Investigating the world of protein design

Michael Doran

A thesis submitted for the degree of Doctor of Philosophy
September 2015

Division of Mathematical Biology, MRC National Institute for Medical Research, London

CoMPLEX: Maths and Physics in Life Sciences and Experimental Biology, UCL, London

I, Michael Doran, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

In this thesis, methods are developed for the design of protein structures based only on abstract structural descriptions of the protein fold. The design protocol starts with rough alpha-carbon (backbone) models and progress through several stages of high-resolution refinement, sequence design and filtering according to known principles of protein structure, coarse-grained and high-resolution knowledge-based potentials and secondary and tertiary structural prediction methods. Following this protocol led to the identification of protein solubility as a major limiting factor in the progression of designs. To overcome this problem, a systematic analysis was undertaken focusing on the more restricted problem of sequence redesign of the native structure to find common principles that govern viable protein sequences that can then be applied to the design of novel structures.

A factorial design of experiments approach was used to screen a multitude of sequence redesigns for known backbones that each possess a unique set of properties. The behaviour of each redesign was characterised when expressed in *E. Coli* in the hopes of elucidating some common features indicative of a viable sequence. Even with the use of fractional factorial design, the number of experimental designs required was limiting and to test the approach, we adopted an *ab initio* prediction method as a proxy for the "wet" experiments. This allowed us to increase the information gain for each experiment and we found that optimizing towards secondary structure prediction had a negative effect on the predictability of sequences.

Following this, an improved design methodology was developed that uses a genetic algorithm to produce sequence redesigns which look realistic to a number of computational measures such as sequence composition, both *ab initio* and comparative modelling prediction methods, and have a high degree of native sequence recapitulation (a good indicator for a viable design method). Although these state-of-the-art measures all point toward

our sequence designs being on-par with (or better than) native sequences, the problem of solubility and foldedness remained.

Machine learning techniques were used to unravel some of the complex intricacies governing these two properties. Using this approach, we discovered features that a substantial portion of native sequences comply with but were missing from our designs. Using the genetic algorithm design method, we directed our sequences to this area of attribute space in the hopes that this would produce more realistic designs. Unfortunately, solubility of designs still remained a large problem.

We also tested the relatively new technique of convergent peptide synthesis to understand how valuable it could be in a synthetic biology context. After designing a single Leucine-rich repeat, we used peptide chemistry to synthetically build the protein before investigating solubility and foldedness. Upon producing a single repeat of this designed protein, multiple repeats were linked together and the resulting properties characterised. A single 28-mer peptide was produced that was soluble and folded.

Acknowledgements

Firstly, I must thank my supervisor Willie Taylor for all the help that he's given me over the past 4 years. Not only was his support and guidance invaluable, but he's also one of the nicest people that I've ever had the pleasure of meeting. I could not have asked for a better mentor.

I'd also like to thank Michael 'Sid' Sadowski for everything he's done. When I first started in the lab and bombarded him with questions, his patience seemed inexhaustible and he always had a detailed answer for anything we discussed (even for many off-topic questions).

I loved the environment in MathBio throughout my time here so I'd like to thank everybody else too (Jens, Grant, Bhav, Kyriakos, Richard, Enrico, Martin, Alessandro, Asif). Everybody was always so friendly, and willing to help out in any way they could. Lunchtime discussions with everybody were frequently the highlight of my day (tip; if you're in a rush to be somewhere, don't get Bhav started on current economics).

To my parents (Marc and Dianne) and my sister (Joanne); I want you to know how important you were in this piece of work, and my life in general. You've shaped who I am, and I want to make you proud. I think watching the Discovery channel as a kid, with Dad trying to explain everything, definitely helped too. Grandma (Sybil) and Grandad (Harry) also had a huge part in raising me, so I'm grateful to them for everything they've done for me.

My girlfriend (Kate) has been hugely supportive throughout writing up, even going so far as to make me a ton of 'freezer burritos' so I didn't have to leave the house to get food if I was too busy writing. I'm so lucky.

Last, but by far not least, I need to thank my friends. Dabby and Brook are two of the nicest people I've met, and they kept me sane when things seemed to be getting tough. Luke has also made me a lot of food (I'm seeing a pattern here), and is a joy to talk to about anything. I'm sorry I can't list more people but I've a feeling the acknowledgements would be longer than my thesis if I continued, so I'll just summarise and say a massive "thank you" to everybody in my life at the moment. You're all amazing.

Mike
September 2015

Abbreviations

3D	Three Dimensional
ANOVA	ANalysis Of VAriance
CASP	Critical Assessment of protein Structure Prediction
CD	Circular Dichroism
CSPPS	Convergence Solid Phase Peptide Synthesis
DCM	Dichloromethane
DoE	Design of Experiments
DSSP	Dictionary of Secondary Structure of Proteins
EDTA	Ethylenediaminetetraacetic acid
ESPRESSO	EStimation of PRotein ExpreSsion and SOlubility
Fmoc	9-fluorenylmethyl carbamate
GA	Genetic Algorithm
GDT	Global Distance Test
HMMer	Hidden Markov Modeller
Hmsb	2-hydroxy-4-methoxy-5-methylsulfinyl benzyl
HPLC	High Pressure Liquid Chromotography
HSQC	Heteronuclear Single Quantum Correlation
I-TASSER	Iterative Threading ASSEmbly Refinement
KBP	Knowledge-Based Potential
KL	Kullback-Leibler
LRR	Leucine-Rice Repeat
MALDI	Matrix-Assisted Laser Desorption/Ionization
MD	Molecular Dynamics
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
POPS	Parameter OPtimized Surfaces
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
RI	Ribouclease Inhibitor
RMSD	Root-Mean-Square Deviation
RP	Reversed-Phase
rSASA	relative Solvent Accessible Surface Area
SAP	Structural Alignment Program
SEC	Size Exclusion Chromotography
SPPS	Solid Phase Peptide Synthesis
TBIO	Thermodynamically Balanced Inside-Out
tBu	di-tert-butyl dicarbonate
TES	Triethylsilane
TFA	Trifluoroacetic acid
TFE	Tetrafluoroethylene
TM	Template Modelling
TMSBr	trimethylsilyl bromide

Table of contents

List of figures

List of tables

1. Introduction

1.1 Protein Folding

The spontaneous formation of a 3-dimensional structure by a polypeptide chain has given rise to a multitude of protein functions present in nature[1,2]. Though not all proteins adopt stable 3D structures in order to perform their function[3], most do. Experiments by Anfinsen in the 1960's established that the sequence of a protein is sufficient to specify the structure adopted[4]. The precisely ordered 3D states of each protein sequence are global free energy minima and the natively folded conformation is the lowest free energy state[5].

In order for folding to occur, the attractive interactions in the folded state must be sufficient to overcome the large entropic cost of folding. For an energy gap to exist such that one native structure is greatly favoured over another is a non-trivial task, given the weak and relatively unspecific noncovalent van der Waals, hydrogen bonding and hydrophobic interactions that stabilise the folded structure[6,7]. The interplay between the configurational entropy of a polypeptide chain and the energy of its 3D state has led to the conclusion that there are folding pathways or energy "funnels", which systematically allow a protein to find the same stable state (Figure 1). The width of the folding funnel is representative of the configurational entropy of the chain and the number of potential states available, whereas the depth of the funnel is proportional to the free energy function (not including a protein's internal degrees of freedom)[8]. Regardless of starting position on the energy landscape, there are multiple pathways that lead to the final folded and stable conformation at the bottom of the funnel.

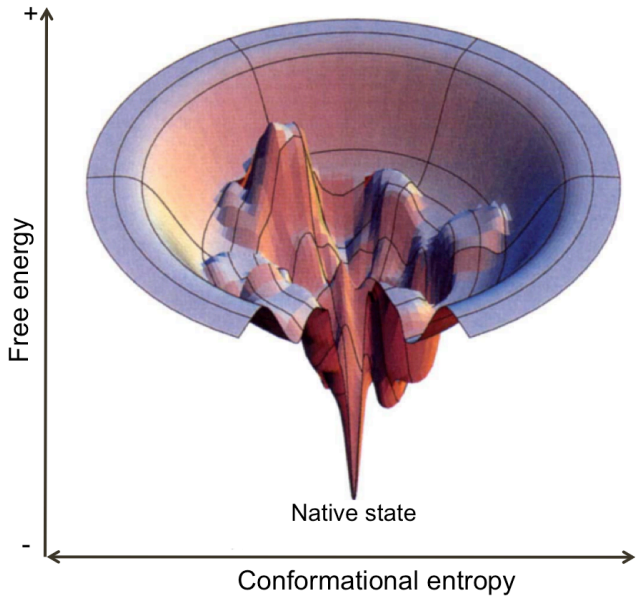

Figure 1. *Diagram of a protein folding funnel*. The width of the funnel represents the conformational entropy of the polypeptide chain and the depth of the funnel indicates the free energy. As conformational entropy is lowered with increasing structure, there is a subsequent drop in free energy. The lowest global minima corresponds to the native folded state. Adapted from Dill and Chan (1997)[204].

1.2 Solubility

A fundamental requirement for proper folding and function of proteins is the existence of a soluble state[9,10]. Many factors contribute to the solubility of a protein, including free folding energy[11], net electrostatic charge[12], and the number of exposed hydrophobic residues on the surface[13]. The aggregation of proteins and peptides increases with the amount of exposed hydrophobic area on the surface[14], and even partially burying these hydrophobic residues can lead to a large increase in solubility[15]. When attempting to express proteins under conditions differing from the native environments for proteins, aggregation can occur and lead to insoluble aggregates[16,17]. Traditionally, solubility of recombinant proteins has been optimised through the use of weak promoters, modified growth media, lower growing temperatures, or solubility tags[18,19]. There has also been some progress made in attempting to predict solubility based on the physicochemical properties of the primary amino acid sequence of the protein, allowing directed mutation to increase expression yields[20–22].

1.3 Fold space

1.3.1 Ideal forms

Proteins can adopt a multitude of different 3D structures (folds), and these can be classified into various families. Many classification methods have been devised for this purpose, which hierarchically cluster structures showing partial or overall similarity[1,2]. However, most of these methods have concentrated on attempting to optimally align 3D structures in space. This allows close relationships between similar proteins to be well defined, but gives the problem that global similarity or tentative relationships between folds can be difficult to interpret[23].

A significant step towards an objective fold definition was presented by Taylor in 2002[24]. The "ideal forms" of protein structures form a model where the hydrogen bonds across a β-sheet impose a layered structure onto the arrangement of secondary structures in the fold. Each layer consists of exclusively α-helices or β-sheets, with seldom more than 4 layers present (Figure 2). Using this method to classify folds is very successful, as it covers a substantial proportion of structures and can allow an unambiguous definition of the structure[25]. The ideal forms presented in the paper enable pathways to be made through secondary structure elements, which produces a topology string unique for that protein. For the purposes of our studies, the terms "fold" and "topology" are used interchangeably. "Topology" does not refer to its mathematical context, where most proteins would be seen as having the same open string topology, but to the relative packing and orientation of secondary structures regardless of twisting or shearing.

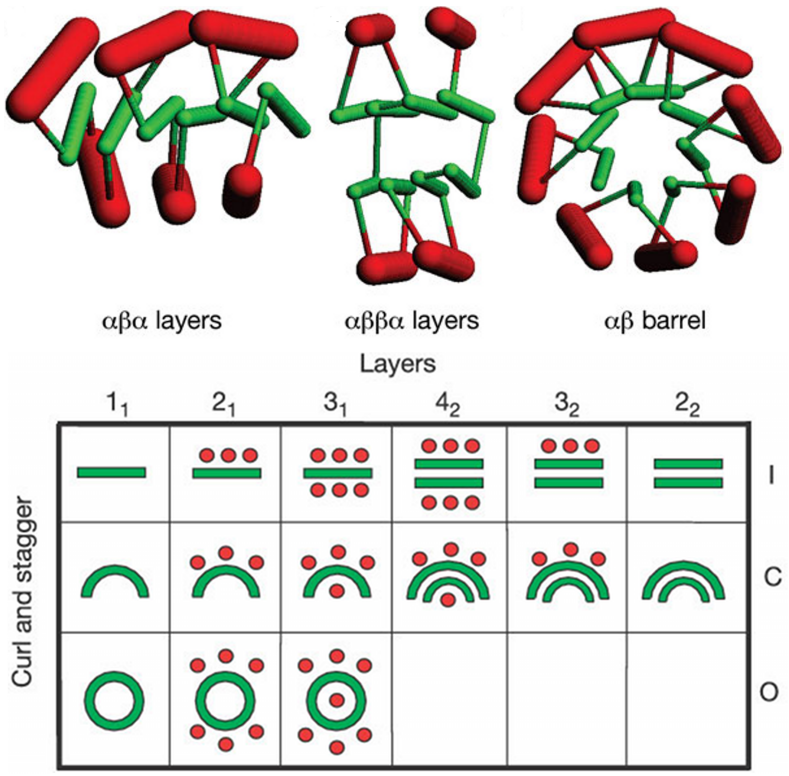

Figure 2. *The "ideal forms" of protein topology*. The 3 models across the top of the diagram highlight the main ideal forms for mixed α/β folds. Below is a representation of the 'periodic table' of these ideal forms. α-helical regions are depicted as red circles, with β-sheets as solid green lines. Each "element" of the periodic table gives an indication of the layering of secondary structures, irrespective of the number of secondary structure elements present in each layer. The curvature of the β-strands can also be represented (I=no curvature, C=partial barrel, O=barrel). Taken from Taylor (2002)[24].

1.3.2 Topology Strings

Using the ideal forms of protein structure, it is possible to create topology strings as unique identifiers for a specific fold[24,26]. The standardised method used in this approach makes it much easier to clearly define the boundaries of structural similarity.

To produce a string representation of a protein, a simple co-ordinate system is used. The layers of secondary structure are classed as "A", "B", "C" and "E", where layers "A" and "C" are helical whereas "B" and "E" are sheets

(shown in Figure 3). Each element in a layer is assigned a number based on its position in the layer, and "+/-" is placed before the element description to give it an orientation (i.e. facing frontwards or backwards). The first element present in the layer is assigned as position "0" and all other positions in the layer are defined relative to position 0. The first strand in the sheet always takes the positive orientation, +B+0, while the first helix is always in the top layer "A"[26]. The string produced via this method is unique for each individual fold, easily visualised and gives a simplistic indication of general structure. It also has the benefit of being able to recognised more global similarities between topologies that would be undetected or wrongly identified by root-mean-square deviation (RMSD)-based superposition methods of comparison.

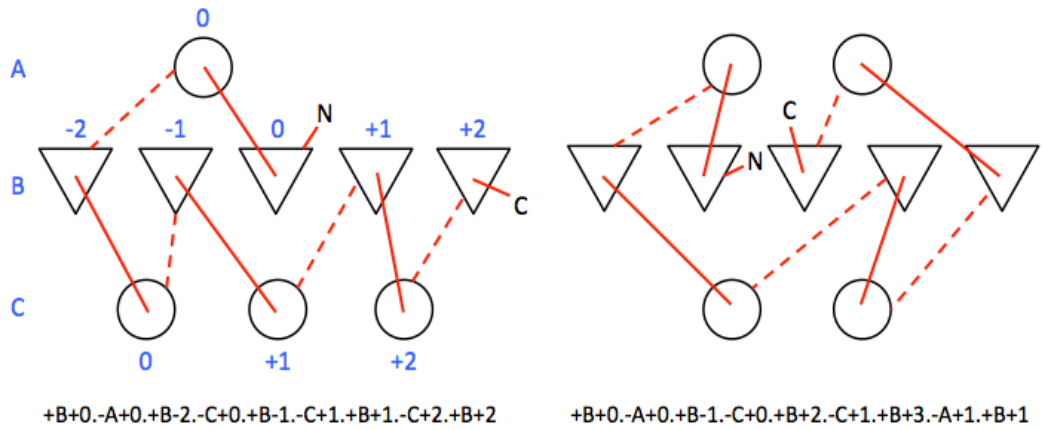

Figure 3. *Two examples of fold topologies with their respective topology strings*. Circles represent α-helices, whereas triangles are β-strands. The two models shown have similar secondary structural features and therefore belong to a similar group in the 'periodic table' of protein folds, but each has a unique descriptive topology string to identify it. Layers "A" and "C" are helical layers and layers "B" and "E" are sheets. Note: Layer "E" isn't shown in these examples but would be present between layers "B" and "C" for a double-layered sheet. A co-ordinate system is used to identify positions within a layer, with the first element to be present in that layer labelled as "0". All other positions are then given a position relative to this. An orientation of "+" or "-" is given to each element to give it an orientation within the structure. Using this method, a pathway can be traced through secondary structures to give a unique descriptive identifier for each protein fold. Adapted from Taylor (2009)[26].

1.3.3 Fold space

Considerable effort has been made towards understanding the space of potential protein structures and though the number of structures in the PDB is around 111,000[27], the rate at which novel folds are being discovered is becoming increasingly rare[28]. It has therefore been proposed that we now possess a structural representation for every basic, or "natural", protein fold[29]. Due to the large amount of known structures available, it has become more and more apparent that there are some very distinct relationships between particular sequences and the secondary structures that they are most likely to adopt[30,31]. Using homology-based template modelling, combined with energy functions to find the lowest energy state, detailed predictive models of protein structure can be produced that possess relatively small root-mean-square deviation (RMSD) from the native conformation[32,33]. Ultimately, this has led to speculation that all natural protein sequences can be modelled through assembly of fragments taken from the current collection[29,34].

However, the number of unique folds in the current database is only a fraction of what should theoretically be possible. Many groups have attempted to estimate the number of unique folds available, ranging from around 400 folds[35] to over 10,000[36]. More likely estimates seem to converge at ~4,000 novel folds, of which ~2,200 should be visible in nature[37,38]. When comparing the possible protein fold space to what have already been observed (using topology strings to identify distinct folds), there is a ratio of 10:1 that exists for unseen:known folds[26] (Figure 4). This suggests that nature is very restricted in the folds that it uses, favouring a specific subset of topologies and leaving protein fold space surprisingly empty.

