates the deconvolution approach to predict the secondary structure information of these chimeras.

The ^{15}N -HSQC spectra of P25K86_CCS suggest that only about 30 amino acids are visible. This indicates that the protein is partially structured with a large proportion of it remaining disordered. This appears to be an example of a protein in a molten globule state, as it lacks tertiary structure but has a significant amount of secondary structure. In globular proteins, lack of tertiary structure interactions indicates an ensemble of fluctuating structures and NMR spectroscopy measurements can reveal this (Kim *et al.* 1999; Bom *et al.* 2010). Folded proteins, in general, have good resonance signals with sharp peaks that are well dispersed in the spectrum whereas proteins in molten globule state shows signal loss with the peaks being broad and less dispersed. The molten globule state has conformational heterogeneity, which is due to the interactions between the residual secondary and side chains causing the cross-peaks to be broader or even disappear completely

(Kamatari *et al.* 2004; Bom *et al.* 2010). Folded proteins exhibiting sharp narrow lines (well-defined frequencies) in the NMR spectra tend to be in the long-lived, excited state whereas broader lines (partially folded/unfolded proteins) result from nuclei with rapid signal decay. On a μ s-ms timescale, a protein that is going through interconversions between different conformations can cause line broadenings (Kwan *et al.* 2011). Despite the exhibition of secondary structure in the far-UV CD spectrum, the presence of few sporadic sharp peaks and a dominance of broad and less dispersed peaks in the NMR spectrum of P25K86_CCS indicate that it is exchanging its conformation between two or more states on a μ s-ms timescale. The presence of broad signals of nearly 30 residues indicates that they are going through rapid internal motion, while the remaining (~70) or a portion of it is giving rise to secondary structure in the far-UV CD spectrum. Hence, there seems to be an exchange between the structured and the unstructured states, which explains the broadening of the signals.

This chimeric protein fulfils all the criteria for it to be considered to exist in a molten globule state as specified by Kuwajima (1996), i.e. the presence of secondary structure; absence of tertiary structure, and presence of loosely packed hydrophobic core. Although no direct experiment was done to prove this latter criterion, the precipitation of the protein during concentration, and the precarious stability of the protein strongly suggest that P25K86_CCS is only partially folded, and suggests that indeed, the protein is in the molten globule state.

3.4.2 What can we learn from this experiment?

The recombinant protein P25K86 was initially produced in *E. coli* cells and purified as a dimer due association of folded monomers in some stepwise manner. Upon mutating one cysteine, which was shown to participate in intermolecular disulfide bond formation, to serine, a soluble monomeric form (P25K86_CCS) was produced that appeared to be more stable and less prone to precipitation.

It is noteworthy that this propensity to form oligomers was also seen for the chimeric protein 1B11 (De Bono *et al.* 2005), which was also created *via* combinatorial assembly of two non-homologous proteins. P25K86 does not have any physiological relevance. The hypothesis that new proteins can form from random combinations of stable folding domains has not been proven. However this work has shown that proteins can be formed from such domains that do not have stable tertiary structure, but form soluble intermediates along the folding path. It is possible that other random mutations might tip the balance so that a folding minima is reached and a new protein is formed.

3.5 Materials and Methods

All reagents were purchased from Sigma unless specifically noted. Common molecular biology materials, techniques and primer sequences are described in Appendix I. Specific materials and methods used in this chapter are described in the following subsections.

3.5.1 Expression and purification of P25K86 & its variants

The P25K86 insert was in the plasmid pSAlect (Lutz *et al.* 2002), flanked between NdeI and SpeI restriction site. The plasmid was digested with NdeI and SpeI and then sub-cloned into the expression plasmid pLAB101 to incorporate a C - terminal (His)₆ - tag for purification via metal affinity chromatography. Electrocompetent DH5 α -E cells were transformed by the plasmid pLAB101-P25K86. After transforming the plasmids (pLAB101-P25K86 - SCC/CSC/CCS) in DH5 α -E cells the clones were screened for inserts with primers, 301_seq.for and Kv β .rev (Appendix I). To test the clones for expression, a 50 mL culture of clones containing the plasmid pLAB101-P25K86-SCC/CSC/CCS was induced with IPTG with a final concentration of 0.5 mM. The proteins were purified and eluted in eight fractions of 0.5 mL each. For the expression of the scale-up culture of P25K86 clone, a 500 mL LB - Carb100 culture was inoculated with 4 mL of an overnight starter culture incubated at 37°C. The OD₆₀₀ was monitored, until it reached 0.6 and then IPTG was added to a final concentration of 0.5 mM. Following the addition of IPTG, the cells were incubated at 28°C for 4 h.

Four hours following induction, the cells were centrifuged at 4000 g for 15 min and the cell pellets were stored at -80°C. After thawing, a cell pellet was

resuspended in 20 ml of column buffer. The column buffer used comprised 40 mM Tris-HCl (pH 8), 300 mM NaCl, 1 mM imidazole, 10% (v/v) glycerol and 1 mM β - mercaptoethanol (BME). Lysozyme to a final concentration of 0.2 mg/mL and 100 μ L of protease inhibitor cocktail (Sigma) were also added. The resuspended cells were then sonicated with amplitude of 50, pulse-on time of 10 s (15 cycles), pulse-off time of 12 s (MISONIX-4000). Following sonication the cells were centrifuged at 20,000 g at 4°C for 40 min. The soluble lysate was filtered through a 0.2 μ m syringe filter and was then allowed to bind, by means of rocking for 2 h at 4°C, with Talon resin (Clontech). Prior to this step, the resin (1 mL bed volume) had been equilibrated with the column buffer by pelleting a 2 mL aliquot of Talon resin (800 g at 4°C for 2 min) and washing it with 2 x 8 mL of column buffer. The resin, after rocking, was pelleted at 1000 g for 2 minutes and washed with 5 mL of column buffer. Both the unbound fraction and resin wash were stored for gel analysis. The resin was then resuspended in 1 mL of column buffer and transferred to a BIO-RAD gravity flow column. The column was washed with 2 x 5 mL of column buffer containing 10 mM imidazole and finally, the (His)₆-tagged protein was eluted by increasing the imidazole to a concentration of 150 mM in 16 x 0.5 mL fractions. Protein fractions were run on 15% SDS-PAGE gels. For the P25K86 variants (SCC, CSC, CCS) all conditions were similar as above except for the culture volumes, which were 50 mL. Every other component of the purification protocol was scaled down accordingly.

3.5.2 Buffer for biophysical experiments

The protein was eluted in Tris buffer but for further biophysical characterisation the buffer was exchanged with 10 mM potassium

phosphate, 50 mM KCl, pH 7.2. To prepare 200 mL, added 0.312 g of $\text{K}_2\text{HPO}_4 \cdot 3\text{H}_2\text{O}$, 0.086 g of KH_2PO_4 and 0.746 g of KCl and added water to the final volume. To achieve the right pH, Henderson–Hasselbalch equation was used.

$\text{pH} = \text{pKa} + \log ([\text{Salt}]/[\text{Acid}])$ - *Henderson–Hasselbalch equation*

Phosphoric acid has multiple dissociation constants, 2.15, 6.86 and 12.32, and for making buffer with pH 7.2, the following equation was used as follows:

$$7.2 = 6.86 + \log [\text{HPO}_4]^{-2}/[\text{H}_2\text{PO}_4]^{-1} \quad [\text{HPO}_4]^{-2} = a; b = [\text{H}_2\text{PO}_4]^{-1}$$

$$7.2 - 6.86 = \log (a/b)$$

$$0.34 = \log (a/b)$$

$$a/b = 10^{0.34}$$

$$a/b = 2.19; a = 2.19b$$

For the solution with a molarity of 10 mM or 0.01 M

$$a + b = 0.01; a = (0.01 - a)2.19$$

$$a = 0.0219 - 2.19a; a + 2.19a = 0.0219; 3.19a = 0.0219; a = 0.00686; b = 0.00314$$

$$\text{OR } a = 0.00686\text{M}; b = 0.00314\text{M};$$

MW of $\text{K}_2\text{HPO}_4 \cdot 3\text{H}_2\text{O}$ (a) = 228.22 g/mol and MW of KH_2PO_4 (b) = 136.09 g/mol; $a = 0.00686 \times 228.22 = 1.56$ g; $b = 0.00314 \times 136.09 = 0.43$ g (This is to prepare 1000 mL). To prepare 50 mM KCl; MW = 74.55 g/mol added 0.746 g in 200 mL water. After making the buffer, the pH was tested using standard procedures and when necessary a base (NaOH) or an acid (HCl) was added to adjust the pH.

3.5.2.1 Exchanging the buffer *via* concentrator

The buffer was exchanged using a Vivaspin6 (GE Healthcare) concentrator as follows.

- 1) The concentrator (MWCO 5,000) was equilibrated with 2 mL of buffer and centrifuged for 10 min at 4000 g.
- 2) The column was equilibrated again at 4,000 g for 15 min with 3 mL buffer.
- 3) Selected elution washes (pure/clean fractions) were pooled together and then loaded in the concentrator. The column was filled with exchange buffer to a final volume of 6 mL and mixed thoroughly. This was spun at 4000 g for 16 min.
- 4) This was repeated 4-5 times and the protein was concentrated to a final volume of 1.5 mL.

3.5.3 Measuring concentration

The protein concentration was measured using SmartSpec (BIO-RAD; at Gill Norris's Laboratory). From a 5 mg/mL BSA standard, five dilutions were made: 0.05, 0.1, 0.2, 0.4 and 0.6 mg/mL (dilutions were done to a final volume of 1 mL). After making the dilutions, 20 μ L of the standards were mixed with 300 μ L of the Bradford reagent and the absorbance measured at 595 nm. The absorbance was recorded on the SmartSpec (BIO-RAD), which also has a built-in feature to prepare standard curves and generate a printed copy of the data. For the blank, 20 μ L of water was used as the dilutions were done in water. To be consistent between each sampling, just one cuvette was used that was washed thoroughly with water and ethanol. On one occasion, the standards were repeated twice and the standard curve plotted with the coefficient of determination (R^2) of 0.977. After this the absorbance of the

sample (20 μ L+300 μ L Bradford reagent) was measured and the estimated concentration was 0.07 mg/mL. This was the standard procedure for all measurements.

3.5.4 Gel filtration & HPLC

For further biophysical analysis, the Tris buffer was exchanged with the phosphate buffer (10 mM potassium phosphate, 10 mM KCl, pH 7.2) using the Vivaspin concentrator (5000 MWCO). The protein was concentrated (1.2 mg/mL) to a final volume of 1.5 mL and 0.2 mL of the protein was injected in size exclusion column (Superdex 75 10/300 GL, GEH). The oligomeric states of P25K86_CCS at room temperature were determined by size exclusion chromatography with a Superdex 75 10/300 GL (GEH) and ÄKTA explorer 10 FPLC (GEH). The column was equilibrated and all runs were performed in 10 mM potassium phosphate, 50 mM KCl, pH 7.2, at a flow rate of 0.6 mL/min in 0.3 mL fractions. The column was calibrated with 2 column volumes of buffer (bed volume = 24 mL). Blue dextran (a large, dyed sugar with a molecular weight of 2×10^6 daltons) was used to measure the void volume V_0 (7.2 mL) of the column. Void volume is the volume of liquid collected from the minute a sample is injected into the column until a non-retained molecule is eluted from the column. Series of known molecular weight standard proteins, i.e. CytoC, Carb Anhydrase, Ovalbumin, BSA, ALDH (Sigma-Aldrich), were injected into the column and the elution volume (V_r) at which each standard protein is eluted was recorded. Retention time or elution volumes were plotted against log MW in linear regression mode.

Protein was further purified and desalted by reverse phase-HPLC. 0.2 ml of protein was injected into a Phenomenex Jupiter® C18 column (4.6x250 mm,

5 μm , 300 \AA) on an Ultimate 3000 HPLC (Dionex Corp.) running the Chromeleon® 6.8 chromatography data system. The program was a linear gradient from 0.1% TFA/ 90% water/ 10% acetonitrile to 0.08% TFA/100% acetonitrile over 40 minutes at a flow rate of 1 ml/min. The main peak at 20.4 min were collected and concentrated to $\sim 30\ \mu\text{l}$ by vacuum evaporation for accurate mass determination by ESI-MS.

3.5.5 Mass spectrophotometry

The concentrated sample was introduced by direct infusion into an Agilent 6520 Q-TOF MS (Agilent Technologies, Germany) with the following source settings: VCap = 3500V, fragmentor = 175V, skimmer = 65V, nebuliser gas = 30 psig, drying gas = 5 l/min, gas temperature = 325°C. The resulting ions were analysed in positive mode with a scan range from 100-1700 m/z , and a data acquisition rate of 5 spectra/s, corrected with internal standards of 121.0508 and 922.0098 Da.

3.5.6 Making the P25K86 mutants

To make the point mutants, genes were ordered from Integrated DNA Technologies (USA). The three single point mutants (SCC, CSC, CCS) came in the plasmid backbone pIDTSMART-AMP. The inserts were flanked by the restrictions site NdeI and SpeI, which are also present in the multiple cloning cassette of the expression plasmid pLAB101. Each parent plasmid containing the point mutant variants had a concentration of 100 ng/ μL . A 1000 fold dilution of the plasmids was done and 3 μL of each was used to transform *E. coli* DH5 α -E (Invitrogen) by electroporation, using a BIO-RAD Gene Pulser II (Refer Appendix I for competent cells protocol). The parameters used were

2.5 kV, 25 μ F, 200 Ω and cuvettes with 0.2 cm gaps (BIO-RAD). The transformed cells were spread on LB-carbenicillin (100 μ g/mL) agar plates and three colonies from each plate, after incubation at 37°C and 16 h, were picked. The colonies were used to inoculate a 5 mL overnight culture in LB-Carb (100 μ g/mL) and following incubation at 37°C, the plasmids were extracted using the EZNA Plasmid Mini Kit (Omega Bio-Tek) by following the manufacturer's guidelines. Under similar conditions an overnight culture containing the pLAB101-P55K173 plasmid, was also set to extract the expression plasmid pLAB101. The plasmid DNA was eluted with 30 μ L elution buffer (10 mM Tris, pH 8.5). Each plasmid (3 μ g) was digested with 20 U of NdeI and SpeI (NEB), 1x BSA, 1x NEBuffer 4 in a total reaction volume of 40 μ L (Fig 3.6A). After incubation at 37°C for 6 h, the restriction enzymes were inactivated at 80°C for 20 minutes. Following this, the samples were loaded on a 1% agarose gel and electrophoresed (Appendix I). The bands corresponding to the three inserts (SCC, CSC, CCS: 336 bp) and the vector backbone pLAB101 (4.3 kb) were excised and the DNA recovered with QIAquick Gel Extraction kit (Qiagen). DNA was eluted from each column in 15 μ L elution buffer. The concentrations of the inserts (SCC: 130 ng/ μ L, CSC: 121 ng/ μ L, CCS: 97 ng/ μ L) and vector (pLAB101: 170ng/ μ L) were measured using spectrophotometer (Eppendorf).

A three-fold molar excess of insert DNA (12 ng) over vector DNA (50 ng), 10 x T4 DNA ligase buffer and 10 Weiss unit of the T4 DNA ligase to a final volume of 30 μ L was set up. A control with only vector was also set up using the same components but without the inserts. After overnight incubation at 16°C, the reactions were heat-inactivated at 65°C for 10 minutes and desalted using MinElute columns (Qiagen). A 50 μ L aliquot of *E. coli* DH5 α -E was transformed with 1.5 μ L of the ligated product by electroporation. Each

aliquot of electroporated cells was recovered in 0.5 mL of SOC at 37°C for 1 h. The recovered cells were plated on carbenicillin (100 µg/mL) agar plates, which were then incubated at 37°C for 16 h.

Three randomly chosen clones from each plate (SCC, CSC, CCS) were screened for the inserts by means of colony PCR. A single colony was picked, resuspended in 10 µL of water, and heated at 95°C for 5 minutes. A 1 µL aliquot of the lysed cells was used as the template for colony PCR, in a reaction with 1x GoTaq buffer (Promega), 0.25 mM of each dNTP, 5 µM of each primer (301_seq.for and Kvβ.rev; Appendix I) and 2.5 units of Taq polymerase (iNtRON), in a total volume of 20 µL. The thermocycling conditions were: 94°C for 2 min; 30 cycles of 94°C for 10 s, 55°C for 20 s, 72°C for 36 s; and one final cycle of 72°C for 5 min. Expected size of the product was 450 bp (Fig. 3.6B)

3.5.7 Immunoblotting

The P25K86 variants were separated using a 15% SDS-PAGE gel. Following electrophoresis the proteins were transferred to a membrane (PVDF, Pierce USA) and surface of the membrane was blocked for 1 h with 5% (w/v) non-fat milk powder in TBS-T (Tris-Buffered Saline and Tween 20). The blocked membrane was incubated with primary antibody (Anti-His (Santa Cruz biotechnology, USA), 1:200 in milk) for 2 h with gentle agitation. Following probing by primary antibody, the membrane was washed three times, for 5 min each, with TBS-T. The membrane was then probed with secondary antibody (anti-Rabbit (Santa Cruz biotechnology, USA), 1:50,000 in milk) sealed in a bag, incubating at room temperature for 1 h with gentle agitation. After incubation, the membrane was washed three times for 5 min each with

TBS-T. The membrane was then sealed in a bag and chemiluminescence detection solutions (1 mL peroxidase and 1 mL luminol) were added and incubated for 5 minutes prior to detection using biomolecular imager (FujiFilm LAS-4000).

3.5.8 Labelling P25K86_CCS

An overnight culture of P25K86_CCS in LB-Carb100 was used to inoculate 1 L of LB-Carb (1% inoculum) and grown overnight at 37°C. Cells were harvested by centrifugation in sterile bottles in a GS3 rotor (Piramoon Technologies Inc, Santa Clara, USA) at 7000 g for 15 minutes at 4°C in a Sorvall Evolution RC. Cells were washed in 200 mL minimal media base to remove any trace of LB media components and re-centrifuged. Cells were re-suspended in 250 mL minimal media base and 10 mL of metal and labelling mix in a 1 L flask and incubated with shaking at 37°C for 1 hour to recover. Protein expression was then induced by addition of IPTG to a final concentration of 0.5 mM. The cells were induced at 28°C for 6 h.

After induction, the cell culture was pelleted and the cell culture was centrifuged at 4000 g for 15 min and the cell pellets were stored at -80°C. The cell pellets were re-suspended in 40 mL of column buffer, which comprised of 10 mM KCl, 10 mM potassium phosphate, and pH 7.2. The cells were lysed with French pressure cell. Talon resin (4 mL bed volume, Clontech) was equilibrated with column buffer that also contained 1 mM imidazole. Following the disruption of cells via French press, the cells were centrifuged at 20,000 g at 4°C for 30 min. The soluble lysate was filtered through a 0.2 µm syringe filter and was then allowed to bind to the equilibrated resin, by pumping it through the column. The unbound fraction

was collected and kept for gel analysis if needed. To elute protein from resin, the column was washed with increasing concentration of imidazole. Two washes each of 10 mM, 20 mM, 40 mM and 150 mM imidazole was done and protein eluted in a final volume of 10 mL. A final wash with 0.5 M imidazole was done to elute any remaining protein.

3.5.8.1 Preparing M9 minimal media base

Minimal Media Base: Sterilise by autoclaving

| | |
|----------------------------------|-----------|
| Na ₂ HPO ₄ | 3.39 g |
| KH ₂ PO ₄ | 1.5 g |
| NaCl | 0.25 g |
| 1M MgSO ₄ | 1 mL |
| 1M CaCl ₂ | 50 µL |
| H ₂ O | to 500 mL |

Metal & labelling Mix: Filter sterilise

| | |
|--|----------|
| Glucose | 0.8 g |
| (¹⁵ NH ₄) ₂ SO ₄ | 0.25 g |
| Thiamine | 5 mg |
| Fe _(III) Cl ₃ | 6.75 mg |
| H ₂ O | to 10 mL |

3.5.8.2 Modified media

A new media recipe was tried with more trace elements as under:

Trace Elements:

Dissolve 5 g of Na₂EDTA and correct pH to 7 in 800 mL of dH₂O. After setting the pH add the following:

| | |
|--------------------------------------|--------|
| FeCl ₃ | 0.5 g |
| ZnCl ₂ | 0.05 g |
| CuCl ₂ | 0.01 g |
| CoCl ₂ .6H ₂ O | 0.01 g |
| H ₃ BO ₃ | 0.01 g |
| MnCl ₂ .6H ₂ O | 1.6 g |

Make up to 1 litre, and sterilise by autoclaving.

Minimal Media Base:

| | |
|----------------------------------|--------|
| Na ₂ HPO ₄ | 6 g |
| KH ₂ PO ₄ | 3 g |
| NaCl | 0.5 g |
| 1M MgSO ₄ | 1 mL |
| 1M CaCl ₂ | 200 µL |
| H ₂ O | to 1 L |

Sterilise by autoclaving

| | |
|---|---------------------------------|
| Thiamine (40 mg/mL) | 1 mL (Filter sterilise) |
| Trace elements | 10 mL |
| Glucose | 0.6% (final) (Filter sterilise) |
| (¹⁵ NH ₄) ₂ SO ₄ (0.1 g/mL) | 10 mL (Filter sterilise) |

3.5.9 Circular dichroism

The protein P25K86_CCS was diluted (5 x) with water to a final concentration of 0.25 mg/mL. The protein was degassed and introduced to a clean 0.1 mm pathlength quartz cell (Hellma® 106-QS, Germany). Buffer was used as a control. Circular Dichroism spectra were measured at 20°C on a Chirascan spectrophotometer (Applied Photo Physics, Surrey, UK). A wavelength range

of 180 nm - 260 nm was measured using a wavelength interval of 1 nm, TOP of 0.5 sec and a bandwidth of 1 nm. Ten replicates were performed per run.

3.5.10 NMR spectroscopy

For the NMR acquisition, the following parameters were used:

^{15}N -HSQC: 2048 x 128 data points, 15.9381 ppm x 44 ppm sweep widths, 1024 scans, 1s delay time.

^{13}C -HSQC: 2048 x 128 data points, 16.0817 ppm x 166.0490 ppm sweep widths, 1s delay time, and 576 scans (for the P25K86_CCS sample) or 992 scans (for the P24K89 sample)

TOCSY: 4096 x 128 data points, 13.9458 ppm x 14 ppm sweep widths, 96 scans, 60 ms mixing time, 1s delay time. The samples were mixed with 10% D₂O before acquisition.

Note: Dr Alexander Goroncy acquired the spectra.

Chapter IV

Improving the folding selection system

Acknowledgements: **Trevor Loo** (from the laboratory of Dr Gill Norris, Institute of Fundamental Sciences, Massey University, Palmerston North) helped in NMR and other biophysical optimisation experiments. **Dr Alexander Goroncy** (Chemistry & Biophysics group, Institute of Fundamental Sciences, Massey University, Palmerston North) performed NMR as described in section 4.3.9. **Laura V. Nigon** (Institute of Natural and Mathematical Sciences, Massey University, Albany) for assisting in sequencing clones from ITCHY libraries.

4.1 Premise of the chapter

This chapter builds on the findings of Chapter 2, where it was found that the pSALect selection system, which was used to create an ITCHY library of trPRAI-Kv β 2 variants, reported false positive clones. The *in vivo* solubility selection system was indeed found to be leaky as only one in six chimeras could be expressed solubly after they were sub-cloned into a protein over-expression vector, pLAB101. This meant that there was room for improvement in the selection system and hence a series of experiments was conducted so as to improve pSALect. The ultimate goal was to isolate more interesting structured chimeras using the newer and improved version of pSALect, and thus to add to the repertoire of folded proteins. In this chapter, the product of the full *trpF* gene, i.e. PRAI, was used instead of the truncated version (trPRAI) that was used in Chapter 2. This was done so as to increase the diversity of variants in the library, minimise any selection biases and, more importantly, to explore new folds.

4.2 Introduction

4.2.1 Switch to pInSAlect

In directed evolution experiments, sampling interesting candidates from a large population of clones harbouring variants can be rate limiting (An *et al.* 2011). To find a soluble, folded and functional protein is the key, and is akin to finding a needle in a haystack. In Chapter 2, the pSAlect selection system was employed. The first generation of the system has head and tail reporters, where the inserts, chimeric variants or proteins-of-interest (POIs) are cloned between the Tat signal peptide (head) and β -lactamase (tail). The translation of the whole construct is required, and only in-frame, folded POIs will be exported to the periplasmic space to give resistance when cells are spread on an agar plate containing ampicillin (Lutz *et al.* 2002).

A second version of the system, pInSAlect, incorporated the *Saccharomyces cerevisiae* VMA split intein into the head and tail reporters. The pInSAlect vector is a strict reading frame selection system and unlike the pSAlect system it is not designed to select for soluble or folded proteins. The VMA intein sequences catalyse the post-translational excision of themselves and the POI. All possible in-frame intein-POI variants are selected (Gerth *et al.* 2004). The vector pInSAlect was used to make a PRAI-Kv β 2 library and the results of this experiment are discussed in this chapter. The plasmid pInSAlect was used as a pre-selection tool, before subcloning the pool of all in-frame chimeras into multiple downstream folding selection systems. Furthermore, the pInSAlect system was used to eliminate out-of-frame clones in a robotic screen via the ESPRIT (expression of soluble proteins by

random incremental truncation) system, which inspired its implementation for this study as well (Yumerefendi *et al.* 2010).

4.2.2 Protein translocation via Tat and Sec pathways

As discussed in Chapter 2, the pSALect system yielded many false positives (five of the six chimeras tested were insoluble). It was hypothesised that the false positives arose when proteins bypassed the Tat export pathway. The translocation of many proteins is via the general secretory (Sec) pathway, which exports newly synthesised polypeptides in largely unfolded states (Blaudeck *et al.* 2003). The proteins are channelled through a protein pore (SecYEG) and the energy to drive this comes from the SecA ATPase, which couples the energy of ATP hydrolysis to protein translocation. The polypeptide is channelled through the Sec pore in a ratchet-like motion and upon reaching the periplasm its signal peptide is cleaved by signal peptidase I and the protein is folded. Owing to its narrow pore size, the Sec pathway is unable to translocate proteins that have already folded (Tullman-Ercek 2006; Berks *et al.* 2000; Auclair *et al.* 2012; Economou & Wickner 1994).

Another pathway capable of exporting proteins that are not compatible with Sec export is the twin-arginine translocation (Tat) pathway (Berks *et al.* 2005). The Tat signal peptide contains a consensus amino acid sequence motif, including two consecutive arginine residues. The translocation components of this pathway are the Tat (A/E) BC proteins, which export pre-folded proteins across energy transducing membranes. The TatBC proteins form a stable unit with TatC, containing the primary binding site for the signal peptide. TatA (or its paralog TatE) acts as a protein-conducting channel, forming an oligomeric ring-like structure. On binding the Tat substrate, TatA

assembles with TatBC, thereby forming the translocon complex. There is a characteristic substrate “proofreading” mechanism during Tat translocation, where only correctly folded proteins are sensed prior to export. This is attained by chaperones such as *Escherichia coli* DmsD and TorD that facilitate cofactor insertion with the protein export via binding the signal peptide and thus preventing premature export (Fisher & Delisa 2004; Fisher *et al.* 2006; Fisher *et al.* 2011). However, proteins without cofactor are also Tat substrates and are transported via the pathway (Berks 1996). This suggests that the use of Tat in folding selection screens can be used for a diverse set of proteins. One example is the *E. coli* alkaline phosphatase, which requires disulfide bonds for proper folding. Strains that lack the mechanism for disulfide bond formation were found to be transport incompetent, suggesting that only proteins that have attained native conformations *in vivo* are exported. This furthers the argument of a “folding quality control” mechanism of the Tat translocon complex (DeLisa *et al.* 2003). As the Tat translocon is essential for protein export and not all the substrates contain cofactors, it has been suggested that the folding quality control mechanism might be occurring within the Tat(A/E)BC components (DeLisa *et al.* 2003). It is possible that proteolytic degradation or chaperone interactions may occur prior to the interaction of the substrate with TatBC. Recently, it has been shown that folding speed and surface hydrophobicity directs Tat substrates to the periplasm, thus substrates that fold fast and with buried hydrophobic residues are favoured (Ribnicky *et al.* 2007; Fisher *et al.* 2011).

4.2.2.1 Tat versus Sec

A signal peptide is required in order to initiate export by either of the two pathways (Sargent *et al.* 2006). Signal peptides consist of three domains: the

positively charged N-terminal “N-region”, the hydrophobic “H-region” and the polar C-terminal (C-region). Signal peptides that target the Tat machinery contain the consensus sequence motif Ser/Thr-Arg-Arg-(X \Rightarrow polar amino acid)-Phe-Leu-Lys (Auclair *et al.* 2012). A contrasting feature between the Sec and Tat signal peptides is the extent of hydrophobicity in the H-region. The presence of more glycine and threonine residues in Sec-specific peptides makes them more hydrophobic than their Tat counterpart. Furthermore, the C-regions of Sec signal peptides are never charged, with a mean charge of +0.03, while Tat often contains basic residues giving it an overall mean charge of +0.5. Tat signal peptides are also longer than Sec signal peptides, with an average length of 38 and 24 amino acids respectively. This is due to an extended N-region in Tat signal peptides (Tullman-Ercek *et al.* 2007; Blaudeck *et al.* 2003; Fisher *et al.* 2006).

4.2.2.2 Export promiscuity

The data in Chapter 2 suggested that Tat-tagged ITCHY variants might have been exported via the Sec pathway, giving rise to false positives. Tullman-Ercek and colleagues studied the export pathway preference of putative Tat signal peptides of *E. coli* by investigating the localisation of three reporter proteins fused to the signal peptides (Tullman-Ercek *et al.* 2007). The three reporter proteins were short-lived green fluorescent protein (GFP), maltose-binding protein (MBP) and alkaline phosphatase (PhoA). The authors distinguished between Tat and Sec export by localisation of fusions using epitope-tagged full-length proteins. In *E. coli*, there are at least 27 Tat-targeting signal peptides and most of these can direct the export of reporter proteins, via both the Tat and the Sec pathways. It was shown that by increasing the charge in the N-terminus of the mature protein, export via the Sec pathway was avoided, without affecting Tat export. In this study, 11 of

the 27 signal peptides exported reporter proteins only via Tat, eight via Sec and eight through both pathways. The eight signal peptides that were routed via Sec had the typical Tat signature motifs and characteristics. It was found that the correlation between charge and “Sec avoidance” improved significantly when the total charge of the signal peptide C-region was considered together with the first few amino acids of the mature protein. Fusions with an overall net charge of $\geq +2$ showed exclusive Tat export whereas fusions with $+1$ charge gave inefficient export via Sec (Fig. 4.1).

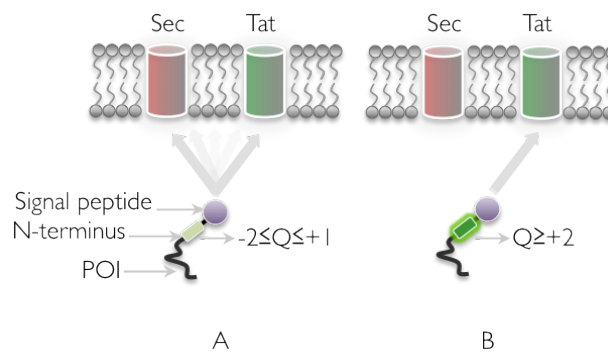


Fig. 4.1. The effect of charge (Q) on export promiscuity. The protein of interest (POI) can be routed either via Sec or Tat by the signal peptide based on the charge density at the N-terminus of the POI. (A) A negative to a $+1$ charge favours translocation via Sec., (B) The presence of a charge greater than $+2$ favours translocation via Tat..

In fact, it was shown that positive charge at the N-terminus of the mature protein alone is sufficient to act as a Sec avoidance signal. A Tat signal peptide MdoD (from glucan biosynthesis protein D) with an uncharged C-region was fused with PhoA and MBP. Site-specific mutagenesis of the first six amino acids of the MdoD mature protein, resulting in a change in the charge from -2 to -1 , had no effect, while increasing the charge to $+2$ completely abolished Sec export of both PhoA and MBP fusions, indicating the effect of positive charge in the N-terminus of the mature protein to avoid the Sec pathway (Tullman-Ercek *et al.* 2007).

With these considerations in mind, a new selection system was developed that increased the positive charge at the N-terminus of the mature proteins. It was hypothesised that this would prevent PRAI-Kv β 2 chimeras from being routed via the Sec pathway, thereby reducing false positives. The new selection system(s) was compared with the pSAlect fold selection system and the results are discussed in section 4.3.6.

4.2.3 Other improvements

4.2.3.1 Improving the desalting step prior to transformation

In many molecular biology protocols, it is desirable, or even essential, to maximise the number of colonies that result from a cloning experiment (i.e. ligation, desalting and transformation). This is especially true for the construction of large libraries in directed evolution experiments. Often, the library screen or selection is very high throughput; it is common to design directed evolution experiments with the capacity to interrogate millions, or even billions, of variants. In these high-throughput cases, it is critical to optimise the library cloning and transformation steps because large and diverse libraries are the most likely to contain variants with improvements in the desired property (Lutz & Patrick 2004; Saraswat *et al.* 2013).

Electroporation is the method of choice for transforming *E. coli* with the products of library ligation reactions. Very high transformation efficiencies can be obtained, resulting in large libraries (Dower *et al.* 1988). However, the high electric field strengths that are used (12-18 kV/cm) mean that efficient transformations require low-conductivity samples, so as to prevent arcing. Therefore, each library ligation reaction must be desalted prior to

electroporation. This purification step is commonly carried out with silica-based microcolumns. A previous study that tested electroporation efficiencies with intact plasmid DNA (rather than with the products of ligation reactions) showed the microcolumn desalting method to be highly effective (Schlaak *et al.* 2005). However, an older study showed that drop dialysis – in which a 5-100 μ L drop is placed on a floating membrane filter – can also result in effective desalting, with DNA recovery rates of 98-99% (Marusyk & Sergeant 1980). In this chapter a rigorous comparison of the two desalting methods is presented and it is shown that drop dialysis is a preferred method for the construction of large libraries.

4.2.3.2 Switching to time-dependent ITCHY

In Chapter 2, THIO-ITCHY was used to make a library of trPRAI-Kv β 2 chimeras. However, it has been predicted, using theoretical models, that ITCHY libraries will have the most unbiased distribution of chimeras if they are created using time-dependent truncation (Ostermeier & Lutz 2003). Both protocols (THIO-ITCHY and time-dependent) begin with the same starting material (a linearised plasmid, with the two genes to be recombined present at each end of the molecule; see Chapter 2, section 2.2.2.1). The points of difference between the THIO and time-dependent ITCHY protocols are:

1. Unlike THIO-ITCHY, time-dependent truncation does not require the use of phosphorothioate dTNPs (α S-dNTPs).
2. In THIO-ITCHY, digestion with exonuclease III (Exo III) is prevented from continuing past the randomly incorporated α S-dNMP, creating an incremental library. In time-dependent ITCHY, small aliquots of the Exo III treated library DNA are removed frequently and quenched by addition to a low pH, high salt buffer.

3. Truncation using α S-dNTPs is less labour intensive when compared to time-dependent truncation, in which multiple time-point sampling is required to attain a comprehensive library. However, time-dependent truncation offers more control over the range of truncation and a higher probability of parental length fusions in the library (Ostermeier & Lutz 2003).

4.2.4 Focus of the chapter

In Chapter 3, a chimeric protein (P25K86) was described. This chapter focuses primarily on the improvements that led to the creation of a new folding selection system, pFoldM. A new library of PRAI-Kv β 2 fusions was also created using plnSAlect instead of pSAlect. This library was created via time-dependent ITCHY and clones from this library were pooled and sub-cloned into the new selection system, pFoldM, to compare with pSAlect. Finally, two interesting proteins were discovered, one of which (P24K89) was characterised biophysically. This protein is closely related to P25K86 and exhibited circular dichroism (CD) spectra and nuclear magnetic resonance (NMR) signals indicative of a partially folded structure. Both proteins are compared in this chapter. Figure 4.2 illustrates the layout of this chapter.

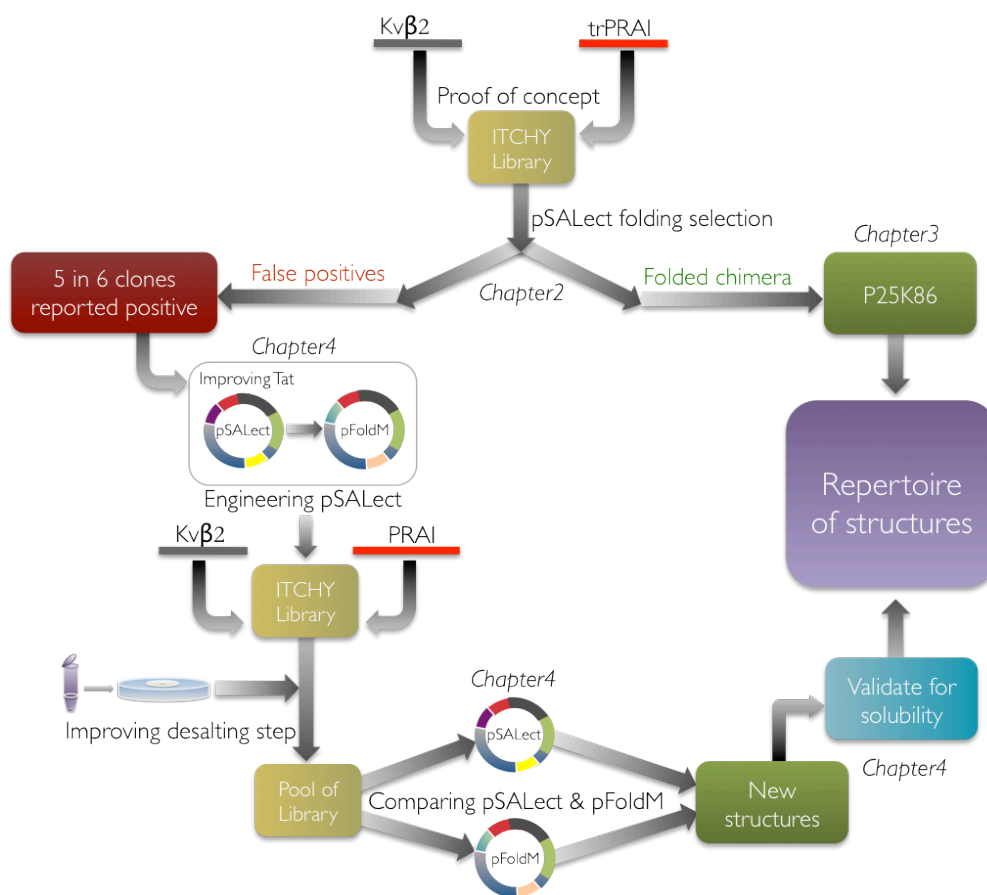


Fig. 4.2. Layout of Chapter 4. In Chapter 2, only one soluble chimeric protein, P25K86, was discovered, along with five other false positives. Chapter 4 involved improving the selection system and ultimately adding new soluble chimeras to the repertoire of structures.

4.3 Results

4.3.1 Attempts to improve the folding selection system

To reengineer Tat and the pSAlect selection system, pSAlect-PRAI was used as the starting template. The pSAlect plasmid encodes the Tat signal peptide of *E. coli* trimethylamine N-oxide reductase (TorA) (Lutz *et al.* 2002). The existing system encodes a signal peptide that lacks three residues (AQA) at its C-terminus. These residues were added back to the Tat sequence, along with positively charged residues at the end of the signal peptide. The existing pSAlect plasmid also contains the chloramphenicol resistance (CmR) cassette. It was also observed in experiments performed by Dr Monica Gerth that replacing this cassette with the kanamycin resistance (KnR) cassette gave more evenly shaped colonies. A two-step strategy to reengineer the pSAlect system was therefore formulated. The first step was to generate an intermediate plasmid, pFoldM, with KnR replacing CmR, and the second step was to reengineer the Tat signal peptide.

4.3.1.1 Step 1 of the two-step strategy to reengineer pSAlect

The KnR cassette was obtained by digesting the plasmid pCM433-KnR (Appendix I) with BglII (Fig. 4.3A). Next, the CmR cassette was removed from the pSAlect-PRAI backbone by whole circle PCR with two primers, one of which contained a BamHI site (Fig. 4.3B). Intramolecular ligation yielded an intermediate plasmid that could still be maintained on ampicillin-containing plates (due to the presence of the Tat-PRAI- β -lactamase fusion). The intermediate plasmid was digested with BamHI and the KnR cassette (from pCM433-KnR) was ligated to it, resulting in pFoldM-PRAI (Fig. 4.3C). The

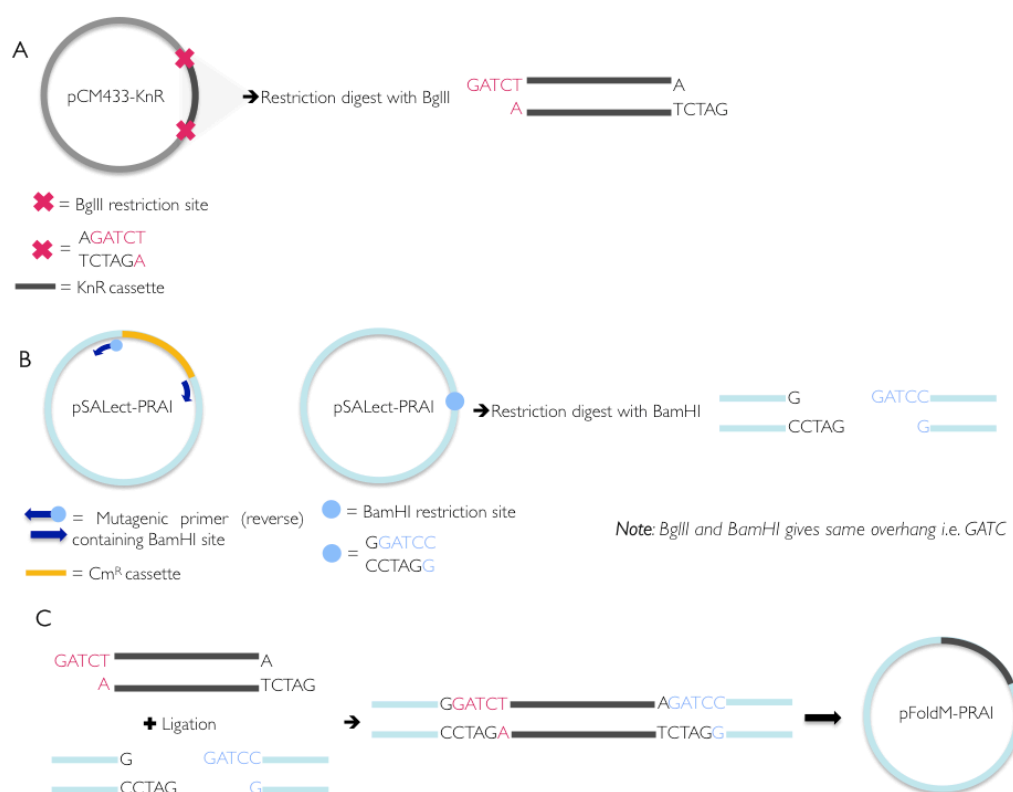


Fig. 4.3. Step I of the strategy to reengineer pSALect. (A) The kanamycin cassette (KnR) was dropped out from the plasmid pCM433-KnR. (B) The chloramphenicol cassette (CmR) was eliminated using whole circle PCR with mutagenic primers containing a restriction site, which upon digestion would create a compatible site to insert the KnR cassette. (C) The KnR cassette was ligated in the intermediate plasmid to give rise to pFoldM-PRAI.

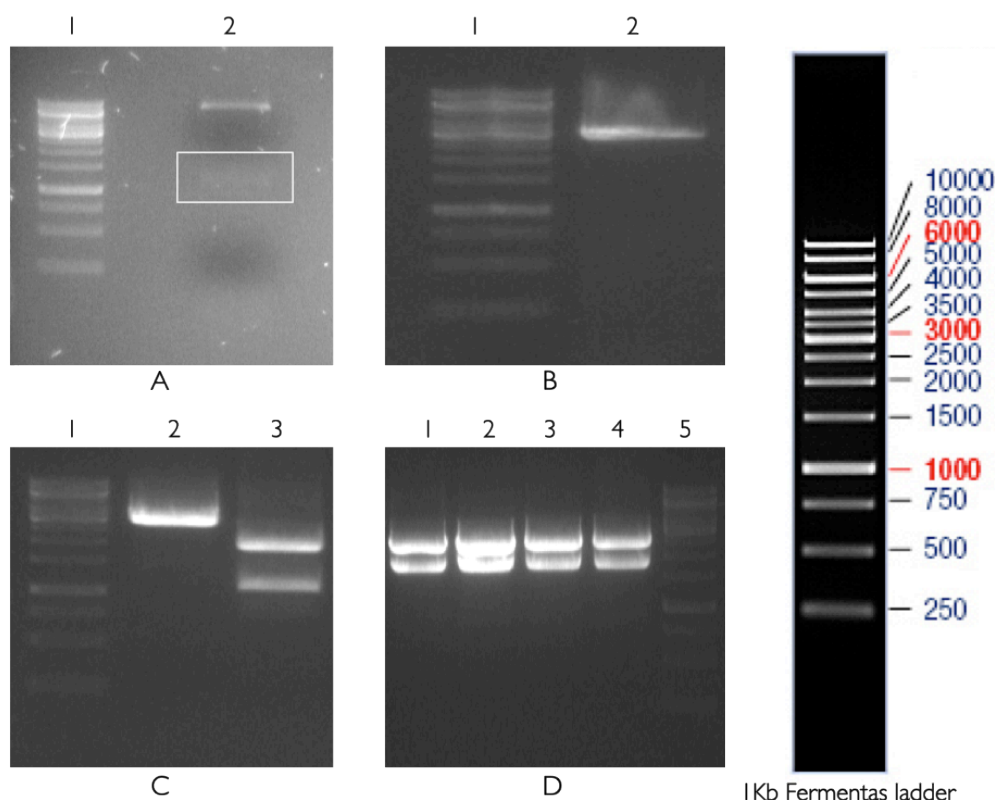


Fig. 4.4. (A) Lane 1: 1 Kb Fermentas ladder; lane 2: restriction digest of plasmid pCM433-KanR (9.3 kb) with BglII. The KnR cassette is 1.2 kb and a faint band corresponding to the size is highlighted with a box. (B) Lane 1: 1 Kb Fermentas ladder; lane 2: a whole circle PCR product with eliminated CmR cassette and newly introduced BamHI site (3.1 kb). The PCR product was self-circularised and ligated, forming an intermediate plasmid. (C) Lane 1: 1 Kb Fermentas ladder; lane 2: to validate the intermediate plasmid, it was linearised with BamHI (3.1 kb); lane 3: a double digest with BamHI and SpeI resulted in fragments of 2 kb and 1.1 kb, which further confirmed the presence of the BamHI site. (D) Lanes 1 to 4: restriction digest of the pFoldM-PRAI plasmids (4.3 kb), which were extracted from four clones after ligating the KnR cassette in the intermediate plasmid and transforming electrocompetent cells. The double digest was done with HindIII and SpeI, which resulted in fragments of sizes 2.6 kb and 1.7 kb; lane 5: 1 Kb Fermentas ladder.

4.3.1.2 Step 2 of the two-step strategy to reengineer pSAlect

To introduce the three residues (AQA) that were missing from the C-region of Tat, as well as positively charged residues at the N-terminus of the mature protein (by using a short sequence KRK and KR, it was hypothesised that it could act as a Sec avoidance signal), two whole circle PCRs with mutagenic

primers were set up (Fig. 4.5B). In the first, pFoldM-PRAI was amplified with a forward primer encoding **Lysine-Arginine-Lysine** (KRK) and a reverse primer encoding AQA (both primers were phosphorylated at their 5' ends). The resulting PCR product was self-circularised (Fig. 4.5C) to form pFoldM-KRK (Fig. 4.6, lane 2). Similarly, pFoldM-KR (Fig. 4.6, lane 3) was formed with a forward primer encoding KR (instead of KRK) and the same reverse primer (encoding AQA). (See section 4.5.1.2.)



Fig. 4.5. Step 2 of the strategy to reengineer pSAlect. (A) Tat signal peptide with positively charged N-terminal “N-region”, the hydrophobic “H-region”, and the polar C-terminal (C-region). The existing pSAlect fold selection system lacks three residues (AQA) and positively charged residues at the N-terminal of the mature protein to avoid routing via the Sec pathway. (B) The intermediate plasmid pFoldM was engineered to introduce AQA and KRK/KR (positively charged residues) using mutagenic primers. (C) The resulting new fold selection system – pFoldM-KRK and pFoldM-KR.

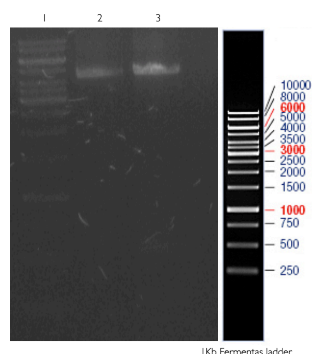


Fig. 4.6. Lane 1: 1 Kb Fermentas ladder; lane 2: PCR product of pFoldM-KRK (~4.3 kb); lane 3: PCR product of pFoldM-KR (~4.3 kb). The PCR products were self-circularised and ligated to form the plasmids pFoldM-KRK/KR.

In pFoldM-KRK/KR, the net charge distribution over the first six residues of the mature protein varies between +1 and +3, while in the existing pSAlect it varies between -0.9 and 0.1 (Table 4.1). An important point to note is that “GENK” is PRAI specific. The new fold selection system was designed with a restriction site *NdeI*, which translates as “HM”, flanking the mature protein. The residues were introduced in accordance with the codon usage in *E. coli* (Sharp *et al.* 1988). The new fold selection systems have a sequence:

- pFoldM-KRK

MNNNDLFQASRRRFLAQLGGLTVAGMLGPSLLTPRRATAAQAKRK[REDACTED]GENK

- pFoldM-KR

MNNNDLFQASRRRFLAQLGGLTVAGMLGPSLLTPRRATAAQAKR[REDACTED]GENK

Table 4.1. Charge distribution in the mature protein.

| Mature protein: First six residues in decreasing order | Charge distribution (pSAlect) at neutral pH | Charge distribution at neutral pH upon adding KRK before H | Charge distribution at neutral pH upon adding KR before H |
|---|--|--|---|
| HMGENK | + 0.1 | + 3.1 | + 2.1 |
| HMGEN | - 0.9 | + 2.1 | + 1.1 |
| HMG | - 0.9 | + 2.1 | + 1.1 |
| HM | + 0.1 | + 3.1 | + 2.1 |
| HM | + 0.1 | + 3.1 | + 2.1 |

4.3.2 Initial tests with pFoldM-KRK and pFoldM-KR

To compare the new folding selection systems with pSAlect, two known chimeras from the trPRAI-Kv β 2 library were sub-cloned into pFoldM-KRK and pFoldM-KR. The two chimeras were P25K86 and P69K149. The former is a highly soluble protein, whereas the latter was a false positive that was insoluble when over-expressed (see Chapter 2, section 2.3.4.2). The expression plasmids pLAB101-P25K86, pLAB101-P69K149 and pFoldM-KRK/KR were digested with NdeI and SpeI and the inserts ligated with pFoldM-KRK/KR, resulting in four new plasmids: pFoldM-KRK-P25K86; pFoldM-KR-P25K86; pFoldM-KR-P69K149; and pFoldM-KR-P69K49 (see section 4.5.2).

It was found that the clones harbouring the plasmid pFoldM-KRK/KR-P69K149 did not grow on the selection plates (LB-carb (100 μ g/mL)), whereas clones harbouring pFoldM-KRK/KR-P25K86 did grow (Fig. 4.7). As P69K149 is an insoluble protein and it came as a pSAlect positive clone, its failure to grow on the selection plate suggested that the new fold selection system might not report false positives.

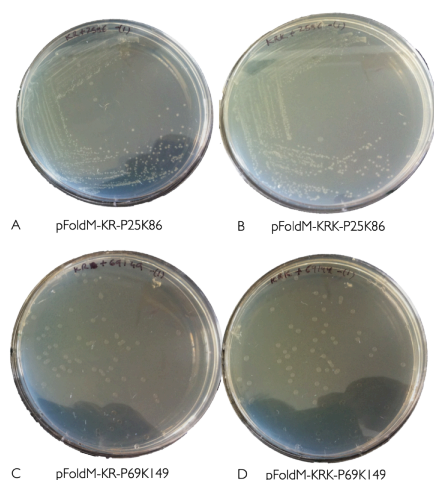


Fig. 4.7. The clones harbouring the plasmids pFoldM-KRK/KR-P25K86 (top; A and B) and pFoldM-KRK/KR-P69K149 (bottom; C and D), were struck on LB-carb (100 μ g/mL) agar plates to test the newly designed pFoldM fold selection system. As P25K86 is soluble while P69K149 not, clones harbouring the pFoldM plasmids with the former insert should grow while the latter should not. (A) pFoldM-KR-P2586 – grew; (B) pFoldM-KR-P2586 – grew; (C) pFoldM-KR-P69K149 – failed to grow; (D) pFoldM-KRK-P69K149 – failed to grow.

Note: Air bubbles in C & D are visual artifacts.

In a further test, the growth of P69K149-containing strains was compared in liquid medium containing various antibiotics (Fig. 4.8). As expected, cells harbouring the plasmid pFoldM-KRK/KR-P69K149 grew in LB-Kan. Clones harbouring the plasmids pLAB101-P69K149 and pSAlect-P69K149, which acted as controls, grew in LB-Carb and LB-Cam respectively. The most interesting finding was when the strains containing pSAlect-P69K149, pFoldM-KRK-P69K149 and pFoldM-KR-P69K149 were grown under the conditions for folding selection (i.e. LB-Carb medium). Clones containing the plasmids pFoldM-KRK-P69K149 and pSAlect-P69K149 grew slightly, suggesting that these systems were allowing some leaky translocation of the insoluble protein, P69K149, via the Sec pathway. The growth of cells containing the plasmid pSAlect-P69K149 was relatively greater than the cells containing pFoldM-KRK-P69K149. Most promisingly, the strain containing pFoldM-KR-P69K149 did not grow, suggesting that pFoldM-KR may be the best system for eliminating false positives (see section 4.5.2).

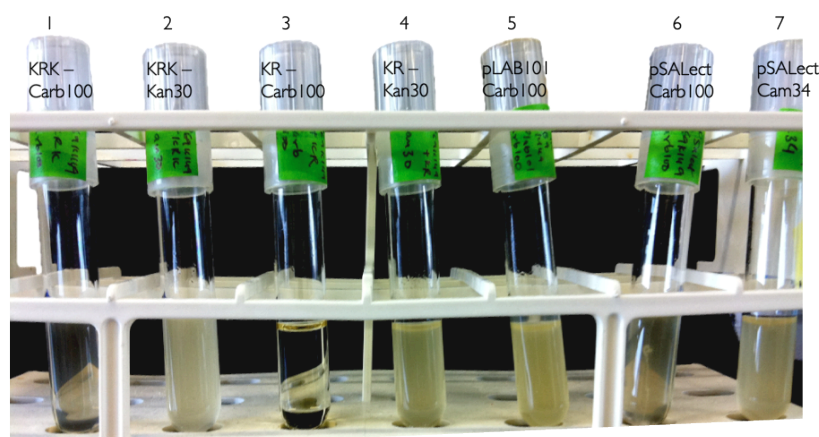


Fig 4.8. Growth of the clones harbouring the plasmids (pFoldM-KRK/KR, pSAlect and pLAB101) with P69K149 insert in liquid medium. The chimera P69K149 expresses as an insoluble protein. Clones harbouring the plasmids pFoldM-KRK/KR with the P69K149 insert were grown in liquid media in the presence of carbenicillin to test and compare the folding quality control feature of Tat in the newly engineered pFoldM selection system with pSAlect. (1) Grew slightly; (2) grew (control); (3) **no growth**; (4) grew (control); (5) grew (control); (6) grew slightly; (7) grew (control).

To further explore the hypothesis that pFoldM-KRK/KR is superior to pSAlect as a folding selection system, one true negative and one true positive protein were tested in three different temperature conditions. The protein cystathionine beta-lyase from *Pelagibacter ubique* strain HTCC1062 (PubMetC) has been studied extensively in our laboratory (Comer 2012). The *metC* gene was chosen as a true negative, as it has been shown to express insolubly unless it is fused to the maltose-binding protein. The *trpF* portion of the bifunctional *E. coli trpCF* gene, coding PRAI, was selected as a true positive because it is highly soluble when it is over-expressed (Patrick & Blackburn 2005). Both genes were cloned into pFoldM-KRK, pFoldM-KR and pSAlect. Each of the six strains were cultured (see section 4.5.2) and ~5000 cells were spread on three agar plates for each culture. The plates were incubated at three different temperatures and colony formation was monitored (Fig. 4.9).

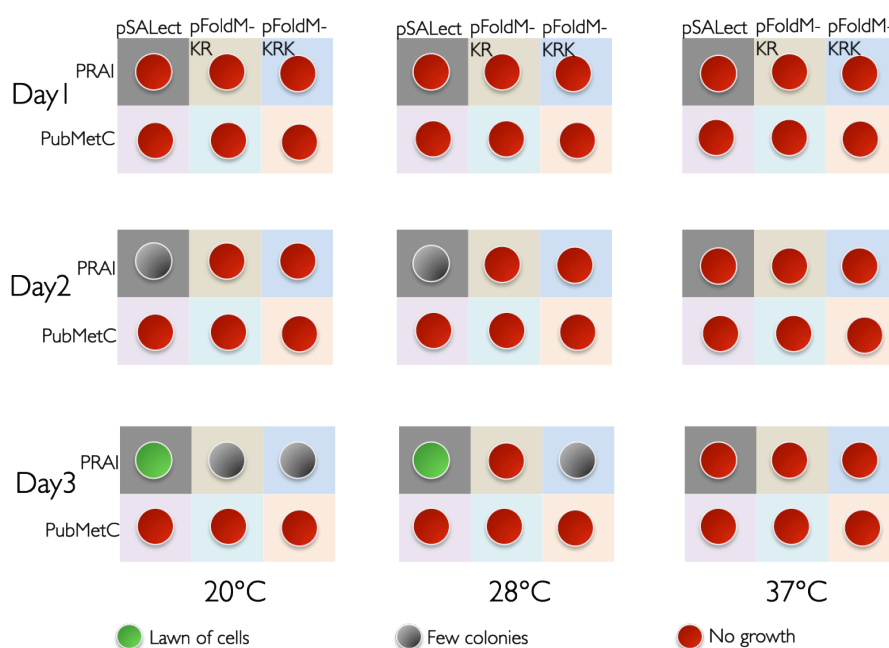


Fig. 4.9. Two known proteins, a true positive (PRAI) and a true negative (PubMetC), were tested for folding selection. Genes encoding PRAI and PubMetC were cloned in pSAlect, pFoldM-KRK and pFoldM-KR. Clones containing the three folding selection plasmids were plated on LB-Carb(100) agar plates and incubated at three different temperatures.

The true positive clone containing pSALect-PRAI grew at both 20°C and 28°C, after two days of incubation, and failed to grow at 37°C. The clones containing pFoldM-KRK-PRAI and pFoldM-KR-PRAI grew at 20°C on the third day after incubation but failed to grow at 37°C, similar to the pSALect fold selection system. At 28°C, the clone containing the plasmid pFoldM-KR-PRAI failed to grow at all, whereas pFoldM-KRK-PRAI grew after three days of incubation. The number of colonies formed by the pSALect-containing strain was 60 times more (1500 colonies; represented as lawn in Fig. 4.9) than the few (25) colonies in both the versions of pFoldM. All clones containing the true negative (encoding PubMetC) failed to grow in all conditions tested. From these results, it appeared that incubation temperature is a critical parameter for modulating the stringency of the folding selection. It also appeared that the pFoldM systems were too stringent; that is, the true positive clone did not give rise to the expected number of colonies. Based on these results (and in contrast to the results in Figs 4.7 and 4.8), pSALect appeared to perform the most reliably.

A new strategy to quantitatively compare pSALect with the pFoldM-KRK/KR folding selection systems was devised. A new ITCHY library was made in the pInSALect vector, and the pool of clones was subsequently subcloned into the three folding selection systems, as shown in Figure 4.10.

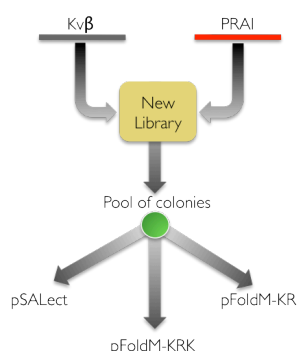


Fig. 4.10. A strategy to make a new library of PRAI-Kvβ2 chimeras was devised. The idea was to sub-clone a single pool of chimeric genes into the three selection systems. This would provide a thorough comparison of the three folding selection systems.

4.3.3 A new library

Using the pInSAlect vector instead of pSAlect, a new single crossover THIO-ITCHY library of PRAI (597 bp) and Kv β 2 (996 bp) chimeras was constructed. First, the *trpF* gene (encoding PRAI) and the Kv β 2 gene were each subcloned into pInSAlect as NdeI/SpeI restriction fragments (see section 4.5.3). Electrocompetent *E. coli* DH5 α -E cells were transformed by the ligated plasmids. Four random clones, two from each plate, were screened (Fig. 4.11) for inserts using primers that anneal in the vector backbone and in the insert (see section 4.5.3).

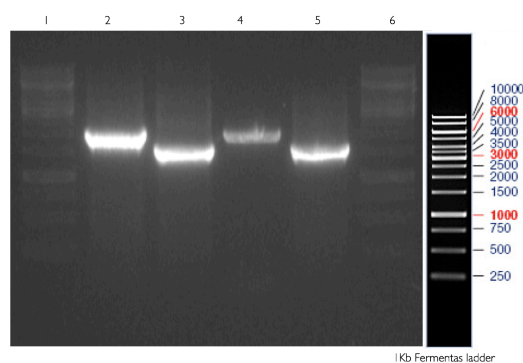


Fig. 4.11. PCR screen of pInSAlect-PRAI and pInSAlect-Kv β 2. Lane 1: 1 Kb Fermentas ladder; lanes 2 and 4: pInSAlect-Kv β 2 (expected size 2 kb); lanes 3 and 5: pInSAlect-PRAI (expected size 1.6 kb); lane 6: 1 Kb Fermentas ladder.

Next, pInSAlect-Kv β 2 and pInSAlect-PRAI were linearised by digesting with NdeI and SpeI respectively. An overlap extension PCR (explained in greater detail in Chapter 2, section 2.3.1.1) to make the recombined fusion product (PRAI-pInSAlect-Kv β 2), spiked with nucleotide analogues (1/8th and 1/10th α S-dNTPs of the total dNTP concentration of 300 μ M), was set up to generate a long PCR product (Fig. 4.12).

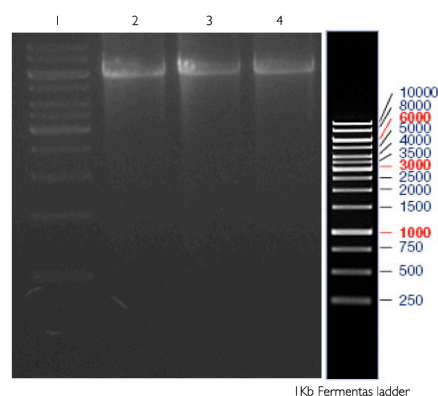


Fig. 4.12. Overlap extension PCR of PRAI-pInSAlect-Kvβ2 fusion (6 Kb). Lane 1: 1 Kb Fermentas ladder; lane 2: fusion product with no αS-dNTPs; lane 3: fusion product with 1/8th αS-dNTPs; lane 4: fusion product with 1/10th αS-dNTPs.

The long PCR product was subjected to the ITCHY truncation protocol and a test library was constructed to estimate the library size and to ensure that crossovers had occurred, by using sequence specific primers (see Chapter 2, section 2.5.3).

4.3.3.1 Test library results – a disappointment

The first THIO-ITCHY library of PRAI and Kvβ2 chimeras was very small. Extrapolating from the test transformation (see section 4.5.3.1), the entire library contained only ~830 variants. The total number of possible variants in a PRAI (597 bp)-Kvβ2 (996 bp) ITCHY library is 594,612, a product of the lengths of the parental genes. The library analysis programme GLUE estimates that in order to sample 95% of all possible variants in this ITCHY library, 1.8×10^6 clones would be required (Firth & Patrick 2008). This suggests that this library represents only 0.1% of the 'all possible' variants. All 16 colonies from this test library plate were picked and screened for inserts with PRAI.for and Kvβ.rev primers (see section 4.5.5). With the exception of one, none of the other clones had the inserts (Fig. 4.13). Two further

attempts to make larger libraries using the THIO-ITCHY protocol also failed. Therefore, an alternative protocol was investigated.