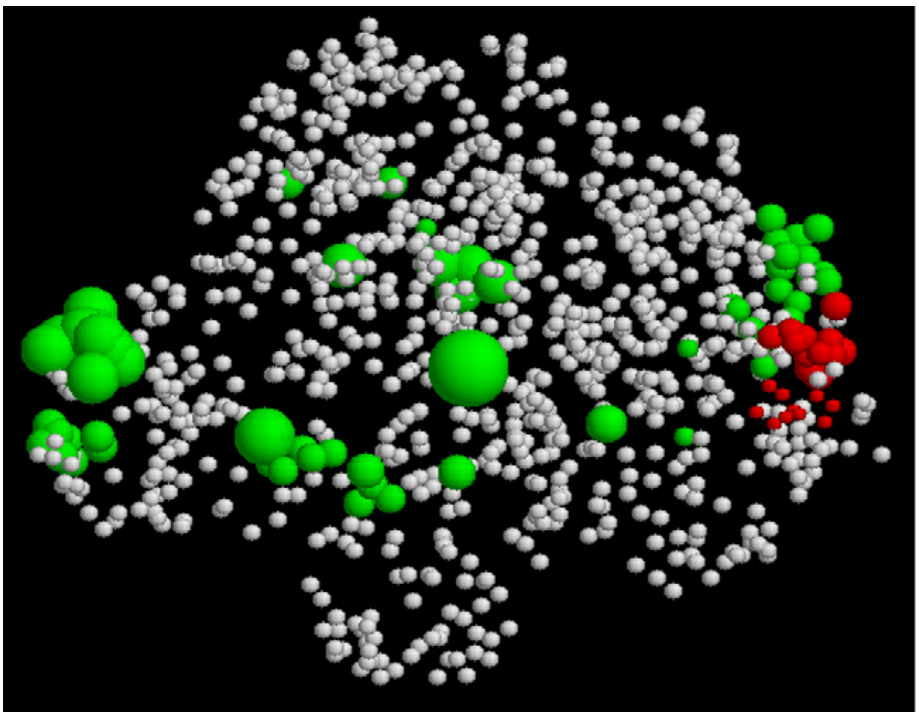

Figure 4. *RMSD projection showing distances between potential folds in 3D space*. Coloured (green and red) spheres represent known folds and white spheres are unobserved folds, with the radius of the sphere indicating the number of proteins known to contain the fold. There is about a 10:1 ratio of unseen:known folds in the projection, indicating that there is a lot of fold space that nature has not yet explored. Taken from Taylor et al (2009)[26].

1.4 Protein structure prediction

1.4.1 Comparative modelling

The tertiary structure of a protein can be predicted from the amino acid sequence using various methods. One of these methods is comparative modelling, which can be implemented with programs such as Rosetta[39] and I-TASSER[40]. Comparative modelling uses a template library of solved protein structures, from NMR and X-ray crystallography techniques, to build a model for a sequence with an unknown 3D conformation[41]. Firstly, the target sequence is aligned with the sequences of proteins that have known structures using an alignment program such as HMMer[42] or PSI-BLAST[43]. Then, incomplete models are produced based on the template structures by copying coordinates (alpha carbons, as well as phi and psi angles) from

aligned regions[44]. Loop regions are inherently more flexible and have greater conformational diversity than defined secondary structures and will not usually be consistently aligned for different proteins. To model the non-conserved loops of proteins *de novo*, Rosetta produces a fragment library (based on real structures) for the loop sequence and then iteratively combines these different fragment conformations to find the lowest energy state and reproduce realistic loop closure[45]. This process is similar to the *ab initio* protocol described in the next section. By stitching together templates from known structures and adding in appropriate loops between secondary structure elements, a coarse-grain model is produced. However, this coarse model may still have poor geometry at segment boundaries with steric clashes, unfavourable hydrogen bonding energies and distorted peptide bonds. To refine the model, a series of various small moves and perturbations are applied to the backbone atoms, with stochastic rotamer packing for side-chains of the residues. Using a Monte-Carlo method of sampling and an all-atom energy function scoring system[46], the energy minimum of the structure is investigated as the local environment of the structure is explored. The schematic of this process is outlined in Figure 5. The Rosetta prediction protocol suggests that a total of 10,000 models are made in this way, with the lowest 10% of models by energy selected and then clustered to give the optimal prediction of tertiary structure[45,47,48].

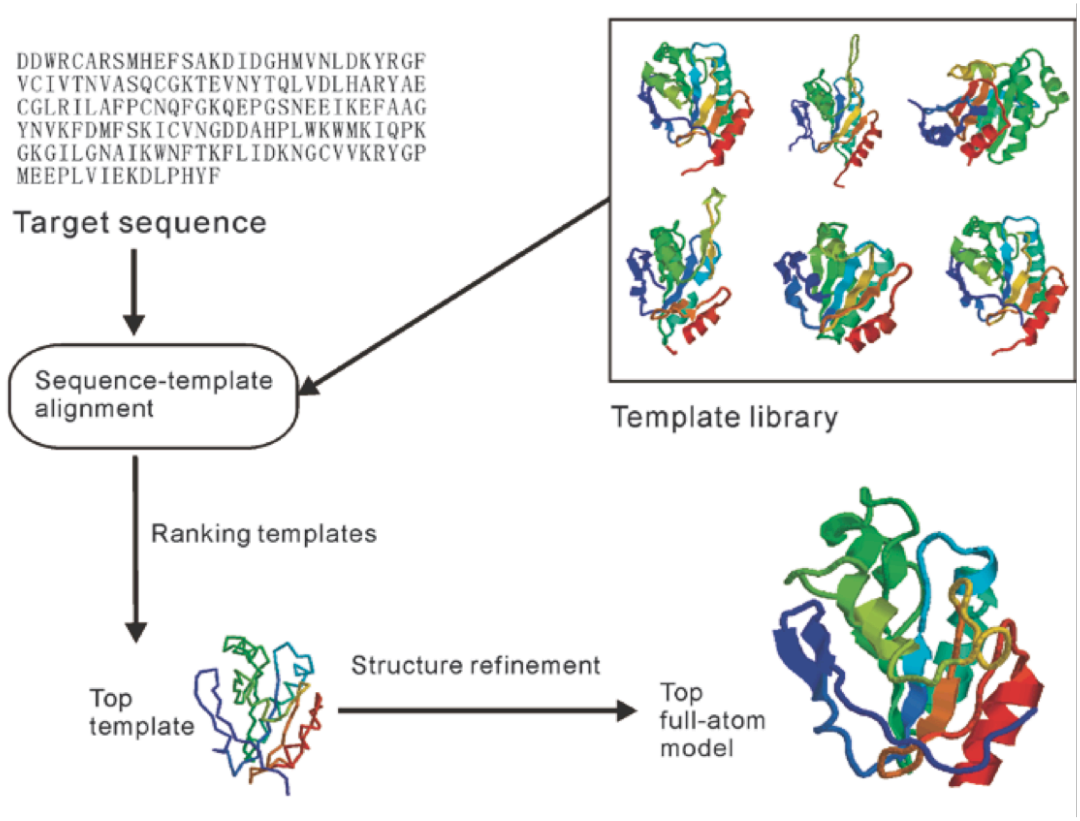

Figure 5. *Schematic representation of the comparative modelling process for protein structure prediction.* The target sequence is aligned with sequences that have a known structure. The templates are then ranked based on their sequence similarity and overlapping regions are copied for the coarse-grain model. Structural refinement is performed on this coarse model by Monte Carlo sampling methods aimed to minimize the energy of the structure. Adapted from Wu and Zhang[44].

1.4.2 Ab initio modelling

Ab initio modelling is another structure prediction technique that can be used, in the absence of templates, to produce a 3D model purely based on fundamental physicochemical properties.

The Rosetta software suite offers an option to predict the structures of proteins *ab initio*, and the basic features will be outlined here. For more comprehensive detail, the published manuals can be consulted[39,49–52].

There are two major tasks that an *ab initio* prediction protocol must perform. It must be able to sample the conformational space of a protein,

and then evaluate these resulting structures to determine which has the lowest energy. Conformational space is explored mainly with knowledge-guided Metropolis Monte Carlo sampling, where a substitution or perturbation is made to the structure and then accepted if the new conformation scores better. To evaluate potential structures, a knowledge-based energy function is used to score each one.

Because the conformational space of even a small sequence is so vast and the energy landscape so rugged, Rosetta starts with a coarse-grained search to find local energy minima. The structure starts as a completely extended chain, with side-chains treated as "super-atom" centroids (Figure 6A). This limits the degrees of freedom in the chain, while still conserving some physicochemical properties of individual residues[53]. By treating side-chains as soft interaction centres, the conformation of a protein is completely specified by the phi (Φ), psi (Ψ) and omega (Ω) torsion angles of the backbone. Next, a position along the chain is randomly selected to undergo change. At this position, a fragment of the same sequence is inserted from a library that has been produced by analysing bond angles between resides in fragments of real structures. If the subsequent conformation results in a lower backbone energy, the change is accepted and if there is a lower energy then the change is discarded. Another position is then randomly selected and the process continues for a given number of iterations (usually 5000) to produce a rough estimate of the global fold (Figure 6B)[49]. Initially, a fragment size of 9 residues is used in the replacement to produce the global fold and then the length is reduced to 3 residues for better local structural refinement.

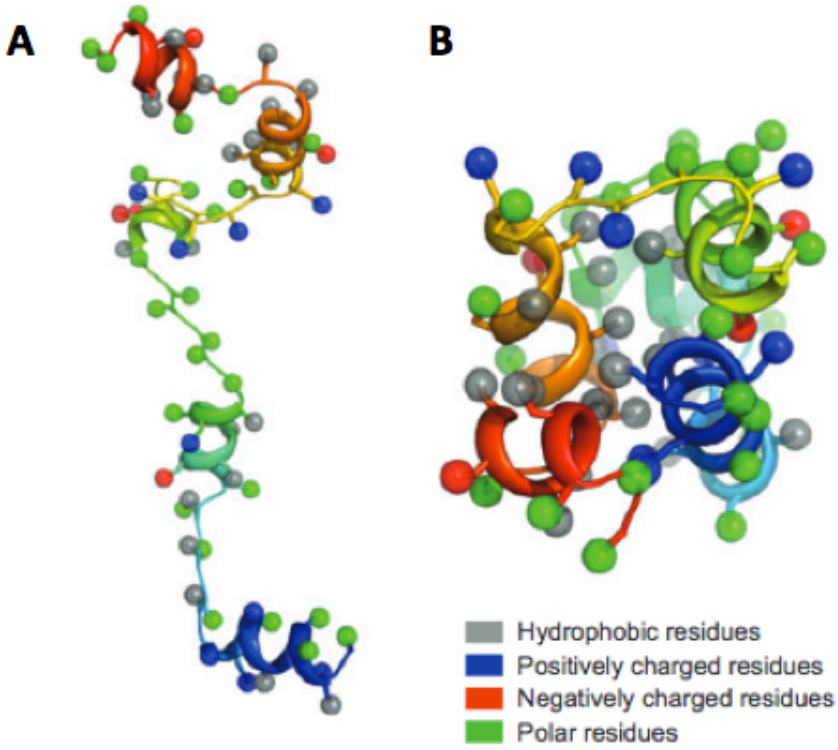

Figure 6. *Representation of the conformational sampling method used in the Rosetta program*. A) The protein is initially treated as an extended chain with side-chains coarse-grained as soft interaction centres (centroids). A position is selected randomly and a structural fragment, containing backbone torsion angles taken from known PDBs, of similar sequence is inserted into that position. If a lower energy is obtained, the conformation of that fragment is adopted. Higher energy changes are discarded and the process continues with another position being selected. B) Fragment selection and replacement proceeds in a Metropolis Monte Carlo method to minimize energy for a number of iterations and converges on a global fold. Because of this coarse-grained low-resolution approach, many local energy minima are found. Picture taken from Das and Baker[49]

This methodology has its own energy function scoring system, based on probability profiles derived from real PDB structures. Solvation effects are modelled as the probability of seeing a particular amino acid with a given number of alpha carbons in the local vicinity. Electrostatic interactions are captured by the probability of observing a given distance between centroids. The radius of gyration is used to model the effect of van der Waals attractive forces and centroid overlap is penalised to reproduce the repulsive component[39,52].

Coarse-graining the energy landscape allows multiple local minima to be found relatively quickly by the sampling technique. Once found, these can be explored in more detail by using a physically realistic, fine-detail atomistic model in order to distinguish between native and non-native states with greater accuracy. Starting from each minima, details of the side-chains are added back onto the structure and a simulated annealing search is performed through all combinations of discrete amino acid rotamers at each position. Monte Carlo sampling occurs again after this, with an assortment of torsion angle perturbations (which do not change the global fold of the protein), one-at-a-time rotamer optimisation and then continuous gradient-based minimization of side-chains and backbone torsion angles finishes the protocol[49].

The energy function used for this fine-detail approach is different from the one used for the low-resolution search. Atom-atom interactions are described using a Lennard-Jones potential, an implicit solvation term calculates the desolvation effects[54] and hydrogen bonds are modelled with an explicit potential[55,56].

By discovering local minima then optimizing with a fine-detail forcefield, the global minima can be explored. Although there are potential problems associated with the inaccuracy of energy functions and the approximations made by them[57], there is a good precedent for using the method for *de novo* structure prediction[53,58–60].

As with the comparative modelling process, it is suggested that a large number of models be made (~10,000) using this method and then the lowest 10% by energy clustered to find an optimal structure prediction[51].

1.5 Protein Design

As protein structure prediction techniques improve and accuracy increases, it opens up greater possibilities for protein design. Structure prediction attempts to determine the structure that a given sequence will adopt. The other side of this coin is being able to design a sequence that will fold into a given structure, i.e. the "inverse folding" problem[61].

Arguably, the field of protein engineering and design began with a series of papers from Gutte *et al.* that aimed to produce peptides with a desired function and a specific tertiary structure[62–64]. Although the peptides seemed to have correct functionality, these experiments did not contribute directly to understanding structural design principles as they proved difficult to characterise. Nevertheless, interest in the field was stimulated and further successes were to follow.

The next big development came in the mid-1980s when Eisenberg attempted to design a simple 16-residue sequence, from first principles, that would form an amphipathic helix and associate into a bundle[65]. This sequence was made by DeGrado using peptide synthesis, and optimised through an incremental design process[66]. To start, a sequence was designed that would fold into a single helix, based on circular dichroism (CD) spectra and association behaviour. Dimers and a tetramer were then made through linking identical helices together and optimising individual residues to produce the final form. The work culminated in the production of a *de novo* designed 4-helix bundle[67].

The first fully automated protein redesign came from Mayo and Dahiyat in 1997, when they successfully produced a sequence that folded into a target ββα backbone structure (taken from the zinc-finger motif of the Zif268 protein)[68]. Kim *et al.* also had success around this time, designing and experimentally characterising right-handed coiled-coils not found in nature[69]. Both of these achievements relied heavily on newly developed

energy functions that model atomistic interactions, a feat only possible with the increased computational power that was available.

Early experiments in protein design focused on simple tertiary structures, in order to reduce the complexity of the problem. By choosing helical bundles (possessing a high degree of symmetry) or a simplified ββα fold, the complexity of secondary structure interactions is reduced and allows for a greater chance of design success. Skipping ahead to today, the best-understood targets for structural design are coiled-coils, which have a high degree of symmetry along with a well-defined parameter set. Association of α-helices to form bundles is programmed at the sequence level by a very well characterised heptad repeat of hydrophobic (H) and polar (P) residues, HPPHPPP[70]. Modelling of a coiled-coil *in silico* can also be performed with a high degree of accuracy by invoking Crick's parametric equations, that use only 4 parameters[71]. Since the principles governing coiled-coil structures are so well understood, they are attractive targets for protein engineering experiments. In fact, large steps have been taken to standardise the design methodology and assessment of potentially viable candidates[72,73]. Complete *de novo* design of backbone and sequence has now been accomplished for helical targets[74–76] and the future potential in this area has only just begun to be explored.

While the successes in structural design have so far been very exciting, ideally we would like to head towards being able to produce more complex protein structures from first principles. Perhaps the most significant step in this direction was by Baker *et al.* in 2003 when they managed to experimentally characterise a completely *de novo* design of a backbone never before seen in nature (Figure 7)[77]. The protein, Top7, was designed with atomic-level accuracy while being highly stable and robust to mutation[78]. Using a similar protocol to the one used to make Top7, Koga *et al.* managed to produce 5 completely *de novo* redesigns of known folds[79].

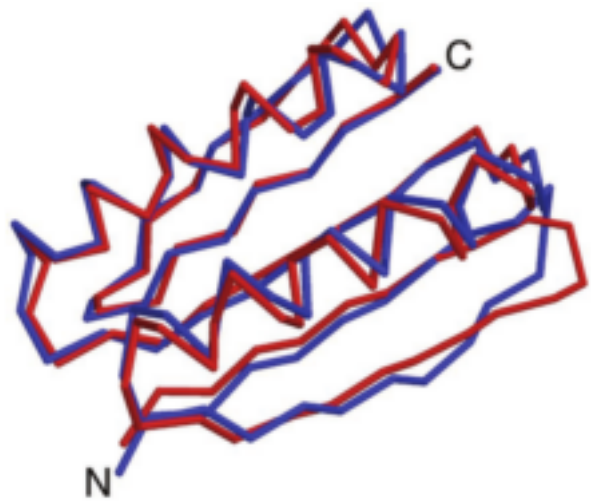

Figure 7. *Top7 computationally designed structure (blue) superposed with the solved x-ray structure (red).* There is very close agreement between the designed and experimentally characterised structures. The fold is a completely novel one, never seen before in nature, and makes Top7 one of the best demonstrations of protein design so far. Taken from Baker *et al.* [77]

Creating proteins purely from first principles still remains a huge challenge and although a number of groups have proposed different methodologies[80–82], experimental success for complete *de novo* design has been mainly limited to the experiments previously outlined.

Rather than complete *de novo* design, a more functional approach can be taken by permuting existing architectures or sequences to produce novel proteins. Hybrid proteins are a good example of using existing folds to create completely new topologies[83,84]. Protein domains are made from modular secondary structure elements with regular geometries[85]. These "building blocks" also form larger supersecondary structures that are present across domain families[86], e.g. the strand-turn-helix repeat of Leucine-rich repeat proteins. Linking these smaller, intrinsically stable subunits together can then create new hybrid folds[87,88]. A good example of this type of fragment recombination is the work of Hocker *et al*, when they successfully created a βα-barrel by combining fragments of two different proteins (CheY and HisF)[89].

Redesigning existing architectures to accommodate different (and sometimes novel) functions has also met with considerable success[90–93]. Natural protein sequence-structure relationships appear to be quite robust to mutation[94] and this property can be exploited in a directed way. To avoid having to design a protein sequence from scratch, designed functional sites

can be incorporated into folds that are already intrinsically stable. By adapting as much information as possible from natural precedent, the chances of success are greatly increased as native sequence-structure relationships are used to support the design process[95]. Perhaps the best successes using this methodology came from a series of papers from the Baker lab, where they successfully designed new enzymes to catalyse the retro-Aldol reaction[96], a Diels-Alder reaction[97], and a new Kemp-elimination reaction[98]. For all of these experiments, an active site engineered for a novel function was grafted onto a fold already found in natural proteins.

1.6 Discussion

The structure and function of a protein are specified directly by the sequence of amino acid residues it is composed from. The folded 3D structures that protein sequences adopt can be categorised into various families, based on common features. At present, the number of different protein folds found in nature appears to be remaining relatively constant, suggesting that there are only a specific number of structure types that have been used for the native landscape. This leads to questions as to why a substantial amount of fold types have not been used. Do these 'dark matter' folds possess some property that makes them difficult to utilise, or has evolution simply not had time to explore all of the possible fold space?

Advances in structure prediction are allowing us to predict accurate 3D models of proteins, solely from their primary sequence. The other side of this coin is; are we able to program a sequence that will adopt a desired structure? There has been success by using fragments of known proteins and repurposing or mutating them to engineer novel functionalities, but ideally we would like to build chosen folds completely *de novo* from first principles. There may be no such fragments that can be repurposed for a novel fold topology, and so utilizing this approach may place boundaries on what is possible. Some new folds may only be accessible through complete *de novo* design, and so it is important to attempt to make advances to this

end. The field of coiled-coil design is the best understood, and huge efforts have been made to produce completely *de novo* designs. That this can be done is hugely exciting and encourages us to delve deeper into more structure types to see if similar success can be found.

In this thesis, I begin by using established design methodologies to try and produce some novel topologies for folds containing both α-helices and β-sheets. This would help to discern how well previous successes generalise to other fold types. I then attempt to probe the general rules that govern solubility and foldedness of protein structures, in the hopes of elucidating general principles that can be used to optimise the design processes.

2. Methods

This section aims to outline the general methods that apply to most chapters of the thesis. Since the exact protocols vary between chapters, it is meant to offer a broad view of the techniques that are consistent across most. For specifics on a method used, the "Methods" section within each chapter should be consulted.

2.1 Computational

2.1.1 Comparative modelling

During our investigations, we rely on comparative modelling in order to produce a 3D prediction of the structure that our designed sequence is likely to adopt. By predicting the structure before experimentally testing our designs, we can computationally screen sequences in order to cut down on experimental time and cost. The predicted model can be compared to the template structure that we are aiming for, to give us an indicator of how well we think that design should perform. Our comparative modelling processes to produce these predictive models were carried out according to the guidelines provided by Baker *et al*[45,47,48]. Because we are using the Rosetta program for both design and prediction, we thought it prudent to double-check our predictions using the I-TASSER modelling suite[40]. I-TASSER and Rosetta have consistently been the highest scoring platforms in recent years at CASP[99] and we found a high level of agreement between the two systems when predicting our designs. All predictive protocols followed the guidelines provided by each program manual and further details should be taken from them if elaboration is required.

2.1.2 Ab initio prediction

We also employ *ab initio* prediction methods quite heavily in this thesis. For some of our protein designs, especially ones with novel folds, there are not any solved structures that can be used as a template. Therefore, to avoid biasing towards native structures, we used an *ab initio* method of prediction.

By doing this, we level the playing field for our novel designs when compared to native comparison predictions. Again, standard protocols for *ab initio* structure prediction were used and the relevant literature can be consulted for specific details[51]. Comparative modelling is also used after *ab initio* predictions, as a final checkpoint for potentially good designs.

2.1.3 Rosetta design

Closely linked to the *ab initio* structure prediction protocol in Rosetta is the protein design capabilities that it possesses. By using its rotamer libraries as a sampling method and the energy function to score the new structure, Rosetta can attempt to find an optimal fit for a non-native amino acid in a given residue position. The protocols provided allow a user to select one or more positions that should be redesigned, as well as specify the types of residues that are permissible at that location (e.g. only allowing the system to choose from hydrophobic resides in the core)[100]. There have already been successes using this technique for protein design including the production of new disease therapeutics[91,101,102], self-assembling peptides[103,104], novel enzymes[96,105] and even completely redesigned sequences[106] or novel folds[77,79].

The specifics of our design protocols are explained in greater depth within the chapters in which they are relevant.

2.1.4 TM-scoring

When predicting the structure of proteins, it is important to have a method of establishing how close any two given structures are to each other. For example, if a reasonable redesign of a native sequence has been found then it would be useful to make a prediction of what the new sequence's structure would look like and then compare this to the structure that the original native sequence adopts. This would give an indication of whether the new design is computationally valid.

Traditionally, root mean square deviation (RMSD) and the Global Distance Test (GDT) are the measures that have been used for this purpose. RMSD measures the average distance between equivalent atoms in the model and the template[107], whereas the GDT measures how many alpha-carbons of the model fall within a defined distance cut-off from the template[108]. Both of these methods have a power-law dependence on protein size, leading to some inconsistencies in how they treat different proteins[109]. They are also very sensitive to outlier regions, such as loops or terminals, which may not be well-predicted[110,111] and so a higher RMSD may be found even if the rest of the model is in strong agreement. Both of these techniques are susceptible to noise and since we are interested in the global folds of the proteins we produce (which may have quite a bit of noise), we used the template-modelling score (TM-score) as a comparison method.

TM-scoring aims to rescale structural modelling errors so that the impact of outlier regions on the overall score is minimized, and is also independent of size[109]. The score is given between 0 and 1, with better templates having higher scores (a perfect model-template comparison will have a score of "1"). Generally speaking, random similarities are below 0.17 TM-score, a significant portion of structure is correctly modelled at 0.35 and the same fold is judged at >0.5[59,112].

2.1.5 Baselines of prediction and TM-scoring

To gain a baseline to test our designs against, we performed a quick *ab initio* structure prediction (1,000 models) on 100 native sequences and then compared the average TM-score of these predicted models to the known structures. For comparison, we also predicted the same number of models for random sequences and then mapped these back onto random structures of equivalent size (Figure 8). For the random sequences mapped onto random structures of equivalent size, the average TM-score across all 1,000 models was around 0.12. This suggests that there is no real correlation

between structures, which is expected. For the native structure predictions, the distribution seems centred about 0.39. This is not perfect, as we would like to have a TM-score >0.50 but this is one of the limitations of purely relying on the *ab initio* structure prediction method. There is still a high degree of structural similarity between the average prediction model and the native structure (>0.35). We now have values to aim for during our design experiments. If our designs are tending towards a value around 0.12, then we know our designs need more improvements. However, the closer we get to being at a >0.35 level, the more native-like our designs will be.

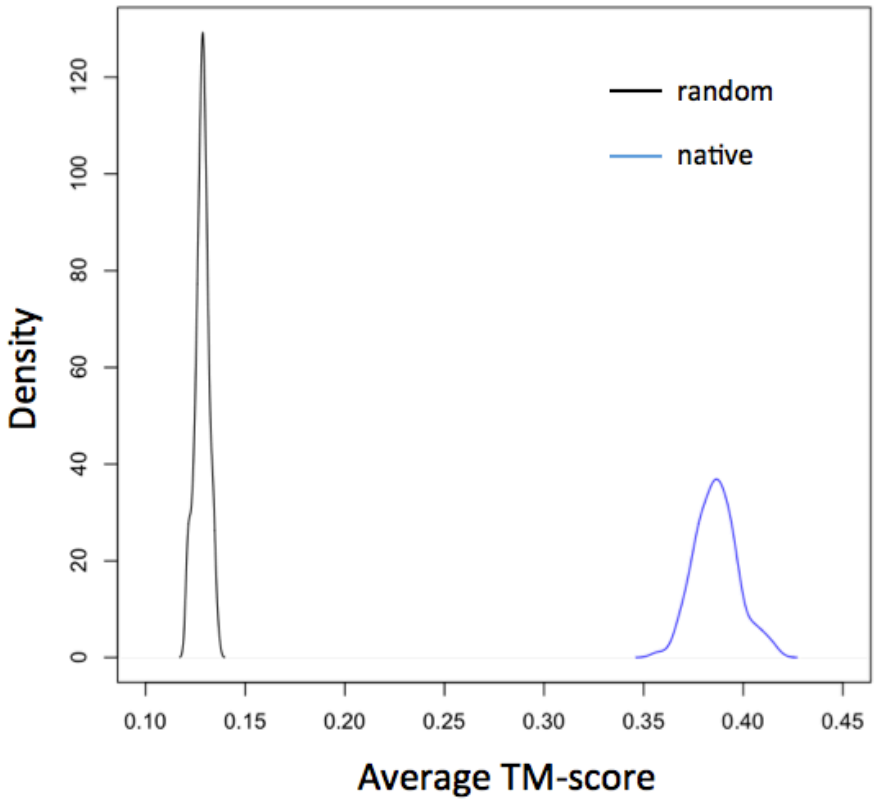

Figure 8. *The density of average TM-score over 1,000 predicted models for 100 native and random structure-sequence combinations*. The ab initio protocol was used to make predictions for a native sequence-structure pairing or a random sequence predicted, and then mapped back onto a random structure of equivalent size. The random pairings had an average TM-score about 0.12, while the natives had a much higher score of 0.39. This gives us a baseline to aim for when designing new sequences. A higher average TM-score after prediction will mean that our sequences are more native-like.

To offer a structural view, examples are provided of structures at 0.12 and 0.35 TM-score to the original (Figure 9). The 0.12 TM-score structure (Fig 9B) has hardly any similarity to the original 3CHY structure (Fig 9A). With the 0.35 TM-score model (Fig 9C), global similarities can be seen between the two. Three helices are relatively well formed, with some packing of the beta strands in the centre of the structure. However, the termini of the protein are not predicted well (blue and red sub-structures in the diagram) and do not exist as the tightly packed helices they are supposed to. Again, the structures are not perfect due to the limitations of the *ab initio* prediction process. The models produced should still be sufficient enough to offer a quick insight into how our designs are doing when compared to native sequences.

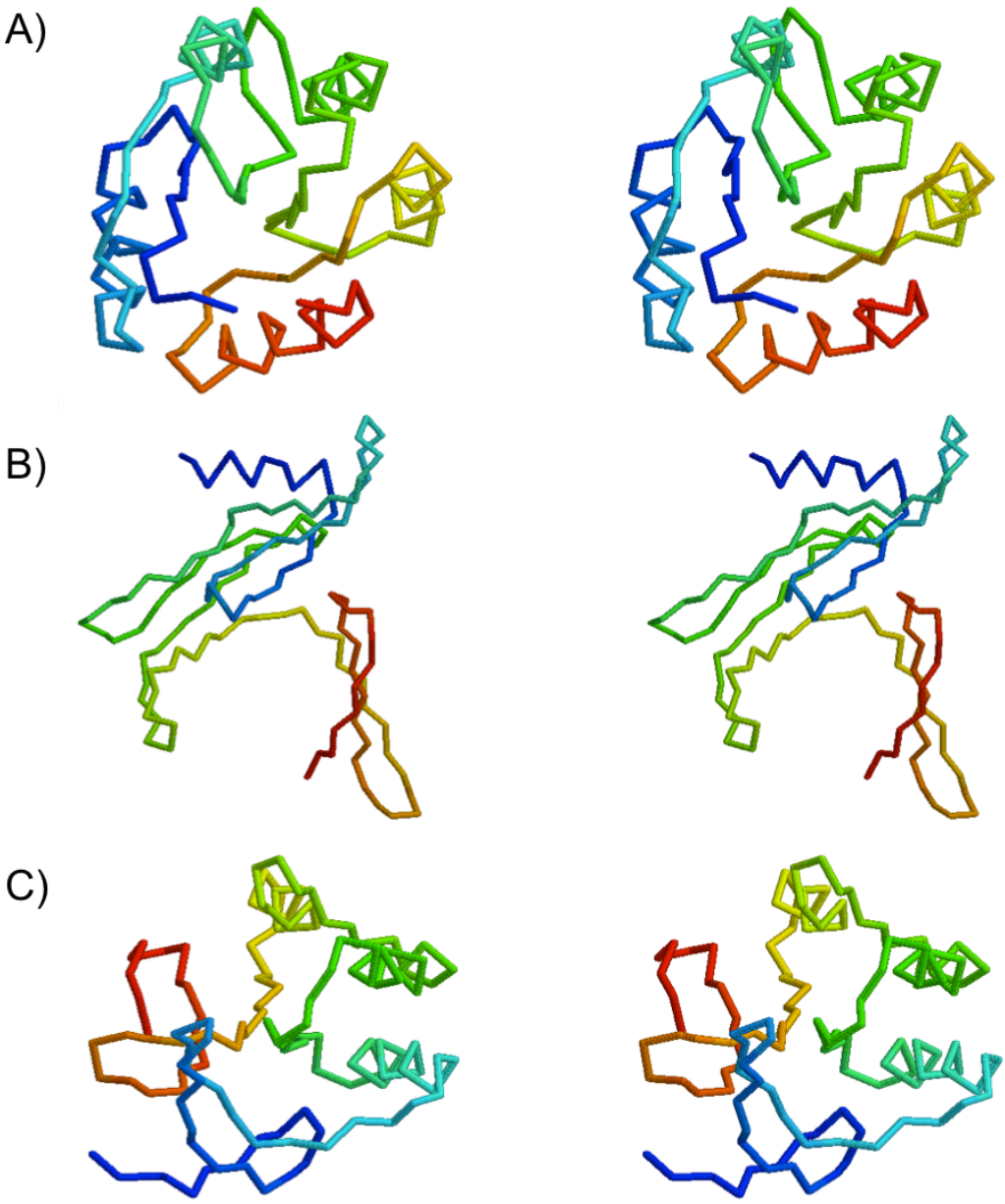

Figure 9. *Illustration of different TM-scores*. A) The original 3CHY structure with well defined global fold and substructures. B) A model that has 0.12 TM-score to the 3CHY template. There are no real commonalities between the prediction and the template structure with substructures and the global fold missing. C) A 0.35 TM-score model. There are quite a few common substructures such as well-formed helices and some packing of the beta strands into the core of the protein. However, the terminal ends (represented by red and blue sections) are not well formed. A higher TM-score is indicative of a higher structural agreement between model and template.

In all chapters of this thesis, the *ab initio* structure prediction offers an initial stage for optimisation, as we can run this relatively quickly, and then full comparative modelling is used to ensure a greater level of accuracy.

2.1.6 Definition of core residues (POPS)

Our definitions for "core" positions were any residue that has an alpha-carbon with a relative surface accessible surface area (rSASA) of 0.30 (or 30%). To obtain the rSASA for each residue position, we used the program POPS[113]. The 30% exposure cut-off seems to pick out the correct core residues, even ones packed on the inside of helices (Figure 10).

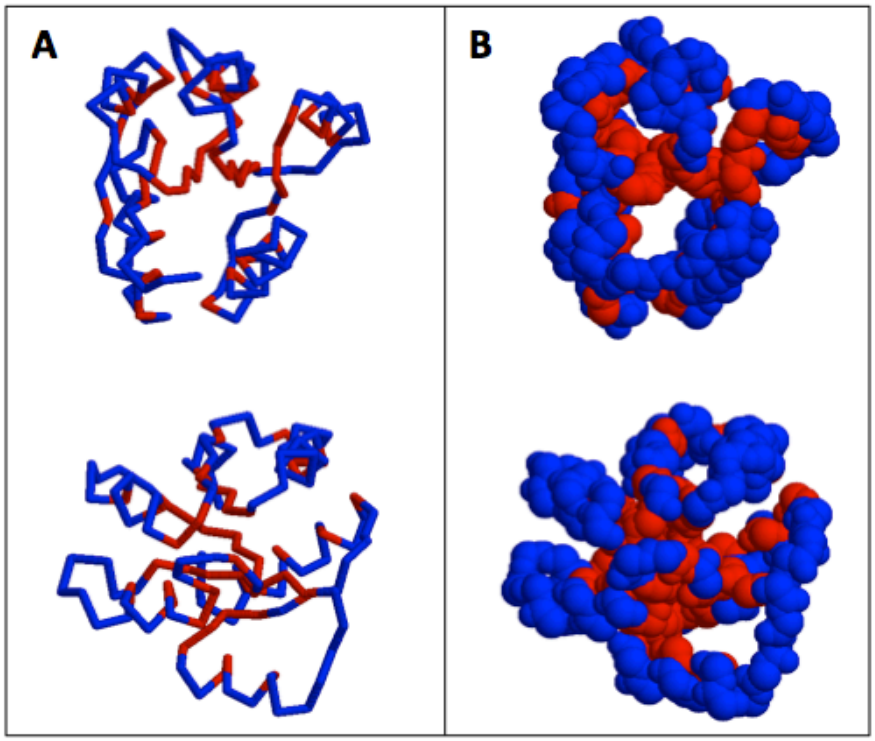

Figure 10. *Diagram showing core and surface residues picked out on the 3HCY fold*. A 30% rSASA was used to define the boundary between core and surface residues. Residue positions less than 30% exposed were classed as "core" (red) and the remaining positions are "surface" (blue). This exposure definition picks out residues in the middle of the protein as "core", with the positions on the inside of helices also chosen. A) Shows the backbone view of the 3CHY fold to enable a better view inside the protein. B) Shows the backbone atoms as shells, to better illustrate the packing.

2.2 Experimental

2.2.1 Gene synthesis

Constructs taken through for experimental testing were first generated as 60-mer overlapping primers (designed with an in-house program) and pieced together using the Thermodynamically Balanced Inside-Out (TBIO) gene synthesis method[114] with KOD polymerase (*New England Biolabs, UK*). The constructs also had specific sequences on the termini, to allow them to be used for ligation independent cloning. These sequences were 5'-TACTTCCAATCCATG-3' added to the 5' end of the upstream primer and 5'-TATCCACCTTTACTGTTA-3' added to the 5' end of the downstream primer. The necessity of these is explained further in Section 2.2.3. Codon bias for expression in *E. Coli* was taken into account while designing primers for our designed proteins.

2.2.2 Agarose gel electrophoresis

DNA preparations were analyzed by electrophoresis on a 2% agarose gel made with Tris-Acetate-EDTA (TAE) buffer containing 0.2 mg/ml ethidium bromide for 30 minutes at 180 V. DNA loading buffer was added to DNA at a ratio of 1:5. DNA was visualized by exposure to UV radiation and the size of DNA fragments was determined through comparison with standard markers (Bioline Hyperladder I).

2.2.3 Ligation Independent Cloning (LIC)

Genes were transformed into E. Coli BL-21 gold (DE3) cells (*Agilent Technologies, UK*) through the use of ligation independent cloning (LIC). This technique utilizes the 3'->5' exonuclease activity of the T4 DNA polymerase to generate complementary base overhangs in vector and insert. The vector and insert are then mixed, added to competent cells, and heat-shocked for a very efficient transformation process.

We used a pNIC28-Bsa4 vector that possesses Kanamycin resistance, a N-terminal His_6 purification tag, and a TEV protease cleavage site (*Structural Genomics Consortium, UK*). The vector also contains the SacB gene surrounded by two BsaI cleavage sites (Figure 11A) and a T7 promoter upstream of this. The SacB protein converts sucrose to a toxic product, meaning that when it is present in a cell, that cell will die. The idea is to use BsaI restriction to release the SacB gene and insert a desired fragment in its place (Figure 11B). If a plasmid is not cut effectively and the SacB gene not removed, growing cells in the presence of sucrose will select these cells out of the population. 5 µg of vector was treated with 2.5 µl of BsaI (10 U/ µl) in 5µl of 10X New England Biolabs buffer 3 (NEB3) for 1 hour at 50 ^{o}C, and linearization of the plasmid was confirmed by running on a 1% agarose gel.

Next, the vector was treated with T4 DNA polymerase (*Qiagen, UK*) in the presence of dGTP. Because of the 3'->5' exonuclease activity of this enzyme, bases are removed from both 3' ends until the first guanine (G) is reached. When a guanine is reached, the polymerase activity of the enzyme will take over due to an excess of this base in the reaction buffer (Figure 11C). 600 ng of the BsaI-digested vector was incubated in a reaction mixture (2 µl 10X NEB3 buffer, 0.5 µl 100 mM dGTP, 1µl 100 mM DTT, 0.2 µl 100X BSA, 0.4 µl of 3 U/µl T4 DNA polymerase) for 30 min at 22 ^{o}C to allow T4 treatment to occur, before being heated to 75 ^{o}C for 20 min to inactivate the polymerase. 0.2 pmol of our synthesized genes also underwent this protocol, with exactly the same volumes and conditions, but with dCTP in place of dGTP. This gives us complementary overhangs that can be annealed before transformation occurs (Figure 11D). 1 µl of the treated vector mixture was added to 2 µl of the insert mixture, with addition of 1 µl EDTA (25 mM), and incubated for 15 min at 22 ^{o}C to allow annealing.