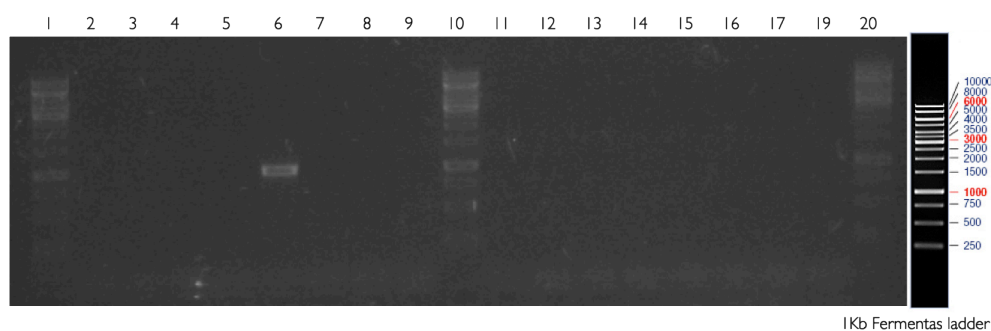


Fig. 4.13. PCR screen of 16 clones from the test library. None of the clones have insert except one (lane 6), with an estimated size of 1 kb. Lanes 1, 10 and 20: 1 Kb Fermentas ladder.

4.3.4 Time-dependent ITCHY of PRAI-Kv β 2

To create a library using time-dependent ITCHY, a long fusion PCR product, as discussed in section 4.3.3, was made. Unlike THIO-ITCHY, the time-dependent truncation protocol does not require the use of α S-dNTPs. The PCR product was treated with Exo III, whose rate of truncation was controlled by the addition of NaCl. Truncation using Exo III is temperature and NaCl dependent. The rate of truncation at 22°C can be varied by using the equation: $\text{rate (bp/min)} = 47.9 \times 10^{(-0.00644 \times N)}$, where N = concentration of NaCl in mM (0-150 mM). The rate of Exo III digestion in this test was ~30 bases/minute, which was achieved by adding 30 mM NaCl to the reaction. This rate equation is valid for a DNA concentration of ~30 ng/ μ L and a ratio of Exo III to DNA of 100 units/ μ g DNA.

Small aliquots were removed frequently from the reaction tube and quenched in another tube, containing buffer PB (5 M GuHCl, 30%

isopropanol) from Qiagen™. Blunt ends were generated with a single-strand nuclease (mung bean nuclease) and a DNA polymerase (T4 DNA polymerase). The library DNA was loaded on an agarose gel and separated for size selection by electrophoresis (Fig. 4.14). Size selection was followed by intramolecular ligation to recylise the plasmid. The recircularised library DNA was then used to transform *E. coli* DH5a-E cells (see section 4.5.6.1).

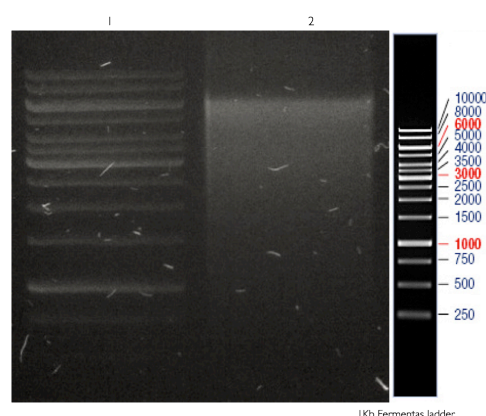


Fig. 4.14. Library size selection on a 0.8% agarose gel. A portion of the gel is cut in the defined size range, in this case between 4.5 and 6 kb.

4.3.4.1 Optimised time-dependent ITCHY protocol

It was speculated that the poor library sizes obtained with the THIO-ITCHY protocol were due to DNA being lost during the post-ligation, pre-transformation purification step. To test this hypothesis, triplicate ligation reactions from the time-dependent ITCHY protocol were desalted with microcolumns and membrane filters (drop dialysis), and compared. A control ligation was heat-inactivated, but not purified further (see section 4.5.7). Figure 4.15 shows the mean number of colonies that resulted when 6.7% of each desalted sample (i.e. 2 μ L of a 30 μ L total volume) was used to transform *E. coli* by electroporation. On average, microcolumn purification

yielded 6.4-fold more colonies than heat inactivation (without further desalting). However, drop dialysis yielded the greatest number of colonies (4.8-fold more than microcolumn purification).

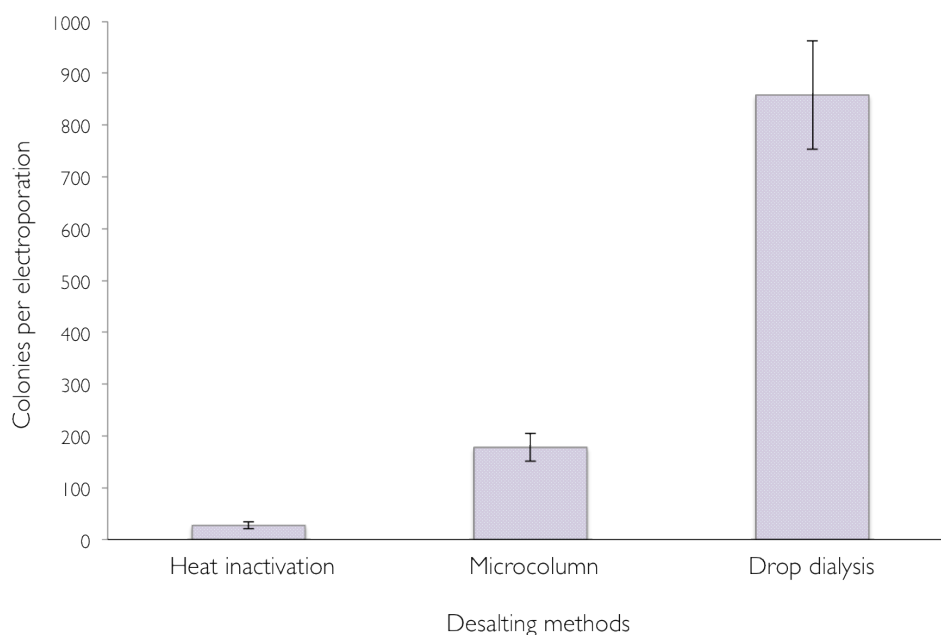


Fig. 4.15. Comparison of DNA desalting methods in the construction of ITCHY libraries. Intramolecular ligation reactions of the library DNA were either heat inactivated only, purified using a silica-based microcolumn, or purified using drop dialysis. Aliquots of the desalted reaction products were used to transform *E. coli* by electroporation and the number of colonies on dilution plates was used to estimate the total number of transformants resulting from each electroporation. Colony counts plotted are the mean \pm SEM from three independent ligation reactions.

Had the desalted samples from the triplicate ligations been pooled, there would have been 84 μ L of each sample (heat inactivated; microcolumn purified; and dialysed) remaining. Transforming more aliquots of electrocompetent *E. coli* with all of this material would have yielded libraries with total sizes of 1.3×10^3 variants (heat inactivated ligations), 8.0×10^3 variants (microcolumn purifications) and 3.9×10^4 variants (drop dialysis). In this example, none of the three desalting methods would have given a library that contained >95% of all possible variants (i.e. 1.8×10^6 clones, as

predicted using GLUE). However, the library from drop dialysis will include ~6.3% of all possible variants, which is certainly preferable to the other alternatives: ~1.3% of all possible variants when microcolumn purification is used; and only 0.2% of all possible variants when the ligation is heat inactivated but not desalted.

4.3.5 A library that pushed things forward

A new pInSAlect-PRAI-Kv β 2 library was constructed using the optimised time-dependent protocol. The protocol yielded 190,000 transformants, which GLUE estimates is likely to contain ~27% of the possible 594,612 variants. As crossovers can occur between any two bases, two thirds of all ITCHY library members will contain frameshifts, which result in premature termination or non - functional progeny. This meant that the library was expected to contain ~63,000 variants that were in-frame (see section 4.5.8).

The pInSAlect-mediated selection for chimeras in the correct reading frame was performed by plating the library on carbenicillin containing agar plates. A number of cells corresponding to ~5-6 times the library size was plated, in order to ensure that every possible clone was represented. In one such selection attempt, ~900,000 cells were plated and ~21,000 clones passed the selection process, giving a survival percentage of 2.3%. This was only 33% of the 63,000 in-frame variants but still an improvement on previous failed attempts (see section 4.3.3.1). The clones from the reading frame selection library were pooled together and several aliquots were made and stored at -80°C. In addition, to check the distribution of crossovers in the library, 12 clones from the pre-selection and 22 clones from post-selection library were sequenced (Fig. 4.16) (see section 4.5.9).

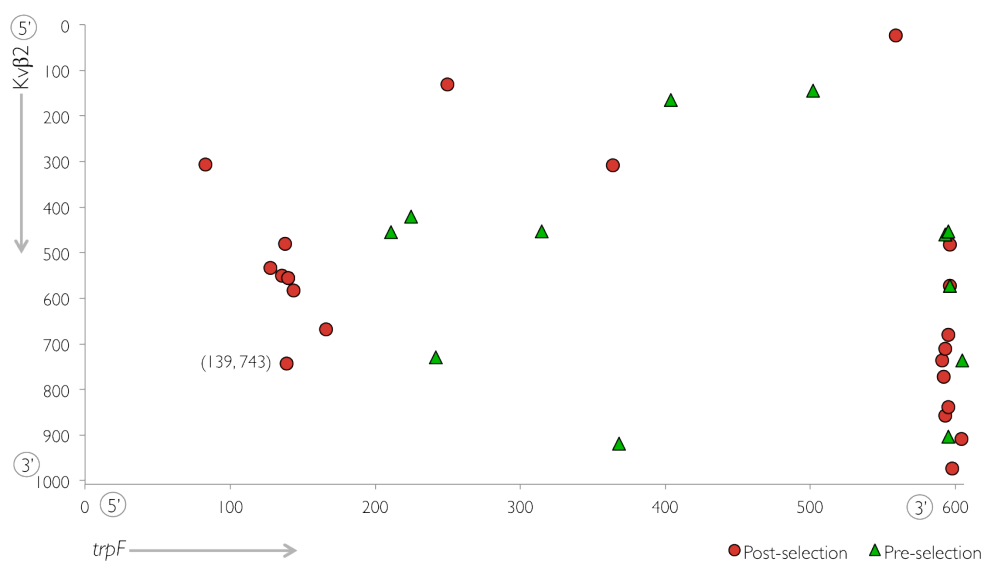


Fig. 4.16. A crossover plot of the distribution of pre-selection (green triangles) and post-selection (red dots) library members in sequence space. On the X-axis is the *trpF* gene (coding PRAI) and on the Y-axis is Kvβ2. The coordinates (x, y) represent (last base of *trpF*, first base of Kvβ2), i.e. if a clone has the coordinate (139, 743), it means that a fragment corresponding to the first 139 bp of *trpF* has been fused to a fragment corresponding to the Kvβ2 gene from its 743rd base to its 3' end.

4.3.6 Comparing the three folding selection systems

This library provided the raw material to compare the pSAlect folding selection system with both versions of pFoldM. First, one aliquot of the pooled, in-frame clones of the library was thawed and plasmids were extracted. The pool of plasmids was digested with NdeI and SpeI to drop out the inserts of variable sizes, which were separated via gel electrophoresis. The vector backbone (pInSAlect) appeared on the gel as a bright band of 4.5 kb and the inserts, due to their variable sizes, as a smear (Fig. 4.17A). The fragments between 300 bp and 1.6 kb were excised and purified. The parent plasmids pSAlect-PRAI, pFoldM-KRK-PRAI and pFoldM-KR-PRAI were digested with NdeI and SpeI and the vector backbones were extracted by separating

on agarose gels. The pool of inserts (chimeras of variable sizes) was ligated into the vectors pSAlect, pFoldM-KRK and pFoldM-KR, as shown in Fig. 4.17B (see section 4.5.11).

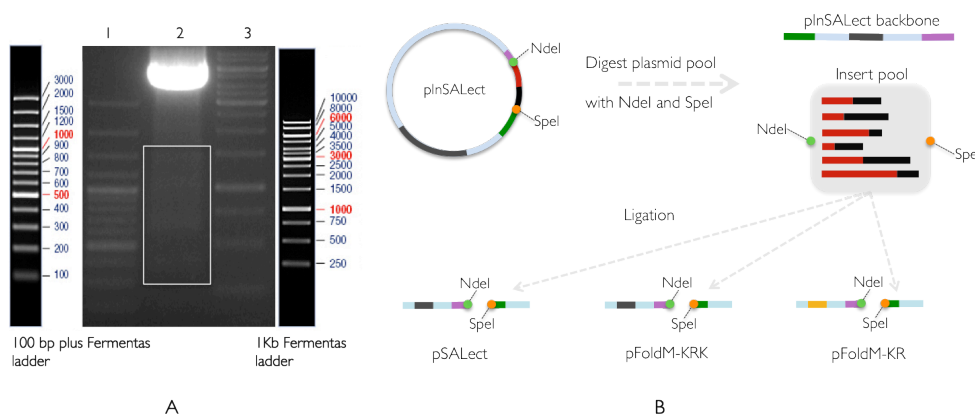


Fig. 4.17. Strategy for subcloning the pool of chimeric inserts into the three folding selection systems. (A) Plasmids from an aliquot of the ITCHY library were digested with NdeI and SpeI. Lane 1: 100 bp ladder. Lane 2: A distinct bright band representing the vector backbone pInSAlect (4.5 kb) can be seen on the gel. The smear indicated by a rectangle represents inserts of variable sizes. The smear in the size range 300 bp to 1.6 kb was excised and purified. The pool of inserts was then ligated into the three fold selection systems. Lane 3: 1 Kb ladder. (B) Graphical representation of the subcloning strategy.

The ligated products were used in a variety of folding selection tests. In each test, a 0.5 μ L sample of a desalted ligation reaction was used to transform a fresh aliquot of electrocompetent *E. coli* cells. Aliquots of the transformed cells were spread on plates containing chloramphenicol in the case of pSAlect, and kanamycin in the case of pFoldM. These were the pre-selection clones. For folding selection, further aliquots of the freshly transformed cells were spread directly on LB-carbenicillin plates. The ratio of colonies on the carbenicillin plate to colonies on the chloramphenicol or kanamycin plate should give an insight into the stringency of the selection for folded chimeras.

4.3.6.1 Is temperature the key to filtering out false positives?

Tests of the three folding selection systems were conducted at two different growth temperatures: 28°C and 37°C. As shown in Fig. 4.18, dramatic differences in survival rates (i.e. in the stringencies of the folding selections) were observed at the two temperatures.

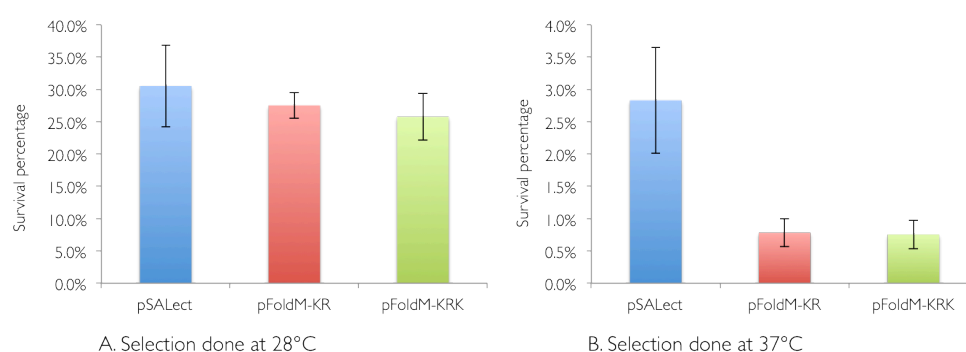


Fig 4.18. Folding selections using pSALect and both versions of pFoldM at 28°C (A) and 37°C (B). Survival percentages plotted are the mean \pm SEM from three independent selection experiments.

The survival rate decreased by 90% in pSALect and by 97% in both pFoldM-KRK and pFoldM-KR when selection was carried out at 37°C instead of 28°C. An interesting observation is that there was no significant difference in survival rates when plates were incubated at 28°C. Also, colonies took 48 h to form at 28°C, whereas at 37°C the colonies appeared after 16 h. Hence, clones that survive the selection at 37°C might contain more stable folded proteins than those selected at 28°C. That is, the survival rate may indicate the true rate of positives (folded chimeras) in the PRAI-Kv β 2 ITCHY library. The selection experiment at 37°C was repeated and showed similar trends as previously (Fig. 4.19). The decrease from pSALect to pFoldM-KR (68%) and the decrease from pSALect to pFoldM-KRK (79%) do not appear to be significantly different. This suggests that in terms of survival rates, pSALect >

pFoldM-KR = pFoldM-KRK. Nonetheless, it is quite apparent that increasing the charge at the N-terminal of the mature proteins improves stringency in selection. Whether this stringency correlates to the sampling of folded structured proteins is still open for debate.

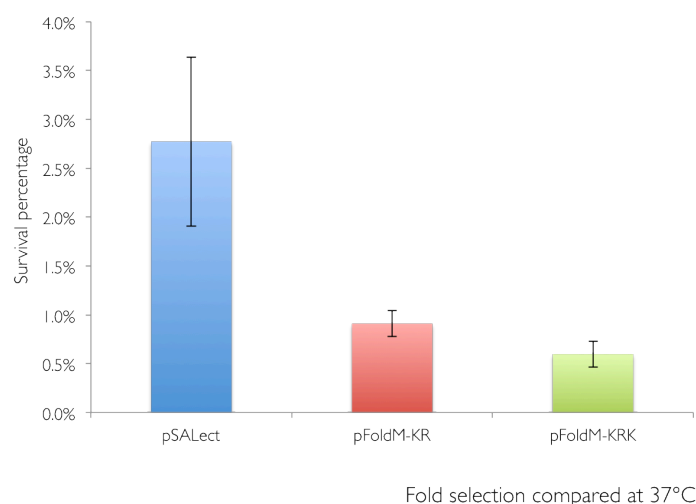
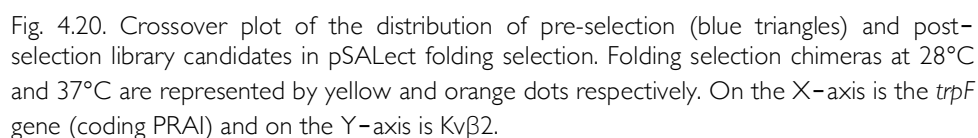


Fig. 4.19 Selection of folded chimeras using pSALect and both versions of pFoldM at 37°C. Survival percentages plotted are the mean \pm SEM from three independent selection experiments.

4.3.6.2 Crossover distribution of PRAI-Kv β 2 in the three-selection systems

Random clones were picked and sequenced, and the crossover plots for all three folding selection systems were generated. The crossover locations in 64, 48 and 43 sequences were analysed from pSALect, pFoldM-KR and pFoldM-KRK respectively (Figs 4.20, 4.21 and 4.22). In terms of sequence crossover diversity, 65% of the sequences from pSALect and 59% from pFoldM-KR are present in pFoldM-KRK. Many of the sequences from each selection system have been sampled more than once during post-selection. In terms of unique crossovers, 53% from pSALect, 47% from pFoldM-KR and



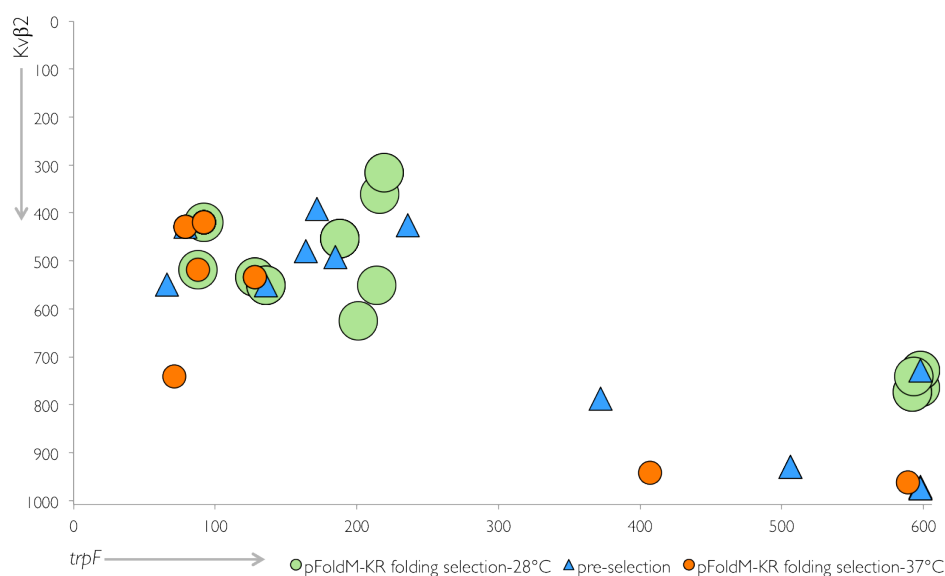


Fig. 4.21. Crossover plot of the distribution of pre-selection (blue triangles) and post-selection library candidates in pFoldM-KR folding selection. Folding selection chimeras at 28°C and 37°C are represented by green and orange dots respectively. On the X-axis is the *trpF* gene (coding PRAI) and on the Y-axis is Kvβ2.

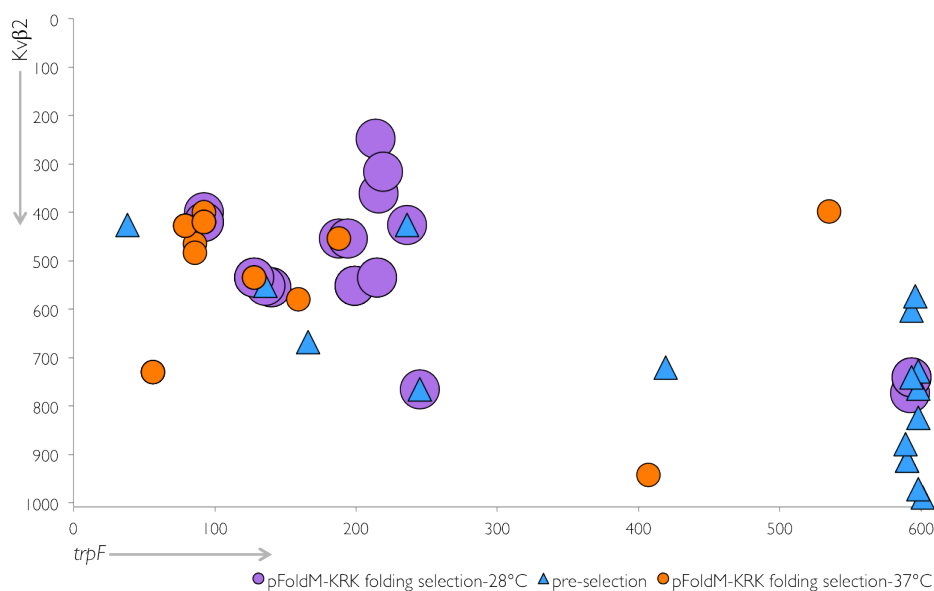


Fig. 4.22. Crossover plot of the distribution of pre-selection (pInSAlect positive – blue triangles) and post-selection library candidates in pFoldM-KRK fold selection. Fold selection chimeras at 28°C and 37°C are represented by purple and orange dots respectively. On the X-axis is the *trpF* gene (coding PRAI) and on the Y-axis is Kvβ2.

4.3.7 An overall ITCHY comparison

Switching from THIO-ITCHY to the time-dependent truncation protocol did improve the library size and provided the raw material to compare the three folding selection systems. However, the pattern of the crossovers (Fig. 4.23) indicated a bias towards longer truncations via time-dependent ITCHY, compared with a bias towards shorter truncations via THIO-ITCHY. Additionally, upon selection, crossovers predominantly occur within 50 bp of the 3' end of *trpF*, or are loosely clustered within 70 to 250 bp of the 5' end of *trpF*, thus forming hotspots in selection space (Fig. 4.24). This spectrum of crossovers was consistent in all three folding selection systems.

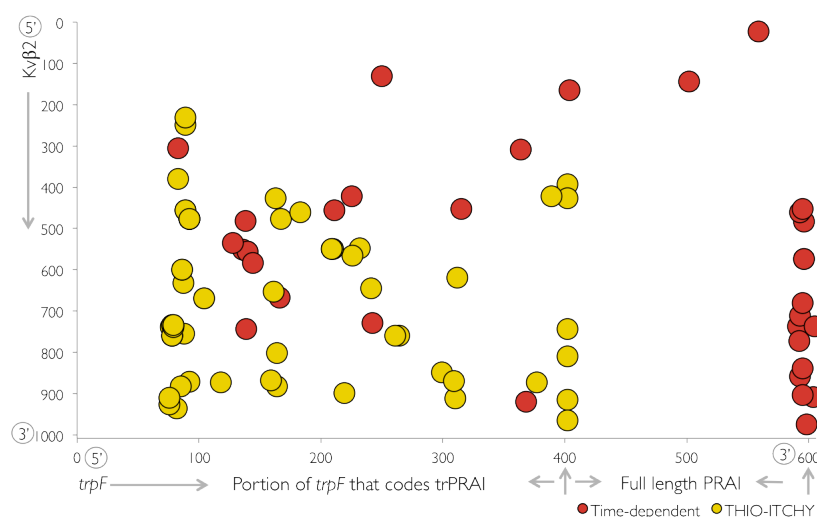


Fig. 4.23. Crossover distribution of PRAI-Kvβ2 chimeras constructed via THIO (yellow circles) and time-dependent (red dots) ITCHY. On the X-axis is the *trpF* gene (coding PRAI and a portion of it, i.e. trPRAI) and on the Y-axis is Kvβ2. These crossovers represent the total of both pre- and post-selection clones.

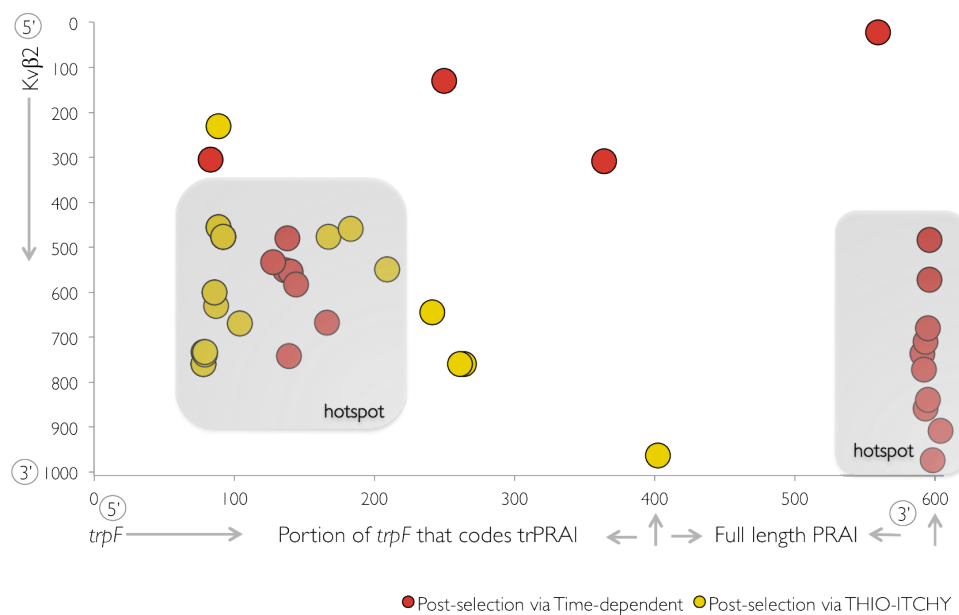


Fig. 4.24. Crossover distribution of PRAI-Kvβ2 chimeras constructed via THIO (yellow circles) and time-dependent (red dots) ITCHY. On the X-axis is the *trpF* gene (coding PRAI and a portion of it, i.e. trPRAI) and on the Y-axis is Kvβ2. These crossovers represent clones from the post-selection library at 28°C.

4.3.8 Solubility validation

In order to compare the pSAlect and pFoldM-KRK folding selection systems, 10 positive clones were picked for solubility validation (Table 4.2). Four out of the 10 clones were selected from both the pSAlect and pFoldM-KRK libraries. Out of these four clones, one clone that dominated the sequence analysis was P31K196 (i.e. 31 amino acids of PRAI fused to 196 amino acids of Kv β 2), with the corresponding gene crossover coordinates (92, 420). This sequence was identified in 23 of the 64 post-selection clones from the pSAlect library, and in five of the 43 post-selection clones from the pFoldM-KRK library. Five other clones selected from the pFoldM-KRK library and one clone that was only pSAlect positive were also chosen (Table 4.2).

Table 4.2. Ten chimeras that were expressed and tested for solubility. The number of times each sequence was sampled during the selection is indicated by its frequency. The total number of selected sequences analysed were 64 from the pSAlect library, and 43 from the pFoldM-KRK library.

| No. | Protein | Frequency pSAlect | Frequency pFoldM-KRK | Temp. selected in °C) | Expression status |
|-----|---------|----------------------|-------------------------|--------------------------|----------------------|
| 1 | P63K185 | 4 | 2 | Both 28 & 37 | Insoluble |
| 2 | P43K158 | 5 | 4 | Both 28 & 37 | Insoluble |
| 3 | P31K196 | 23 | 5 | Both 28 & 37 | Insoluble |
| 4 | P27K193 | 1 | 3 | Only 37 | Insoluble |
| 5 | P82K81 | 0 | 1 | Only 28 | Insoluble |
| 6 | P31K203 | 0 | 2 | Both 28 & 37 | Insoluble |
| 7 | P46K152 | 0 | 1 | Only 28 | Insoluble |
| 8 | P19K93 | 0 | 4 | Only 37 | Soluble |
| 9 | P29K181 | 0 | 1 | Only 37 | Insoluble |
| 10 | P24K89 | 4 | 0 | Only 37 | Soluble |

The 10 chimeras were subcloned into the expression vector pLAB101 (see section 4.5.12). The proteins were over-expressed using IPTG induction, and fractions were run on SDS-PAGE gels. Out of the 10 clones that were

expressed, only two appeared in the soluble fraction. Both of these proteins, P24K89 and P19K93, were identified in selections that were at 37°C. The protein P24K89 was from the pSALect library, while P19K93 was from the pFoldM-KRK library. Another pFoldM-KRK positive from 37°C, P29K181, was insoluble when over-expressed. While the sample sizes remain small, it does seem noteworthy that the only two soluble chimeras were selected at 37°C. While the sample size remains small, this result is consistent with the earlier observations (Fig. 4.18) that performing folding selections at the higher temperature increases stringency, and perhaps removes many of the false positives that are observed in more permissive, lower temperature selections. However, in order to draw any solid conclusions a larger sample size and more solubility validation screens would be required.

On the other hand, pSALect and pFoldM-KRK each yielded one true positive, and several false positives. Overall, there does not seem to be a significant difference in the ability of the two systems to select true positives. It was also noteworthy that the two soluble chimeras (P24K89 and P19K93) were similar in size and sequence to the previously characterised chimera, P25K86 (see Chapter 3). These three chimeras differ from the rest in having a small portion of PRAI and also a narrow range from the C-terminal end of Kvβ2. Furthermore, four clones with small portions of PRAI, i.e. P31K196, P27K193, P31K203 and P29K181, were all found to be insoluble. Another clone P82K81, which contains a portion of Kvβ2, almost similar in size to the soluble chimeras (P24K89, P25K89 and P19K93) but having a higher proportion of PRAI, expresses insolubly. This suggests that in order for these chimeras to be soluble, a portion from the C-terminal end of Kvβ2 and a portion from the N-terminal end of PRAI or subsets thereof may be required.

4.3.8.1 P19K93: a quirky selection

Protein purification trials (see section 4.5.12) indicated that P19K93 was soluble but had poor yields, while P24K89 was soluble and could be purified with high yields (Fig. 4.25). Sequencing of the gene encoding P19K93 revealed that the crossover had occurred within a codon, resulting in an amber stop codon (TAG) at the crossover site. All selection and expression experiments had been performed in *E. coli* strain DH5 α -E, which carries the *supE44* allele. Strains that carry the *supE44* tRNA insert glutamine instead of the amber stop codon (UAG), thereby allowing the translation to continue (Loomis *et al.* 2001). Therefore, in P19K93 the cells produced a full-length protein. Poor protein yields is caused due to competition of the release factor 1 (RF1)-dependent stop codon (UAG) and aminoacylated suppressor tRNA. To study the suppression efficiency in *in vitro* translation systems caused by amber suppression codons, an mRNA encoding esterase (Est2) from *Alicyclobacillus acidocaldarius* that contains catalytically essential serine (serine 155) codon (ACG) was replaced by an amber codon (UAG). It was shown that the increased suppressor tRNA concentration in the cells inhibits protein production (Agafonov *et al.* 2005).

The volume of the P19K93 expression culture was increased to 1 L, in the hope that protein yields would be sufficient for biophysical characterisation. However, the yields of purified P19K93 remained low (Fig. 4.25D). Therefore, no further attempts to optimise the expression and purification of this protein were made and more focus was given to P24K89.