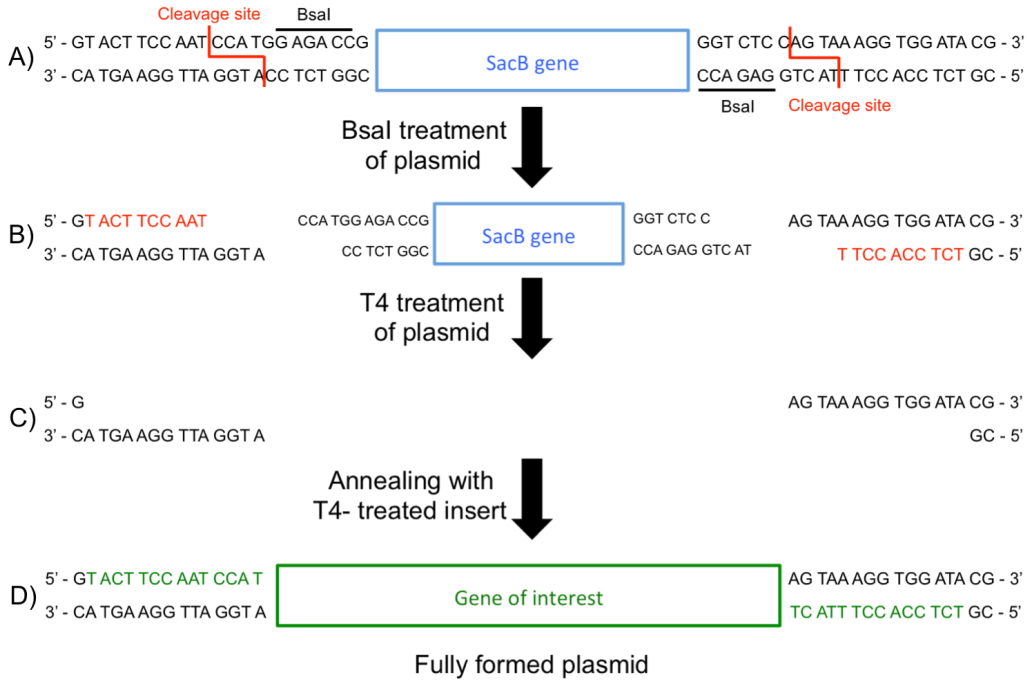

Figure 11. *LIC protocol for preparing vectors to be used in transformations*. A) The pNIC-Bsa4 vector has a SacB gene surrounded by two BsaI cleavage sites. This SacB gene produces a protein that converts sucrose to a product that is toxic to the cells. B) Upon BsaI treatment, the SacB gene is released, linearizing the plasmid and leaving the ends open for T4 polymerase treatment. In the presence of dGTP, the 3'-5' exonuclease activity of the T4 enzyme will remove bases from the 3' end. This stops at the first guanine base (G) because there is an excess of this base in the reaction buffer and so the polymerase activity of the enzyme will take over. Bases that will be removed are marked in red. C) After T4 treatment of the vector, we are left with overhangs that can be used to anneal to a different insert that has complementary sequences. D) Complementary T4-treatment of an insert can generate these overhangs and allow plasmid formation and transformation into cells without the need to ligate the different fragments together.

After annealing, a standard transformation protocol was used to put the constructs into BL21-gold (DE3) cells. 1 µl of the annealing mixture was added to 50 µl of competent cells and incubated on ice for 30 min. The cells were then heat-shocked for 45 seconds at 42 ^{o}C before quickly being places back on ice for a further 2 min. Cells then recovered in 200 µl of SOC medium (2% (w/v) tryptone, 0.5% yeast extract, 0.05% NaCl, 20 mM glucose) at 37 ^{o}C for 1 hour before being plated on agar plates containing 5% sucrose and 50 µg/ml Kanamycin. If cells had not taken up the plasmid, they would not have possessed Kanamycin resistance and so would be killed

off. Similarly, any cells that had taken up uncut plasmid where the SacB gene was still present would be sensitive to the sucrose in the agar plates. This allowed us to pick only colonies that had taken up the plasmid with the correct inserts, and reduced the number of false-positives.

A single bacterial colony from each agar plate was picked and taken into 10 ml of Luria Bertani medium (LB - 1% (w/v) tryptone, 0.5% yeast extract, 0.5% (w/v) NaCl), again containing 50 µg/ml Kanamycin. This was incubated at 37 ^{o}C, 200 rpm for ~5 hours. After making bacterial stocks for future use (Section 2.2.4), plasmid was then recovered from cells using a QIAGEN Plasmid Mini Kit (following the protocols outlined by the manufacturer) and sequences verified externally by the GATC company to ensure that the correct inserts were present.

2.2.4 Bacterial stocks

To ensure we had consistent stocks for future use, frozen bacterial stocks were made. 500 µl of bacterial culture was added to 500 µl of 50% glycerol in a 2 ml cryovial and gently mixed. This was then stored at -80 ^{o}C.

2.2.5 Protein expression

10 ml of LB culture media (1% (w/v) tryptone, 0.5% yeast extract, 0.5% (w/v) NaCl) containing 50 µg/ml was inoculated with transformed BL21-gold (DE3) cells containing the appropriate sequence verified plasmid. This was left to grow at 37 ^{o}C, 200 rpm to an OD_{600} of 0.6 and then protein expression was induced with 1 mM IPTG overnight at 25 ^{o}C, 200 rpm[115].

In the case of proteins for NMR experiments, cells were grown in PG minimal medium (50 mM Na_2HPO_4, 50 mM KH_2PO_4, 2 mM $MgSO_4$, 0.2X trace metals, 0.5% glucose) containing isotope labelled ammonium sulphate-$^{15}N_2$ (*Cambridge Isotope Laboratories, UK*) in place of LB media.

2.2.6 Protein purification

Cells were pelleted by centrifugation at 3,000 RPM for 10 minutes, the supernatant removed and the pellet resuspended in 20 mM Tris-HCl (pH 7.7) at 10 ^{o}C for 15 min. Once pellets were fully resuspended, cells were lysed using 60 mM Tris-HCl (pH 7.7), 90 mM Imidazole (pH 8), 900 mM NaCl, 0.6% OTG, 30% glycerol, 30 mM $MgSO_4$, 4 U/ml DNase I, 3 mM TCEP and 0.3 mM PMSF for 15 min at 4^{o}C with shaking. The resulting lysates were then spun at 5,900 rpm for 1 hour at 4 ^{o}C in order to pellet cell debris and insoluble protein aggregates. After centrifugation, the proteins were purified on a Ni^{2+}-NTA affinity column as per the manufacturers instructions (*Qiagen, UK*).

2.2.7 Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis

5 µl of NuPAGE LDS Sample Buffer (*Invitrogen*) was added to 15 µl aliquots of protein sample, and heated at 90 ^{o}C for 5 min. Protein samples and Mark 12 protein standards (*Invitrogen*) were loaded onto a NuPAGE 4-12% Bis-Tris Gels (*Invitrogen*). Gels were run at 180 V for 35 min in NuPAGE MES SDS Running buffer (*Invitrogen*). Protein bands were visualised by staining Coomassie blue (*Invitrogen*).

An example of what we would class as "soluble" protein is given in Figure 12. In the whole cell lysate of Design 1, there is a clear band at the appropriate size to indicate that our design is expressed. The band indicating our expressed protein is still present in the soluble fraction after centrifugation. This suggests that the protein doesn't precipitate out of solution, and is therefore soluble. For Design 2, the whole cell fraction again shows that our protein is expressed in the cells. However, the band is lacking from the soluble fraction for this design and this indicates that this design is not soluble, forming aggregates that precipitate out of solution.

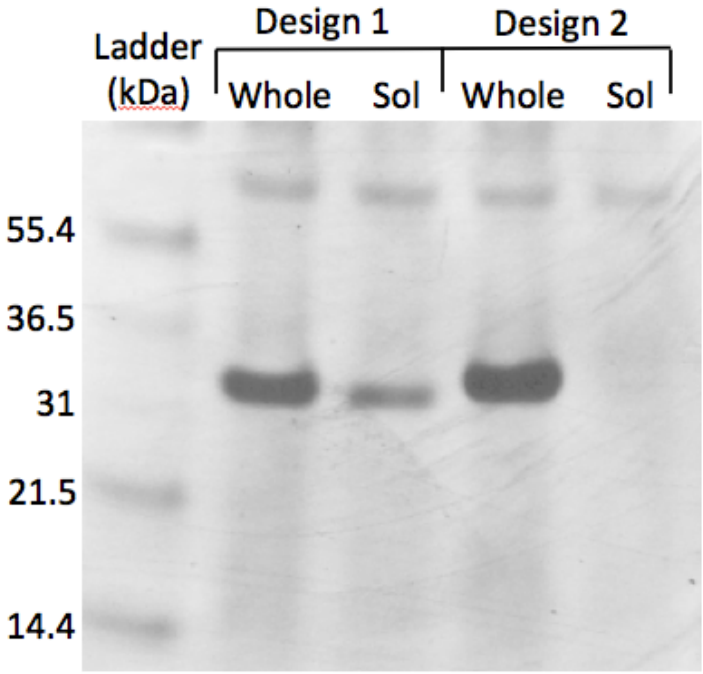

Figure 12. *Example of what we define as "soluble" and "insoluble" proteins, shown on an SDS-PAGE gel with Coomassie Blue staining*. Design 1 is a soluble protein, with a band of appropriate size present in both the whole cell and soluble fractions. This indicates that it does not form aggregates and therefore doesn't precipitate out of solution. Design 2 shows an insoluble protein where the protein is expressed, as seen in the "whole" cell lysate but not soluble. The lack of a band in the "soluble" lane of the gel suggests that the protein forms aggregates, precipitates out of solution and then is removed by the centrifugation step.

2.2.8 Size exclusion chromatography (SEC)

Apparent soluble designs were purified and tested for foldedness using a size exclusion chromatography (SEC) column with pore size 10 kDa or 20 kDa, depending on the backbone. Superdex 75 (*GE Healthcare, UK*) columns were equilibrated with buffer (20 mM MES pH 6.5, 50 mM NaCl, 1 mM TCEP) on an AKTA purifier 10 (*GE Healthcare, UK*) with a flow rate of 1 ml/min. 5 ml of sample was injected onto the column and eluted in 1 ml fractions, with absorbance tracked at a wavelength of 280 nm. Samples were collected and analysed via SDS-PAGE (Section 2.2.7).

In the event of all proteins being eluted in the void zone, a new buffer containing 50 mM NaCl, 400 mM Urea and 200 mM $MgSO_4$ was used with

the same columns (after washing and re-equilibration) in order to disrupt any potential aggregates that had formed[116].

3. Design of novel folds

3.1 Introduction

As highlighted previously, there are many unexplored regions in protein fold space[26]. Given the reported successes in protein design using the Rosetta program, we attempted to make some novel folds not yet found in nature. Firstly, we identified some fold topologies that are not found in solved PDB structures. To do this, we converted every structure in the PDB into a topology string in order to give us a set of known full folds. Then, we split each global topology into smaller fragments by progressively removing a secondary structure element from each terminal of the protein. By doing this, we had a representative set of full folds, as well as substructures contained within them. We then chose two ideal forms[24], a 2-4-0 and a 2-4-2 structure (Figure 13), and combinatorially generated all possible topologies for these forms. By filtering out known topologies from the selection, as well as ones with uncommon features (crossing loops, left-handed connections, parallel connections, layer skips etc), we identified potentially realistic novel fold topologies.

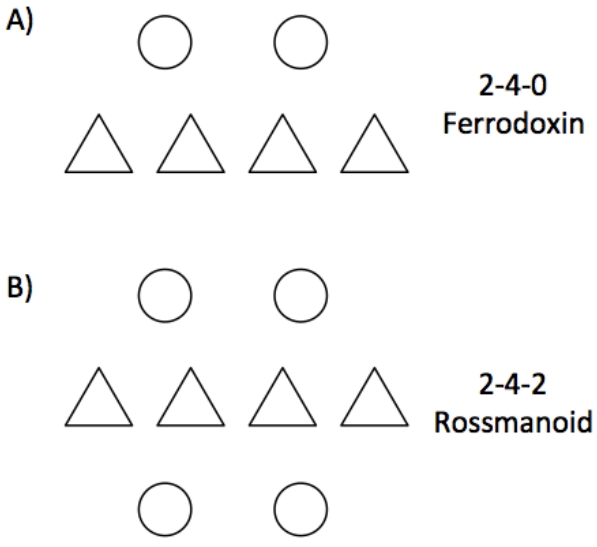

Figure 13. *Ideal protein forms used for investigation into novel folds.* A) A ferrodoxin-like, 2-4-0, structure. B) A rossmanoid fold with a 2-4-2 format. All possible pathways through the forms were made. Then folds that are found in the PDB were removed, along with any topologies with uncommon features (crossing loops, layer skips etc). This left us with a list of novel folds that could be investigated. The representations shown are empty forms, with no topologies shown and no orientation implied for any of the secondary structure elements.

Once novel topologies were found, we employed a previously published backbone construction program to create the scaffold of the fold[117]. We then used Rosetta with a flexible backbone design protocol to find an optimal sequence for this structure and attempted to experimentally characterise them through solubility screening and nuclear magnetic resonance (NMR) studies.

3.2 Methods

3.2.1 Novel topology identification

We began our search for novel folds by first generating a large number of ideal forms according to prescribed lattice structures. These are essentially the endpoints of secondary structure elements according to fold type[24]. We then compared each of these forms to solved native structures in the PDB, in order to understand which fold topologies are represented naturally. The methods to do this have already been outlined in previous papers[24], so only a general overview will be given here. Firstly, the native structure was reduced to linear segments to give a "stick-like" representation of the fold[118]. Graph matching was then used to pre-filter the closest forms[119], followed by a more in-depth double-dynamic superposition algorithm to score fits between native stick models and ideal forms[120]. Once the best scoring form had been found, connectivity between secondary structure elements allowed us to obtain the global topology string for the native fold. Global topology strings for folds were also split into smaller sub-topologies to ensure maximal coverage. There are some potential problems with using this method for β proteins with internal repetition (β-trefoils, β-prisms etc) or structures possessing orthogonal secondary structure elements. However, these are relatively rare and our ideal form choices of 2-4-0 and 2-4-2 are very well represented using this method.

Next, we combinatorially generated all possible topology strings contained within our 2-4-0 and 2-4-2 ideal forms. Strings that were found in native

proteins were then eliminated, along with any that possessed uncommon features such as crossing loops, left-handed connections, parallel connections, or layer skips. This left us with a number of novel topologies that have so far not been discovered in nature. From these, we chose 8 topologies for both the 2-4-0 (Figure 14) and 2-4-2 (Figure 15) folds.

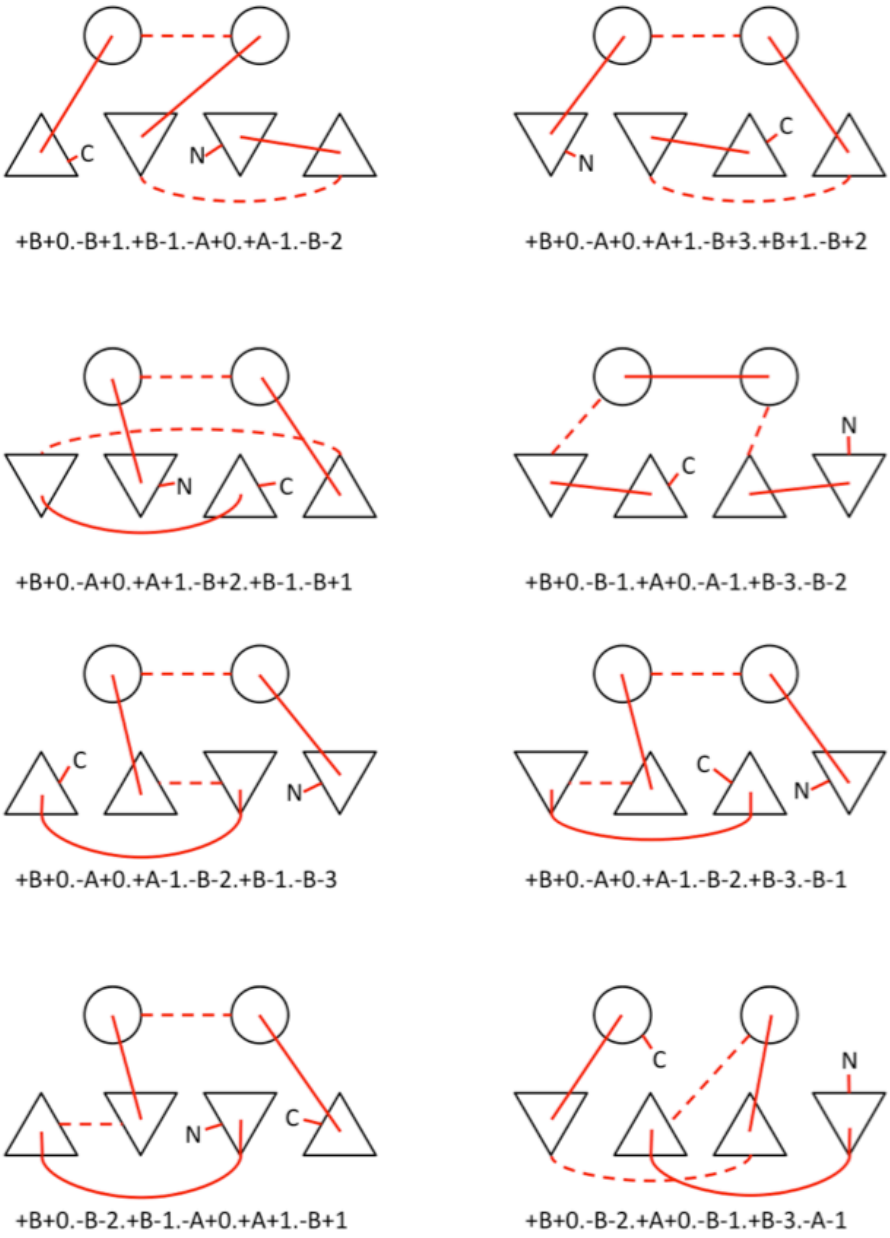

Figure 14. *Topologies chosen for the 2-4-0 fold*. We chose 8 topologies that were not observable in nature and attempted to design and experimentally test them

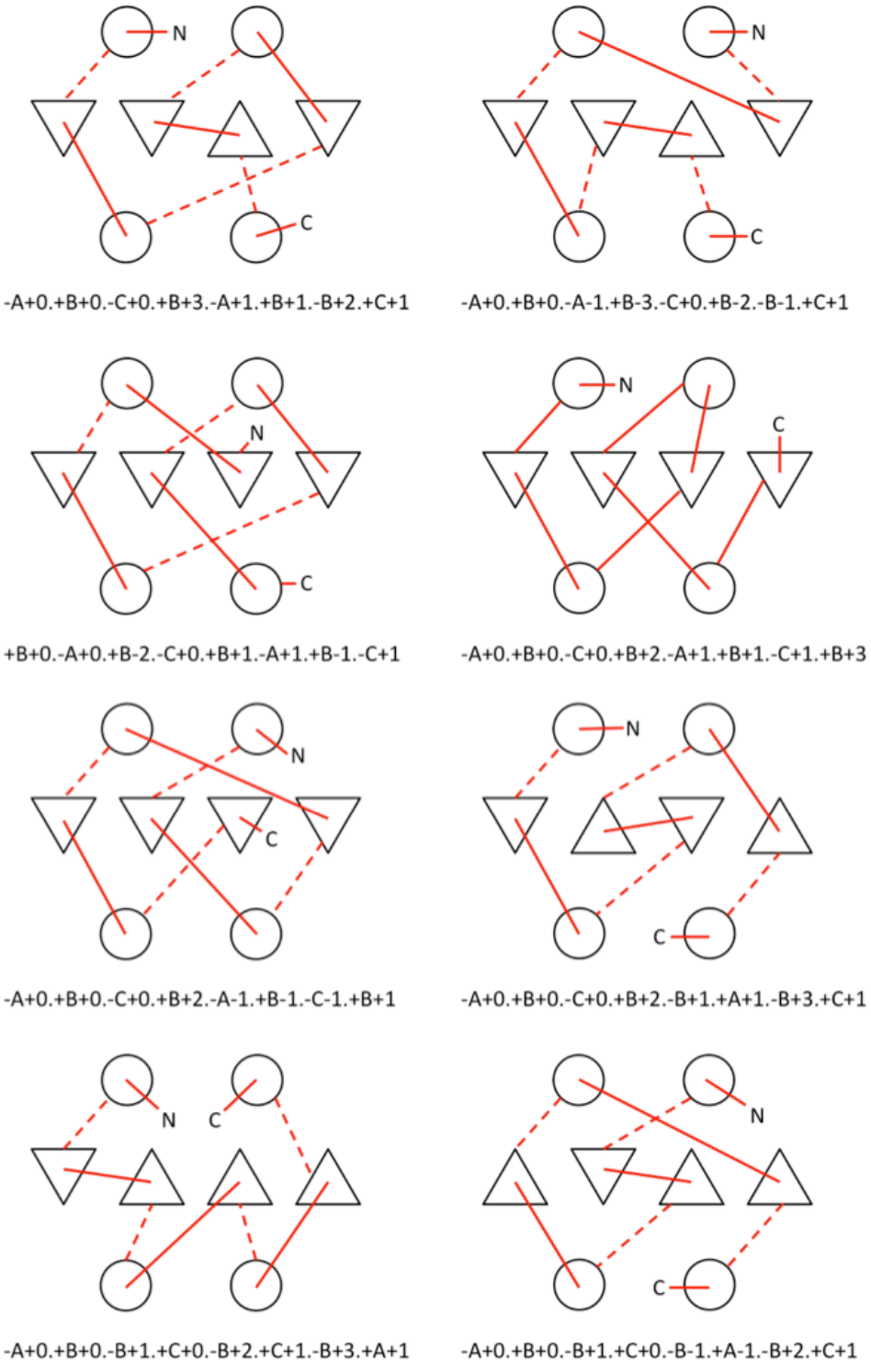

Figure 15. *Topologies chosen for the 2-4-2 fold.* Again, 8 topologies unobservable in nature were chosen for our design studies.

3.2.2 Backbone construction

After identifying some novel topologies that we wanted to try and make, we used a previously published *de novo* backbone construction method to generate realistic structures[117]. In this, rough models are constructed using distance geometry[121], with alpha-carbon models produced using idealized secondary structure elements and random walks for loop regions. Coarse-grain refinement of the structure is performed through modelling C_α-C_α "virtual bonds", pseudo-hydrogen bonding potentials and a structural alphabet of angles between sets of backbone atoms[122]. Finally, mainchain atoms are added and optimised with a potential energy function combined with a Monte Carlo sampling method that allows slight adjustments to backbone torsion angles.

3.2.3 Flexible design process

Once we had constructed viable backbones, we attempted to find a suitable sequence that would be able to fit onto the scaffold. We did this by using a flexible backbone design protocol that iterates between steps of stochastic rotamer packing and backbone relaxations. Previous papers have already established that flexible design is a more successful method than keeping the structure permanently fixed during the design process[123,124].

Our flexible design protocol is separated into two stages. Firstly, we have a fixed-backbone step that takes a stochastic approach, optimally packing rotamers into a given structure without allowing minimal movement of the protein backbone[100]. Once a suitable structure has been designed onto this fixed structure, the backbone is relaxed using the RosettaRelax protocol[125]. This allows small adjustments to the backbone torsion angles and attempts to repack the whole protein to find the lowest energy state. By iterating 1,000 times between the fixed and relaxation protocols, we hoped to produce a much more robust design method that would result in lower energies and more stable sequences.

In an attempt to direct our design process better and speed up the conformational search through rotamers, we applied some constraints to the amino acid selections possible at various locations. For residues with less than 30% rSASA exposure of alpha carbons, we allowed only hydrophobic residues to be selected in the hope of producing a more realistic core. Cysteine was also excluded from all positions, to avoid complications with unwanted disulphide bonds forming. Using these simple constraints, in line with established methods, we generated 3,000 sequences per backbone that were then filtered based upon various criteria (described below) to ensure the most viable were taken forward for experimental testing.

However, it has been suggested that the Rosetta fixed backbone design protocol has a tendency to create a "patchy" surface on the protein, due to potential inadequacies of the potential energy function[126]. Traditionally, this has been solved by manually curating designs after-the-fact by visual inspection and manually changing residues[127]. Some papers have even suggested that up to one third of surface residues need to be re-optimised through human intervention[128]. Rather than manually curate our design process, we wanted to automate it as much as possible and so attempted to bias the amino acid selection on the surface by including a knowledge-based potential (KBP). This consisted of searching the PDB of known structures and gaining an average percentage representation for each amino acid on the surface (except Cys). We used a set of 2,000 structures under 250 residues to gain these values (shown in Table 1).

Residue(s)	% representation on surface
E	12
K	11
R, D	9
S, T, A	8
G	7
N, P, Q	6
H	3
F, I, L, M, V, W, Y	1

Table 1. *Percentage representation for each residue (except Cys) on the surface of native proteins.* A set of 2,000 native structures under 250 residues was used to obtain idealised percentage representations of each residue on the surface. This was then used in an attempt to bias our designs to produce a more realistic surface.

At each surface residue position in our designed backbone scaffold, we gave the fixed backbone design a choice of 3 different amino acids in that position. These 3 amino acids chosen for selection were decided based on their percentage representation in real proteins. For example, a residue that composes 10% of the surface of real proteins will have a 10% of being selected as one of the 3 residues available for design. In restricting selections this way, we hoped to recapitulate a more realistic surface while still offering a reasonable variety at each position. Another 3,000 sequences for each backbone were designed using this method and taken through to the filtering stage.

An attempt to rectify the need for manual curation of designs was published at a later date[127], but we had already attempted to solve this problem ourselves by this time.

3.2.4 Design selection filters

After we had made our designs, they were subjected to a number of filters to ensure that they looked as realistic as possible before we attempted to experimentally test them. Backbone torsion and side chain angles were checked with ProCheck[129] to ensure that the structures of the designs were within normal ranges. Packing was checked with RosettaHoles[130] and any designs possessing voids on the interior of the protein were discarded.

We also filtered out any designs that were compositionally much different to native proteins. Sorting amino acids by their physicochemical properties, using the colour wheel outlined by Taylor in 1997[131], we created 3 equally sized groups. The "phobic" group contained Ala, Val, Ile, Leu, Met and Phe. The "positive" group was Arg, Lys, His, Trp, Tyr, Asn and Gln. Finally, the "negative" group consisted of Glu, Asp, Ser, Thr, Gly, Pro and Cys (Figure 16). Although the labels are not directly representative of every amino acid in the category (eg Pro being in the "negative" group), it still offers a coarse-grain filter for weeding out bad designs and the labels are only used for the sake of convenience.

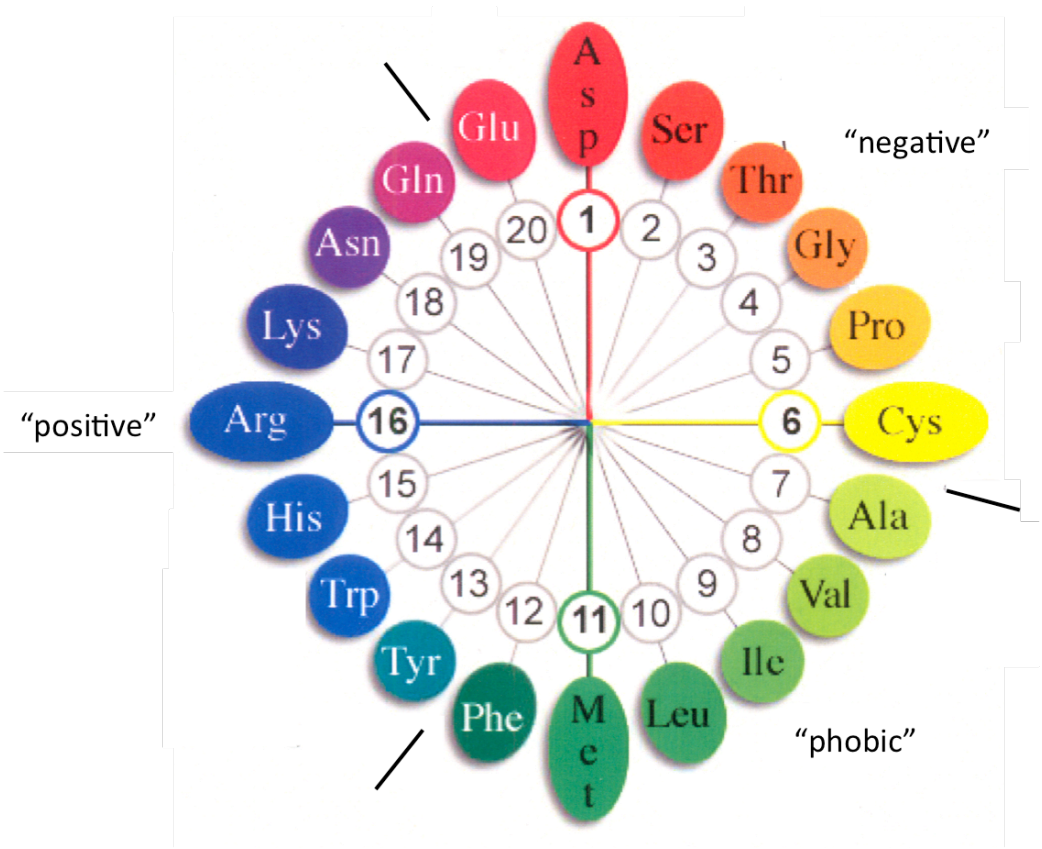

Figure 16. *Colour wheel used for amino acid category definitions*. Residues are defined by their physicochemical properties, as defined by Taylor (1997)[131]. These are then split into 3 almost equally sized groups, with labels given to each group. These residue groups were then used to filter out designs that were too far away from native compositions.

Using these categories, we then used a set of 2,000 proteins under 250 residues to discover what the representation for each grouping of residues was in native proteins for the core and surface. Once we had composition profiles for 2,000 native proteins, we set some restrictions that our designs needed to fall into. The boundaries that 99% of real protein compositions lay between are given in Table 2.

	Core (%)	Surface (%)
Negative	0 – 30	18 – 52
Positive	0 – 33	26 – 55
Phobic	21-86	12-40

Table 2. *Compositional boundaries for 99% of native sequences analysed, used as a filter for potential designs*. 2,000 native protein compositions were analysed using the groupings outlined previously. Any designed protein that fell outside these ranges of values was automatically discarded as too compositionally different from natives. This acted as a quick and simple selection filter for the number of designs we produced.

Any designs that had compositions outside of these range of values were automatically discarded as unrealistic. Although this filter may be open to criticism and potentially better groupings could have been used, we felt that this offered a robust and easy method for discarding any designs that strayed too far from a native-like composition.

On the sequences remaining after these sets of filters, we performed *ab initio* structure prediction. The average TM-score to the template over 1,000 models was used as an indicator of how suitable the sequences were for the structure. The top 100 sequences for each backbone were taken through to a full comparative modelling stage, and the 6 with the highest TM-score to template were then chosen to be tested experimentally.

3.2.5 Protein production techniques

Designed sequences were expressed by following the experimental procedures outlined in the main methods (Section 2.2), with the difference that the designs were fused to a GB1 solubility tag in an attempt increase the yield of soluble protein[132]. Adding a solubility tag can protect the peptide during its folding, increasing the chances of it adopting a stable fold. It has also been shown that previously insoluble protein domains can exist in

solution when a GB1 tag is added[133], and it is our hope that this system will provide the best chance for our novel proteins to exist.

When soluble designs were discovered, we purified directly with a His-Ni^{2+} column and then went straight to NMR analysis. There is no need to cleave the solubility tag, as clear spectra can be obtained with it still fused to the target protein[134].

3.2.6 NMR

Cells were grown in PG minimal media in the presence of isotope labelled ammonium sulphate as outlined in Section 2.2.5. After protein purification, each protein was exchanged overnight into a buffer suitable for NMR studies (20 mM MES pH 6.2, 100 mM NaCl, 1mM EDTA, 3 mM NaN_3, 0.5 mM TCEP).

NMR spectra were acquired at 298 K with a Bruker Advance II (operating at a nominal ^1H frequency of 700 MHz). A 2D ^1H-^{15}N heteronuclear single quantum coherence (HSQC) spectroscopy data set was obtained for each protein put forward into NMR trials. 1D and 2D spectra were obtained at a variety of temperatures (20 ^oC, 25 ^oC, 30 ^oC, 35 ^oC and 40 ^oC) to fully investigate any potential structure. All data were processed by using NMRpipe/NMRDraw[135] and analysed using CCPN Analysis, version 2[136].

3.3 Results

To begin, we performed some computational analysis on the designs that our process produced. This helped us to decide on which designs to test experimentally for each backbone, as well as obtain some insight into how our protocol was functioning.

3.3.1 Over-compaction of structure

One of the fears we had for our design process was that it could potentially result in a gradual over-compaction of our model. For instance, if a small residue such as Ala is chosen at a location and then the structure is relaxed and repacked, the repacking process could result in local steric constraints that only allowed Ala (or small residues) to be picked at this location in subsequent design iterations. This could lead to unrealistic sequences being produced as more constraints were inadvertently added as the sequence design protocol progressed. If this were happening, the radius of gyration for our structure would get smaller as we advanced through more iterations of sequence design. By checking the radius of gyration through design iterations, we were able to establish that this is not the case (Figure 17). Although there is an initial compaction after the first round of design, this is only ~1% of the total radius. For comparison, when the normal backbone relaxation and repacking protocol is run on solved crystal structures, a compaction of between 5-7% is seen. After this initial compaction, the radius of gyration remains constant for 2,000 iterations and so we can safely say that our sequence design protocol is not over-compacting the structure.

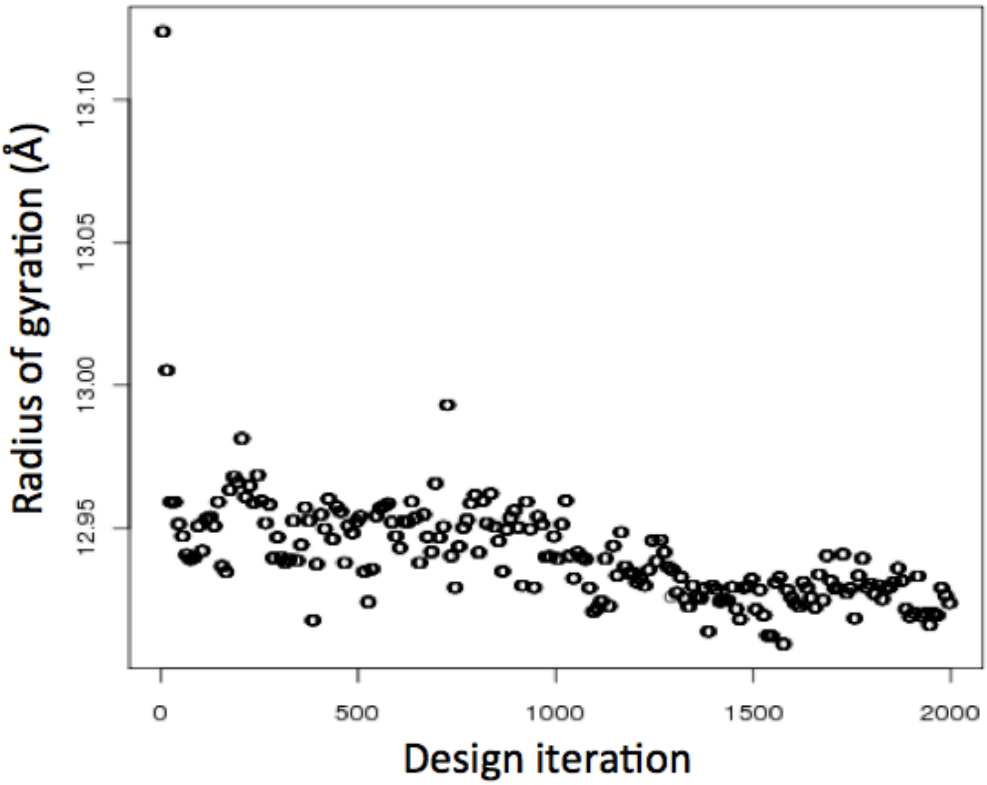

Figure 17. *Radius of gyration for a designed protein remains steady as the design process iterates*. There is an initial compaction of ~1% in the first few design iterations, but this is a negligible amount. Subsequent design iterations see the radius of gyration remain constant, indicating that our design process is not over-compacting the structure.

3.3.2 Predicted solubility of designed proteins

Because our KBP surface selection method was untested, we decided to make one batch using it (36,000 designs leading to 96 experimentally tested proteins), and another batch of designs with no restrictions on surface residue selection. To try and gain some insight into how our surface biasing was affecting compositions, we performed some pre-analysis.

Firstly, we used the ESPRESSO program[137] to predict how soluble our potential designs were. This uses a combination of 43 property-based indicators, as well as motif-based predictors, that can help determine the solubility of a protein in *E. Coli* expression systems[138]. We chose 250 protein sequences from natives, and the same number from both an unrestricted

and KBP design method. Using ESPRESSO, we gained an indication about how our designs were performing. The native proteins seem to have a solubility score of about 0.75, and our unrestricted surface design protocol tends to have a slightly lower score with a peak at about 0.63 (Figure 18). Although the unrestricted design protocol has a lower score, the sequences generated by it are still predicted to be viable as any score above 0.5 is classed as "soluble". Using the KBP-based surface selection, we saw an increase in predicted solubility with the peak being around 0.85. Our distribution of solubility scores seems to be broader than native and unrestricted design sequences, meaning we are potentially covering a wider range of theoretical solubilities.