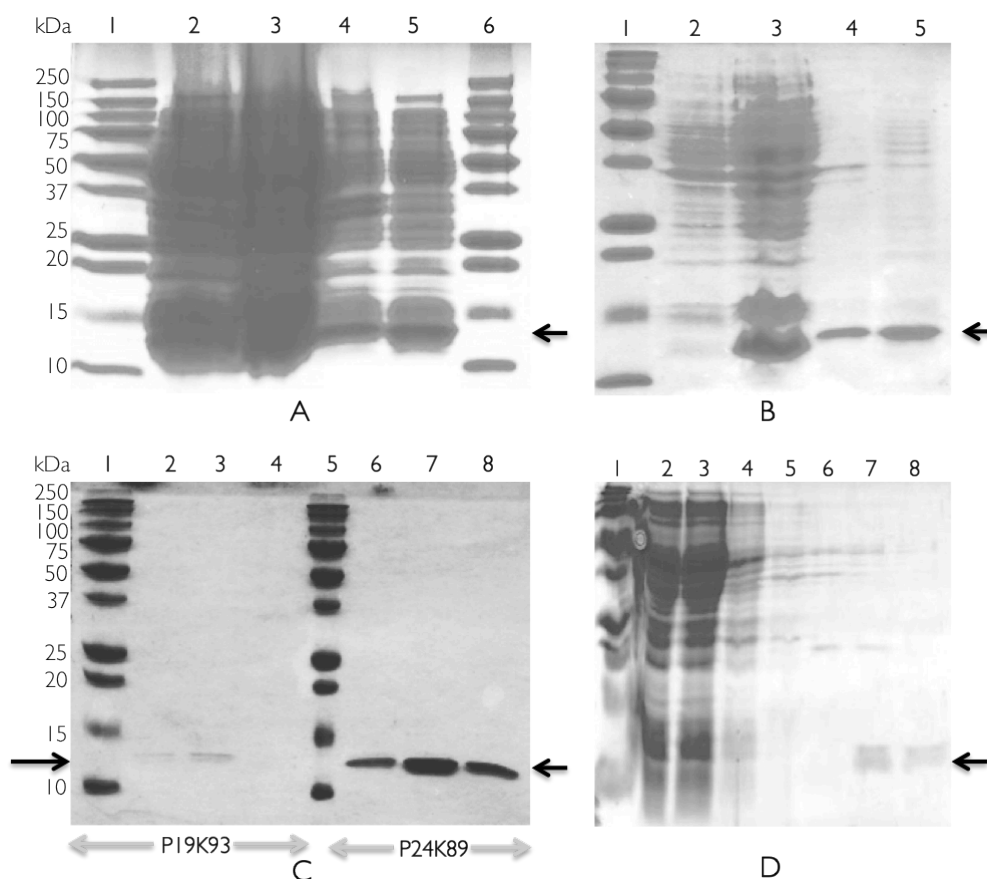


Fig. 4.25. (A) SDS-PAGE gel of P19K93. Lane 1: ladder; lane 2: pre-induction; lane 3: induction after 4 h at 28°C; lane 4: insoluble fraction; lane 5: soluble fraction; lane 6: ladder. (B) SDS-PAGE gel of P24K89. Lane 1: ladder; lane 2: pre-induction; lane 3: induction after 4 h at 28°C; lane 4: insoluble fraction; lane 5: soluble fraction. (C) SDS-PAGE gel of the first three elution fractions of P19K93 (lanes 2, 3 and 4) and P24K89 (lanes 6, 7 and 8). Elution was done with 150 mM imidazole. (D) SDS-PAGE gel of the elution washes of the scaled up culture (1 L) of P19K93. Lane 1: ladder; lane 2: soluble fraction; lane 3: unbound fraction; lane 4: elution with 10 mM imidazole; lane 5: elution with 20 mM imidazole; lane 6: elution with 40 mM imidazole; lanes 7 and 8: elution with 0.5 M imidazole.

4.3.8.2 The protein P24K89

The chimeric protein P24K89 has a molecular weight of 12 kDa and comprises 24 residues from PRAI fused to 89 residues from Kv β 2. It is almost identical to the previously characterised chimera, P25K86 (Fig. 4.26). Most interestingly, P24K89 was purified as a monomer (Fig. 4.25C). This is in contrast to P25K86, which was shown to form disulfide-linked dimers (see Fig.

2.11B). In Chapter 3, it was shown that Cys56 of P25K86 was responsible for dimerisation. Despite the presence of the equivalent cysteine in P24K89 (Cys55), P24K89 was never observed in a dimeric state.

CLUSTAL 2.1 multiple sequence alignment

```

P25K86      MGENKVCGLTRGQDAKAAAYDAGAIYGGRRQAKLQAIARLGCCTLPQLAIACLRNE
P24K89      MGENKVCGLTRGQDAKAAAYDAGAMEG-RRQAKLQAIARLGCCTLPQLAIACLRNE
*****; * *****

P25K86      GVSSVLLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRSTSGHHHHH
P24K89      GVSSVLLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRSTSGHHHHH
*****

P25K86      H
P24K89      H
              *
```

Fig. 4.26. Clustal multiple sequence alignment (Thompson & Higgins 1994) of P25K86 and P24K89. The three cysteines are indicated by green rectangles in P25K86.

4.3.8.2.1 Purification attempts for biophysical characterisation

Experiments to characterise P24K89 were based on those used for P25K86 (see Chapter 3, section 3.3.3.1). In order to obtain labelled protein for NMR experiments, cells expressing P24K89 were cultured in M9 minimal medium containing $[^{15}\text{NH}_4]_2\text{SO}_4$ (see section 4.5.14). The yields of purified protein from these experiments were low (Fig. 4.27). The fractions containing the purest protein (lanes 8, 9 and 10 in Fig. 4.27) were pooled together and concentrated using a Vivaspın concentrator (see section 4.5.13). However, the protein was prone to precipitation and/or adhering to the concentrator's membrane. The final yield after concentration was only 70 μg of purified protein, from a 1 L culture.

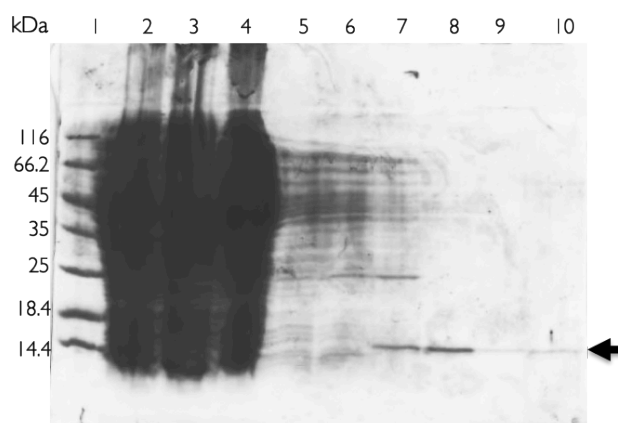


Fig. 4.27. SDS-PAGE gel of P24K89 grown in M9 minimal medium. Lane 1: ladder; lane 2: pre-induction; lane 3: induction after 16 h at 25°C; lane 4: soluble lysate; lane 5: elution with 20 mM imidazole; lane 6: elution with 40 mM imidazole; lanes 7, 8 and 9: elution with 150 mM imidazole; lane 10: elution with 500 mM imidazole.

Faced with such low yields, it was decided to culture P24K89-expressing cells in rich medium (LB) instead. As hoped, this led to higher levels of P24K89 expression (Fig. 4.28). The overall yield of purified protein, from a 1 L culture, was 400 mg (10 mL of protein was purified at a concentration of 40 mg/mL).

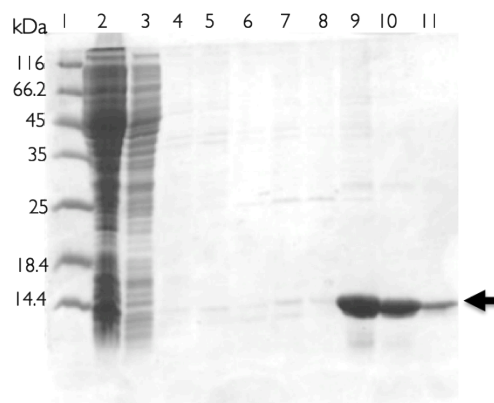


Fig. 4.28. SDS-PAGE gel of P24K89 grown in LB. Lane 1: ladder; lane 2: soluble lysate; lanes 3 and 4: elution with 10 mM imidazole; lanes 5 and 6: elution with 20 mM imidazole; lanes 7 and 8: elution with 40 mM imidazole; lanes 9: elution with 0.5 M imidazole.

4.3.8.2.2 CD spectrum of P24K89

Far-UV circular dichroism (CD) spectroscopy was used to analyse the secondary structure of P24K89 and compare its spectrum to P25K86_CCS (see Fig. 3.12C). The far-UV CD spectrum of P24K89 (Fig. 4.29) displays secondary structure and it is broadly similar to the spectrum of P25K86_CCS. Similar to P25K86_CCS, the protein P24K89 has a higher percentage of α -helices compared to β -sheet (Table 4.3). This is obvious as they both have similar sequences except for I24M, Y26E, G27del and S56C55 (Fig. 4.26). There are two negative bands at 208 nm and 222 nm, and a positive band at 190 nm (Fig. 4.29).

| Method | SELCON3 | CONTIN | CDSSTR | K2D |
|-------------|---------|--------|--------------|------|
| NRMSD | 0.40 | 0.22 | 0.001 | 0.48 |
| Helix | 0.50 | 0.73 | 0.65 | 1.00 |
| Strand | 0.08 | 0.13 | 0.08 | 0.00 |
| Turns | 0.19 | 0.15 | 0.11 | — |
| Disordered | 0.38 | 0.00 | 0.18 | — |
| Random Coil | — | — | — | 0.00 |

Table 4.3. Deconvolution of P24K89, using the DichroWeb application (Whitmore & Wallace 2004). Four algorithms (methods) were used to calculate the amount of secondary structure present, all of which show a similar trend. It is evident from the data that this chimeric protein contains a higher percentage of α -helices over β -sheets no matter what method is used to deconvolute.

Upon deconvolution using the CDSSTR algorithm (Sreerama & Woody 2000; Compton & Johnson 1986) in the DichroWeb application (Whitmore & Wallace 2004), it can be inferred that P24K89 has 65% α -helices, 8% β -sheets, 11% turns and 18% of it remains disordered. However, what is interesting is the higher percentage of α -helices and the lower percentage of β -sheets compared to P25K86_CCS. The turns and disordered region in P24K89 remain more or less similar to P25K86_CCS. This algorithm gave the smallest

NRMSD value, and despite the large range of NRMSD values, all four algorithms gave a good level of agreement in predicting the secondary structure elements present in P24K89. All four algorithms report a high percentage of α -helices over β -sheets. In fact, on averaging SELCON3 and CONTIN the amount of α -helices comes to 62%, which is close to the CDSSTR value. The neural network based algorithm K2D estimated the highest value for α -helices (100%), with no other secondary structure elements present. If this is true, i.e. P24K89 is an all-helical protein, it would mean that the protein is structured with no disordered regions. This is certainly not the case as it becomes obvious from the NMR spectroscopy data (see section 4.3.9) and purification experiments that the protein has only ~30 residues visible and that it precipitates quite rapidly. Overall, it seems that the protein has some secondary structure, with unstructured regions as well.

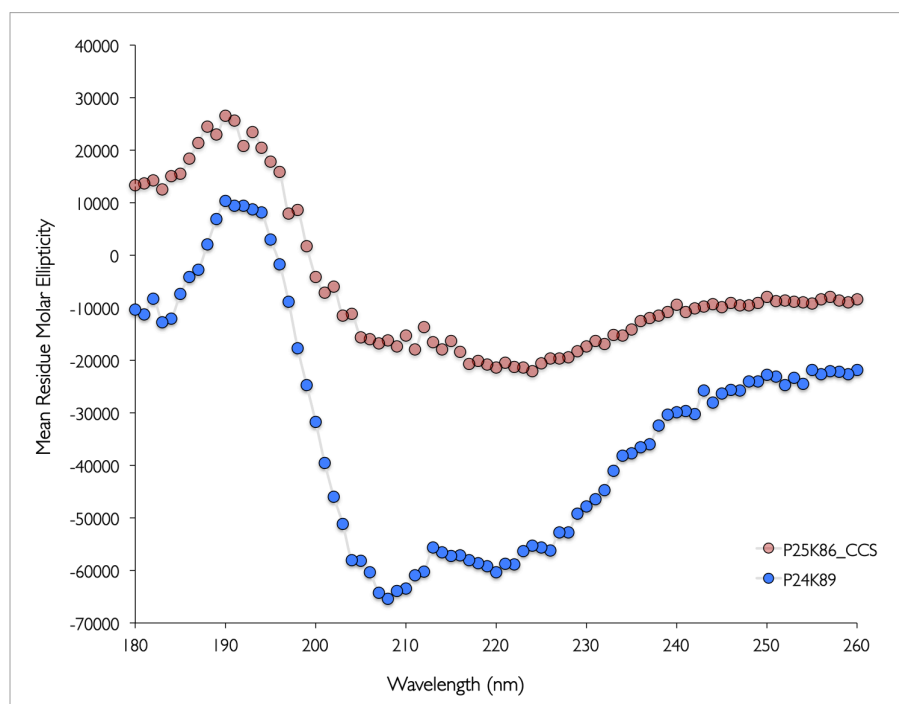


Fig. 4.29 Far-UV CD spectrum of P24K89 (blue) and P25K86_CCS (red) at 20°C. Both proteins have a spectrum that has a higher percentage of α -helices. The trace represents the mean of 10 scans.

4.3.8.2.3 Impediments in NMR trials

As attempts to express P24K89 in M9 minimal medium with $[^{15}\text{NH}_4]_2\text{SO}_4$ resulted in poor yields of protein, obtaining two-dimensional NMR spectra using natural abundance ^{13}C in P24K89 was considered instead. To prepare the sample for NMR experiments, one of the pure fractions of P24K89 (Fig. 4.28, lane 10) was further purified via size-exclusion chromatography (SEC). However, following injection of the protein sample, no signal was detected. This was despite the presence of 0.4 M salt in the buffer (see section 4.5.12). This indicated that the protein might be adhering to the column. In order to test this hypothesis, an organic solvent, acetonitrile, was injected in the mobile phase. Organic solvents such as acetonitrile lower the polarity of the mobile phase, thus increasing the elution strength (Buszewski & Noga 2012). Following the injection of acetonitrile, fractions were collected and the protein appeared in the column wash fractions (Fig. 4.30). This suggested that P24K89 was binding to the column.

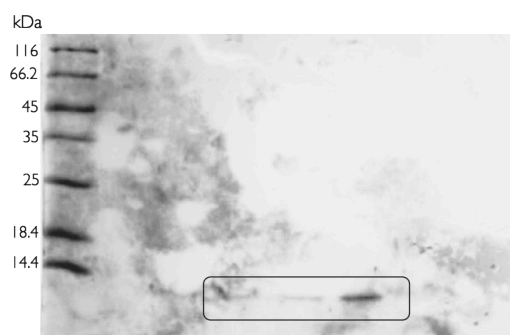


Fig. 4.30. SDS-PAGE gel of P24K89 treated with acetonitrile. As the protein was sticking to the SEC column, on injecting acetonitrile in the column the protein came off (shown inside the rectangular box).

4.3.9 Support from our collaborators

Considering the impediments while labelling both P25K86_CCS and P24K89, it was decided to scale up the cultures (4 L) of both clones to gather two-dimensional NMR spectra using natural abundance ^{13}C . The proteins were purified and supplied to our collaborators in the Chemistry & Biophysics group, Institute of Fundamental Sciences, Massey University, Palmerston North. At the time of writing, Dr Alexander Goroncy has succeeded in collecting HSQC spectra on P25K86_CCS and P24K89, interpretation of this data is reported here.

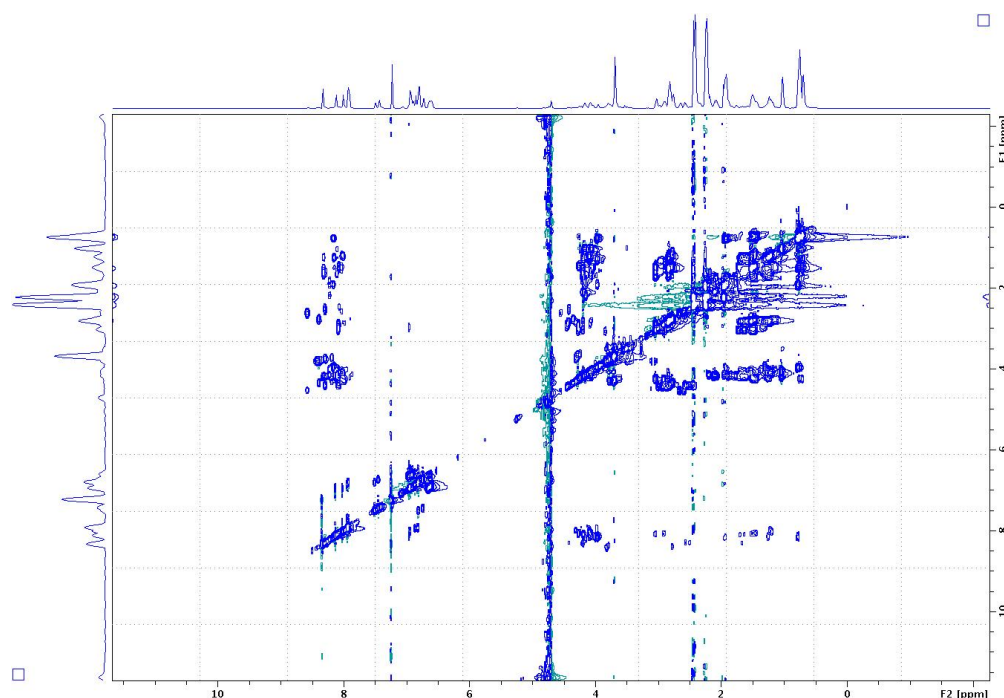


Fig. 4.31. ^1H - ^1H TOCSY of P24K89. On the x-axis is the proton NMR region of the peptide bond (NH) in parts per million (ppm) and on the y-axis is the chemical shift of the protons (αH) in the range 0.5-8.5 ppm.

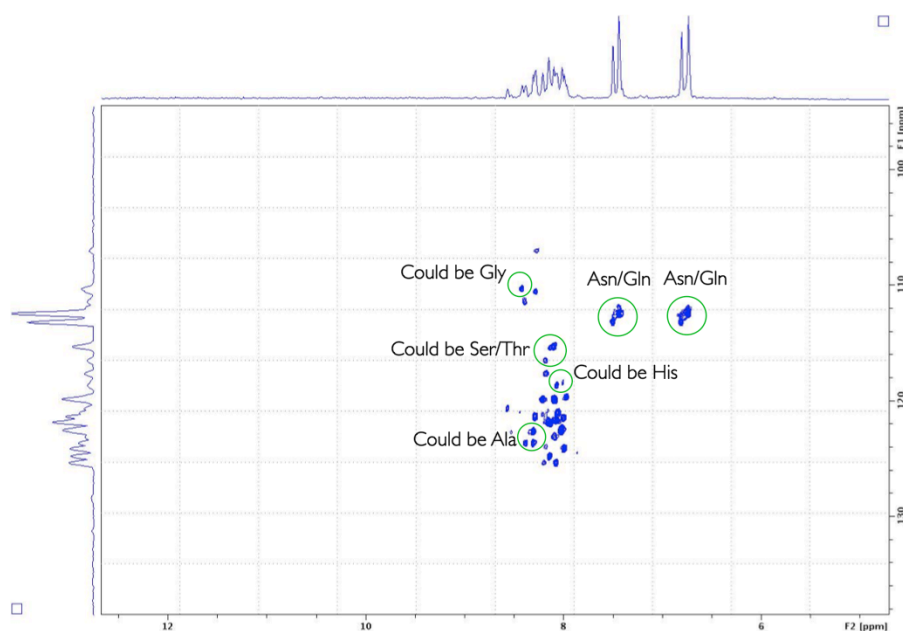


Fig. 4.32. ^{15}N -HSQC of P24K89. On the x-axis is ^1H in parts per million (ppm) and on the y-axis is ^{15}N . Assigned peaks (probable) of seven amino acids are highlighted in green circles.

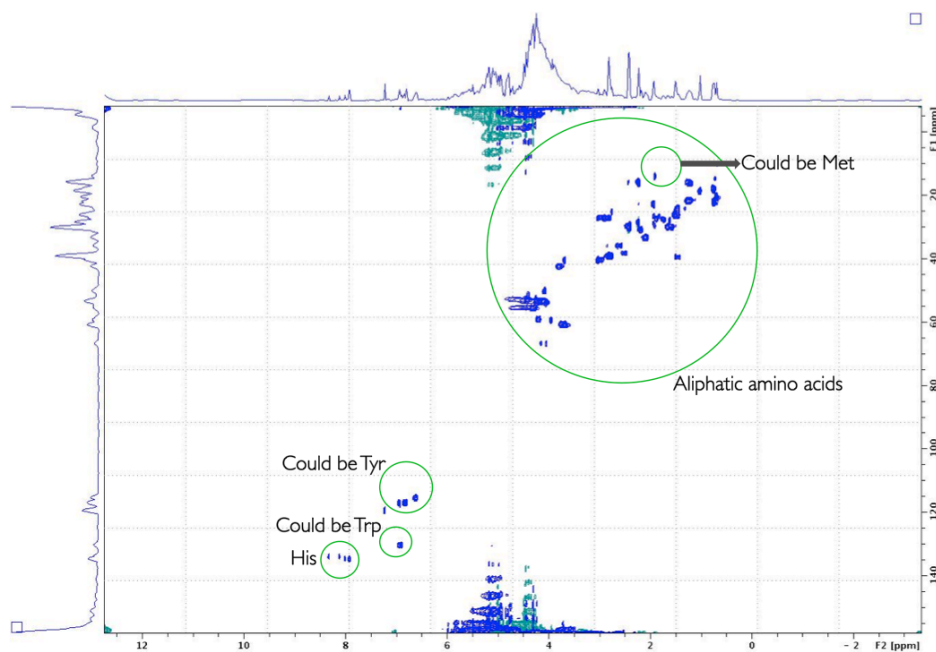


Fig. 4.33. ^{13}C -HSQC of P24K89. On the x-axis is ^1H in parts per million (ppm) and on the y-axis is ^{13}C . Three aromatic amino acid residues and the aliphatic region, along with a probable peak of methionine, which might be visible, are highlighted in green circles.

The fact that the backbone (NH- α H) or the fingerprint region is present in the ^1H - ^1H TOCSY of P24K89 (Fig. 4.31) along with peaks, although broad in the ^{15}N -HSQC (Fig. 4.32) and ^{13}C -HSQC (Fig. 4.33) spectra, suggests that some of the residues (~30) might be visible. The spectrum looks very similar to the spectrum of P25K86_CCS. An attempt was also made to assign certain peaks by using the statistics calculated for selected chemical shifts from atoms in the 20 common amino acids (http://www.bmrb.wisc.edu/ref_info/statsel.htm). Peaks were assigned for Asn, Gln, Gly, Ser, Thr, His and Ala in the ^{15}N -HSQC of P24K89 based on data from average chemical shifts and are circled in green (Fig. 4.32). The peaks for Asn/Gln, Gly, Ser/Thr, Ala are (7.5/6.7, 112/111), (8.2, 109.6), (8.3/8.2, 116.2/115.3) and (8.2, 123.5) respectively. The Histidine-epsilon may be visible at peak (7.9, 119.6). The ^{13}C -HSQC of P24K89 (Fig. 4.33) indicates the presence of few peaks in the aromatic and aliphatic regions of the spectra. The C-epsilon of His at 137.67 and C-delta of Trp at 131 may be visible. There is also a possibility of the C-epsilon of Tyr being visible at 117.9. The nine residues, which might be visible, represent in total 60 amino acids of the 120, which is nearly 50% of the entire chain. However, approximately 30 amino acids were visible by NMR spectroscopy, which constitutes 25% of the chimeric protein. Despite the small amount of structural information gathered in the study, it seems that the protein is only partially structured, which is evident from the CD (secondary structure; primarily α -helical) and NMR (visibility of ~30 peaks; primarily broad) spectrum. Data from this experiment indicate that the protein is present in a molten globule state (for a detailed discussion on molten globule state, refer to section 3.4.1 in Chapter 3).

4.3.10 Comparing P24K89 and P25K86_CCS

The ^{13}C -HSQC spectrum of P24K89 is quite similar to that of P25K86_CCS. This is not surprising, as both proteins have an almost identical amino acid sequence. However, the ^{13}C -HSQC spectrum of P24K89 shows more peaks, suggesting that there are more amino acids visible than for P25K86_CCS. Exactly what stretch of the protein chain these amino acids belong to is hard to predict, but from the overlay of ^{13}C -HSQC spectra of P24K89 and P25K86_CCS (Fig. 4.34) it can be seen that there are more aromatic residues visible in P24K89. In addition, the fact that the ^{15}N -HSQC spectra of P24K89 and P25K86_CCS (Fig. 4.35) are not very similar suggests that the visible regions are not the same.

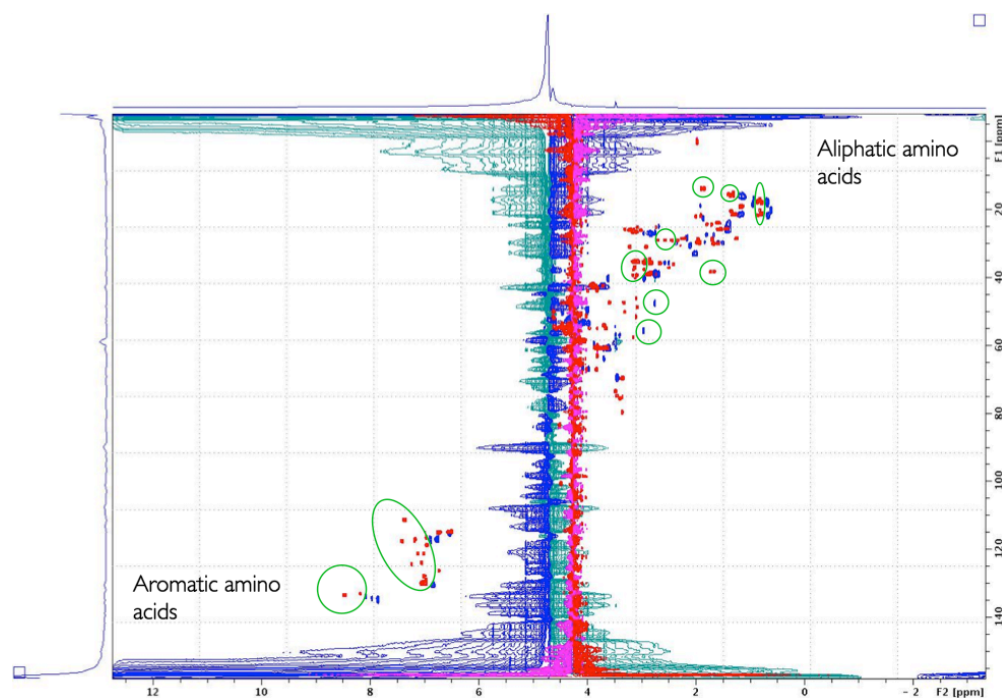


Fig. 4.34. Overlay of the ^{13}C -HSQC of P24K89 and P25K86_CCS. On the x-axis is ^1H in parts per million (ppm) and on the y-axis is ^{13}C . The P25K86_CCS peaks are represented by blue colour whereas P24K89 by red. The spectra for both the proteins look quite similar, however regions of difference are circled in green. These differences are discussed in section 4.3.10. The data was acquired and prepared by Dr Alexander Goroncy.

The presence of few histidine peaks in both the proteins suggests that perhaps the C-terminal region is visible (Fig. 4.33 and Fig. 3.16). In the ^{13}C -HSQC spectrum of P25K86_CCS (Fig. 4.34) few additional peaks can be seen at (2.7, 47) and (2.9, 55), whereas in P24K89 a variety of peaks can be seen around (0.6, 20), (1.5, 12.5), (1.7, 39), (1.8, 14), (2.4, 30), (3, 35) and (7.2, 119). Based on the data by BMRB of average chemical shifts, it can be suggested that perhaps the (1.8, 14) peak (Fig. 4.33) is indicative of the additional methionine (M24). If true, the C-epsilon (17.06 with a standard deviation of 1.4) and H-epsilon (1.89 with a standard deviation of 0.4) might be visible. Furthermore, the aromatic region of P24K89 contains five instead of three histidine-epsilon peaks, which suggests that perhaps P24K89 has more visibility of amino acid residues in the C-terminal region.

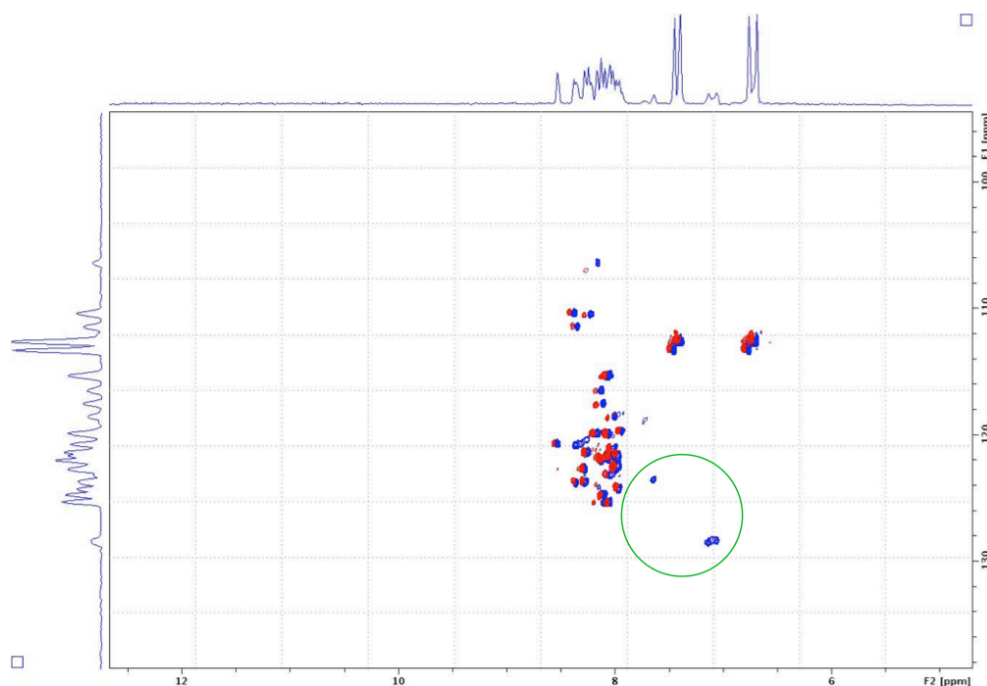


Fig. 4.35. Overlay of the ^{15}N -HSQC of P24K89 and P25K86_CCS. On the x-axis is ^1H in parts per million (ppm) and on the y-axis is ^{15}N . The P25K86_CCS peaks are represented by blue colour whereas P24K89 by red. The spectra for both the proteins look quite similar except the region circled in green. A few peaks that may be visible are annotated for P25K86_CCS and P24K89 and can be seen in Fig. 3.15 (A) and Fig. 4.32 respectively. The data was acquired and prepared by Dr Alexander Goroncy.

These results suggest that both the chimeric proteins are partially structured and are going through conformational exchange between different states. Nonetheless, it seems that P24K89 has more amino acids visible and perhaps different structured regions than P25K86_CCS.

4.4 Discussion

4.4.1 An improved selection for protein folding?

Based on the experimental evidence that only one in six of the proteins identified via the pSALect folding selection system was a true positive (Chapter 2), an investigation to improve the system was carried out. A number of improvements were investigated, ultimately resulting in the discovery of two new PRAI-Kv β 2 chimeras, P19K93 and P24K89.

As attempts to construct a new PRAI-Kv β 2 library via THIO-ITCHY were futile, the use of time-dependent ITCHY did show significant improvements. Switching to the time-dependent truncation protocol yielded a library that contained 230-fold more clones (190,000 vs. 830) than the THIO-ITCHY. This is interesting because the first library that resulted in the discovery of P25K86 was created using THIO-ITCHY and had 52,000 variants, which is nearly four-fold less than achieved here using time-dependent ITCHY. The library described in this chapter was created using the pInSALect vector backbone, time-dependent ITCHY and drop-dialysis to desalt the ligation mixture prior to electroporation. Therefore, the combination of these three optimisation steps helped to push the study forward in terms of finding more interesting chimeras.

To further investigate the effect of the choice of desalting method, a rigorous experiment was conducted to compare the commonly used silica-based microcolumn purification with drop dialysis using a mixed cellulose ester membrane. The data (see Fig. 4.15) showed that drop dialysis is a highly effective method for desalting DNA, confirming earlier results. The

intramolecular ligation of the PRAI-Kv β 2 library DNA yielded ~3-5 times more transformants when desalted by drop dialysis, compared with the more commonly used silica-based microcolumn purification. In this study, a single brand of microcolumn (the EZNA MicroElute Cycle Pure Kit from Omega Bio-Tek) was tested. However, in preliminary desalting tests with intact plasmid DNA (rather than with library ligations), it was shown that this brand of microcolumn and its associated purification protocol yielded identical results to a well-known but more expensive alternative (Qiagen's QIAQuick PCR Purification Kit). Therefore, drop dialysis was identified as the superior protocol for desalting ligation reactions, regardless of the microcolumn to which it is compared (Marusyk & Sergeant 1980; Schlaak *et al.* 2005; Saraswat *et al.* 2013).

Furthermore, a previous study showed that varying the membrane filter, from one with an average pore diameter of 0.01 μ m to one with an average pore size of 0.05 μ m, does not change the effectiveness of the drop dialysis protocol, although the use of membranes with very small pore sizes can increase the time required for complete removal of buffer salts (Marusyk & Sergeant 1980). The membranes used in this experiment (average pore diameter of 0.025 μ m) allowed for rapid dialysis (30 min), while minimising the likelihood that any ligation products were lost. Also, drop dialysis requires less hands-on time than does microcolumn purification. However, desalting via the membrane filter discs does require careful pipette handling as the samples are loaded onto the membrane.

Therefore, the switch from THIO-ITCHY to time-dependent ITCHY, with the use of drop dialysis as the means to desalt DNA during the final steps of the protocol, showed significant improvement in the diversity of the ITCHY library.

However, an interesting pattern of the crossover distribution that fits very well with the theoretical distribution by Ostermeier was also observed. Ostermeier has constructed models for theoretical predictions of the distribution of truncation lengths in an ITCHY library. The models predict that a time-dependent ITCHY library will have the most uniform distribution of crossovers but will have a bias towards parental-length fusions, while THIO-ITCHY libraries will have a bias against longer truncations (Ostermeier 2003). This was clearly seen in the experiments presented in this chapter (see Fig. 4.23). A high proportion of crossovers were accumulated towards the 3' end of the *trpF* and Kv β 2 (time-dependent), while crossovers were biased towards shorter truncations via THIO-ITCHY. Regardless of any putative biases, the THIO-ITCHY protocol failed to yield suitably sized libraries. As a large library of diverse clones is usually required in directed evolution experiments, the choice of time-dependent should not be preferred over THIO-ITCHY, and vice versa, and the researcher should choose whichever protocol yields maximum diversity based on the outcome of the trials.

4.4.2 Comparing the three folding selection systems

Although the survival rates via pSAlect, pFoldM-KR and pFoldM-KRK were not significantly different at 28°C, a dramatic difference in survival rate was seen when selection was carried out at 37°C. Therefore, an increase in the selection temperature alone in the pSAlect system will result in lower survival rates, but if the overall survival rates are seen in the light of positive charge at the N-terminus of the mature protein (pFoldM-KR/KRK), then it becomes evident that the combination of both temperature and charge density is causing the selection to be more stringent. The selection systems pFoldM-KRK and pFoldM-KR, which have a higher positive charge as

compared to pSALect, were found to be the more stringent, in terms of survival percentage.

Interestingly, out of the 10 proteins that were validated for solubility, the two that were sampled at 37°C only, appeared in the soluble fraction. The protein P19K93 is a pFoldM-KRK positive, while P24K89 was selected via pSALect. Given the small sample size, if more clones were sampled from pFoldM-KRK at 37°C, clones like P24K89 (a pSALect positive) could well emerge with other more interesting candidates *via* this system. Considering that P24K89 was sampled *via* pFoldM-KRK, one in five (20%) proteins were soluble *via* pSALect and two in eight (25%) were soluble *via* pFoldM-KRK at 37°C. Although, this is not a significant improvement, it has not been detrimental.

An advanced robotic technique, ESPRIT, can be used to screen ITCHY variants in a more efficient way (An *et al.* 2011). In this technique, a target protein is truncated and is then screened by means of an arraying robot for soluble/folded proteins. A library of truncated variants is subcloned into a high-level expression plasmid, where the POI is sandwiched between an N-terminal hexahistidine tag and a C-terminal biotin acceptor peptide. Clones with both these tags displayed are then validated for solubility by Ni²⁺NTA purification from liquid cultures. Efficient *in vivo* biotinylation of the POI acts as proxy for soluble proteins and these are handled and screened by a robot (Yumerefendi *et al.* 2010).

An experiment could be developed where an ITCHY library of PRAI-Kvβ2 sandwiched between these two tags (hexahistidine and biotin tag) could be screened robotically via the ESPRIT system. Folded protein from this library

could then be compared with pFoldM and pSALect (*in vivo* screening systems).

4.4.3 Comparing and contrasting the three soluble chimeras

The experiments in this study yielded three novel, soluble PRAI-Kv β 2 chimeras: P25K86, P19K93 and P24K89. The lengths of these proteins are 111, 112 and 113 amino acids, respectively, giving them an average molecular mass of 13 kDa. An obvious question then is, are these chimeras the result of a sampling coincidence or are they an outcome of a genuine selection process? It was found that more regions of sequence space were sampled in this study (Figs 4.20-4.22) and it is significant that these three true positives are very closely related clones. All three contain subdomain size fragments from both PRAI and Kv β 2 and they also disfavour larger chunks of both proteins. Looking at the crossover plots it can be seen that crossovers between the two parental gene sequences accumulate towards the 5' and 3' ends of the *trpF* and 3' end of Kv β 2, representing selection hotspots *via* pSALect or pFoldM-KRK. This also suggests that subdomain size portions of both parental gene sequences are preferred when recombined via non-homologous recombination.

During the course of this study, P19K93 was abandoned due to its poor yields (see section 4.3.8.1), which was because the crossover resulted in an amber stop codon. As experiments were done in an *E. coli* strain (DH5 α -E) that inserts glutamine (Gln) instead of an amber stop codon, an experiment could be devised to point mutate TAG (stop codon) to the bona fide CAG (Gln). This experiment might help produce more of this protein and thus enable the exploration of its biophysical attributes as well.

The CD spectrum data for both P25K86_CCS and P24K89 suggests an overall higher percentage of α -helices over β -sheets, with 15-30% remaining disordered. However, the acetonitrile experiment (see section 4.3.8.2.3), in the case of P24K89, where the protein was sticking to the column and addition of DTT and increased salt concentration to detect P25K86 (see section 3.3.1.1) while performing size exclusion chromatography, suggests that both these proteins are less structured. This appears to support the predictions from my data, as there are only ~30 peaks per spectrum, suggesting that only certain regions of these chimeras are visible but that the majority of the proteins remains unstructured. This is an example of a protein in molten globule state, where parts of the protein chain are more organised than others. For a detailed discussion on molten globule state, refer to section 3.4.1.

A big challenge in this study was labelling the proteins for NMR experiments and the continuous lack of growth in minimal media. Some suggestions for future work are therefore addressed here. The chimeric proteins can be co-expressed using a Takara (Clontech, USA) chaperone plasmid set, which constitutes of five plasmids to express multiple molecular chaperones that enable optimal protein expression. Any one of these five plasmids (pG-KJE8, pGro7, pKJE7, pG-Tf2 and pTf16), referred to as the “chaperone team” by Takara (Clontech, USA), can be co-expressed with the target protein or in this case, the chimeras. This can be achieved by first transforming *E. coli* with any one of the chaperone plasmids, followed by re-transformation with the expression plasmid bearing the target protein (P25K86, P24K89 or P19K93). Another method that could be trialed in order to improve the yields in minimal media is to grow the cells in BioExpress®1000 cell growth media from Cambridge Isotope Laboratories, USA. This proprietary media is

considered to have all the necessary substrates (complex mixture of glucose, amino acids, peptides, vitamins, minerals and cofactors) for optimal growth and protein expression. Two further parameters could also be trialed to enhance the yields of these chimeric proteins in minimal media. Firstly, by codon optimising the DNA sequence (PRAI is bacterial but Kv β 2 is mammalian) of these chimeras for expression in *E. coli*. Secondly, by venturing different strains, for example BL21 (DE3) Rosetta® from Merck Millipore, USA. This strain, which is commercially available, is widely used as a host background for protein expression and contains tRNAs with optimised human codons so as to get a better codon usage in *E. coli*.

Combined, these observations suggest that proteins translocating via the Tat pathway need not be fully folded (Rocco *et al.* 2012). Interestingly, it has been shown using the Tat transport machinery in thylakoids (known as cpTat) that the cpTat system can transport unstructured peptides by themselves, or even if they are combined with folded protein domains. It was also found that certain chimeric precursor proteins stalled during translocation, indicating the limitations of the cpTat machinery. The translocation process stalled and even got terminated completely when the length of the protein substrate exceeded 20-30 nm (Cline & McCaffery 2007). As the ITCHY chimeras are flanked at the C-terminal by a folded domain (β -lactamase) and both P25K86 and P24K89 are partially structured, it fits well with the findings of Cline & McCaffery (2007), which showed cpTat can transport certain precursors that contain both folded and unstructured protein. In fact, it was also shown that unstructured peptides as long as 120 residues can also be exported via cpTat. However, the study by Cline & McCaffery (2007) in plant thylakoids together with another study that showed Tat transport of naturally

unfolded FG repeats from the yeast Nsp1P nuclear pore protein are considered “rare exceptions” (Richter *et al.* 2007; Rocco *et al.* 2012).

These attempts to improve the folding selection system have increased my understanding of the *in vivo* solubility screen systems that exist in the biological market. Despite the Sec avoidance strategy, which was implemented in the pSAlect system that resulted in pFoldM-KRK, a leaky selection persists and the development of a leak-proof system is still open for challenge. Nonetheless, using these engineering experiments, ITCHY, as a means of mimicking the subdomain assembly model, was applied *in vitro*. The discovery of two interesting chimeras (P25K86 and P24K89) using high-throughput engineering experiments widens the possibilities of exploring the protein structure space, and perhaps offers close encounters with these *never born proteins* that may be trapped in an ensemble of fluctuating (structured and unstructured) states. Further developments and suggestions are discussed in Chapter 5.

4.5 Materials and methods

All reagents were purchased from Sigma-Aldrich unless stated otherwise. Common molecular biology materials, techniques and primer sequences are described in Appendix I. Specific materials and methods used in this chapter are described in the following subsections.

4.5.1 Construction of pFoldM-KR and pFoldM-KRK

4.5.1.1 Step I of the strategy

A restriction digest with BglII restriction enzyme was performed to drop the kanamycin cassette out of the plasmid pCM433-KnR. The plasmid pCM433-KnR (2.5 µg) was digested with 10 units of BglII (NEB), 1× NEBuffer 3 in a total reaction volume of 30 µL. After incubation at 37°C for 6 h, the reaction was loaded on a 1% agarose gel and electrophoresed at 70-90 v for 30-40 min in 1× TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 8.0). The band corresponding to the KnR cassette (KnR cassette: 1219 bp; vector backbone, 8081 bp) was excised and the DNA recovered with the QIAQuick Gel Extraction Kit (Qiagen). DNA was eluted from each column in 30 µL elution buffer.

In order to introduce the BamHI site and eliminate the CmR cassette from the pSAlect vector, a whole circle PCR was set up with 1× Phusion HF buffer, 0.2 mM dNTPs, 5 µM pFoldM.for, 5 µM pFoldM.rev, 10 ng of template pSAlect-PRAI, 1 units of Phusion polymerase in a final volume of 50 µl. The primers used were phosphorylated at their 5' ends (for primer sequence, refer to Appendix I). The PCR cycling conditions were: 98°C for 30 s; 30 cycles

of 98°C for 10 s, 67°C for 20 s, 72°C for 52 s; and then one final cycle of 72°C for 5 min. The size of the PCR product was 3140 bp. It was purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from the column with 40 µL elution buffer (EB). The purified product was then treated with DpnI (NEB) to digest any of the pSALect-PRAI PCR template. DNA (1 µg) and 20 units of DpnI were incubated in 1x NEB Buffer 4 (final volume of 20 µL) at 37°C for 2 h, followed by heat inactivation at 80°C for 20 min. The digested sample was PCR purified and then a ligation reaction was set up to self-circularise the DNA. Approximately 60 ng of DNA, 1x T4 DNA ligase buffer (Fermentas), and 10 Weiss unit of T4 DNA ligase (Fermentas) were mixed in a final volume of 20 µL and incubated at 22°C for 1 h. The ligase was inactivated by heating at 65°C for 10 min, then DNA was purified using the QIAQuick PCR Purification Kit and eluted from the column with 20 µL EB. A 5 µL aliquot of the DNA was used to transform 50 µL electrocompetent *E. coli* DH5α-E. The cells were allowed to recover from electroporation by adding 500 µL SOC and incubating at 37°C for 1 h. The cells were then spread on agar plates with LB and carbenicillin (100 µg/mL). After overnight incubation at 37°C, colonies were picked and used to inoculate LB medium containing carb-100 (5 mL). The plasmids were extracted using the QIAGEN Plasmid Mini Kit (Qiagen), which acted as an intermediate plasmid. The intermediate plasmid was diagnosed by restriction mapping to check the newly introduced BamHI site in the backbone.

To diagnose, the intermediate plasmid for restriction site (1 µg) was digested with 10 units of BamHI-HF (NEB), 1x NEBuffer 4 in a total reaction volume of 10 µL. For double digest, 10 units of BamHI-HF (NEB) and 10 units of SpeI (NEB), 1x NEBuffer 4, 1x BSA in a final volume of 10 µL was incubated at 37°C for 2 h. The reaction was loaded on a 1% agarose gel and electrophoresed at

70-90 v for 30-40 min and the right size band (linearised plasmid: 3140 bp; double digest: one fragment of 1106 bp and the other 2034 bp) was verified.

Ligation reaction to make pFoldM-PRAI was performed using a three-fold molar excess of insert DNA (58.2 ng of KnR insert; size: 1219 bp) over vector DNA (50 ng of vector; size: 3140 bp), 1× Quick Ligation Buffer (NEB) and 2000 unit of the Quick Ligase (NEB) to a final volume of 21 µL. After 15 min of incubation at 25°C, the reaction was purified using the QIAQuick PCR Purification Kit (Qiagen). DNA was recovered from each column with 30 µL elution buffer. An aliquot (1.5 µL) of the purified ligation product was used to transform electrocompetent *E. coli* DH5α-E (Invitrogen) by electroporation, using a BIO-RAD Gene Pulser II. The parameters used were 2.5 kV, 25 µF, 200 Ω and cuvettes with 0.2 cm gaps (BIO-RAD). The transformed cells were spread on LB-kanamycin (30 µg/mL) agar plates and the number of colonies formed was counted after incubation (37°C, 16 h). To screen the clone for the presence of the KnR cassette, plasmids were extracted from four random clones using the QIAGEN Plasmid Mini Kit (Qiagen) and then plasmids were digested with HindIII and SpeI. For double digest, 10 units of HindIII (NEB) and 10 units of SpeI (NEB), 1× NEBuffer 4, 1× BSA in a final volume of 10 µL was incubated at 37°C for 2 h. The reaction was loaded on a 1% agarose gel and electrophoresed at 70-90 v for 30-40 min and the right size band was verified. The verified clones were stored at -80°C for future use (Appendix I).

4.5.1.2 Step II of the strategy

In order to introduce AQA and increase the positive charge at the N-terminal of the mature protein by the addition of lysine (K) and arginine (R), two PCRs were performed. One to make a +3 version (KRK) and the other a +2 version

(KR). For making KRK, a whole circle PCR was set up with 1x Phusion HF buffer, 0.2 mM dNTPs, 5 μ M pFoldM.KRK.for, 5 μ M pFoldM.AQA.rev, 10 ng of template pFoldM-PRAI, 1 unit of Phusion polymerase in a final volume of 50 μ l. The primers used were phosphorylated at their 5' ends (for primer sequence, refer to Appendix I). The PCR cycling conditions were: 98°C for 30 s; 30 cycles of 98°C for 10 s, 64°C for 10 s, 72°C for 67 s; and then one final cycle of 72°C for 5 min. The size of the PCR product was 4377 bp. For making KR, all conditions were similar except the primer used was pFoldM.KR.for. It was purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted from the column with 30 μ l EB. The purified product was then treated with DpnI to digest any of the pFoldM-PRAI PCR template. DNA (1 μ g) and 20 units of DpnI were incubated in 1x NEB Buffer 4 (final volume of 20 μ L) at 37°C for 2 h, followed by heat inactivation at 80°C for 20 min. The digested sample was PCR purified and then a ligation reaction was set up to self-circularise the DNA. Approximately 60 ng of DNA, 1x T4 DNA ligase buffer (Fermentas), and 10 Weiss unit of T4 DNA ligase (Fermentas) were mixed in a final volume of 20 μ L and incubated at 22°C for 1 h. The ligase was inactivated by heating at 65°C for 10 min, then DNA was purified using the QIAQuick PCR Purification Kit and eluted from the column with 20 μ L EB. A 5 μ L aliquot of the DNA was used to transform 50 μ L electrocompetent *E. coli* DH5 α -E. The cells were allowed to recover from electroporation by adding 500 μ L SOC and incubating at 37°C for 1 h. The cells were then spread on agar plates with LB and kanamycin (30 μ g/mL) or Kan-30. After overnight incubation at 37°C, random colonies were picked and used to inoculate LB medium containing Kan-30 (5 mL), followed by incubating at 37°C for 16 h. Freezer stocks for preservation at -80°C were made. The randomly chosen clones were sequenced by the Massey Genome Service, Massey University, Palmerston North. The primer used for sequencing was pSAlect.for.

4.5.2 Sub-cloning P25K86 and P69K149 in both versions of pFoldM

The P25K86 and P69K149 inserts were dropped out from the expression vector pLAB101 by digesting with NdeI and SpeI and ligated in both versions of pFoldM, which is also digested with similar restriction enzymes. Ligation reactions to make pFoldM-KRK/KR-P25K86 and pFoldM-KRK/KR-P69K149 were performed using a three-fold molar excess of insert DNA (14 ng of P25K86; size: 336 bp and 26 ng of P69K149; size: 657 bp) over vector DNA (50 ng), 1× Quick Ligation Buffer (NEB) and 2000 units of the Quick Ligase (NEB) to a final volume of 21 µL. After 25 min of incubation at 25°C, the reactions were purified using the QIAQuick PCR Purification Kit (Qiagen). DNA was recovered from each column with 30 µL elution buffer. Aliquots (1.5 µL) of the purified ligation products were used to transform electrocompetent *E. coli* DH5α-E (Invitrogen) by electroporation using a BIO-RAD Gene Pulser II. The transformed cells were spread on LB-kanamycin (30 µg/mL) agar plates and the number of colonies formed was counted after incubation (37°C, 16 h). Two colonies from each ligation plate were re-suspended in 10 µL water, from which 5 µL was used as a template for a colony PCR, to screen the inserts, and the rest was used to streak LB-Kan (30 µg /mL) plates. The struck plates were incubated at 37°C for 16 h. After confirming the colonies by PCR screen, a single colony from the LB-Kan (30 µg/mL) plates was re-struck on LB-carbenicillin (100 µg/mL) agar plates for fold selection and incubated at 28°C. In addition, a quick selection experiment to test the growth of the P69K149 chimera in liquid medium, was set up. An overnight culture of pFoldM-KRK/KR-P69K149 in LB-Kan (30 µg/mL) and LB-Carb (100 µg/mL) was incubated in a moving shaker at 37°C. To compare the growth in liquid medium with the pSAlect fold selection system, overnight culture of

pSAlect-P69K149 in LB-Cam (34 µg/mL) and LB-Carb (100 µg/mL) was incubated under similar conditions. A positive control, pLAB101-P69K149 that has the insert in the expression vector, was also grown in LB-Carb (100 µg/mL), incubated at 37°C overnight in a shaking incubator.

To further test the growth of PRAI and PubMetC in pFoldM-KRK/KR and pSAlect, overnight cultures of the clones containing the plasmids pFoldM and pSAlect, with the inserts coding PRAI and PubMetC, were grown in LB-Kan (30 µg/mL) and LB-Cam (34 µg/mL) respectively. Fresh cultures of LB with the appropriate antibiotic (5 mL) were inoculated with 0.2 mL of overnight cultures and growth monitored via optical density measurements of the cultures at 600 nm (OD₆₀₀). All cultures were diluted 100-fold to plate approximately 5000 cells on the LB-Carb (100 µg/mL) agar selection plate. Three agar plates for each culture, for incubating at three different temperature conditions (20°C, 28°C and 37°C), were plated with ~5000 cells.

4.5.2.1 Screening the clones

Randomly chosen clones were screened for the presence of P25K86 and P69K149 inserts by means of colony PCR. A single colony was picked, resuspended in 10 µL of water, and 5 µL of that was heated at 95°C for 5 min. A 1 µL aliquot of the lysed cells was used as the template for colony PCR, in a reaction with 1x GoTaq buffer (Promega), 0.25 mM of each dNTP, 5 µM of each primer (pSAlect.for and Kvβ.rev) and 2.5 units of Taq polymerase (iNtRON), in a total volume of 20 µL. The thermocycling conditions were: 94°C for 2 min; 30 cycles of 94°C for 10 s, 58°C for 20 s, 72°C for 70 s; and one final cycle of 72°C for 5 min.

4.5.3 Sub-cloning PRAI and Kv β 2 in the vector backbone pInSAlect

PRAI and Kv β 2 inserts were dropped out from their parent vector pSAlect by digesting with NdeI and SpeI and ligated in pInSAlect, which is also digested with similar restriction enzymes. Ligation reactions to create pInSAlect-Kv β 2 and pInSAlect-PRAI were performed using a three-fold molar excess of insert DNA (32 ng of Kv β 2; size: 996 bp and 20 ng of PRAI; size: 597 bp) over vector DNA (50 ng), 1 \times Quick Ligation Buffer (NEB) and 2000 units of the Quick Ligase (NEB) to a final volume of 21 μ L. After 25 min of incubation at 25°C, the reactions were purified using the QIAQuick PCR Purification Kit (Qiagen). DNA was recovered from each column with 30 μ L elution buffer. Aliquots (1.5 μ L) of the purified ligation products were used to transform electrocompetent *E. coli* DH5 α -E (Invitrogen) by electroporation, using a BIO-RAD Gene Pulser II. The transformed cells were spread on LB-Cam (34 μ g/mL) agar plates and the number of colonies formed was counted after incubation (37°C, 16 h). The colonies were screened for the presence of the desired inserts by colony PCR with one primer specific to the pInSAlect backbone (pSAlect.for; Appendix I), and another specific to the 3' end of the inserts (Kv β .rev). The verified clones were stored at -80°C for future use.

4.5.3.1 Constructing the test library

The long PCR product was subjected to ITCHY protocol and a test library was constructed to estimate the coverage or library size and to ensure that crossovers had occurred, using sequence specific primers (see Chapter 2, section 2.5.2). For constructing the test library, 2 μ L of the library DNA was transformed in a 50 μ L aliquot of electrocompetent *E. coli* DH5 α -E cells. After

electroporation, the cells were recovered in 0.5 mL SOC medium followed by incubation at 37°C for 1 h and then various volumes of the cell suspension was plated on LB-Cam (34 µg/mL) agar (pre-selection). Two volumes of the cell suspension, 50 µL and 100 µL, were plated on LB-Cam (34 µg/mL) agar plates. After incubating at 37°C for 16 h, the 50 µL plate had five colonies, while the 100 µL had only 16 colonies. The colony forming units (cfu) in this particular library was 0.1 cfu/µL of the cell suspension plated, i.e. if all the recovery culture (550 µL) were to be plated, a total of 55 colonies would have formed.

4.5.4 THIO-ITCHY library

For a detailed description of the THIO-ITCHY protocol, refer to Chapter 2, section 2.5.2.

4.5.5 Screening the clones

Randomly chosen clones from the test library and the scaled-up library were screened for the presence of hybrid PRAI-Kvβ2 inserts by means of colony PCR. A single colony was picked, resuspended in 10 µL of water, and heated at 95°C for 5 min. A 1 µL aliquot of the lysed cells was used as the template for colony PCR, in a reaction with 1x GoTaq buffer (Promega), 0.25 mM of each dNTP, 5 µM of each primer (PRAI.for and Kvβ.rev) and 2.5 units of Taq polymerase (iNtRON), in a total volume of 20 µL. The thermocycling conditions were: 94°C for 2 min; 30 cycles of 94°C for 10 s, 58°C for 20 s, 72°C for 70 s; and one final cycle of 72°C for 5 min. The PCR products were loaded on a 1% agarose gel and electrophoresed at 90 v for 30-40 min in 1×

TAE buffer. After this, the PCR products with the inserts (chimeras of variable sizes) were sent for sequencing (Appendix I).

4.5.6 Time-dependent ITCHY library

The time-dependent ITCHY was performed by following the optimised protocol.

4.5.6.1. Steps of the optimised protocol

1. Linearise the plasmids pInSAlect-Kv β and pInSAlect-PRAI by digesting with NdeI and SpeI respectively.
2. Set up an overlap extension PCR using 10 ng of the linearised plasmids.
3. Digest the PCR product with DpNI, followed by PCR purification and estimate the concentration (140 ng/ μ L).
4. Take 4 μ g of PCR purified product, dilute in 1 \times NEBuffer 1 and 30 mM NaCl to a final volume of 120 μ L (Tube-R/Reaction tube).
5. Equilibrate 300 μ L of PB buffer (Qiagen: 5 M GuHCl, 30% isopropanol) on ice in a tube (Tube-Q/Quenching tube).
6. Immediately take 20 μ L from tube-R and dilute in 140 μ L of PB buffer (Qiagen). This acts as T = 0 s control (Tube-A).
7. To the remaining 100 μ L in tube-R, add 3.5 μ L of Exo III (100 U/ μ L) and maintain the temperature of the tube at 22°C (a PCR machine can be used to maintain the temperature).
8. At 30 s intervals remove 1 μ L of reaction mix from tube-R and add to the quenching tube-Q. Continue for 40 min until 80 μ L of the reaction mix has been transferred/quenched.

9. The remaining 20 μL ($T = 40$ min control) from tube-R is transferred to 140 μL of buffer PB (Tube-B).
10. Clean up $T = 0$ s (A), $T = 40$ min (B), truncated DNA (Q) by PCR purification and elute in 30 μL (tubes A and B) and 50 μL (tube Q) respectively.
11. The truncated library DNA (50 μL) is diluted in $1\times$ mung bean nuclease buffer and 4 units of mung bean (10 U/ μL) is mixed thoroughly. Incubate at 37°C for 30 min and follow with PCR purification, eluting in 65 μL EB.
12. Dilute 64 μL of library DNA in $1\times$ T4 DNA polymerase buffer (NEBuffer 2) with 2.5 mM dNTPs and 1.5 units of T4 DNA polymerase (3 U/ μL) in a final volume of 100 μL . Incubate at 12°C for 20 min and quench the reaction by EDTA (10 mM) and heating to 75°C for 20 min.
13. Load the reaction on 0.8% agarose gel for size selection by separating via electrophoresis (see Fig. 4.14).
14. Cut the appropriate size band (in this case between 4.5 kb and 6 kb; the size of the vector pInSAlect being 4.5 kb) and purify using the QIAQuick Gel Extraction Kit (Qiagen) following the manufacturer's guidelines and eluting in 60 μL of elution buffer. Estimate the concentration and set up ligations.
15. To set up self-circularisation (intramolecular ligation) of the library DNA, maintain an overall DNA concentration of <3 ng/ μL . Incubate the ligations at 16°C for 16 h, followed by heat inactivation at 65°C for 10 min.
16. Desalt the ligations by using drop dialysis instead of the routinely used microcolumns. Drop dialysis was found to yield 3-5 times more transformants than microcolumn purification (Saraswat *et al.* 2013).
17. The DNA is now ready to be transformed into the desired host.

4.5.7 Column versus drop dialysis

A test time-dependent ITCHY library was constructed using the protocol as outlined in section 4.5.6.1. In the final stage of library construction, three identical intramolecular ligations (90 μ L) were set up, each of which comprised: blunt-ended DNA from the ITCHY protocol (180 ng); 1 \times Fermentas T4 DNA ligase buffer; and 30 units of T4 DNA ligase (Fermentas). The ligation reactions were incubated at 16°C for 16 h, and then heat inactivated (65°C, 10 min). After heat inactivation, each of the three reactions was split into three 30 μ L aliquots. One aliquot was desalted using a microcolumn (EZNA MicroElute Cycle Pure Kit; Omega Bio-Tek) according to the manufacturer's guidelines. The desalted DNA was eluted from the column with 30 μ L elution buffer (10 mM Tris, pH 8.5). The second aliquot was desalted using drop dialysis. A standard petri dish was half-filled with 30 mL deionised (Milli-Q) water. A mixed cellulose ester membrane filter (MF-Millipore, 0.025 μ m pore size, 25 mm diameter) was floated on the water. The 30 μ L aliquot of the ligation reaction was pipetted onto the membrane, covered with the lid of the petri dish, and left to dialyse for 1 h. After dialysis, the desalted sample was recovered from the top of the membrane and the volume of the sample was adjusted to 30 μ L with water. The third 30 μ L sample from each ligation was not desalted further. Aliquots (2 μ L) of each desalted library ligation, and the heat inactivated controls, were used to transform 50 μ L aliquots of *E. coli* DH5 α -E (Invitrogen), using electroporation. SOC medium (500 μ L) was added to each aliquot of cells immediately after pulsing. The transformed cells were allowed to recover at 37°C with shaking, for 1 h, before aliquots were spread on LB-agar plates containing chloramphenicol (34 μ g/ml). Colonies were counted after 16 h incubation at 37°C.

4.5.8 Scaling up and harvesting the library

For constructing the big library, electrocompetent *E. coli* DH5 α -E cells were transformed by the re-circularised plasmids (1 μ L), and in total 25 transformations were pooled together and plated on LB agar containing chloramphenicol (LB-Cam34 μ g/mL). Pre-selection was done on LB agar containing chloramphenicol (Cam-34 μ g/mL), while frame selection was done on carbenicillin (Carb-100 μ g/mL) containing plates. The colonies from the big plate were picked and screened for inserts as discussed in section 4.5.5. To make -80°C freezer stocks of the entire pre-selection library for future use, the clones were scraped off with a glass spreader using 20 mL of LB chloramphenicol liquid media and harvested. After harvesting, the cells were pelleted by centrifuging at 3000 g at 4°C for 15 min, and the supernatant removed. The pelleted cells were then resuspended in a smaller volume of LB chloramphenicol liquid media and aliquoted for freezer stocks. While making freezer stocks of the pre-selection library, a 1000-fold dilution of the resuspended library was made and its optical density at 600 nm was measured to record the number of cells present in the library.

4.5.9 Frame selection

For the selection of in-frame chimeras via plnSALect, an aliquot of the frozen pre-selection library was used to inoculate 40 mL of LB-Cam (34 μ g/mL) liquid media and grown at 37°C until it reached an optical density (OD) of 0.4. The optical density was measured using the BioPhotometer (Eppendorf). Cells that represented six times the pre-selection library coverage were then plated on LB agar containing carbenicillin (Carb-100 μ g/mL). The plates were then incubated at 28°C for 24 h. Random clones from the selection plate

were picked for sequencing in order to determine the spectrum of crossover locations in the trPRAI - Kv β 2 library.

4.5.10 Sequencing

Randomly chosen clones from the pre-selection library, and those that survived the frame selection, were sequenced by the Massey Genome Service, Massey University, Palmerston North. The primer used for sequencing was PRAI.for (Appendix I).

4.5.11 Comparing the fold selection system

An aliquot of the in-frame pooled library was thawed and plasmids were extracted by using E.Z.N.A.® Plasmid Mini Kit (Omega). The pool of plasmid was digested with 20 U of NdeI and SpeI (NEB), 1× BSA, 1× NEBuffer 4 in a total reaction volume of 40 μ L. After incubation at 37°C for 6 h, the restriction enzymes were inactivated at 80°C for 20 min. Following this, the samples were loaded on a 1% agarose gel and electrophoresed at 50 v for 60 min in 1× TAE buffer. The smear between 300 bp and 1.6 kb was excised and purified using E.Z.N.A.® Gel Extraction Kit (Omega). In addition, the parent plasmids pSAlect-PRAI, pFoldM-KRK-PRAI and pFoldM-KR-PRAI were digested with NdeI and SpeI and the vector backbone of the corresponding size (pSAlect: 3.1kb, pFoldM-KRK: 3.7 kb and pFoldM-KR: 3.7 kb) was gel extracted in a similar fashion. Three ligation reactions with three-fold molar excess of insert over vector DNA were set up. Each ligation reaction had 32 ng of insert DNA, 38 ng of vectors, 1× T4 ligase buffer (Fermentas), 10 Weiss unit of T4 DNA ligase (Fermentas) in a final volume of 20 μ L. The ligation reactions were incubated at 16°C for 16 h followed by heat killing at 65°C for

10 min. The ligations were then purified using the E.Z.N.A.® Cycle Pure Kit (Omega) and eluted in 30 µL EB. Aliquots (0.5 µL) of the purified ligation products were used to transform electrocompetent *E. coli* DH5α-E (Invitrogen) by electroporation, using a BIO-RAD Gene Pulser II. The cells were recovered in 500 µL SOC and incubated at 37°C for 1 h. The transformed cells were spread on LB-chloramphenicol (34 µg/mL) agar plates in the case of pSALect and kanamycin (30 µg/mL) in the case of pFoldM. Pre-selection or naïve plates were incubated at 37°C, whereas fold-selection plates were incubated at 28°C as well as 37°C. The colonies from each plate were counted and the clones were screened for the presence of the desired inserts by colony PCR. For colony PCR, one primer specific to the pSALect/pFodM backbone and another specific to the 3' end of the inserts (Kvβ.rev) were used. Clones with inserts were sent for sequencing to Macrogen (Korea), Appendix I. The primer used for sequencing of the inserts was PRAI.for. The verified clones were stored at -80°C for future use.