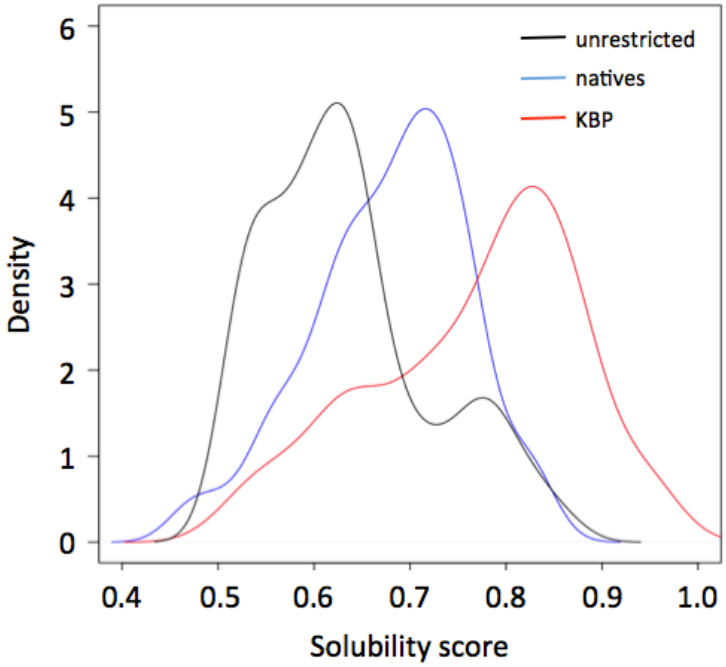

Figure 18. *Distribution of solubility scores for surface design methods.* Using ESPRESSO, we gained solubility predictions for 250 native proteins and the same number from each of our surface design processes. The solubility of native proteins seems to centre around 0.75, with the "unrestricted" surface design slightly lower at 0.62. Our KBP designs have a higher predicted solubility than both, at 0.85.

3.3.3 Surface compositions of designed proteins

We also checked how the two surface design methods affected the general composition of the proteins produced. Using the same amino acid

categories as in our composition filters ("hydrophobic", "positive" and "negative"), we checked the percentage representation of each group and compared them to native sequences (Figure 19).

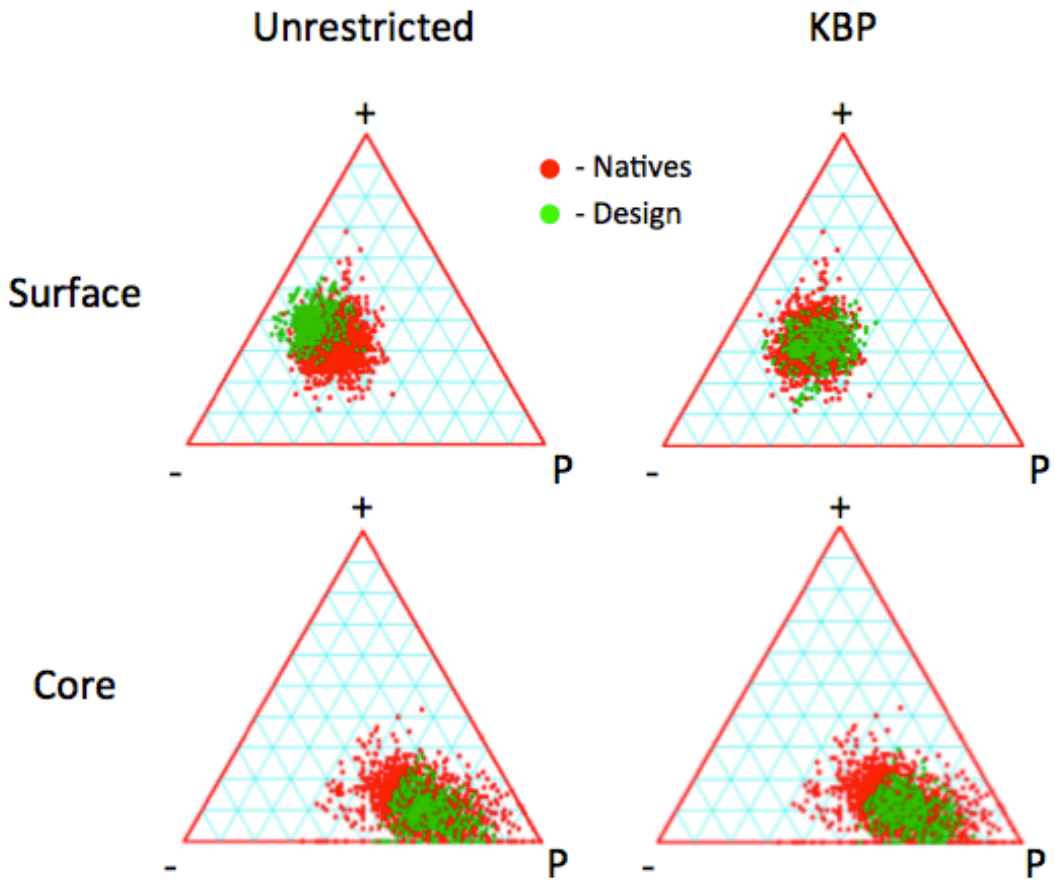

Figure 19. *Compositional analysis of our unrestricted and KBP surface design protocols, compared to native sequences.* Amino acids were categorised into groups based on their physicochemical properties ("P" = hydrophobic, "+" = positive, "-" = negative). Compositions for an unrestricted surface design process or using a KBP surface design were compared to native sequences. The unrestricted design process did have quite good overlap with natives, with solvent-exposed hydrophobic residues reduced. Using the KBP, more hydrophobic residues appeared to be present on the surface but the overall compositions of native sequences were more closely mirrored. Core compositions for both remained much the same, which was expected, as there should not have been any differing selection criteria for those residues.

In general, both surface design methods appeared to produce many native-like compositions. The unrestricted surface protocol generated designs that had much less hydrophobic surfaces when compared to native sequences. However, this may not be a problem, as most native proteins must

accommodate interaction sites (possibly containing hydrophobic residues) on their surface in order to perform their function. Because our designs are merely scaffolds with no defined function, we can solely concentrate on reducing potential aggregation sites. There could be a worry that having oppositely charged residues on the surface could lead to increased aggregation, which could affect the potential solubility of the protein. However, in terms of actual charged residues (Asp/Glu/Lys/Arg), there was no perceivable overrepresentation produced by the unrestricted protocol when compared to the range of natives. Using a KBP surface design method produced a greater diversity of surface compositions, which appeared to closely mirror the distribution of natives.

Both design protocols produced an interesting selection of designs that we took forward to experimental testing. The core compositions for both methods look very native-like and appeared to stay the same between protocols, which is to be expected, as the selection criteria for those residues did not change between methods. Any designs that did not fall inside the boundaries outlined in the methods were filtered out, but it is encouraging that both methods are producing a high proportion of designs that are in the native-like region of composition space.

3.3.4 Structure prediction

After passing through our selection filters, the final steps were to do a full tertiary structure prediction using a purely *ab initio* method first and then checking the highest scoring sequences with full comparative modelling. The predicted models produced for some of our designs seemed to have quite a high similarity to the target structure. Some predictions had close to 0.6 TM-score to the template we were aiming to design. A particularly high scoring design, 0.58 TM-score between prediction and template, is seen in Figure 20. Although the score isn't perfect (a TM-score of 1 is the maximum), the comparison between models appears to be very good. The global fold is predicted very well, with all secondary structure elements well

formed and in the correct positions. However, there are many differences when atomistic detail is taken into account. The strands are consistent in both models, but both helices have quite a large shift in the prediction compared to what we were expecting. This has led to a reduced score, as there are still quite a few disparities between the two models. Although we would prefer to have an atomistic level of design, where all atoms are in the exact place we designed them to be, we are still encouraged with the overall agreement in prediction. There was no discernable difference in predictability for either surface design protocol.

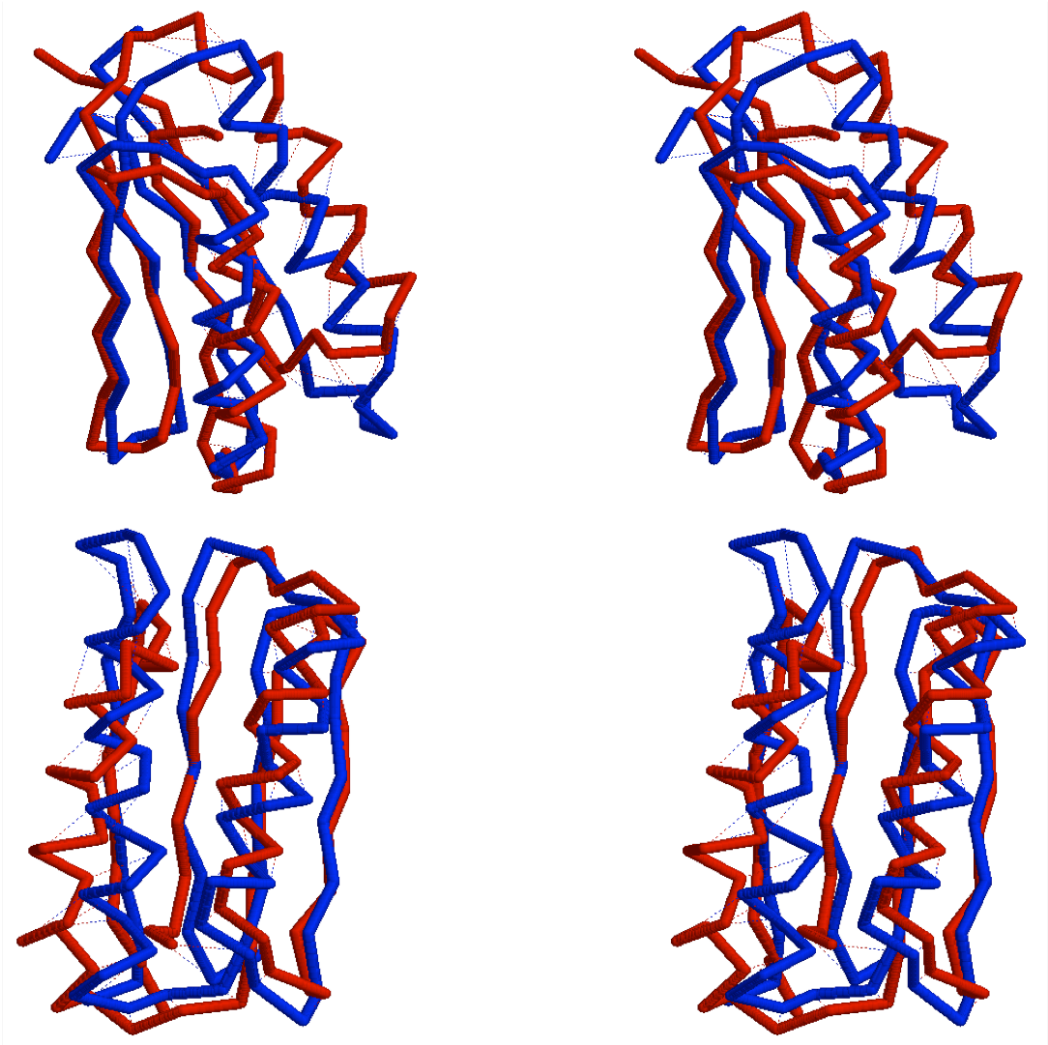

Figure 20. *Stereo images showing the overall agreement of the predicted structure of a designed sequence and the template we were designing for*. The predicted structure for some of our designs appeared to be very good, with a good agreement in global structure when mapped back onto the scaffold structure we were attempting to design. However, there are a few major differences in the atomistic detail of the models. Although the strands are well predicted, there is a shift in both helices and this has led to a slightly lower TM-score of 0.58.

3.3.5 Experimental testing results

In total, we took forward 96 designs using the KBP-based surface design protocol and the same number made using an unrestricted surface design method. These sequences had compositions that were comparable to native sequences, along with good predicted solubility and structure prediction. Each design was expressed with a GB1 solubility tag in BL21-gold *E. Coli* cells and screened for solubility. Out of a total of 192 designs, 24 were soluble from the KBP group and 18 were soluble from the "unrestricted" group. The soluble designs were then purified using a His_6 tag on a Ni^{2+} column and NMR was performed. Protein expression and solubility screening protocols are given in Section 2.2, and NMR conditions are given in Section 3.2.6.

Unfortunately, all of the designs that appeared to be soluble did not possess tertiary structure that was visible through NMR studies (Figure 21). When tertiary structure is present, each backbone N and H atom should have a defined local chemical environment and therefore a particular chemical shift associated with that environment. This would show as a well-defined peak on the HSQC contour plot. Although we see some well-defined peaks in the design-GB1 fusion (shown in green on the contour plot), most of these seem to come from the well-folded GB1 solubility tag (shown in magenta). This means that the peaks from our design are mostly in the large middle portion of the spectrum, suggesting that our protein does not possess defined tertiary structure. Were our designs to have more tertiary structure interactions, they would be visible as defined peaks in the "fingerprint region" surrounding the centre.

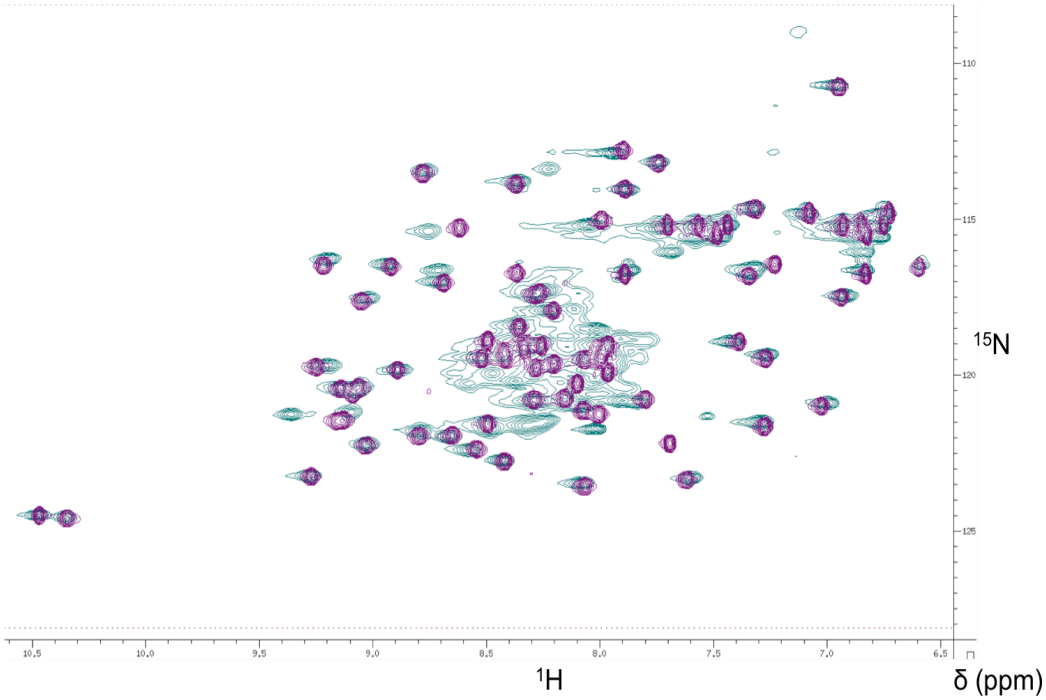

Figure 21. 2D-HSQC *NMR spectra of a design-GB1 fusion protein (green) and the GB1 solubility tag by itself (magenta)*. The design-GB1 spectrum shows some well-defined peaks, meaning there are well-defined local chemical environments and therefore tertiary structure. Unfortunately, the vast majority of these peaks appear to be from the GB1 solubility tag. This suggests that our designed protein is in the large unfolded region in the middle of the spectrum.

It is possible that our designs exist as molten globules, where the protein does have formed secondary structures but they are not as closely packed as they are in a fully folded globular state[139]. The lack of fixed tertiary structure in these molten globules can result in an ensemble of native-like structures that intra-convert on a very short time scale[140], potentially giving a HSQC spectra similar to the one that we have seen for our designs. To investigate this further, we could conduct circular dichroism (CD) experiments, which show the proportion of α-helices and β-strands in a given protein. If we do appear to have formed secondary structure elements but no tertiary interactions discernable via NMR, this could suggest that we are indeed making molten globules. If secondary structures are not visible through CD, this would indicate that our protein is completely unfolded.

Another possibility for investigating molten globules is by using a hydrophobic fluorescent probe, such as 8-anilino-1-naphthalenesulfonic acid (ANS). These probes have a much higher affinity for the molten-globular states of proteins as compared with the unfolded or fully compacted states[141]. After establishing a baseline level of fluorescence from the probe in the presence of our designed proteins, we could denature the proteins and then perform a similar experiment. If fluorescence decreases from the denatured proteins, we can infer that there must have been a molten globular state present that is then unfolded. However, if the readings from the probe remain constant then there is a good chance that the protein already exists in an unfolded state.

Both of these techniques could give us information about how much structure is actually present in our designs, which could potentially lead to success by rounds of iterative redesign. However, this would most likely involve manually specified target residues or other techniques that would be specific to each individual protein. One of the aims for this thesis is to investigate generally applicable design techniques that can be used for a variety of fold types, and so we chose instead to focus the next sections on attempting to uncover some general rules to improve the chances of success for a broader scope of proteins. Future experiments in the lab may focus on attempting to investigate potential molten globules, but we decided at this point to explore broader concepts in successful protein design.

3.4 Discussion

Though there have already been numerous reported successes for protein design, including the design of novel folds, we found it very hard to replicate the results some groups have seen. We attempted to experimentally test a total of 16 designs for 12 novel backbones and while some of them were soluble, none showed any compact tertiary structure. Unfortunately, the solubility data is not useful for investigating as the solubility tag used may skew any information found. There have been reports that solubility tags

can frequently cause false positives when looking at the solubility of a fused protein[142–144]. Even if our designed proteins are unfolded, their precipitation can be prevented by the presence of the folded and highly soluble GB1 tag. The unfolded proteins can form the "core" of the aggregate with a surrounding "shell" of soluble GB1, maintaining the aggregate in solution. The solubility tag may be cleaved off, but it is our understanding that if our designs are unfolded under the ideal conditions provided by tagging, then this would still be the case without a tag.

Interestingly, the solubility predictor we used seemed to show no correlation with experimental results. All the designs chosen to be taken forward to lab testing were defined as "soluble" but only a small number of these appeared to confirm this prediction, even with the use of a solubility tag. Solubility prediction is a notoriously difficult field to tackle, and our results seem to suggest that there is a long way to go before reliable indicators are found. For this reason, we decided to exclude solubility predictors from further experiments, as their usefulness is not particularly apparent.

Our designs looked realistic to our eyes, with native-like sequence compositions and good predictability. Given the successes seen by previous studies, along with the encouraging computational analysis, we would have expected our designs to perhaps possess more structure. However, we are using a backbone construction method that is so far experimentally untested. Coupled with a sequence design protocol that we do not know for certain works, we appear to be competing on two fronts. Trying to design a novel backbone and then a novel sequence to adopt that particular structure are two confounding problems. In order to separate these problems, it may be best firstly to have an established sequence design protocol that we have demonstrated successfully and then trying to use this to design sequences for a novel scaffold.