4.5.11.1 Plating experiment

In order to obtain a sizeable number of colonies on pre-selection and post-selection agar plates, a 100-fold dilution of the recovery culture (section 4.5.11) was plated for making a pre-selection mini-library and a two-fold dilution was plated to make the selection library. The rate of survival (%), an indicator for possible folded chimeras, was calculated by using the formula: $(\text{colony count}^S \times 0.02) / \text{colony count}^N \times 100$; where S = Selection plate and N = Naïve (pre-selection) plate. The experiments were conducted in two different temperature conditions, 28°C and 37°C, in order to see if any difference in survival rate emerged.

4.5.12 Expression and purification of chimeric proteins

As the ITCHY inserts in pSAlect and pFoldM are flanked between the restriction sites NdeI and SpeI, the ITCHY clones were digested with the restriction enzymes NdeI and SpeI and then sub-cloned into the expression plasmid pLAB101 to incorporate a C - terminal (His)₆-tag for purification via metal affinity chromatography. The plasmid pLAB101 (1.5 µL) harbouring the variants was used to transform 50 µL aliquots of *E. coli* DH5α-E (Invitrogen) by electroporation using a BIO-RAD Gene Pulser II. The cells were allowed to recover from electroporation by adding 500 µL SOC and incubating at 37°C for 1 h. The transformed cells were spread on LB-carbenicillin (100 µg/mL) agar plates and the number of colonies formed was counted after incubation (37°C, 16 h). The colonies were screened for the presence of the desired inserts by colony PCR with one primer specific to the pLAB101 backbone (301_seq.for; Appendix I) and another specific to the 3' end of the inserts (Kvβ.rev; Appendix I). The verified clones were stored at -80°C for future use.

A 50 mL culture of *E. coli* DH5α-E cells, harbouring pLAB101-(inserts) in LB - Carb-100, was inoculated with 4 mL of an overnight starter culture incubated at 37°C. The OD₆₀₀ was monitored, until it reached 0.6 and then IPTG was added to a final concentration of 0.5 mM. Following the addition of IPTG, the cells were incubated at 28°C for 4 h. Four hours following induction, the cells were centrifuged at 4000 g for 15 min and the cell pellets were stored at -80°C. After thawing, a cell pellet was resuspended in 10 ml of column buffer. The column buffer used comprised 10 mM potassium phosphate, 50 mM KCl, pH 7.2 and 10% (v/v) glycerol. Lysozyme to a final concentration of 0.2 mg/mL and 100 µL of protease inhibitor cocktail (Sigma) were also added. The resuspended cells were then sonicated with amplitude of 50, pulse-on time

of 10 s (15 cycles), pulse-off time of 12 s (MISONIX-4000). Following sonication, the cells were centrifuged at 20,000 *g* at 4°C for 40 min. The soluble lysate was filtered through a 0.2 µm syringe filter and was then allowed to bind, by means of rocking for 2 h at 4°C, with Talon resin (Clontech). Prior to this step, the resin (0.5 mL bed volume) had been equilibrated with the column buffer by pelleting a 1 mL aliquot of Talon resin (800 *g* at 4°C for 2 min) and washing it with 2 x 8 mL of column buffer. The resin, after rocking for 2 h, was pelleted at 1000 *g* for 2 min and washed with 5 mL of column buffer. Both the unbound fraction and resin wash were stored for gel analysis. The resin was then resuspended in 1 mL of column buffer and transferred to a BIO-RAD gravity flow column. The column was washed with 2 x 5 mL of column buffer containing 10 mM imidazole and finally, the His₆-tagged protein was eluted by increasing the imidazole to a concentration of 150 mM in 8 x 0.5 mL fractions. Protein fractions were run on 15% SDS-PAGE gels.

4.5.13 Exchanging the buffer via concentrator and dialysis

The buffer was exchanged using a Vivaspin6 (GE Healthcare) concentrator as follows.

- 5) The concentrator (MWCO 5,000) was equilibrated with 2 mL of buffer and centrifuged for 10 min at 4000 *g*.
- 6) The column was equilibrated again at 4000 *g* for 15 min with 3 mL buffer.
- 7) Selected elution washes (pure/clean fractions) were pooled together and then loaded in the concentrator. The column was filled with exchange buffer to a final volume of 6 mL and mixed thoroughly. This was spun at 4000 *g* for 16 min.

- 8) This was repeated 4-5 times and the protein was concentrated to a final volume of 1.5 mL.

For dialysis, a 500 MWCO membrane (Spectrum lab) was used and protein was dialysed against water for three days at 4°C.

4.5.14 Labelling P24K89

An overnight culture of P24K89 in LB-Carb100 was used to inoculate 1 L of LB-Carb (1% inoculum) and grown overnight at 37°C. Cells were harvested by centrifugation in sterile bottles in a GS3 rotor (Piramoon Technologies Inc, Santa Clara, USA) at 7000 g for 15 min at 4°C in a Sorvall Evolution RC. Cells were washed in 200 mL minimal media base to remove any trace of LB media components and re-centrifuged. Cells were re-suspended in 250 mL minimal media base and 10 mL of metal and labelling mix in a 1 L flask and incubated with shaking at 37°C for 1 h to recover. Protein expression was then induced by addition of IPTG to a final concentration of 0.5 mM. The cells were induced at 28°C for 6 h.

After induction, the cell culture was pelleted and the cell culture was centrifuged at 4000 g for 15 min and the cell pellets were stored at -80°C. The cell pellets were re-suspended in 40 mL of column buffer, which comprised 10 mM KCl, 10 mM potassium phosphate, and pH 7.2. The cells were lysed with French pressure cell. Talon resin (4 mL bed volume, Clontech) was equilibrated with column buffer that also contained 1 mM imidazole. Following the disruption of cells via French press, the cells were centrifuged at 20,000 g at 4°C for 30 min. The soluble lysate was filtered through a 0.2 µm syringe filter and was then allowed to bind to the

equilibrated resin, by pumping it through the column by means of a peristaltic pump. The unbound fraction was collected and kept for gel analysis if needed. To elute protein from resin, the column was washed with increasing concentrations of imidazole. Two washes each of 10 mM, 20 mM, 40 mM and 150 mM imidazole were performed and protein eluted in a final volume of 10 mL. A final wash with 0.5 M imidazole was performed to elute any remaining protein.

4.5.15 Circular dichroism

Far-UV circular dichroism (CD) spectroscopy was used to monitor the secondary structure of P24K89. The protein P24K89 was dialysed against water to a final concentration of 0.2 mg/mL. The protein was degassed and introduced to a clean 0.1 mm pathlength quartz cell (Hellma® 106-QS, Germany). Buffer was used as a control. Circular dichroism spectra were measured at 20°C on a Chirascan spectrophotometer (Applied Photo Physics, Surrey, UK). A wavelength range of 180-260 nm was measured using a wavelength interval of 1 nm, TOP of 0.5 s and a bandwidth of 1 nm. Ten replicates were performed per run.

4.4.16 Nuclear Magnetic Resonance

For the NMR acquisition, the following parameters were used:

¹⁵N-HSQC: 2048 x 128 data points, 15.9381 ppm x 44 ppm sweep widths, 1024 scans, 1 s delay time.

¹³C-HSQC: 2048 x 128 data points, 16.0817 ppm x 166.0490 ppm sweep widths, 1 s delay time, and 576 scans (for the P25K86_CCS sample) or 992 scans (for the P24K89 sample).

TOCSY: 4096 x 128 data points, 13.9458 ppm x 14 ppm sweep widths, 96 scans, 60 ms mixing time, 1 s delay time. The samples were mixed with 10% D₂O before acquisition.

Note: Dr Alexander Goroncy acquired the spectra.

Chapter V

Conclusions

5.1 The findings

In this study, I tested the subdomain assembly model by employing a method, Incremental Truncation for the Creation of Hybrid enzYmes (ITCHY), which mimicked non-homologous recombination. In this model, small polypeptides that contribute a basic unit of structure or function are assembled combinatorially to form new domains. To test the model, two distantly related $(\beta\alpha)_8$ barrel proteins, one from bacteria (PRAI) and another from rat (Kv β 2), were randomly recombined in an attempt to construct a library of shuffled subdomain fragments. The search for novel folds resulted in two chimeric proteins, P25K86_CCS and P24K89, that appear to be partially structured.

A total of 240,000 chimeras were created by non-homologous recombination and screened for solubility using the pSALect system. Of the pSALect-positive clones, 16 were tested and two were found to be soluble when over-expressed and purified. While it was shown that pSALect and its engineered variants, pFoldM-KR and pFoldM-KRK, were prone to yielding false positives (sections 4.3.6.1, 4.3.6.2, 4.3.7 and 4.3.8), this nevertheless suggests that the probability of gaining a soluble chimera from non-homologous recombination of PRAI and Kv β 2 is significantly greater than 1 in 10^{-5} (i.e. higher than 2 in 240,000, but lower than 2 in 16).

Data from CD spectroscopy combined with HSQC spectra suggest that both proteins are only partially structured and are present in a molten globule state. The molten globule state is heterogeneous, as parts of the structure can be more organised than others (Szilágyi *et al.* 2007). However, it is considered by some to be a productive on-pathway folding intermediate, where the native state is formed from the molten globule state (Ptitsyn *et al.* 1990; Maki *et al.* 1999; Jamin & Baldwin 1998; Szilágyi *et al.* 2007). On the other hand, molten globule is believed by others to be a state where the folding intermediate may not be productive and is kinetically trapped, forming a misfolded species (Dill & Chan 1997; Szilágyi *et al.* 2007). In the light of the current data and experimental conditions, it can perhaps be inferred that these chimeric proteins are kinetically trapped in an ensemble of fluctuating (structured and unstructured) states.

This research also demonstrates that non-homologous recombination may result in proteins that have not yet been sampled by evolution. Newly folded polypeptides, which have never been synthesised in nature, have the potential to expand the functional space. The human-mediated exploration

of sequence space that has islands of soluble proteins can inform our understanding of natural fold evolution. Urvoas *et al.* (2012) noted: *"if folds result from selection pressure for interactions, emerging proteins may not be optimized for stability and possibly have unstable/fluctuating conformations"*(Urvoas *et al.* 2012).

5.2 Combinatorial assembly: a way to innovate

It has been suggested that protein domains are too complicated to have evolved *de novo* in the primordial environment. Domains may have arisen from fusion or exchange of short polypeptide segments, which may have acted as cofactors in RNA-based life (Ponting & Russell 2002; Söding & Lupas 2003). Non-homologous recombination could have facilitated the shuffling of secondary structure elements in primordial combinatorial libraries. The modularity that exists in present day proteins does suggest the role of duplications and recombination events in the course of evolution (Go 1983). Smaller primordial minigenes could have resulted from random sequence combination events. Their transitional products, short peptides, could have had a partial biological activity and partial stability. A combinatorial assembly of these minigenes could then have given rise to larger domains with more-refined activities.

Other researchers have discussed the possibility that short peptides with partial structures could contribute to partial function in the course of evolution (Lupas *et al.* 2001; Söding & Lupas 2003; Carny & Gazit 2005). It has been shown that even simple peptides with barely two residues (His-Ser and even Gly-Gly) can show proteolytic and catalytic activity (Li *et al.* 2000; Plankensteiner *et al.* 2002). It is speculated that combinations of short peptides (products of minigenes) could have resulted in a pool of longer

chains from which a few chains passed through a stringent selection process to possibly fold and survive the prebiotic environmental conditions. The chains that survived the selection conditions underwent further elongation to form stable domains by combinatorial assembly (Luisi 2007). This is possibly how extant diversity in protein sequence, structure and function space arose, and this is how we can also explore the unexplored protein fold space. Another study, where an artificial protein was synthesised to induce apoptosis in cancer cell lines by means of combinatorial assembly of naturally occurring peptide motifs, adds to this idea of assembling proteins from short polypeptides (Saito *et al.* 2004). This study showed that combinatorial assembly could give rise to functional proteins that exhibited a strong inhibitory effect on the growth of human hepatoma cell lines.

It has also been shown that some protein domains, under certain physiological conditions, may be present in partially structured or even unstructured states (Campbell *et al.* 1998). An example is the VH domain or variable heavy chain of the large polypeptide subunit of the antibody. The variable heavy (VH) and the variable light (VL) are non-covalently linked. When expressed individually, the functional VH domain of the mouse anti-human ferritin monoclonal antibody F11 was found to adopt a partially structured state with distorted secondary and destabilised tertiary structures (Martsev *et al.* 2002).

In this study, two (P25K86 and P24K89) out of 16 chimeric proteins were found to be soluble, and when P25K86 was scanned for NADPH-binding it was found to not bind NADPH. This indicated a lack of NADPH binding in this chimeric protein. However, there exists an opportunity to explore potentially 10,400 (in a total of 80,000 in-frame chimeras two out of 16 were soluble, i.e.

nearly 13% of the clones, which is an extrapolation) soluble chimeras that may have a stable scaffold and some novel function. But can new functions be engineered on artificially created scaffolds?

5.3 Engineering functions on artificially created folds

Both of these artificial proteins, P25K86 and P24K89, have no known function, but the question remains as to whether a function can be established for these proteins or any other stable scaffolds?

Recent research findings suggest that this might be possible. In a directed evolution experiment, random polypeptides of 140 amino acids were evolved for solubility and functionality for six generations. The selection was based on esterase activity and it was found that with only 10 randomly chosen polypeptides from each generation, a significant increase in the activity was observed (Yamauchi *et al.* 2002).

Similarly, in an experiment to probe enzyme-like activity from a combinatorial library of *de novo* proteins, a protein S-824 (α -helical) was discovered that showed esterase activity. The library was constructed using binary patterning of polar and non-polar residues in order to find folded proteins. Although the level of activity in this protein, as compared to naturally occurring enzymes, was found to be relatively low, this study demonstrates the potential to establish a function on a protein fold that is the result of a *de novo* combinatorial library (Wei & Hecht 2004).

In another human-mediated exploration of protein sequence space, sets of artificial ATP-binding proteins from a library of random sequences were

discovered (Keefe & Szostak 2001). This experiment demonstrated how functional proteins could be selected from random sequences.

In a recent elegant study, *de novo* designed proteins from a combinatorial library were probed for function *in vivo*, in a set of 27 *E. coli* auxotrophic strains. Four strains were rescued by novel proteins (102-residue 4-helix bundles) made in this library. The four strains were $\Delta serB$ (encodes phosphoserine phosphatase), $\Delta gltA$ (encodes citrate synthase), $\Delta ilvA$ (encodes biosynthetic threonine deaminase) and Δfes (encodes enterobactin esterase). This study suggests that artificial proteins can sustain growth (Fisher *et al.* 2011).

5.4 In the future

This proof-of-concept study was done to test whether the approach using ITCHY could be scaled up for an all-on-all library. The all-on-all library is a global or proteome-wide random recombination to expand the idea of subdomain assembly using a complete set of *E. coli* open reading frames or the ASKA collection. The ASKA collection is a complete set of individual genes that encodes proteins of predicted open reading frames (ORF) that are all histidine-tagged (Kitagawa *et al.* 2005). The ASKA collection is cloned into an expression vector (pCA24N) as an Sfi I restriction fragment and all plasmids (5272 clones) have been pooled together by Patrick *et al.* 2007. A mammoth experiment, where the ASKA ORFs are excised from pCA24N and subcloned into two pInSALect plasmids, could be undertaken, just like what was done in this study. The two-pInSALect plasmids must have unique restriction sites that occur at a very low frequency in the entire *E. coli* genome to minimise the digest of ASKA ORFs. The pooled plasmids of the two subcloned

pInSAlect libraries could then be linearised by digestion with these unique restriction sites. The optimised ITCHY protocol (Saraswat *et al.* 2013), created in this study, could then be applied to create an all-on-all library. Due to the limitations posed by Tat mediated fold selection demonstrated in this study (i.e. reporting of false positives), the ESPRIT (Expression of Soluble Proteins by Random Incremental Truncation) system could be applied. This robotic screening method uses *in vivo* biotinylation as a proxy for solubility (An *et al.* 2011).

Alternatively, an experiment could be designed to screen for new phenotypes for both of the two (P24K89 and P25K86) artificial proteins. This could be done in a similar way as in a recently published study, where an experiment was conducted to probe the biological function of three *de novo* designed proteins. These proteins were screened for binding to 10,000 compounds displayed on microarrays. All three proteins fold into 4-helix bundles and they have high sequence and structure similarity. Despite this, they bind different compounds and targets, and exhibited selective ligand binding (Cherny *et al.* 2012). Another approach would be to select both P25K86 and P24K89 for a specific function, such as esterase activity (Wei & Hecht 2004), as discussed earlier. A directed evolution experiment (rounds of random mutagenesis and selection) could potentially yield a desired function in both proteins. An iterative approach using Rosetta (Chaudhury & Gray 2008) can be applied to predict mutations that can improve the energy or favour a stabilised scaffold of these two chimeric proteins. In addition, isomerisation of phosphoribosylanthranilate (PRA) to carboxyphenylamino-deoxyribulose-5-phosphate (CdRP) by PRAI in the tryptophan biosynthesis pathway could be used to explore whether these partially structured proteins can be recruited for a function delivered by a $(\beta\alpha)_8$ barrel. Both

P24K89 and P25K86 lack the phosphate-binding motif, which is located in the β -strands 7 and 8 of their parent protein PRAI (Wilmanns *et al.* 1991). However, the C-terminal end of Kv β 2 with a predominant α -helical insertion comprises an NADPH-binding subdomain (Gulbis *et al.* 2000). A *trpF* knock-out strain could be transformed with plasmids containing the inserts (P25K86 and P24K89), followed by plating on a selective media that lacks tryptophan.

Alternatively, a new truncation library could be constructed to hunt for soluble proteins using robotic screening via the ESPRIT (Expression of Soluble Proteins by Random Incremental Truncation) system (An *et al.* 2011). An experiment could be designed to first screen soluble proteins and compare the effectiveness of robotic (expensive) over *in vivo* (economical) selection screens, as discussed earlier. From the subset of soluble proteins, a screen for new functions in order to understand protein evolution by combinatorial assembly of subdomains could be developed further.

As nature does not have predetermined destinations in fold or function space, nor does it manoeuvre in restricted mutational space, no single pathway to new folds exists. Evolution is a step-by-step sequence of events and every step acts upon products of the preceding steps. There are and there will be many routes for exploring and expanding the protein universe.

Appendix I

General Materials and Methods

AI.1 Reagents

All chemicals used in this research were purchased from Sigma-Aldrich, unless otherwise stated. All solutions were prepared using autoclaved MilliQ® water.

AI.2 Growth media and antibiotics

E. coli strain, DH5 α -E (Invitrogen) was grown in LB medium (ForMedium; 10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) at 37°C (Bertani 1951) and was agitated in a shaking incubator at 180 rpm. For making agar plates bacteriological agar (ForMedium) was added to a final concentration of 1.5% (w/v). All Antibiotics, except chloramphenicol (Duchefa Biochemie), were bought from Melford. Stock solutions were made by adding ethanol in chloramphenicol and water in other antibiotics. Final concentrations of antibiotics used in this study were: carbenicillin, 100 μ g/mL; chloramphenicol,

34 µg/mL; kanamycin, 30 µg/mL.

AI.3 Storing strains

To make freezer stocks, a single bacterial colony was inoculated overnight in selective LB. Glycerol was added to a final concentration of 15% (v/v) to 600 µL of the culture and then the suspension was stored at -80°C. For plating experiments, agar plates were stored at 4°C.

AI.4 Bacterial strain

For this study all experiments were conducted in DH5α-E (Invitrogen), which has the Genotype: $F^- \phi 80lacZ\Delta M15 \Delta(lacZYA-argF)U169 \text{ } recA1 \text{ } endA1 \text{ } hsdR17(r_k^-, m_k^+) gal^- \text{ } phoA \text{ } supE44 \lambda^- \text{ } thi-1 \text{ } gyrA96 \text{ } relA1$.

AI.5 Plasmids

Purified plasmid DNA was stored at -20°C

| Plasmid | Description | Reference |
|----------------|--|----------------------------|
| pDEP-Kvβ2 | <i>Amp^R</i> ; plasmid-encoded Kvβ2; used to get the insert. | Dr. Wayne Patrick |
| pDEP-trPRAI | <i>Amp^R</i> ; plasmid-encoded trPRAI (subdomain of <i>trpF</i> (PRAI)). | This study |
| pSALect-PRAI | <i>Amp^R & Cam^R</i> ; plasmid-encoded <i>trpF</i> (PRAI); used for making ITCHY long PCR product and folding selection. | This study |
| pSALect-trPRAI | <i>Amp^R & Cam^R</i> ; plasmid-encoded trPRAI; used for making ITCHY long PCR product and folding selection. | This study |
| pSALect-Kvβ2 | <i>Amp^R & Cam^R</i> ; plasmid-encoded Kvβ2; used for making ITCHY long PCR product and folding selection. | This study |
| pLAB101 | <i>Amp^R</i> ; plasmid used to express ITCHY chimeras. | This study |
| pMS401 | <i>Amp^R</i> ; plasmid-encoded <i>trpF</i> (PRAI); used to express PRAI. | (Patrick & Blackburn 2005) |

| | | |
|--------------------|---|-----------------------------------|
| pMS501 | <i>Amp^R</i> ; plasmid-encoded <i>trpF</i> (PRAI); Intermediate plasmid with NdeI site eliminated. | This study |
| pLAB101-P25K86-SCC | <i>Amp^R</i> ; plasmid-encoded P25K86; used to express P25K86. | This study |
| pLAB101-P25K86-CSC | <i>Amp^R</i> ; plasmid-encoded P25K86_CSC; used to express P25K86_CSC. | This study |
| pLAB101-P25K86-CCS | <i>Amp^R</i> ; plasmid-encoded P25K86_CCS; used to express P25K86_CCS. | This study |
| pIDTSMART-AMP-SCC | <i>Amp^R</i> ; plasmid-encoded P25K86_SCC; plasmid was synthesized with cysteine (Cys7) point mutation. | Integrated DNA Technologies (IDT) |
| pIDTSMART-AMP- CSC | <i>Amp^R</i> ; plasmid-encoded P25K86_CSC; plasmid was synthesized with cysteine (Cys46) point mutation. | Integrated DNA Technologies (IDT) |
| pIDTSMART-AMP-CCS | <i>Amp^R</i> ; plasmid-encoded P25K86_CCS; plasmid was synthesized with cysteine (Cys56) point mutation. | Integrated DNA Technologies (IDT) |
| pLAB101-P55K173 | <i>Amp^R</i> ; plasmid-encoded P55K173; used to express P55K173. | This study |
| pCM433-KnR | <i>Kan^R</i> ; plasmid used to get the KnR or <i>Kan^R</i> cassette. | Dr. Monica Gerth |
| pFoldM-PRAI | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded <i>trpF</i> (PRAI); used for making the stringent folding selection system. | This study |
| pFoldM-KRK-P25K86 | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded P25K86; used for comparing the stringency of folding selection. | This study |
| pFoldM-KR-P25K86 | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded P25K86; used for comparing the stringency of folding selection. | This study |
| pFoldM-KRK-P69K149 | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded P69K149; used for comparing the stringency of folding selection. | This study |
| pFoldM-KR-P69K149 | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded P69K149; used for comparing the stringency of folding selection. | This study |
| pSALect-P69K149 | <i>Amp^R</i> & <i>Cam^R</i> ; plasmid-encoded P69K149; used for folding selection. | This study |
| pLAB101-P69K149 | <i>Amp^R</i> ; plasmid-encoded P69K149; used to express P69K149. | This study |
| pFoldM-KR | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded PRAI; used for folding selection. | This study |

| | | |
|--------------------|---|---------------------|
| pFoldM-KRK | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded PRAI; used for folding selection. | This study |
| pFoldM-KRK-PubMetC | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded <i>metC</i> (PubMetC) from <i>Pelagibacter ubique</i> ; used for folding selection. | This study |
| pFoldM-KR-PubMetC | <i>Amp^R</i> & <i>Kan^R</i> ; plasmid-encoded <i>metC</i> (PubMetC) from <i>Pelagibacter ubique</i> ; used for folding selection. | This study |
| pInSAlect-PRAI | <i>Amp^R</i> & <i>Cam^R</i> ; plasmid-encoded PRAI; used for making time-dependent ITCHY library | This study |
| pInSAlect-Kvβ2 | <i>Amp^R</i> & <i>Cam^R</i> ; plasmid-encoded Kvβ2; used for making time-dependent ITCHY library | This study |
| pFoldM-KR-PRAI | <i>Amp^R</i> & <i>Cam^R</i> ; plasmid-encoded PRAI; used for comparing the stringency of folding selection. | This study |
| pFoldM-KRK-PRAI | <i>Amp^R</i> & <i>Cam^R</i> ; plasmid-encoded PRAI; used for comparing the stringency of folding selection. | This study |
| pUC19 | <i>Amp^R</i> ; 2.7-kb control plasmid for transformation | Invitrogen |
| pInSAlect | <i>Amp^R</i> & <i>Cam^R</i> ; Plasmid for in-frame selection. | (Gerth et al. 2004) |

AI.6 Oligonucleotides

Primers were ordered from Integrated DNA Technologies (Coralville, Iowa). All primers were resuspended in TE buffer (10 mM Tris-HCl, pH 8.0; 1 mM EDTA) to a final concentration of 100 μM. The working stock had a concentration of 10 μM. The primers were stored in -20°C

| Primer | Sequence (5' → 3') | Reference |
|-----------------|--|------------|
| pFoldM.rev | GGATCCTTTTTTTAAGGCAGTTATTGGTGC | This study |
| pFoldM.for | TTTAGCTTCCTTAGCTCCTGAAATCTCG | This study |
| pSAlect.for | CTTTACACTTTATGCTTCCGG | This study |
| trPRAI_Nsil_rev | CAGCAGACTAGTGGCATGCATGGATCCCCTCCACCCTG | This study |
| Kvβ.rev | GATCGCAGCACATATGCTCCAGTTTACAGGAATCTGG | This study |

| | | |
|----------------|--|------------|
| PRAl.for | ACTGACTGCATATGGGTGAGAATAAAGTATGTGGCCTG | This study |
| pMS_Nde_EL.for | GTGCACTCTCAGTACAATCTGCTC | This study |
| pMS_Nde_EL.rev | CGTATGCGGTGTGAAATACCGCAC | This study |
| pMS501_Spe.for | GGCGGTACTAGTGGACATCACCATCACCATCACTAATTTCTG | This study |
| pMS501_Nde.rev | GGCGGTCATATGTTATTCCTCCTTATTTAATC | This study |
| 301_seq.for | TTATCAGACAATCTGTGTGG | This study |
| pFoldM.KRK.for | AAACGTAAACATATGGGTGAGAATAAAGTATGTGG | This study |
| pFoldM.AQA.rev | CGCCTGCGCCGCGAGTCGCACGTCGCGG | This study |
| pFoldM.KR.for | AAACGTCATATGGGTGAGAATAAAGTATGTGG | This study |

AI.7 Software

To align, modify, view and annotate DNA and protein sequences, MacVector (MacVector, Inc) was used. Molecular structures were visualized using MacPymol (Schrödinger LLC).

AI.8 Agarose gel electrophoresis

The agarose powder (Axygen) was dissolved in 1× TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 8.0) and microwaved for 2-3 min. For staining ethidium bromide (BIO-RAD), was added to a final concentration 0.5 ug/mL. Agarose gels were viewed using a UV transilluminator (part of the Gel Doc™ 2000 Gel Documentation System, BIO-RAD). For gel extraction of DNA 1× SYBR Safe Gel DNA stain (Invitrogen) was used. When DNA was extracted from gels, Blue Light Transilluminator (Invitrogen) was used.

AI.9 SDS-PAGE

SDS-PAGE gels in this study were prepared in accordance with the discontinuous buffer system (Laemmli 1970). The constituents used to make polyacrylamide gels were 40% acrylamide: *N,N'*-methylene-bis-acrylamide

(29:1) solution (BIO-RAD), 1.5 M Tris-HCl (pH 8.8), 0.5 M Tris-HCl (pH 6.8), 10% (w/v) sodium dodecyl sulfate, 10% (w/v) ammonium persulfate and *N,N,N',N'*-tetramethylethylenediamine (~6.67 M). Resolving gel consisted of 12% acrylamide : *N,N'*-methylene-bis- acrylamide (29:1) solution, 375 mM Tris-HCl (pH 8.8), 0.1% sodium dodecyl sulfate, 0.05% ammonium persulfate and 6.67 mM *N,N,N',N'*-tetramethylethylenediamine. Stacking gel consisted of 4% acrylamide : *N,N'*-methylene-bis-acrylamide (29:1) solution, 125 mM Tris-HCl (pH 6.8), 0.1% sodium dodecyl sulfate, 0.05% ammonium persulfate and 6.67 mM *N,N,N',N'*-tetramethylethylenediamine.

Polyacrylamide gels were cast by following the manufacturer's guidelines (BIO-RAD). Protein samples were mixed with equal volume of 2 × Loading Buffer (100 mM Tris- HCl (pH 8.8), 4% (w/v) sodium dodecyl sulfate, 20% (v/v) glycerol, 0.2% (w/v) bromophenol blue, 200 mM β-mercaptoethanol). The sample and the buffer mix were heated at 95°C for 5 min prior to electrophoresis. The gel was run in 1× SDS-PAGE running buffer (25 mM Tris, 250 mM glycine (pH 8.3), 0.1% (w/v) sodium dodecyl sulfate), at 200 V until the tracking dye reached the bottom of the gel. The gels were stained in Coomassie Blue (2.5 g/L (w/v) Coomassie R250 (BIO-RAD) in 4 volumes of water:5 volumes of methanol:1 volume of acetic acid). The gels were rocked for 30 min followed by destaining in 6 volumes of water: 3 volumes of methanol: 1 volume of acetic acid. The gels were destained for an hour by gently rocking.

AI.10 Electrocompetent cells

Electrocompetent cells were prepared as described in the laboratory Manual (Sambrook *et al.* 2001). Briefly, *E. coli* strain DH5α-E (Invitrogen) was grown overnight in 5 mL LB incubated at 37°C. The overnight culture was inoculated in 250 mL SOB medium (20 g/L tryptone, 5 g/L yeast extract, 10 mM NaCl, 2.5

mM KCl) (Hanahan 1983). The culture was incubated at 37°C to an OD₆₀₀ ~0.4. The cells were divided equally in pre-chilled centrifuge tubes, and allowed to incubate on ice for another 1 h. All following steps were carried out at 4°C. The cells were pelleted at 3,000 × g for 15 min, at 4°C (Heraeus Multifuge 1S-R) and the pellet was washed thoroughly with 10% (v/v) glycerol and re-pelleted and washed for another 5 times. The final cell pellet was weighed, and resuspended with 0.1 mL of 10% glycerol per 0.1 g of wet cell weight. The cells were aliquoted (50 µL) and stored at - 80°C for future use.

AI.11 Sequencing

All plasmids were sequenced by either Massey Genome Service (Palmerston North) or Macrogen Inc (South Korea). Both house the automated capillary-based ABI3730 DNA Analyzers (Applied Biosystems) for sequencing

Appendix II

Statement of contributions



MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

**STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Mayank Saraswat

Name/Title of Principal Supervisor: Dr. Wayne M. Patrick

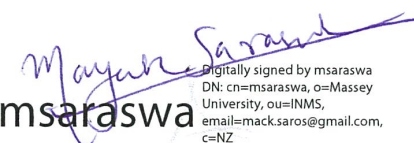
Name of Published Research Output and full reference:

Saraswat, M., Grand, R.S. & Patrick, W.M., 2013. Desalting DNA by Drop Dialysis Increases Library Size upon Transformation. *Bioscience, Biotechnology, and Biochemistry*, 77(2), pp.402–404.

In which Chapter is the Published Work: Chapter 4

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 75%
and / or
- Describe the contribution that the candidate has made to the Published Work:


 Digitally signed by msaraswa
 DN: cn=msaraswa, o=Massey
 University, ou=INMS,
 email=mack.saros@gmail.com,
 c=NZ
 Date: 2013.06.25 14:32:49 +12'00'
 msaraswa
 Candidate's Signature

25 June, 2013

Date


 W.M. Patrick
 Principal Supervisor's signature

25th June, 2013

Date

Appendix III

Publication arising from this
work



Note

Desalting DNA by Drop Dialysis Increases Library Size upon Transformation

Mayank SARASWAT,¹ Ralph. S. GRAND,¹ and Wayne M. PATRICK^{1,2,†}¹*Institute of Natural Sciences, Massey University, Private Bag 102 904, Auckland 0745, New Zealand*²*Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054, New Zealand*

Received October 4, 2012; Accepted November 18, 2012; Online Publication, February 7, 2013

[doi:10.1271/bbb.120767]

It is often desirable to obtain gene libraries with the greatest possible number of variants. We tested two different methods for desalting the products of library ligation reactions (silica-based microcolumns and drop dialysis), and examined their effects on final library size. For both intramolecular and intermolecular ligation, desalting by drop dialysis yielded approximately 3–5 times more transformants than microcolumn purification.

Key words: directed evolution; drop dialysis; library construction; ligation; microcolumn

In many molecular biology protocols, it is useful, or even essential, to maximize the number of colonies that result from a cloning experiment (*i.e.*, ligation, desalting, and transformation). This is especially true in the field of directed evolution, in which the Darwinian principles of random mutagenesis and selection are used to identify proteins with new or improved properties. The first step in a directed evolution experiment is to introduce molecular diversity into parental gene sequences. Many random mutagenesis protocols have been developed,^{1,2} including methods for random point mutagenesis (*e.g.*, error-prone PCR), random homologous recombination (*e.g.*, DNA shuffling), and random non-homologous recombination (*e.g.*, Incremental Truncation for the Creation of Hybrid enzYmes, ITCHY). The next step is to capture the diversity that has been generated, by cloning the pool of mutagenized DNA molecules into an appropriate expression vector. Transformation of a suitable host, typically *Escherichia coli*, with the cloned DNA yields a library that can be stored for use in downstream screening or selection steps. Often, the library screen or selection is very high-throughput; it is common to design directed evolution experiments with the capacity to interrogate millions, or even billions, of variants. In these high-throughput cases, it is critical to optimize the library cloning and transformation steps, because large, diverse libraries are the most likely to include variants with improvements in the desired property.³

Electroporation is the method of choice for transforming *E. coli* with the products of library ligation reactions. Very high transformation efficiencies can be obtained,⁴ resulting in large libraries. However, the high electric field strengths that are used (12–18 kV/cm) mean that efficient transformation requires low-conductivity samples, to prevent arcing. Hence, each library

ligation reaction must be desalted prior to electroporation. This purification step is commonly carried out with silica-based microcolumns. A previous study that tested electroporation efficiencies with intact plasmid DNA (rather than with the products of ligation reactions) showed the microcolumn desalting method to be highly effective.⁵ However, an older study showed that drop dialysis, in which a 5–100 μ L drop is placed on a floating membrane filter, can also result in effective desalting, with DNA recovery rates of 98–99%.⁶ Here we present a rigorous comparison of the two desalting methods, and show that drop dialysis is preferred in the construction of large libraries.

First, two parental sequences were randomly recombined using ITCHY: (i) the *trpF* portion of the bi-functional *E. coli trpCF* gene (GenBank accession no. NP_415778); and (ii) a cDNA encoding residues 36–367 of the $\beta 2$ subunit of the *Rattus norvegicus* voltage-gated potassium channel (Kv $\beta 2$; GenBank accession no. NM_017304). The TrpF and Kv $\beta 2$ proteins share the same ($\beta\alpha$)₈ barrel fold, but their sequences are highly divergent (<10% sequence identity). The genes were each cloned into vector pInSAlect⁷ and then recombined onto the same linearized vector molecule by PCR, as described previously.⁸ Our library was constructed using ITCHY with time-dependent truncation.⁹

At the final stage of library construction, three identical intramolecular ligations (90 μ L) were set up, each of which comprised blunt-ended DNA from the ITCHY protocol (180 ng), 1 \times Fermentas T4 DNA Ligase Buffer, and 30 units of T4 DNA ligase (Fermentas, Vilnius, Lithuania). The ligation reactions were incubated at 16 °C for 16 h, and then heat inactivated (65 °C, 10 min). After heat inactivation, each of the three reactions was split into three 30- μ L aliquots. One aliquot was desalted using a microcolumn (EZNA MicroElute Cycle Pure kit; Omega Bio-Tek, Norcross, GA, USA) following the manufacturer's guidelines. The desalted DNA was eluted from the column with 30 μ L of elution buffer (10 mM Tris, pH 8.5). The second aliquot was desalted by drop dialysis. A standard Petri dish was half-filled with 30 mL of deionized (Milli-Q) water. A mixed cellulose ester membrane filter (pore size 0.025 μ m, diameter 25 mm, MF-Millipore, Billerica, MA, USA) was floated on the water. The 30- μ L aliquot of the ligation reaction was pipetted onto the membrane, covered with the lid of the Petri dish, and left to dialyze for 1 h. After dialysis, the desalted sample was recovered

[†] To whom correspondence should be addressed. Tel: +64-3-4797897; Fax: +64-3-4797866; E-mail: wayne.patrick@otago.ac.nz
Abbreviations: ITCHY, incremental truncation for the creation of hybrid enzymes; PCR, polymerase chain reaction

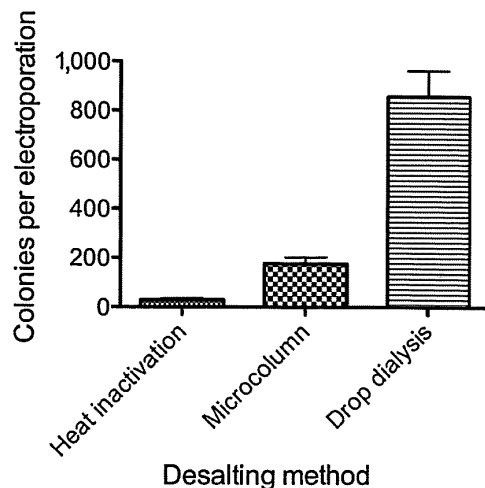


Fig. 1. Comparison of DNA Desalting Methods in the Construction of ITCHY Libraries.

Intramolecular ITCHY ligation reactions were heat inactivated only, purified with a silica-based microcolumn, or purified by drop dialysis. Aliquots of the desalted reaction products were used to transform *E. coli* by electroporation, and the numbers of colonies on dilution plates were used to estimate the total number of transformants resulting from each electroporation. Colony counts plotted are mean \pm SEM for three independent ligation reactions.

from the top of the membrane, and the volume of the sample was adjusted to 30 μ L with water. The third 30- μ L sample from each ligation was not desalted further. Aliquots (2 μ L) of each desalted library ligation and the heat inactivated controls were used to transform 50- μ L aliquots of *E. coli* DH5 α -E (Invitrogen, Carlsbad, CA, USA), by electroporation. SOC medium (500 μ L) was added to each aliquot of cells immediately after pulsing. The transformed cells were allowed to recover at 37 $^{\circ}$ C with shaking for 1 h, and then aliquots were spread on LB-agar plates containing chloramphenicol (34 μ g/mL). Colonies were counted after 16 h of incubation at 37 $^{\circ}$ C.

Figure 1 shows the mean numbers of colonies that resulted when 6.7% of each desalted sample (2 μ L of a 30 μ L total volume) was used to transform *E. coli* by electroporation. On average, microcolumn purification yielded 6.4-fold more colonies than heat inactivation (without further desalting). However, drop dialysis yielded the greatest number of colonies (4.8-fold more than microcolumn purification).

Had the desalted samples from the triplicate ligations been pooled, there would have been 84 μ L of each sample (heat inactivated, microcolumn purified, and dialyzed) remaining. Transforming more aliquots of electrocompetent *E. coli* with all of this material would have yielded libraries with total sizes of approximately 1.3×10^3 variants (heat-inactivated ligation), 8.0×10^3 variants (microcolumn purification), and 3.9×10^4 variants (drop dialysis) respectively. The total number of possible variants in an ITCHY library is given by the product of the lengths of the two parental genes. In this case, the number of possible variants is $597 \text{ bp} \times 996 \text{ bp} = 594,612$. Our library analysis program, GLUE,¹⁰ estimates that 1.8×10^6 clones would be required in our ITCHY library in order to sample 95% of all possible variants. In this example, none of the three desalting methods led to a library of that size. However,

Table 1. Comparison of Methods for Desalting the Products of Intermolecular (Vector + Insert) Ligation Reactions

| Desalting method | Colonies per electroporation* |
|------------------|-------------------------------|
| Microcolumn | 870 \pm 50 |
| Drop dialysis | 2,930 \pm 540 |

*Mean \pm SEM for three independent ligation reactions.

our analysis with GLUE suggests that the library from drop dialysis will include approximately 6.3% of all possible variants, which is certainly preferable to the other alternatives: 1.3% of all possible variants when microcolumn purification is used; and only 0.2% of all possible variants when the ligation is heat-inactivated but not desalted.

In a second experiment, we tested the effect of varying the desalting method on the outcome of intermolecular (vector + insert) ligation. This mimics the construction of an error-prone PCR library. A 338-bp DNA fragment was ligated with the 4.3-kb expression vector pLAB101¹¹ after each had been digested with restriction enzymes NdeI and SpeI (both New England Biolabs, Ipswich, MA, USA). Three ligation reactions were performed in a total volume of 60 μ L per reaction. Each reaction contained 150 ng of vector DNA, 35 ng of insert DNA (a 3-fold molar excess of insert over vector), 1 \times T4 DNA Ligase Buffer (Fermentas), and 20 units of T4 DNA ligase (Fermentas). The reactions were incubated at 16 $^{\circ}$ C for 18 h and then heat inactivated (65 $^{\circ}$ C, 10 min). Each of the three ligation reactions was split into two 30- μ L aliquots. One of the two 30- μ L aliquots was desalted with a microcolumn, and the other was desalted by drop dialysis (see above). Aliquots (2 μ L) of the desalted reactions were used to transform *E. coli* DH5 α -E by electroporation, as described above. Dilutions of the transformed cells were spread on LB-agar plates containing carbenicillin (100 μ g/mL) and colonies were counted after 16 h of incubation at 37 $^{\circ}$ C. Desalting the intermolecular ligation reactions by drop dialysis yielded 3.4-fold more colonies than microcolumn purification (Table 1). This is similar to the 4.8-fold improvement observed for intramolecular ITCHY ligations (Fig. 1).

Our data indicate that drop dialysis is a highly effective method for desalting DNA, confirming earlier results.⁶ In both intramolecular and intermolecular ligation tests, we obtained approximately 3–5 times more transformants when we desalted by drop dialysis, as compared with the more commonly used silica-based microcolumn purification. In this study, we tested a single brand of microcolumn (the EZNA MicroElute Cycle Pure kit from Omega Bio-Tek). However, in preliminary desalting tests with intact plasmid DNA (rather than with library ligations), conducted as described previously,⁵ we found that this microcolumn and its associated purification protocol yielded identical results to a well-known but more expensive alternative (the QIAquick PCR Purification Kit, Qiagen, Valencia, CA). Hence, drop dialysis remains the superior protocol for desalting ligation reactions, regardless of the microcolumn to which it is compared. Further, a previous study found that varying the membrane filter from one with an average pore diameter of 0.01 μ m to one with an average pore diameter of 0.05 μ m did not change the

effectiveness of the drop dialysis protocol, although the use of membranes with very small pore diameters can increase the time required for complete removal of buffer salts.⁶⁾ The membranes used in our experiments (average pore diameter, 0.025 μm) allow for rapid dialysis while minimizing the likelihood that ligation products are lost.

Finally, drop dialysis requires less hands-on time than microcolumn purification. The membrane filter discs do require careful pipette handling as the samples are loaded onto the membrane. They are also more expensive than microcolumns (NZ\$7.48 per membrane versus NZ\$3.40 per microcolumn at the time of writing), although expert users can reduce the cost by desalting 2–4 reactions on a single membrane simultaneously. Overall, the extra care and costs required (as compared with microcolumn purification) are likely to be warranted for practitioners of directed evolution, for whom the largest possible libraries are often desirable.

Acknowledgments

We gratefully acknowledge financial support for this study from the Marsden Fund. M.S. and R.S.G. were also supported by Massey University Doctoral Scholarships.

References

- 1) Lutz S and Patrick WM, *Curr. Opin. Biotechnol.*, **15**, 291–297 (2004).
- 2) Otten LG and Quax WJ, *Biomol. Eng.*, **22**, 1–9 (2005).
- 3) Patrick WM, Firth AE, and Blackburn JM, *Protein Eng.*, **16**, 451–457 (2003).
- 4) Dower WJ, Miller JF, and Ragsdale CW, *Nucleic Acids Res.*, **16**, 6127–6145 (1988).
- 5) Schlaak C, Hoffmann P, May K, and Weimann A, *Biotechnol. Lett.*, **27**, 1003–1005 (2005).
- 6) Marusyk R and Sergeant A, *Anal. Biochem.*, **105**, 403–404 (1980).
- 7) Gerth ML, Patrick WM, and Lutz S, *Protein Eng. Des. Sel.*, **17**, 595–602 (2004).
- 8) Gerth ML and Lutz S, *J. Mol. Biol.*, **370**, 742–751 (2007).
- 9) Ostermeier M and Lutz S, *Methods Mol. Biol.*, **231**, 129–141 (2003).
- 10) Firth AE and Patrick WM, *Bioinformatics*, **21**, 3314–3315 (2005).
- 11) Gerth ML, Nigon LV, and Patrick WM, *Protein J.*, **31**, 359–365 (2012).

Appendix IV

Supplementary data

AIV.1 Protein sequence of chimeras

> P25K86

MGENKVCGLTRGQDAKAAYDAGAIYGRRQQAQKLKELQAIAERLGCTLPQLAIWCLRNEG
VSSVLLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS

>P28K132

MGENKVCGLTRGQDAKAAYDAGAIYGGLMIGVGAMTWSPLACGIVSGKYDSGIPPYSRASLK
GYQWLKDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIWCLRNEGVSLLGASNAEQL
MENIGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS

>P29K122

MGENKVCGLTRGQDAKAAYDAGAIYGGLILACGIVSGKYDSGIPPYSRASLKGYQWLKDKILS
EEGRRQQAQKLKELQAIAERLGCTLPQLAIWCLRNEGVSLLGASNAEQLMENIGAIQVLPK
LSSSIVHEIDSILGNKPYSKKDYRS

>P55K173

MGENKVCGLTRGQDAKAAYDAGAIYGGLIFVATSPRCVNVEQAQEVMAAAPLQYVGIMEAYS
VARQFNLIPIEQAEYHMFQREKVEVQLPELFHKIGVGAMTWSPLACGIVSGKYDSGIPPYSR
ASLKGQWLKDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIWCLRNEGVSLLGASNA
EQLMENIGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS

>P69K149

MGENKVCGLTRGQDAKAAYDAGAIYGGGLIFVATSPRCVNVEQAQEVMAAAPLQYVGVRNH
DIADVVDKAMFQREKVEVQLPELFHKIGVGAMTWSPLACGIVSGKYDSGIPPYSRASLKGYQ
WLKDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIWCLRNEGVS SVLLGASNAEQLMENI
GAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS

>P88K179

MGENKVCGLTRGQDAKAAYDAGAIYGGGLIFVATSPRCVNVEQAQEVMAAAPLQYVGVRNH
DIADVVDKAKVLSLAHVQLHGNEEQLYLKELQAIAERLGCTLPQLAIWCLRNEGVS SVLLGAS
NAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS

AIV.2 DNA and protein sequence of Kv β 2 and trPRAI

>kv β 2

CATATGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGGGG
CTTGGAACATGGGTGACCTTCGGAGGCCAGATCACAGATGAGATGGCAGAGCACCTAAT
GACCTTGGCCTATGACAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCAGCTGG
CAAGGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGGTGGAGACGGTCCAGCC
TTGTCATCACCAAGATCTTCTGGGGCGGAAAGGCAGAGACCGAGAGAGGCCTTTCCC
GGAAGCACATAATCGAAGGACTGAAAGCTTCCCTGGAGAGGCTGCAGCTGGAGTACGTG
GATGTGGTTTTTGCCAACCGCCCAGACCCCAACACACCCATGGAAGAGACTGTGCGGGCC
ATGACCCATGTCATCAACCAAGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCC
ATGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGC
GAGCAAGCGGAATACCACATGTTCCAGAGGGGAGAAGGTAGAGGTCCAGCTGCCAGAGCT
GTTCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCCTGCGGCATTGTCTC
AGGGAAGTATGACAGTGGCATCCCACCCTACTCCAGAGCCTCTCTGAAGGGCTACCAGTG
GTTGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAAGCCAACTGAAGGAAC
TGCAGGCCATTGCAGAGCGCCTAGGCTGCACCCTCCCACAGCTGGCCATAGCCTGGTGC
CTGAGGAACGAGGGCGTCAGCTCCGTGCTCCTGGGTGCTTCCAATGCAGAACAACTTATG
GAGAACATTGGAGCAATACAGGTCCTTCCAAAATTGTCGTCCTCCATCGTCCACGAGATCG
ACAGCATTCTGGGCAATAAACCTACAGCAAAAAGGACTATAGATCCACTAGT

>kv β 2

HMLQFYRNLGKSGLRVSLGLGTWVTFGGQITDEMAEHLMTLAYDNGINLFDTAEVYAAG
KAEVVLGNIIKKKGWRRSSLVITTKIFWGGKAETERGLSRKHIIIEGLKASLERLQLEYVD
VWFANRPDPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYSVARQFNLIPIICE
QAEYHMFQREKVEVQLPELFHKIGVGAMTWSPLACGIVSGKYDSGIPPYSRASLKGYQWL
KDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIWCLRNEGVS SVLLGASNAEQLMEN
IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRSTS

>trPRAI

ATGGGTGAGAATAAAGTATGTGGCCTGACGCGTGGGCAAGATGCTAAAGCAGCTTATGAC
GCGGGCGCGATTTACGGTGGGTTGATTTTTGTTGCGACATCACCGCGTTGCGTCAACGTT
GAACAGGCGCAGGAAGTATGGCTGCGGCACCGTTGCAGTATGTTGGCGTGTTCCGCAA

TCACGATATTGCCGATGTGGTGGACAAAGCTAAGGTGTTATCGCTGGCGGCAGTGCAACT
GCATGGTAATGAAGAACAGCTGTATATCGATACGCTGCGTGAAGCTCTGCCAGCACATGTT
GCCATCTGGAAAGCATTAAAGCGTCGGTGAAACCCTGCCCCGCCGAGTTTCAGCACGTT
GATAAATATGTTTTAGACAACGGCCAGGGTGGAGCGGGATCC

>trPRAI

MGENKVCGLTRGQDAKAAVDAGAIYGGLIFVATSPRCVNVEQAQEVMAAAPLQYVGVFRN
HDIADWDKAKVLSLAQVQLHGNEEQLYIDTLREALPAHVAIWKALSVGETLPAREFQHV
DKYVLDNGQGGAGS

AIV.3 Multiple sequence alignment of all chimeras

Source: <http://www.ebi.ac.uk/Tools/msa/clustalo/>

CLUSTAL O(1.2.1) multiple sequence alignment

```

trPRAI      -----
KVB         HMLQFYRNLGKSGLRVSCILGLTWTFTGGQITDEMAEHLMTLAYDNGINLFDTAEVYAAG
P88K179     -----
P25K86      -----
P28K132     -----
P69K149     -----
P29K122     -----
P55K173     -----

trPRAI      -----MG--ENKVCGLTRGQD--AKAAY-DAGAIYGG
KVB         KAEVVLGNIIKKKGWRRSSSLVITTKIFWGGKAETERGLSRKHIEGLKASLERLQLEYVD
P88K179     -----MGE--NKVCGLTRGQD--AKAAY-DAGAIYGG
P25K86      -----MGE--NKVCGLTRGQD--AKAAY-DAGAIYGG
P28K132     -----MGE--NKVCGLTRGQD--AKAAY-DAGAIYGG
P69K149     -----MGE--NKVCGLTRGQD--AKAAY-DAGAIYGG
P29K122     -----MGE--NKVCGLTRGQD--AKAAY-DAGAIYGG
P55K173     -----MGE--NKVCGLTRGQD--AKAAY-DAGAIYGG
                  *   :   ***   :   ***   *

trPRAI      LIFVATSPRCVNVEQAQE-----VMAAAPLQYVGVFNRHDIADVVDKAKVLSL
KVB         VVFANRPDPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYSVARQFNLIPIICE
P88K179     LIFVATSPRCVNVEQA-----QE-----VMAAAPLQYVGVFNRHDIADVVDKAKV---
P25K86      -----
P28K132     LM-----
P69K149     LIFVATSPRCVNVEQA-----QE-----VMAAAPLQYVGVFNRHDI-----D
P29K122     LI-----
P55K173     LIFVATSPRCVNVEQA-----QE-----VMAAAPLQYVGIMEAYSVARQFNLIPIICE

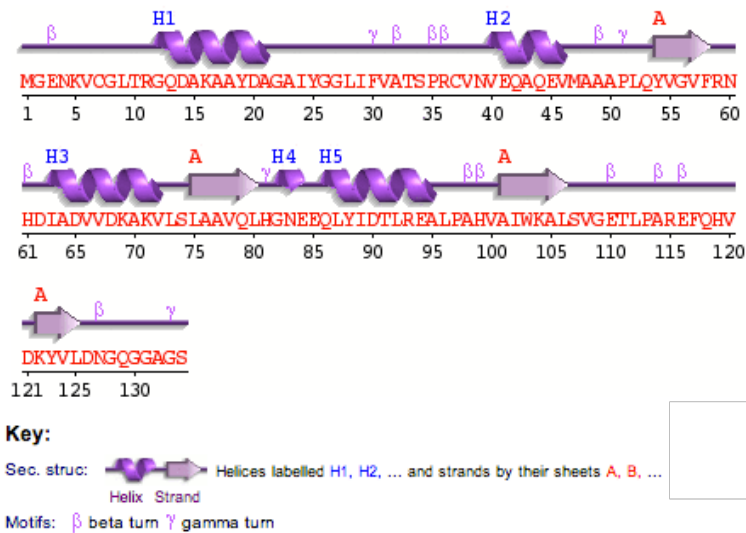
trPRAI      AAVQLHG--NEEQLYIDTLREALPAHVAIWKALSVGETLPA-----REFQHV
KVB         QAEYHMFQREKVEVQLPELFHKIGVGAMTWSPLACGIVSGKYDSGIPPYSRASLKGQWL
P88K179     -----L
P25K86      -----
P28K132     -----IGVGAMTWSPLACGIVSGKYDSGIPPYSRASLKGQWL
P69K149     VVDKAMFQREKVEVQLPELFHKIGVGAMTWSPLACGIVSGKYDSGIPPYSRASLKGQWL
P29K122     -----LACGIVSGKYDSGIPPYSRASLKGQWL
P55K173     QAEYHMFQREKVEVQLPELFHKIGVGAMTWSPLACGIVSGKYDSGIPPYSRASLKGQWL

trPRAI      DKYV-LDNGQGGAGS-----
KVB         KDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
P88K179     SLAAVQLHGNEEQLYLQELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
P25K86      -----RRQQAQKLKELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
P28K132     KDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
P69K149     KDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
P29K122     KDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
P55K173     KDKILSEEGRRQQAQKLKELQAIAERLGCTLPQLAIAWCLRNNEGVSSVLLGASNAEQLMEN
                  .

trPRAI      -----
KVB         IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRST
P88K179     IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS--
P25K86      IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS--
P28K132     IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS--
P69K149     IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS--
P29K122     IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS--
P55K173     IGAIQVLPKLSSSIVHEIDSILGNKPYSKKDYRS--

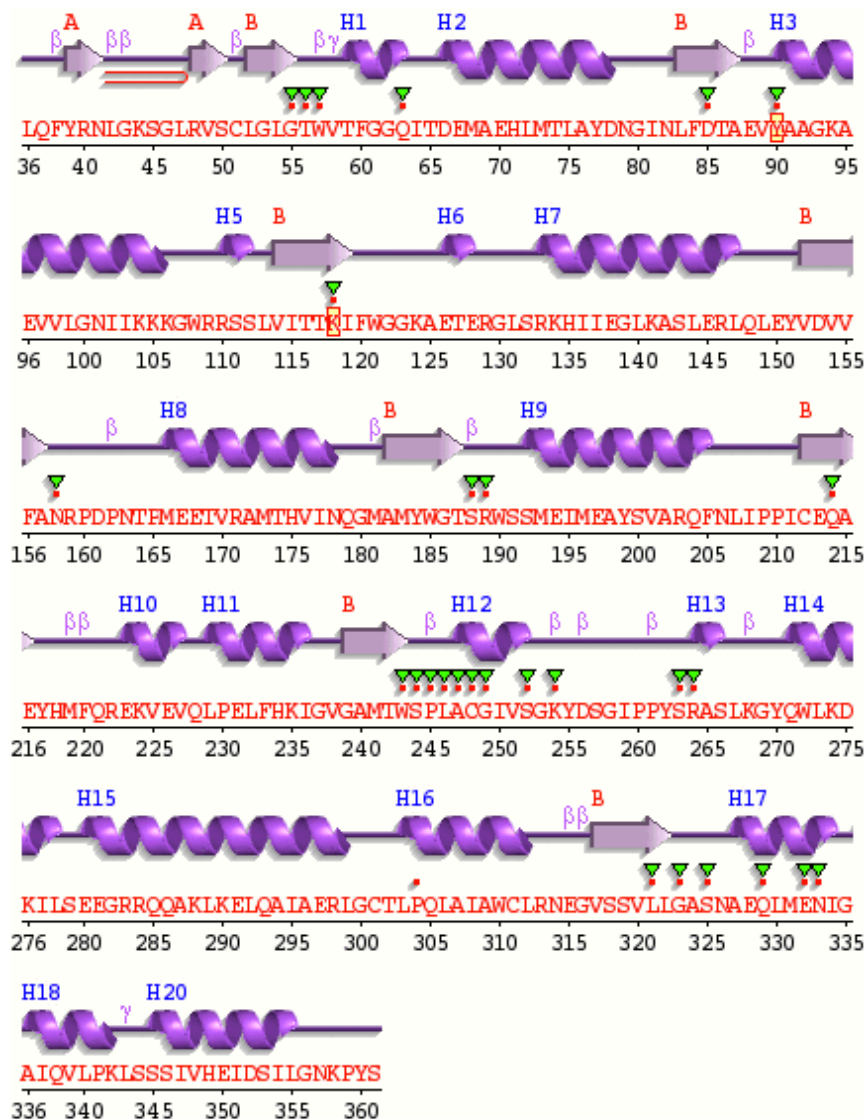
```



AIV.4 Secondary structure elements in trPRAI; PDB-2KZH





Source: <http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=2kzh&template=protein.html&r=wiring&l=1&chain=A>


On a per residue basis, the protein has 32% α -helices and 15% β -sheets. Forty-three residues contribute to α -helices and 20 residues contribute to β -sheets in a chain of 134 residues.


AIV.5 Secondary structure elements in Kv β ; PDB-1EXB**Key:**

Sec. struc:  Helices labelled H1, H2, ... and strands by their sheets A, B, ...
 Strand

Motifs: β beta turn γ gamma turn  beta hairpin

CSA annotation:  catalytic residue

Residue contacts:  to ligand

PDB SITE records:  AC1

Source: <http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1exb&template=protein.html&r=wiring&l=1&chain=A>

On a per residue basis, the protein has 54% α -helices and 16.2% β -sheets. One hundred and seventy-five residues contribute to α -helices and 53 residues contribute to β -sheets in a chain of 326 residues.

AIV.6 Deconvolution of Kv β 2

| Method | SELCON3 | CONTIN | CDSSTR | K2D |
|-------------|---------|--------|--------|------|
| NRMSD | 0.19 | 0.21 | 0.001 | 0.36 |
| Helix | 0.48 | 0.63 | 0.57 | 1.00 |
| Strand | 0.25 | 0.08 | 0.20 | 0.00 |
| Turns | 0.10 | 0.30 | 0.07 | — |
| Disordered | 0.22 | 0.00 | 0.17 | — |
| Random Coil | — | — | — | 0.00 |

Table AIV.1. Deconvolution of Kv β using the DichroWeb application (Whitmore & Wallace, 2004). Four algorithms (methods) were used to calculate the amount of secondary structure present, all of which show a similar trend, i.e. a higher percentage of α -helices over β -sheets. What is interesting is that the predicted secondary structure elements *via* the CDSSTR method (lowest NRMSD) closely matches the original structure (PDB: 1exb) on a per residue basis – see section AIV.5

AIV.7 Deconvolution of trPRAI

| Method | SELCON3 | CONTIN | CDSSTR | K2D |
|-------------|---------|--------|--------|------|
| NRMSD | 0.10 | 0.09 | 0.03 | 0.08 |
| Helix | 0.21 | 0.22 | 0.21 | 0.26 |
| Strand | 0.33 | 0.32 | 0.33 | 0.43 |
| Turns | 0.11 | 0.11 | 0.11 | — |
| Disordered | 0.36 | 0.34 | 0.35 | — |
| Random Coil | — | — | — | 0.32 |

Table AIV.2. Deconvolution of trPRAI using the DichroWeb application (Whitmore & Wallace, 2004). Four algorithms (methods) were used to calculate the amount of secondary structure present, all of which show a similar trend, i.e. a higher percentage of β -sheets over α -helices. Unlike for Kv β , the predicted secondary structure elements *via* the CDSSTR method (lowest NRMSD) does not match the original structure (PDB: 2kzh) on a per residue basis – see section AIV.4. In fact, there are more α -helices than β -sheets in the original structure. Note that this data was not generated by the author of this thesis, but was provided by Dr Wayne Patrick who has published it previously (Patrick & Blackburn, 2005).

AIV.8 NADPH-binding subdomain experiment

In order to test the NADPH-binding subdomain of Kv β 2 in the clone P25K86, a wavelength scan of the protein was performed from 240 nm to 400 nm. It was speculated that if the chimera bind NADPH, there should be a peak at 340 nm (NADPH absorbs light at 340 nm) in addition to the peak at 280 nm. This was a qualitative test and was made simply to check the presence of a peak. The scan suggested that there was no NADPH binding, hence no indication of function.

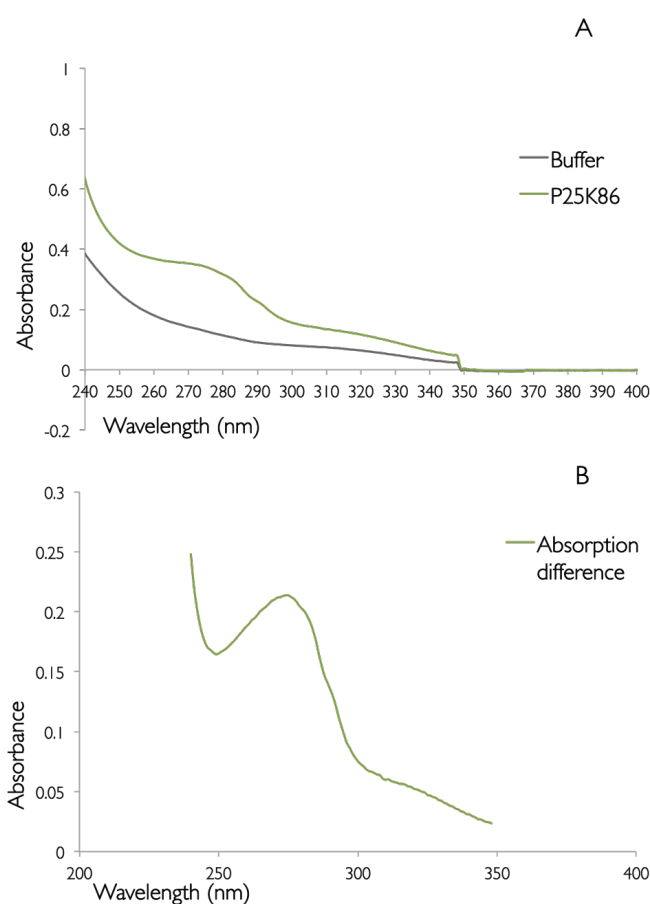


Fig. AIV.1. Scan of P25K86 from 240 to 400 nm. For this qualitative experiment, 50 μ L of the protein sample was used. For the blank, buffer containing 40 mM TrisHCl, pH 7.2, 300 mM NaCl, 10mM imidazole, 10% v/v glycerol and 1 mM β -mercaptoethanol (BME) were used. (A) A bump at 280 nm is indicative of protein in the sample, however the flat line at 340 nm indicates no NADPH binding. The step at 350 nm is the point at which the spectrophotometer (CARY-VARIAN) changes from tungsten halide to the deuterium lamp. There is no evidence for a peak at 340 nm. Broad, featureless absorption is consistent with light scattering. (B) The difference of the absorption between buffer and the protein sample.