As the next stage in our investigations into protein design, we attempted to gain more insight into sequence design for known backbones using a factorial analysis approach. This would allow us to choose some factors that we believed to be important for sequence design and attempt to understand whether they really have an impact on solubility and foldedness.

Because we were planning on testing a large number of designs in the next stage, we decided to drop the GB1 solubility tag fusion. This would speed up our screening process for designs, as less time would be spent chasing up false positives on the solubility front. There is a danger that we will lose false negatives (sequences which may have been folded with the solubility tag that we otherwise miss), but the rate of these is thought to be far less than the rate of false negatives. We were willing to sacrifice the few potential designs folded with tags in order to better explore our factors in a shorter amount of time.

Acknowledgements

These experiments were performed with the help of Michael 'Sid' Sadowski (*de novo* design) and TJ Ragan (experimental characterisation).

4. Factorial analysis of sequence design

4.1 Introduction

Aiming to design a backbone completely *de novo* and then finding a suitable sequence that would allow the protein to adopt that specific 3D conformation proved to be very challenging, even given the successes reported by the Baker group[77,79,91,104]. In order to simplify the problem slightly, we decided to remove the backbone construction protocol and only focus on the sequence design algorithm to see if we could garner some insight into what would make a good sequence design.

To do this, we chose a factorial design of experiment (DoE) approach that would allow us to see which factors were indicative of good design and give us something to potentially optimise towards. We selected four native proteins (1CC7, 1Q5V, 3CHY, 1E5D) with solved crystal structures and attempted to redesign the sequence of these proteins to recapitulate the adopted conformations. The 1CC7/1Q5V are homologues belonging to the ferredoxin family, and 3CHY/1E5D are homologues with a Rossmanoid-like structure (Figure 22).

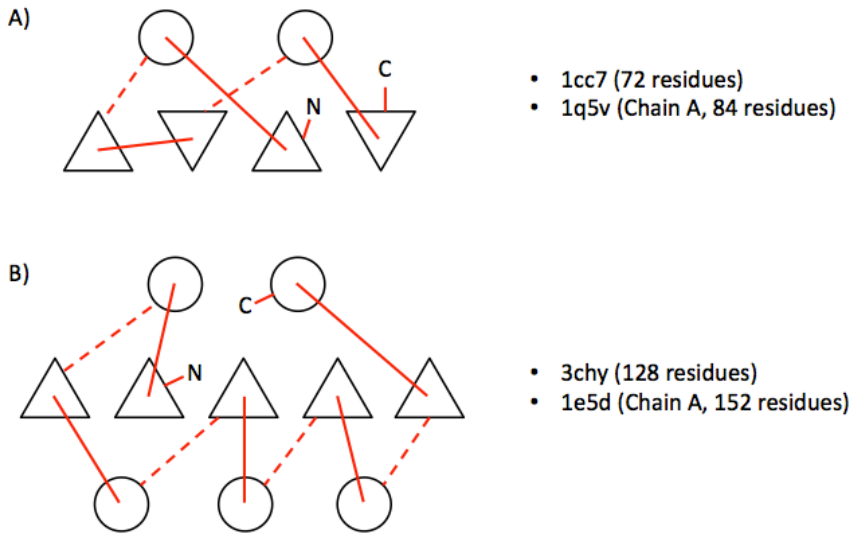

Figure 22. *Diagram showing the topologies of the folds selected for redesign.* A) The ferredoxin fold adopted by 1CC7 and 1Q5V. B) 3CHY and 1E5D have a Rossmanoid-like structure. Solid lines indicate connections in the foreground, with dashed lines representative of background loops.

4.2 Methods

4.2.1 Full factorial design of experiments

Factorial experiments can be used to investigate the impact of continuous or categorical factors on a response variable when each factor has varying levels. They can give insight into the main effects of each factor by themselves, as well as whether there are interactions between factors (i.e. if the effect of one factor depends on another factor).

For example, if we have a 2 factor (A and B) system with 2 levels (+ and -), then each cell takes a different combination of factors at each level (A_iB_j), with a different response in each cell (y_c). This is shown in Figure 23.

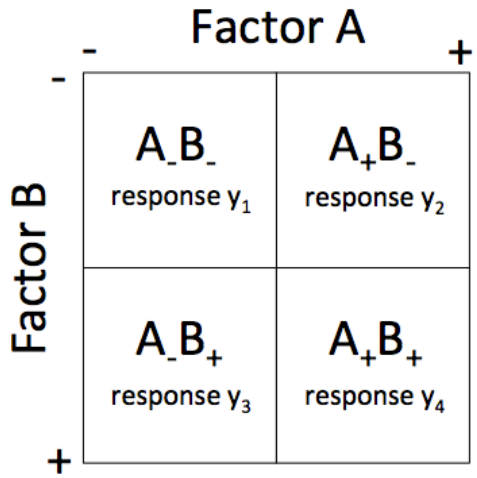

Figure 23. *Diagram illustrating the potential experimental conditions in a 2-factor system that has two levels for each factor.* Both factors A and B can be at either the '+' or '-' level, giving 4 potential cells, with a response for each (y).

If we assume that the response variable output (y) is linearly dependent on the effects of each factor/level combination, then the model will follow the equation:

$$y_c = b_0 + b_aA + b_bB + b_{ab}AB$$

where y is the output, b_0 is the intercept, A and B are coded variables indicating the level of our factors (+1 or -1), and b_i is the weighting coefficient for each term, which implies how significant the effect of each

factor is. As interactions are symmetrical, there is need for only one interaction term.

We can expand this into a system of linear equations;

$$y_1 = b_0 + b_a(-1) + b_b(-1) + b_{ab}(-1)(-1)$$
$$y_2 = b_0 + b_a(+1) + b_b(-1) + b_{ab}(+1)(-1)$$
$$y_3 = b_0 + b_a(-1) + b_b(+1) + b_{ab}(-1)(+1)$$
$$y_4 = b_0 + b_a(+1) + b_b(+1) + b_{ab}(+1)(+1)$$

This can be represented as the following matrix;

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & (-1) & (-1) & (+1) \\ 1 & (+1) & (-1) & (-1) \\ 1 & (-1) & (+1) & (-1) \\ 1 & (+1) & (+1) & (+1) \end{bmatrix} . \begin{bmatrix} b_0 \\ b_a \\ b_b \\ b_{ab} \end{bmatrix}$$
$$y = X . b$$

We can then find a solution for vector *b* by using a least-squares model to minimize the residuals between our input correlation matrix and the correlation matrix reproduced by the factors.

$$b = (X^T X)^{-1} X^T y$$

In this way, we can gain weighting coefficients to describe the magnitude of effect for each of our individual factors and their interactions.

4.2.2 Analysis of variance (ANOVA)

Another technique for measuring the effects of various factors is to do analysis of variance (ANOVA). This compares the variability *between* groups to the variability *within* groups to give an idea if there are any significant factor (or interaction) effects.

If we have a number (n) of individual responses (x) in each cell, as well as each individual response value in each cell (x), there is a mean for each combination of factors ($\bar{x}$) and a marginal mean for each factor at the same level. For example, when factor B is always at the "+" level, the marginal mean for that given row will be $\bar{x}_{A\bullet B+}$ (i.e. the mean of $\bar{x}_{A-B+} + \bar{x}_{A+B+)}$ The symbol "•" denotes that the preceding factor can be at either at the "+" or "-" level. This is illustrated in Figure 24.

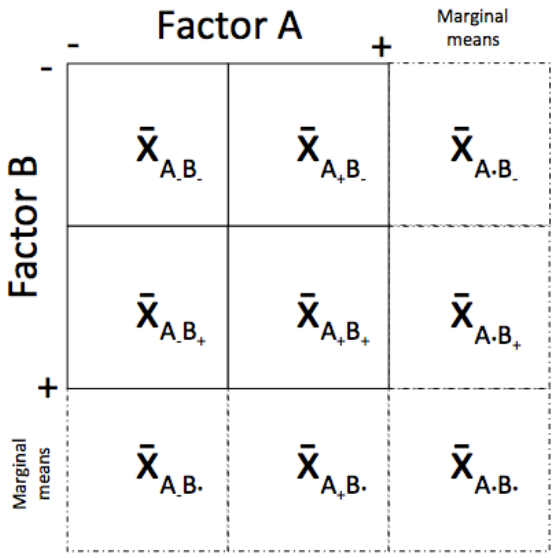

Figure 24. *Visual explanation of marginal means.* Marginal means are present for each factor when it remains at the same level, e.g. when factor B is always at the "+" level.

From these means, we can calculate the variance to compare between groups and pick out any significant effects.

Definitions;

$SS_{error} = \Sigma(x - \bar{x})^2$

$SS_A = n_{i\bullet}\,\Sigma(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$ where i is factor A at both +1 and -1

$SS_B = n_{\bullet j}\,\Sigma(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2$ and j is factor B at both +1 and -1.

$SS_{AB} = n_{ij}\,\Sigma(\bar{x}_{ij} - \bar{x}_{\bullet\bullet})^2 - SS_A - SS_B$

$SS_{total} = SS_{error} + SS_A + SS_B + SS_{AB}$

df_A = (number of levels of factor A) – 1

df_B = (number of levels of factor B) – 1

$df_{AB} = (df_A)(df_B)$

df_{error} = n – number of cells

$df_{total} = n - 1$

Using the appropriate degrees of freedom (df) value for each square sum (SS), we can obtain the corresponding mean square (MS) term.

$MS_t = SS_t / df_t$ where *t* is "A", "B" or "AB".

$MS_{error} = SS_{error} / df_{error}$

From these mean square terms, we can calculate an F-ratio which is a ratio of the *between* group variability to the *within* group variability. If the F-ratio is above the critical F-value for the given degrees of freedom, then the effect shown by that factor is statistically significant.

$F_t = MS_t / MS_{error}$

4.2.3 Scalability

This approach to investigate multiple factors is not just restricted to 2x2 factorials. It can be expanded to approach problems with any number of factors and levels, although doing so can increase the time it takes and cost for experimentation.

4.2.4 Factorial analysis applied to sequence design

As mentioned previously, we have chosen a factorial approach to investigate the impact that various factors have on producing viable protein sequence designs. The factors we have chosen to use in our experiments are buried (core) sequence divergence, surface sequence divergence, secondary structure prediction accuracy and Rosetta energy score. We will have two levels of each factor, labeled as "bad" (-) or "good" (+). This means that we have 2^4 (= 16) cells in our experiment, and we have chosen to do 6 replicates contained within each cell. It was our hope that we could make 96 proteins (per backbone) with redesigned sequences, and screen them for solubility

and foldedness levels to provide some suggestions for which factors would signify a viable protein design.

To begin, we produced a broad range of designs (~36,000 for each backbone) using the RosettaDesign flexible backbone protocol[145] , and scored them on our various measures.

4.2.5 Definition of factors and levels

4.2.5.1 Rosetta energy

When producing protein designs with Rosetta, the program automatically scores the design with its own internal scoring system. Among other things, this involves weighting together measures such as the Lennard-Jones attractive/repulsive potentials, Lazaridis-Karplus solvation energy, short/long range hydrogen bonding and many others (for a comprehensive scoring overview, see the relevant papers[46,48,49]). This single "Rosetta energy score" indicates how stable a potential design should be, with a more negative score corresponding to a lower free-energy state.

Our 36,000 designs almost produced a normal distribution of Rosetta energy scores (Figure 25A) and so we split this in half, with the more negative 50% defined as "good" (+) and the more positive 50% defined as "bad" (-).

4.2.5.2 Secondary structure prediction accuracy

We thought that secondary structure prediction accuracy, obtained from the primary sequence of each protein, would be important in scoring our designs as it should be beneficial to have sets of residues which have a strong predilection for adopting the correct secondary structure element. Since we are designing new sequences for an already known structure (solved by X-ray crystallography), we know the secondary structure

definitions for each position (obtained by using DSSP[146]). To give a score of how likely residues are to adopt these conformations at each location, we ran PsiPred[147] in single sequence mode, which calculates secondary structure predictions based on the intrinsic properties of the residues and then compared this output to DSSP. Note; PsiPred is much more reliable when presented with a multiple sequence alignment with which it can perform machine learning techniques, but single sequence mode was used to avoid biasing our designs towards known native structures.

Our theory is that sequences more likely to adopt the desired secondary structure element will be predicted as such by PsiPred, and match DSSP more frequently. We ran this comparison across all our potential designs and represented secondary structure prediction accuracy as a percentage of the whole protein. Again, this produced a normal distribution (Figure 25B) and so we classed the highest scoring 50% as "good" (+) and the lowest scoring 50% as "bad" (-).

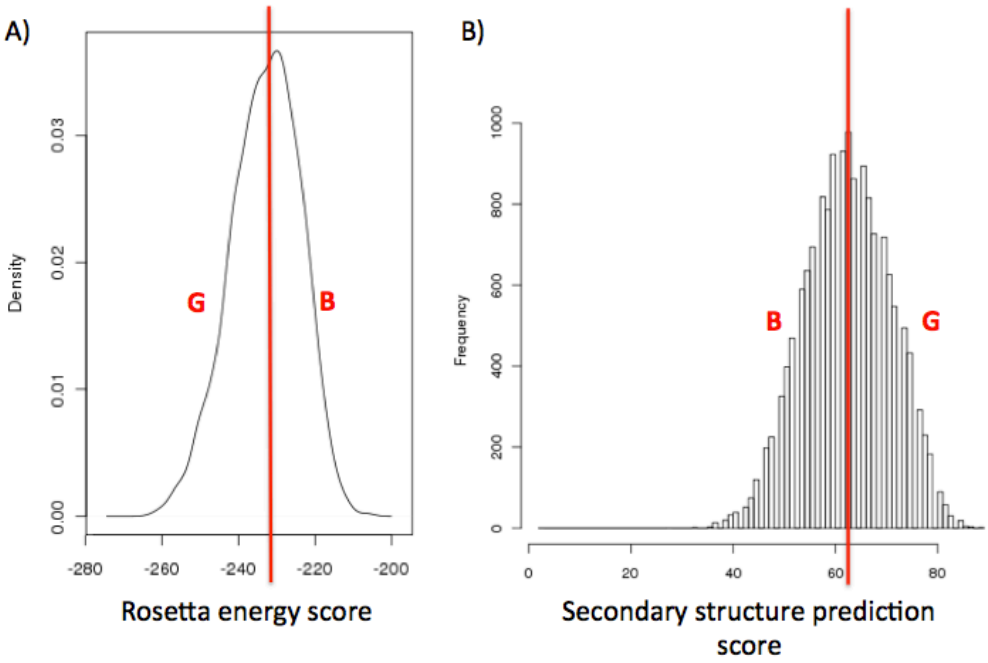

Figure 25. *Density plot showing the distribution of the Rosetta energy and secondary structure prediction accuracy factors.* A) Each design is scored by the Rosetta Design protocol, taking into account numerous different energy terms. The more stable 50% of our designs were classified as "good" (+), the lower 50% as "bad" (-). B) Secondary structure prediction accuracy was obtained by comparing single sequence PsiPred output against known DSSP secondary structure classifications. The highest scoring 50% were "good" and the lower 50% were "bad".

4.2.5.3 Core and surface composition

For our final factors, we thought that the composition of a protein, both in the core and on the surface, could be an important factor in producing a viable protein design. In order to investigate this, we needed to come up with a new metric to test as a factor in our factorial experiment. We took a subsample of 800 proteins under 200 residues from the PDB and found the percentage representation of each amino acid in the core of each protein and on the surface. "Core" residues were classified as residue positions that had C-alphas with a relative solvent accessible surface area (rSASA) of less than 0.3 on a skeletal Cα backbone. We then averaged each amino acid frequency over the entire set to produce an "ideal" protein composition for proteins under 200 residues. Each design was put through the same

protocol and compared to the idealized composition model extracted from real proteins using a modified Kullback-Leibler (KL) divergence for discrete random variables[148].

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

where "*P*" is the frequency at which an amino acid ("*i*") is found in the design and "*Q*" is the frequency from the ideal composition of the PDB.

To simplify this in a worked example, let us imagine the core of a protein is composed of only two amino acids. In native proteins of this world, 50% of the core is Valine (Val) and 50% is Isoleucine (Ile). One of our designs is significantly off the mark in percentage composition, with 95% Val and 5% Ile. In this case, the KL-divergence would be;

$$D_{KL} = 0.95 \ln \frac{0.95}{0.5} + 0.05 \ln \frac{0.05}{0.5} = 0.4946$$

However, one of our other designs is closer to the idealized composition with 80% Val and 20% Ile;

$$D_{KL} = 0.8 \ln \frac{0.8}{0.5} + 0.2 \ln \frac{0.2}{0.5} = 0.1927$$

The composition that is compositionally closer to the "native" proteins has a lower score, indicating less divergence from the observed model.

The KL divergence of each design will always be non-negative due to Gibb's inequality. This states that for probability distributions *P* and *Q*, the following statement is true (when $p_i=q_i$) since p_i and q_i are positive numbers less than one;

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

By using KL divergence as a measurement for our designs, we obtained two more distributions that could be split to produce our levels (Figure 26). Less divergence from the ideal composition (a lower KL-score) was considered to be "good" and more divergence was "bad".

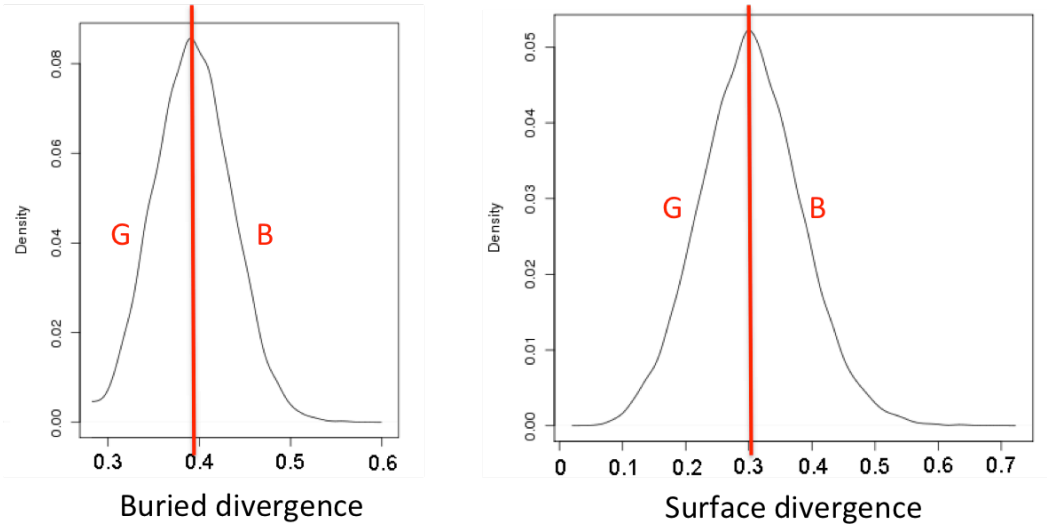

Figure 26. *Density plot showing the distribution of composition divergences.* An idealised composition model was constructed using 800 proteins under 200 residues from the PDB, with each design scored against this model. More divergence from this model is classed as "bad" (-), whereas less divergence is "good" (+).