AIV.9 SDS-PAGE gel of P25K86

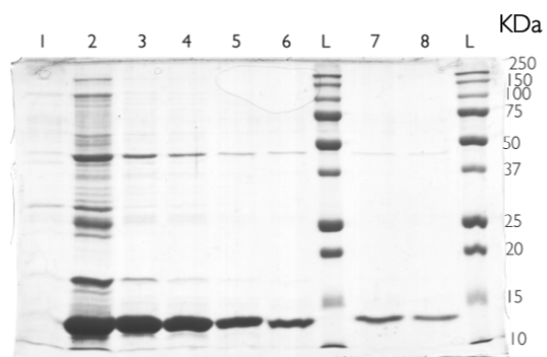


Fig. AIV.2. SDS-PAGE of P25K86. Eight fractions labeled 1 to 8 are shown at the top of the gel and precision plus protein ladder (BIO-RAD), which was used as protein standard is indicated by L.

References

- Agafonov, D.E. et al., 2005. Efficient suppression of the amber codon in E. coli in vitro translation system. *FEBS letters*, 579(10), pp.2156–60.
- Ali, M.H. & Imperiali, B., 2005. Protein oligomerization: how and why. *Bioorganic & medicinal chemistry*, 13(17), pp.5013–20.
- Aloy, P. & Russell, R.B., 2004. Ten thousand interactions for molecular biologist. *Nature biotechnology*, 22(10), pp.1317–1321.
- Alva, V. et al., 2010. A galaxy of folds. *Protein Science*, 19(1), pp.124–130.
- An, Y. et al., 2011. ORF-selector ESPRIT: a second generation library screen for soluble protein expression employing precise open reading frame selection. *Journal of structural biology*, 175(2), pp.189–97.
- Andrade, M.A. et al., 1993. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein engineering*, 6(4), pp.383–390.
- Apic, G. & Russell, R.B., 2010. Domain recombination: a workhorse for evolutionary innovation. *Science signaling*, 3(139), p.pe30.
- Arguello, J.R. et al., 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila. *PLoS Genetics*, 2(5), pp.745–754.
- Arguello, J.R. et al., 2007. Origination of chimeric genes through DNA-level recombination. In V. J.-N. (Lyon), ed. *Gene and Protein Evolution*. Genome Dyn. Basel, Karger, pp. 131–46.
- Auclair, S.M., Bhanu, M.K. & Kendall, D.A., 2012. Signal peptidase I: cleaving the way to mature proteins. *Protein science: a publication of the Protein Society*, 21(1), pp.13–25.

- Bagos, P.G. et al., 2010. Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics (Oxford, England)*, 26(22), pp.2811–7.
- Bain, A.D., 2003. Chemical exchange in NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43(3-4), pp.63–103.
- Bantscheff, M. et al., 2007. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4), pp.1017–1031.
- Batas, B., Schiraldi, C. & Chaudhuri, J.B., 1999. Inclusion body purification and protein refolding using microfiltration and size exclusion chromatography. *Journal of Biotechnology*, 68(2-3), pp.149–158.
- Berks, B.C., 1996. A common export pathway for proteins binding complex redox cofactors? *Molecular microbiology*, 22(3), pp.393–404.
- Berks, B.C., Palmer, T. & Sargent, F., 2005. Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Current opinion in microbiology*, 8(2), pp.174–81.
- Berks, B.C., Sargent, F. & Palmer, T., 2000. MicroReview The Tat protein export pathway. , 35, pp.260–274.
- Bharat, T.A.M. et al., 2008. A $\beta\alpha$ -barrel built by the combination of fragments from different folds. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29), pp.9942–9947.
- Bieri, M. et al., 2011. Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. *The FEBS journal*, 278(5), pp.704–715.
- Billeter, M., Wagner, G. & Wüthrich, K., 2008. Solution NMR structure determination of proteins revisited. *Journal of Biomolecular NMR*, 42(3), pp.155–158.
- Blake, C.C.F., 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature*, 273(5660), p.267.
- Blaudeck, N. et al., 2003. Genetic Analysis of Pathway Specificity during Posttranslational Protein Translocation across the Escherichia coli Plasma Membrane. , 185(9), pp.2811–2819.

- Bogarad, L.D. & Deem, M.W., 1999. A hierarchical approach to protein molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6), pp.2591–5.
- Bom, A.P.D.A. et al., 2010. The p53 core domain is a molten globule at low pH: functional implications of a partially unfolded structure. *The Journal of biological chemistry*, 285(4), pp.2857–2866.
- De Bono, S. et al., 2005. A segment of cold shock protein directs the folding of a combinatorial protein. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5), pp.1396–1401.
- Bostrom, M. et al., 2005. Why forces between proteins follow different Hofmeister series for pH above and below pI. *Biophysical Chemistry*, 117(3), pp.217–224.
- Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, 238(1), pp.115–134.
- Brosius, J., 2003. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica*, 118(2-3), pp.99–116.
- Bukhari, S.A. & Caetano-Anollés, G., 2013. Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. *PLoS computational biology*, 9(3), p.e1003009.
- Buszewski, B. & Noga, S., 2012. Hydrophilic interaction liquid chromatography (HILIC)--a powerful separation technique. *Analytical and Bioanalytical Chemistry*, 402(1), pp.231–47.
- Campbell, I.D. et al., 1998. NMR supplement Equilibrium NMR studies of unfolded and partially folded proteins. *Nature*, 5(july), pp.499–503.
- Caras, I.W. et al., 1987. Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins. *Nature*, 325(6104), pp.545–549.
- Carny, O. & Gazit, E., 2005. A model for the role of short self-assembled peptides in the very early stages of the origin of life. *The FASEB journal official publication of the Federation of American Societies for Experimental Biology*, 19(9), pp.1051–1055.

- Chaudhury, S. & Gray, J.J., 2008. Conformer Selection and Induced Fit in Flexible Backbone Protein-Protein Docking Using Computational and NMR Ensembles. *Journal of Molecular Biology*, 381(4), pp.1068–1087.
- Cheng, S. et al., 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 91(12), pp.5695–9.
- Cherny, I. et al., 2012. Proteins from an Unevolved Library of de novo Designed Sequences Bind a Range of Small Molecules. *ACS synthetic biology*, 1(4), pp.130–8.
- Choi, I.G. & Kim, S.H., 2006. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38), pp.14056–61.
- Choi, J.H. & Lee, S.Y., 2004. Secretory and extracellular production of recombinant proteins using *Escherichia coli*. *Applied Microbiology and Biotechnology*, 64(5), pp.625–635.
- Chothia, C. & Gough, J., 2009. Genomic and structural aspects of protein evolution. *The Biochemical journal*, 419(1), pp.15–28.
- Cline, K. & McCaffery, M., 2007. Evidence for a dynamic and transient pathway through the TAT protein transport machinery. *The EMBO journal*, 26(13), pp.3039–49.
- Cochran, A.G., Skelton, N.J. & Starovasnik, M.A., 2001. Tryptophan zippers: stable, monomeric beta -hairpins. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10), pp.5578–5583.
- Coco, W.M. et al., 2002. Growth factor engineering by degenerate homoduplex gene family recombination. *Nature biotechnology*, 20(12), pp.1246–1250.
- Comer, N., 2012. *Characterisation of a Multifunctional Enzyme from the Marine Bacterium Pelagibacter ubique* Natasha Comer. Massey University.
- Compton, L.A. & Johnson, W.C., 1986. Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Analytical biochemistry*, 155(1), pp.155–167.
- Lo Conte, L. et al., 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Research*, 25(1), pp.236–239.

- DeLano, W.L., 2002. The PyMOL Molecular Graphics System. *Schrödinger LLC* www.pymol.org, Version 1., p.<http://www.pymol.org>.
- DeLisa, M.P., Tullman, D. & Georgiou, G., 2003. Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), pp.6115–20.
- Denesyuk, A.I. et al., 2002. Functional attributes of the phosphate group binding cup of pyridoxal phosphate-dependent enzymes. *Journal of Molecular Biology*, 316(1), pp.155–172.
- Dill, K.A., 1999. Polymer principles and protein folding. *Protein Science*, 8(6), pp.1166–1180.
- Dill, K.A. & Chan, H.S., 1997. From Levinthal to pathways to funnels. *Nature structural biology*, 4(1), pp.10–19.
- Ding, Y., Zhou, Q. & Wang, W., 2013. Origins of New Genes and Evolution of Their Novel Functions. *Annu. Rev. Ecol. Evol. Syst.*, 43(1), pp.345–363.
- Doolittle, W.F., 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, 14(8), pp.307–311.
- Dorit, R.L., Schoenbach, L. & Gilbert, W., 1990. How Big Is the Universe of Exons ? *Science*, 250(4986), pp.1377–82.
- Dower, W.J., Miller, J.F. & Ragsdale, C.W., 1988. High efficiency transformation of *E. coli* by high voltage electroporation. *Nucleic Acids Res*, V(13), pp.6127–6145.
- Economou, A. & Wickner, W., 1994. SecA promotes preprotein translocation by undergoing ATP-driven cycles of membrane insertion and deinsertion. *Cell*, 78(5), pp.835–843.
- Edwards, A.J. & Reid, D., 2001. Introduction to NMR of proteins. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, Chapter 17, p.Unit 17.5.
- Eisenbeis, S. et al., 2012. Potential of fragment recombination for rational design of proteins. *Journal of the American Chemical Society*, 134(9), pp.4019–22.

- Fersht, A.R., 1995. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proceedings of the National Academy of Sciences of the United States of America*, 92(24), pp.10869–10873.
- Firth, A.E. & Patrick, W.M., 2008. GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in randomized libraries. *Nucleic acids research*, 36(Web Server issue), pp.W281–5.
- Fisher, A.C. & Delisa, M.P., 2004. MINIREVIEW A Little Help from My Friends : Quality Control of Presecretory Proteins in Bacteria. , 186(22), pp.7467–7473.
- Fisher, A.C., Kim, W. & Delisa, M.P., 2006. Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway. *Protein science : a publication of the Protein Society*, pp.449–458.
- Fisher, A.C., Rocco, M.A. & Delisa, M.P., 2011. *Heterologous Gene Expression in E. coli, Methods in Molecular Biology* T. C. Evans, & M.-Q. Xu, eds., Totowa, NJ: Humana Press.
- Fisher, M.A. et al., 2011. De novo designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth. *PloS one*, 6(1), p.e15364.
- Foster, M.P., McElroy, C.A. & Amero, C.D., 2007. Solution NMR of large molecules and assemblies. *Biochemistry*, 46(2), pp.331–340.
- Frenkel, Z.M. & Trifonov, E.N., 2008. From protein sequence space to elementary protein modules. *Gene*, 408(1-2), pp.64–71.
- Gerth, M.L. & Lutz, S., 2007. Non-homologous Recombination of Deoxyribonucleoside Kinases from Human and Drosophila melanogaster Yields Human-like Enzymes with Novel Activities. *Journal of Molecular Biology*, 370(4), pp.742–751.
- Gerth, M.L., Patrick, W.M. & Lutz, S., 2004. A second-generation system for unbiased reading frame selection. *Protein engineering, design & selection : PEDS*, 17(7), pp.595–602.
- Gilbert, W., 1978. Why genes in pieces? *Nature*, 271(5645), p.501.

- Go, M., 1983. Modular structural units, exons, and function in chicken lysozyme. *Proceedings of the National Academy of Sciences*, 80(7), pp.1964–1968.
- Greenfield, N.J., 1996. Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Analytical biochemistry*, 235(1), pp.1–10.
- Greenfield, N.J., 2006. Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols*, 1(6), pp.2876–2890.
- Gulbis, J.M. et al., 2000. Structure of the cytoplasmic beta subunit-T1 assembly of voltage-dependent K⁺ channels. *Science*, 289(5476), pp.123–127.
- Halaby, D.M., Poupon, A. & Mornon, J., 1999. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein engineering*, 12(7), pp.563–571.
- Höcker, B., 2013. Engineering chimaeric proteins from fold fragments: “hopeful monsters” in protein design. *Biochemical Society transactions*, 41(5), pp.1137–40.
- Höcker, B. et al., 2001. Stability, catalytic versatility and evolution of the ($\beta\alpha$)₈-barrel fold. *Current Opinion in Biotechnology*, 12(4), pp.376–381.
- Höcker, B., Claren, J. & Sterner, R., 2004. Mimicking enzyme evolution by generating new ($\beta\alpha$)₈-barrels from ($\beta\alpha$)₄-half-barrels. *Proceedings of the National Academy of Sciences of the United States of America*, 101(47), pp.16448–16453.
- Hong, P., Koza, S. & Bouvier, E.S.P., 2012. A Review Size-Exclusion Chromatography for the Analysis of Protein Biotherapeutics and Their Aggregates. *Journal of Liquid Chromatography Related Technologies*, 35, pp.2923–2950.
- Huse, M. & Kuriyan, J., 2002. The conformational plasticity of protein kinases. *Cell*, 109(3), pp.275–282.
- Jamin, M. & Baldwin, R.L., 1998. Two forms of the pH 4 folding intermediate of apomyoglobin. *J. Mol. Biol.*, 276(2), pp.491–504.

- Janecek, S. & Baláz, S., 1993. Evolution of parallel beta/alpha-barrel enzyme family lightened by structural data on starch-processing enzymes. *Journal of Protein Chemistry*, 12(5), pp.509–514.
- Johnson, W.C., 1999. Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins*, 35(3), pp.307–312.
- Kamatari, Y.O. et al., 2004. High-pressure NMR spectroscopy for characterizing folding intermediates and denatured states of proteins. *Methods*, 34(1), pp.133–143.
- Keefe, A.D. & Szostak, J.W., 2001. Functional proteins from a random-sequence library. *Nature*, 410(6829), pp.715–718.
- Kegel, A. et al., 2006. Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis*. *Nucleic Acids Research*, 34(5), pp.1633–1645.
- Kelly, S.M., Jess, T.J. & Price, N.C., 2005. How to study proteins by circular dichroism. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1751(2), pp.119–139.
- Kelly, S.M. & Price, N.C., 2000. The use of circular dichroism in the investigation of protein structure and function. *Current protein & peptide science*, 1(4), pp.349–384.
- Keren, H., Lev-Maor, G. & Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, 11(5), pp.345–55.
- Kim, S., Bracken, C. & Baum, J., 1999. Characterization of millisecond time-scale dynamics in the molten globule state of alpha-lactalbumin by NMR. *Journal of molecular biology*, 294(2), pp.551–560.
- Kitagawa, M. et al., 2005. Complete set of ORF clones of *Escherichia coli* ASKA library (A complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Research*, 12(5), pp.291–299.
- Kitamura, K. et al., 2002. Construction of block-shuffled libraries of DNA for evolutionary protein engineering: Y-ligation-based block shuffling. *Protein engineering*, 15(10), pp.843–853.

- Koide, S., 2009. Generation of new protein functions by nonhomologous combinations and rearrangements of domains and modules. *Current opinion in biotechnology*, 20(4), pp.398–404.
- Kolkman, J.A. & Stemmer, W.P., 2001. Directed evolution of proteins by exon shuffling. *Nature biotechnology*, 19(5), pp.423–428.
- Kolodny, R. et al., 2013. On the universe of protein folds. *Annu Rev Biophys*, 42, pp.559–582.
- Koning, A. D. et al., 2000. Lateral Gene Transfer and Metabolic Adaptation in the Human Parasite *Trichomonas vaginalis*. *Mol. Biol. Evol.*, 17(11), pp.1769–1773.
- Koretke, K.K. et al., 2000. Evolution of two-component signal transduction. *Molecular Biology and Evolution*, 17(12), pp.1956–1970.
- Kurata, S. et al., 2004. Reevaluation and Reduction of a PCR Bias Caused by Reannealing of Templates. *Applied and Environmental Microbiology*, 70(12), pp.7545–7549.
- Kuwajima, K., 1996. The molten globule state of alpha-lactalbumin. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 10(1), pp.102–109.
- Kwan, A.H. et al., 2011. Macromolecular NMR spectroscopy for the non-spectroscopist. *The FEBS journal*, 278(5), pp.687–703.
- Lawrence, J.G., 1997. Selfish operons and speciation by gene transfer. *Trends in microbiology*, 5(9), pp.355–9.
- Lee, J. & Blaber, M., 2011. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1), pp.1317–1321.
- Li, Y. et al., 2000. Dipeptide seryl-histidine and related oligopeptides cleave DNA, protein, and a carboxyl ester. *Bioorganic & Medicinal Chemistry*, 8(12), pp.2675–2680.
- Long, M., 2001. Evolution of novel genes. *Curr Opin Genet Dev*, 11(6), pp.1317–1321.

- Long, M., Deutsch, M., et al., 2003. Origin of new genes: Evidence from experimental and computational analyses. *Genetica*, 118(2-3), pp.171–182.
- Long, M., Betrán, E., et al., 2003. The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics*, 4(11), pp.865–875.
- Long, M. & Langley, C.H., 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science (New York, N.Y.)*, 260(5104), pp.91–95.
- Loomis, W.P. et al., 2001. A tripeptide sequence within the nascent DaaP protein is required for mRNA processing of a fimbrial operon in *Escherichia coli*. *Molecular microbiology*, 39(3), pp.693–707.
- Ludwig, C. & Viant, M.R., 2010. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical analysis : PCA*, 21(1), pp.22–32.
- Luger, K. et al., 1989. Correct folding of circularly permuted variants of a beta alpha barrel enzyme in vivo. *Science*, 243(4888), pp.206–210.
- Luger, K., Szadkowski, H. & Kirschner, K., 1990. An 8-fold beta alpha barrel protein with redundant folding possibilities. *Protein Engineering*, 3(4), pp.249–58.
- Luisi, P.L., 2007. Chemical aspects of synthetic biology. *Chemistry & biodiversity*, 4(4), pp.603–21.
- Lupas, A.N., Ponting, C.P. & Russell, R.B., 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? *Journal of Structural Biology*, 134(2-3), pp.191–203.
- Lutz, S., Ostermeier, M., Moore, G.L., et al., 2001. Creating multiple-crossover DNA libraries independent of sequence identity. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20), pp.11248–11253.
- Lutz, S., Fast, W. & Benkovic, S.J., 2002. A universal, vector-based system for nucleic acid reading-frame selection. *Protein engineering*, 15(12), pp.1025–30.

- Lutz, S., Ostermeier, M. & Benkovic, S.J., 2001. Rapid generation of incremental truncation libraries for protein engineering using α -phosphothioate nucleotides. *Nucleic Acids Research*, 29(4), p.e16.
- Lutz, S. & Patrick, W.M., 2004. Novel methods for directed evolution of enzymes: quality, not quantity. *Current opinion in biotechnology*, 15(4), pp.291–7.
- Makalowski, W. & Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16), pp.9407–9412.
- Maki, K. et al., 1999. Effects of proline mutations on the folding of staphylococcal nuclease. *Biochemistry*, 38(7), pp.2213–2223.
- Mansell, T.J. et al., 2010. A rapid protein folding assay for the bacterial periplasm. *Protein science : a publication of the Protein Society*, 19(5), pp.1079–90.
- Martsev, S.P. et al., 2002. Partially structured state of the functional VH domain of the mouse anti-ferritin antibody F11. *FEBS Letters*, 518(1-3), pp.177–182.
- Marusyk, R. & Sergeant, A., 1980. A Simple Method for the Dialysis of Small-Volume Samples. *Analytical Biochemistry*, 105, pp.403–404.
- McClintock, B., 1984. The significance of responses of the genome to challenge. *Science (New York, N.Y.)*, 226(4676), pp.792–801.
- Michnick, S.W. & Arnold, F.H., 1999. “Itching” for new strategies in protein engineering. *Nature biotechnology*, 17(12), pp.1159–1160.
- Moran, J. V., DeBerardinis, R.J. & Kazazian, H.H., 1999. Exon shuffling by L1 retrotransposition. *Science*, 283(5407), pp.1530–1534.
- Morgante, M. et al., 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature genetics*, 37(9), pp.997–1002.
- Nishigaki, K. et al., 1998. Y-ligation: an efficient method for ligating single-stranded DNAs and RNAs with T4 RNA ligase. *Molecular diversity*, 4(3), pp.187–190.

- Onuchic, J.N. & Wolynes, P.G., 2004. Theory of protein folding. *Current Opinion in Structural Biology*, 14(1), pp.70–75.
- Ostermeier, M., 2003. Theoretical distribution of truncation lengths in incremental truncation libraries. *Biotechnology and bioengineering*, 82(5), pp.564–77.
- Ostermeier, M. & Benkovic, S.J., 2001. EVOLUTION OF PROTEIN FUNCTION BY DOMAIN SWAPPING. *ADVANCES IN PROTEIN CHEMISTRY*, 55, pp.29–76.
- Ostermeier, M. & Lutz, S., 2003. The creation of ITCHY hybrid protein libraries. *Methods in molecular biology (Clifton, N.J.)*, 231(6), pp.129–41.
- Ostermeier, M., Nixon, A.E. & Benkovic, S.J., 1999. Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorganic & Medicinal Chemistry*, 7(10), pp.2139–2144.
- Ostermeier, M., Shim, J.H. & Benkovic, S.J., 1999. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature biotechnology*, 17(12)Ostermeier, M., Shim, J.H. & Benkovic, S.J., 1999. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature biotechnology*, 17(12), pp.1205–1209.), pp.1205–1209.
- Pace, C.N., Grimsley, G.R. & Scholtz, J.M., 2009. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *The Journal of biological chemistry*, 284(20), pp.13285–13289.
- Palmer, T. & Berks, B.C., 2003. Moving folded proteins across the bacterial cell membrane. *Microbiology*, 149(Pt 3), pp.547–556.
- Patel, P.H. & Loeb, L.A., 2000. Multiple amino acid substitutions allow DNA polymerases to synthesize RNA. *The Journal of biological chemistry*, 275(51), pp.40266–40272.
- Patrick, W.M. et al., 2007. Multicopy suppression underpins metabolic evolvability. *Molecular Biology and Evolution*, 24(12), pp.2716–2722.
- Patrick, W.M. & Blackburn, J.M., 2005. In vitro selection and characterization of a stable subdomain of phosphoribosylanthranilate isomerase. *The FEBS journal*, 272(14), pp.3684–97.
- Paulding, C.A., Ruvolo, M. & Haber, D.A., 2003. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5), pp.2507–2511.

- Pennisi, E., 1998. How the genome readies itself for evolution. *Science (New York, N.Y.)*, 281(5380), pp.1131,1133–1134.
- Petrounia, I.P. & Arnold, F.H., 2000. Designed evolution of enzymatic properties. *Current Opinion in Biotechnology*, 11(4), pp.325–330.
- Petsko, G.A. & Ringe, D., 2004. *Protein Structure and Function* illustrate., New Science Press.
- Plankensteiner, K., Righi, A. & Rode, B., 2002. Glycine and diglycine as possible catalytic factors in the prebiotic evolution of peptides. *Origins of life and evolution of the biosphere the journal of the International Society for the Study of the Origin of Life*, 32(3), pp.225–236.
- Ponting, C.P. & Russell, R.B., 1995. Swaposins: Circular permutations within genes encoding saposin homologues. *Trends Biochem. Sci.*, 20(5), pp.179–180.
- Ponting, C.P. & Russell, R.R., 2002. The natural history of protein domains. *Annual review of biophysics and biomolecular structure*, 31, pp.45–71.
- Popławski, T. & Błasiak, J., 2003. Non-homologous DNA end joining. *Postepy Biochemii*, 51(3), pp.122–135.
- Provencher, S.W. & Glöckner, J., 1981. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, 20(1), pp.33–37.
- Ptitsyn, O.B. et al., 1990. Evidence for a molten globule state as a general intermediate in protein folding. *FEBS letters*, 262(1), pp.20–24.
- Ranjbar, B. & Gill, P., 2009. Circular dichroism techniques: biomolecular and nanostructural analyses- a review. *Chemical biology drug design*, 74(2), pp.101–120.
- Remmert, M. et al., 2010. Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. *Molecular Biology and Evolution*, 27(6), pp.1348–1358.
- Ribnicky, B., Van Blarcom, T. & Georgiou, G., 2007. A scFv antibody mutant isolated in a genetic screen for improved export via the twin arginine transporter pathway exhibits faster folding. *Journal of molecular biology*, 369(3), pp.631–9.

- Richter, M. et al., 2010. Computational and experimental evidence for the evolution of a (beta alpha)₈-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *Journal of Molecular Biology*, 398(5), pp.763–773.
- Richter, S. et al., 2007. Functional Tat transport of unstructured, small, hydrophilic proteins. *The Journal of biological chemistry*, 282(46), pp.33257–64.
- Van Rijk, A.A.F., De Jong, W.W. & Bloemendal, H., 1999. Exon shuffling mimicked in cell culture. *Proceedings of the National Academy of Sciences of the United States of America*, 96(14), pp.8074–8079.
- Van Rijk, A.F. et al., 2000. Characteristics of super alphaA-crystallin, a product of in vitro exon shuffling. *FEBS Letters*, 480(2-3), pp.79–83.
- Rocco, M.A., Waraho-Zhmayev, D. & DeLisa, M.P., 2012. Twin-arginine translocase mutations that suppress folding quality control and permit export of misfolded substrate proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 109(33), pp.13392–7.
- Rosas-Magallanes, V. et al., 2006. Horizontal transfer of a virulence operon to the ancestor of *Mycobacterium tuberculosis*. *Molecular Biology and Evolution*, 23(6), pp.1129–1135.
- Ruepp, A. et al., 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*, 407(6803), pp.508–513.
- Russell, R.B., 2007. Classification of protein folds. *Molecular Biotechnology*, 36(3), pp.1317–1321.
- Russell, R.B., 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of Molecular Biology*, 279(5), pp.1211–1227.
- Sadowski, M.I. & Taylor, W.R., 2010. Protein structures, folds and fold spaces. *Journal of physics Condensed matter an Institute of Physics journal*, 22(3), p.033103.
- Saito, H. et al., 2004. Synthesis of functional proteins by mixing peptide motifs. *Chemistry & Biology*, 11(6), pp.765–773.

- Salem, G.M. et al., 1999. Correlation of observed fold frequency with the occurrence of local structural motifs. *Journal of Molecular Biology*, 287(5), pp.969–981.
- Saraswat, M., Grand, R.S. & Patrick, W.M., 2013. Desalting DNA by Drop Dialysis Increases Library Size upon Transformation. *Bioscience, Biotechnology, and Biochemistry*, 77(2), pp.402–404.
- Sargent, F., Berks, B.C. & Palmer, T., 2006. Pathfinders and trailblazers: a prokaryotic targeting system for transport of folded proteins. *FEMS microbiology letters*, 254(2), pp.198–207.
- Schaeffer, R.D. & Daggett, V., 2011. Protein folds and protein folding. *Protein engineering, design & selection : PEDS*, 24(1-2), pp.11–19.
- Schlaak, C. et al., 2005. Desalting minimal amounts of DNA for electroporation in E. coli: a comparison of different physical methods. *Biotechnology letters*, 27(14), pp.1003–5.
- Setiyaputra, S., Mackay, J.P. & Patrick, W.M., 2011. The structure of a truncated phosphoribosylanthranilate isomerase suggests a unified model for evolution of the ($\beta\alpha$)₈ barrel fold. *Journal of Molecular Biology*, 408(2), pp.291–303.
- Shanmugaratnam, S., Eisenbeis, S. & Höcker, B., 2012. A highly stable protein chimera built from fragments of different folds. *Protein engineering, design & selection : PEDS*, 25(11), pp.699–703.
- Shapiro, J.A., 1997. Genome organization, natural genetic engineering and adaptive mutation. *Trends in Genetics*, 13(3), pp.98–104.
- Sharp, P.M. et al., 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Research*, 16(17), pp.8207–8211.
- Sheinerman, F.B., Norel, R. & Honig, B., 2000. Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology*, 10(2), pp.153–159.
- Sieber, V., Martinez, C.A. & Arnold, F.H., 2001. Libraries of hybrid proteins from distantly related sequences. *Nature biotechnology*, 19(5), pp.456–460.

- Söding, J. & Lupas, A.N., 2003. More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays news and reviews in molecular cellular and developmental biology*, 25(9), pp.837–846.
- Sreerama, N. & Woody, R.W., 1993. A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal Biochem*, 209(1), pp.32–44.
- Sreerama, N. & Woody, R.W., 2000. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Analytical biochemistry*, 287(2), pp.252–260.
- Stirling, P.C., Lundin, V.F. & Leroux, M.R., 2003. Getting a grip on non-native proteins. *EMBO reports*, 4(6), pp.1317–1321.
- Stoltzfus, A. et al., 1994. Testing the exon theory of genes: the evidence from protein structure. *Science*, 265(5169), pp.1317–1321.
- Strauch, E. & Georgiou, G., 2007. A bacterial two-hybrid system based on the twin-arginine transporter pathway of *E. coli*. *Protein Science*, 16(5), pp.1001–1008.
- Südhof, T.C. et al., 1985. The LDL receptor gene: a mosaic of exons shared with different proteins. *Science*, 228(4701), pp.815–822.
- Szilágyi, A. et al., 2007. Protein Folding. In *Handbook of Neurochemistry and Molecular Neurobiology*. Springer US, pp. 303–343.
- Thompson, J.D. & Higgins, D.G., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix. *Nucleic acids research*.
- Tolman, J.R. et al., 1995. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 92(20), pp.9279–9283.
- Trevino, S.R., Scholtz, J.M. & Pace, C.N., 2007. Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa. *Journal of Molecular Biology*, 366(2), pp.449–460.

- Tsuji, T., Onimaru, M. & Yanagawa, H., 2001. Random multi-recombinant PCR for the construction of combinatorial protein libraries. *Nucleic acids research*, 29(20), p.E97.
- Tullman-Ercek, D., 2006. *Characterization and Engineering of the Twin-Arginine Translocation Pathway of Escherichia coli*. The University of Texas at Austin.
- Tullman-Ercek, D. et al., 2007. Export pathway selectivity of Escherichia coli twin arginine translocation signal peptides. *The Journal of biological chemistry*, 282(11), pp.8309–16.
- Urvoas, A., Valerio-Lepiniec, M. & Minard, P., 2012. Artificial proteins from combinatorial approaches. *Trends in biotechnology*, pp.1–9.
- Valencia-Burton, M. et al., 2006. Different mating-type-regulated genes affect the DNA repair defects of Saccharomyces RAD51, RAD52 and RAD55 mutants. *Genetics*, 174(1), pp.41–55.
- Wang, W. et al., 2006. High rate of chimeric gene origination by retroposition in plant genomes. *The Plant cell*, 18(8), pp.1791–1802.
- Wei, Y. & Hecht, M.H., 2004. Enzyme-like proteins from an unselected library of designed amino acid sequences. *Protein engineering, design & selection : PEDS*, 17(1), pp.67–75.
- Whitmore, L. & Wallace, B.A., 2004. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic acids research*, 32(Web Server issue), pp.W668–W673.
- Williams, G.J., Nelson, A.S. & Berry, A., 2004. Directed evolution of enzymes for biocatalysis and the life sciences. *Cellular and molecular life sciences CMLS*, 61(24), pp.3034–3046.
- Wilmanns, M. et al., 1991. Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry*, 30(38), pp.9161–9169.
- Wu, L.C., Grandori, R. & Carey, J., 1994. Autonomous subdomains in protein folding. *Protein Sci*, 3(3), pp.1317–1321.
- Wüthrich, K., 2001. The way to NMR structures of proteins. *Nature structural biology*, 8(11), pp.923–925.

- Xu, Y., Long, D. & Yang, D., 2007. Rapid data collection for protein structure determination by NMR spectroscopy. *Journal of the American Chemical Society*, 129(25), pp.7722–7723.
- Yadid, I. & Tawfik, D.S., 2007. Reconstruction of functional beta-propeller lectins via homo-oligomeric assembly of shorter fragments. *Journal of Molecular Biology*, 365(1), pp.10–17.
- Yamauchi, A. et al., 2002. Evolvability of random polypeptides through functional selection within a small library. *Protein Engineering*, 15(7), pp.619–626.
- Yang, S., Valas, R. & Bourne, P.E., 2009. Evolution studied using protein structure. *Structural Bioinformatics, Second Edition*, pp.559–572.
- Yumerefendi, H. et al., 2010. ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. *Journal of structural biology*, 172(1), pp.66–74.
- Zerbe, O. & Jurt, S., 2013. *Applied NMR Spectroscopy for Chemists and Life Scientists*, Wiley-VCH Verlag GmbH.
- Zhang, C.T., 1997. Relations of the numbers of protein sequences, families and folds. *Protein engineering*, 10(7), pp.1317–1321.
- Žídek, L., Štefl, R. & Sklenář, V., 2001. NMR methodology for the study of nucleic acids. *Current Opinion in Structural Biology*, 11(3), pp.275–281.