4.2.6 Selection of designs to experimentally test

Now that we had 36,000 designs all scored and given designations in each factor/level, we needed to pick which 96 of them to put through to experimental trials. To gain the most information, the selection of designs should be maximally spread out across all dimensions of factors. Each level of a factor was binned into 6 equally spaced "compartments" (Figure 27), and we chose the design set that gave us the highest maximum entropy;

$$\sum_i p_{(i)} ln \frac{1}{p_{(i)}}$$

where "*p*" is the number of designs chosen in a bin ("*i*").

For example;

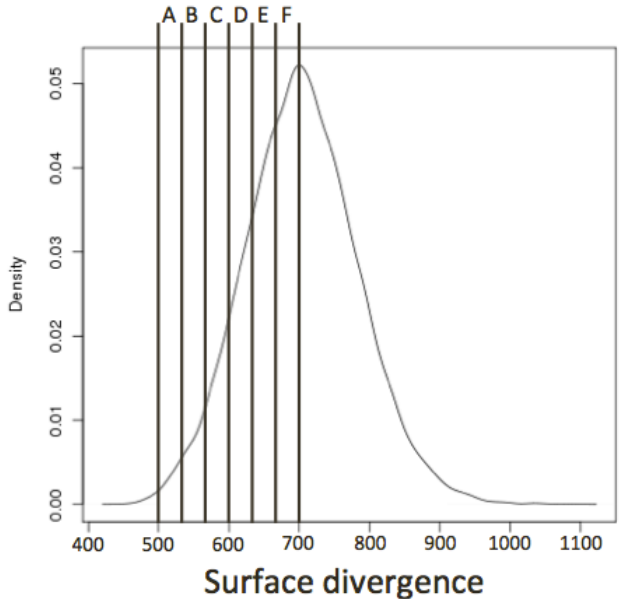

Figure 27. *Illustration of bin sizes for maximum entropy calculations.* Each level of each factor (in this case the "+" portion of surface divergence) was split into 6 equally spaced bins (labeled A-F).

If a set was chosen that had the following selections;

Bin	A	B	C	D	E	F
Number of selections	3	0	0	1	0	2

Then the entropy would be;

$$3 \ln \frac{1}{3} + 1 \ln \frac{1}{1} + 2 \ln \frac{1}{2} = -4.68$$

If a more dispersed selection was made;

Bin	A	B	C	D	E	F
Number of selections	1	1	1	1	1	1

Then the entropy would be;

$$6 \cdot \left[1 \ln \frac{1}{1} \right] = 0$$

Therefore, the more dispersed selection would be put forward as the better selection as it has the highest entropy.

We randomly selected 96 proteins from each backbone set containing 36,000 designs and summed the entropy of those selections across all factors and levels, iterating this process 2,000 times to obtain the set with the highest score.

4.2.7 Fractional factorials

Instead of committing to a full factorial analysis for all of our backbone designs, we decided it would be more practical to do a fractional factorial instead. This means that we can halve the number of experiments that are needed, without losing most of the resolution for single and 2-factor effects.

A full factorial allows 2^k parameters to be investigated, but higher order interactions such as 3-factor and 4-factor are much more uncommon and are usually less meaningful than main and 2-factor effects. If we make the assumption that the contributions of higher order interactions are negligible, then we can conduct a much quicker screening using this fractional factorial without losing much investigative potential.

For example, if we are investigating 3 factors (A, B, C) at 2 levels (+/-) then we can draw this as a cube with each vertex corresponding to an experiment (Figure 28).

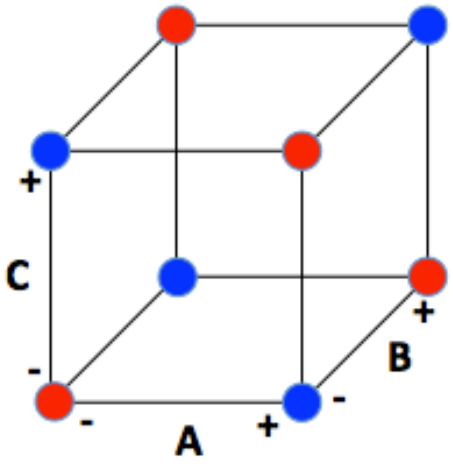

Figure 28. *Illustration of a cube where each vertex corresponds to a set of conditions for each experiment.* Each of the factors (A, B, C) can be present at either a high level (+) or a low level (-). Every vertex of the cube is therefore a specific set of conditions for each factor.

To do a full factorial would require 2^3 experiments, but doing a 2^{3-1} fractional factorial can cut the number of required experiments in half.

By intelligently selecting the experimental conditions to be tested, we can obtain a lot of information with a reduced workload but not lose too much resolution.

For instance, if we choose to test the red vertices on the cube and one factor turns out to have no effect on the response variable, the experiment collapses into a full factorial experiment for the remaining factors. If "A" has no effect on the response, then it will not matter if "A" is at high or low levels, and the cube will therefore collapse into a 2D representation with an experiment performed at each factor combination of "B" and "C" (a full factorial). This means that full factorials are embedded through each level of a fractional factorial design (Figure 29).

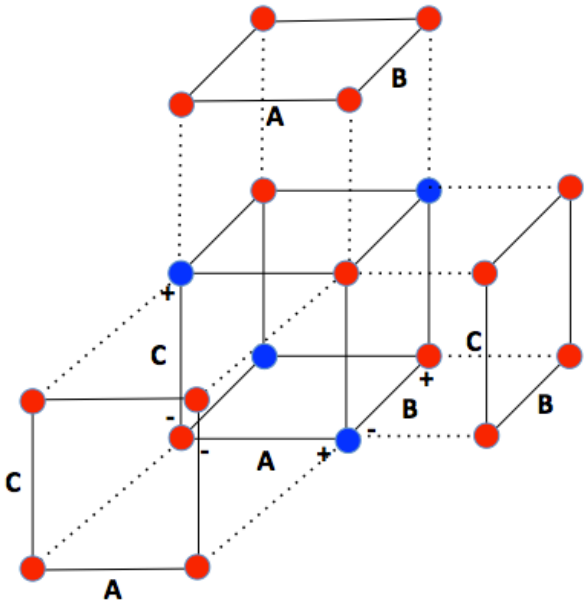

Figure 29. *A fractional factorial experiment has full factorials embedded within it.* If one of the factors (A, B, C) in an experiment has no effect on the response variable, then the cube can collapse into a 2D full factorial experiment for the remaining factors.

To choose the correct bins that produce this type of balanced, collapsible fractional factorial experiment, we need to use an aliasing structure. For our hypothetical 3-factor experiment above, we begin by drawing a design table for a 2-factor experiment, where all levels of factors are present. The levels of "C" are then needed to produce a balanced experiment are defined as the product of "A" and "B" (Table 3).

A	B	C=AB
-	-	+
+	-	-
-	+	-
+	+	+

Table 3. *Aliasing structure for a fractional factorial with 3 factors (A, B, C)*. By performing experiments with the factor conditions outlined in this table, we can produce a fractional factorial experiment. By using an aliasing structure in this way, the effect of Factor C is confounded with the effect of both A and B combined. However, losing this level of resolution allows us to screen potential effects much more quickly.

Representing the 3-factor experiment as a linear representation of the response would be;

$$y_i = b_0 + b_aA + b_bB + b_cC + b_{ab}AB + b_{ac}AC + b_{bc}BC + b_{abc}ABC$$

Therefore, for these 4 experiments, the matrix representation is;

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{matrix} b_0 & b_a & b_b & b_c & b_{ab} & b_{ac} & b_{bc} & b_{abc} \end{matrix} \begin{pmatrix} + & - & - & + & + & - & - & + \\ + & + & - & - & - & - & + & + \\ + & - & + & - & - & + & - & + \\ + & + & + & + & + & + & + & + \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_a \\ b_b \\ b_c \\ b_{ab} \\ b_{ac} \\ b_{bc} \\ b_{abc} \end{pmatrix}$$

Note that there is symmetry down the center of the matrix, so some columns are the same. These are aliases, and it means that there is a confounding effect for each of these pairs. For example, the effects from column b_a cannot be separated from effects due to the 2-factor interaction b_{bc}. There are numerous other aliases that follow;

$b_a = b_{bc}$ $b_b=b_{ac}$ $b_c=b_{ab}$ $b_{abc}=b_0$ (intercept)

We can remove aliases by collapsing columns;

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} + & - & - & + \\ + & + & - & - \\ + & - & + & - \\ + & + & + & + \end{pmatrix} \cdot \begin{pmatrix} b_0 + b_{abc} \\ b_a + b_{bc} \\ b_b + b_{ac} \\ b_c + b_{ab} \end{pmatrix}$$

A 4-factor fractional experiment such as ours will have the design table given in Table 4.

A	B	C	D=ABC
-	-	-	-
+	-	-	+
-	+	-	+
+	+	-	-
-	-	+	+
+	-	+	-
-	+	+	-
+	+	+	+

Aliasing structures;

$b_a=b_{bcd}$ $b_b=b_{acd}$ $b_c=b_{abd}$ $b_d=b_{abc}$

$b_{ab}=b_{cd}$ $b_{ac}=b_{bd}$ $b_{ad}=b_{bc}$

Table 4. *Experiment design table for a fractional factorial experiment with four factors (A, B, C, D) at two different levels (+/-).* By using the combinations of factor levels given in this table, we can reduce the number of experiments needed to investigate the potential effects of our factors. Some of the effects that factor combinations have on the output of the experiment will be aliased with others, as shown below the table.

This means that main factor effects will be aliased with 3-factor effects and 2-factor effects will be aliased with each other. Since 3-factor effects are very rare, the confounding with main effects should not be that significant (a '*sparsity of effects*' assumption). For 2-factor effects, we will be able to tell that there are 2-point interactions between factors but will not be able to discern which specific factors are interacting. If there are significant 2-

factor effects, it is always possible to carry out the other half of the experiments to improve the resolution.

By doing only half of the experimental bins, but still with 6 replicates in each, we hoped to gain some useful information in a much quicker way than using full factorials. Once designs have been make, analysis was carried out the same way, using factorial analysis and ANOVA, to compare the variance within and between groups to see if there are any significant effects.

4.2.8 Experimental testing

Synthetic genes were constructed using the "Thermodynamically Balanced Inside-Out" (TBIO) method[114], with overlapping primers designed using an in-house program. Genes were then inserted into a pNIC vector with a His_6 tag for subsequent Ni^{2+} affinity purification and transformed into a BL21-DE3 (gold) *E. Coli* strain for expression[115]. Cells were lysed and centrifuged at 20,000 rpm and solubility of designs was inferred from their presence in the soluble fraction. Apparent soluble designs were tested for foldedness using a size exclusion chromatography (SEC) column with pore size 10 kDa or 20 kDa, depending on the backbone. After SEC, the plan was to put viable designs through to CD and NMR and see what tertiary structural features we could see. For full experimental protocols, see main methods chapter in Section 2.2.

4.3 Results

4.3.1 1CC7 results (full factorial)

1CC7 was the first backbone that we decided to test, with a full complement of 96 designs experimentally made and tested. Of those 96 designs, 22 were still present in the soluble fraction after centrifugation but upon attempted purification using SEC, all promising soluble designs were eluted in the void zone (Figure 30).

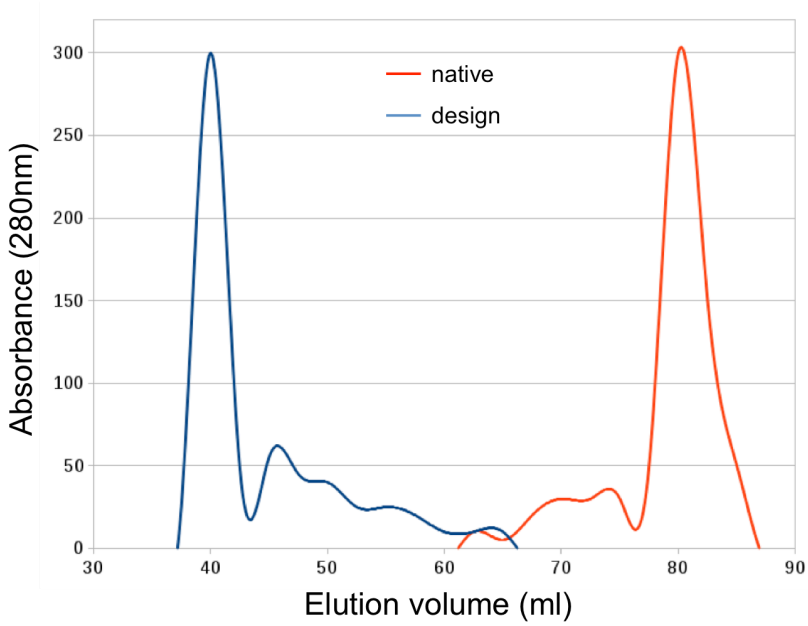

Figure 30. *Elution profile that is representative of the 22 soluble designs that were run through a SEC column of appropriate size.* The native 1CC7 protein (red) is eluted after approximately 80 ml of elution volume, suggesting that it is compact enough to enter the beads of the SEC column. Our designs for the same backbone (blue) elute at a much lower volume, in the so-called "void zone" which indicates that they are too large to enter the pores of the beads and so do not exist in a compact, globular state.

Compact globular proteins would be able to pass through the small pores in the beads of the SEC column, therefore taking longer to elute. This is seen in Figure 30 where the native protein is made using our expression protocol and elutes at around 80 ml. That our designs are eluted much sooner than the native sequence suggests that the designed proteins exist in large soluble aggregates that fall through the column without spending time inside the beads. This could indicate that the proteins therefore do not exist in a defined, compact 3D structure. Another possibility is that our proteins are folded, but the monomers aggregate together. Upon seeing that our proteins were eluted in the void zone, we re-ran the SEC experiments in a buffer that was intended to disrupt any oligomerization (Section 2.2.8). Even with this buffer, proteins eluted in the void zone in the same manner as seen in Figure 30. Because of the oligomer-disrupting buffer had similar results to the standard SEC protocol, we can assume that folded monomers

aggregating together are not very likely. The most probable explanation is that our proteins are unfolded and exist in an extended chain that cannot fit through the pores of the SEC column, thereby eluting much quicker than the folded state.

Analysis was carried out on the data, hoping to find some correlations between factors and increased solubility. The results in Table 5 show that the main effect for each factor, as well as 2-point interactions between factors, are very small. As well as each parameter estimate being small, the F-ratios are also below the critical F-value for our degrees of freedom (1 + 80 degrees gives a critical F-value of 3.96). All parameter estimates and F-ratios for 3-factor and 4-factor interactions were 0. These low values indicate that the data obtained from experiments did not produce any significant signal, and we cannot propose any insight into which factors are important for solubility. Unfortunately, since none of our designs were folded, we could not obtain any information from that data either.

Category	Soluble designs	Category	Soluble designs
----	1	+---	3
---+	3	+--+	1
--+-	0	+-+-	2
--++	1	+-++	1
-+--	1	++--	1
-+-+	0	++-+	3
-++-	2	+++-	0
-+++	1	++++	2

Category levels (+/-); buried divergence, surface divergence, secondary structure prediction accuracy, Rosetta energy

Effect parameter	Estimate of effect coefficient	F-ratio
Intercept	0.22	
Buried divergence (bur)	-0.04	0.87
Surface divergence (surf)	0.02	0.22
Secondary structure (ss)	0.04	0.87
Rosetta energy (energy)	-0.02	0.22
bur*surf	0	0
bur*ss	-0.02	0.22
surf*ss	0.04	0.87
bur*energy	0	0
surf*energy	0.02	0.22
ss*energy	0	0

Table 5. *Results table and analysis for 1CC7 solubility data.* The raw solubility data appears relatively evenly distributed across all combinations of levels of factors. Further analysis of this indicates that parameter estimated for each factor, and 2-point interactions between factors, are very low. The F-ratios obtained for each experimental bin are also below the critical F-value (3.96) and so the parameter estimates are not considered significant. All coefficients and F-ratios for 3-factor and 4-factor interactions were 0.

Because of the lack of information obtained from the full-factorial experiment, we decided to employ a quicker screening method for main factor effects and 2-point interactions when investigating the 3 remaining backbones. This would allow us to halve the number of proteins

experimentally tested, but still give us enough information to investigate potentially interesting features.

4.3.2 3CHY, 1Q5V and 1E5D fractional factorial experiments

Unfortunately, after making the balanced fractional factorial for each remaining backbone, we only managed to obtain a sparse number of soluble designs. For 1E5D we obtained 4 soluble designs, for 1Q5V only 3 were soluble and for 3CHY 2 were soluble. Again, all of these apparently soluble designs were eluted in the void zone upon SEC analysis and so were assumed to be soluble aggregates.

Because of the small number of soluble designs in each batch, the information gained from each of them is not very conclusive. In the fractional factorial analysis for these batches, the effect estimates are very low (<0.001) and not very significant (F-ratio < 1). This is hugely disappointing and from an experiment with this large of a sample size (using established method), we were expecting a much greater insight into the rules of protein design. As a way to still utilize this very efficient "design of experiments" approach, we decided to attempt to use computational analysis as the response variable in a similar framework.

4.3.3 Predictability as a response variable

Given the lack of signal obtained from the experimental approach, we turned to a computational method where the potential of the factorial approach could be investigated using a response variable that could be obtained more rapidly than observed solubility/folding. As a computational proxy for folding, we used the degree to which the native structure could be predicted using *ab initio* prediction with Rosetta[48]. The 1CC7 and 3CHY folds were used for this full factorial investigation, with 300 designed sequences in each experimental bin. We used the RosettaPredict protocol to generate 40 predicted models per designed sequence and then compared

each of these models to the 3D structural template we were aiming to use the TM-score[109]. The average TM-score of each of these models was then taken as a measure of predictability. 40 predicted models for each sequence is a very low number, especially considering that standard Rosetta protocol is to produce 10,000 designs with cluster analysis afterwards to find the most populated states. However, the computational demands of this level of detail are enormously high and are intractable for the number of designs we have made. Our reasoning is that 40 models should be sufficient enough to allow a general insight into how predictable our protein designs should be, with the more predictable ones having a narrower (and therefore more easily populated) energy landscape for a stable state. Having a lower number of models to produce allows us to perform this experiment in a realistic time frame and attempt to garner some information more easily. As stated previously, predictability has already been used as an indicator for more-viable designs[77,79,149] and so we foresee no problems in using this as a proxy to measure our design "fitness".

4.3.3.1 Pre-analysis

In order to test whether our hypothesis was valid: that predictability was correlated with a more viable design, we performed some pre-analysis. Using our accelerated prediction protocol, we produced 40-model sets for three different categories of proteins and then mapped each predicted model onto a given structure. The average TM-score over all models was used as an output of "predictability". 350 individual protein sequences were chosen to fit into the categories; "random", "design" and "natives". The "native" category contained sequences under 150 residues with a known 3D structure, and the predictions were mapped onto the solved PDB. The "design" category included a subset taken from our sequence designs and mapped against the template model we had designed them for. The "random" group was obtained using a random generator of letters that are in the amino acid alphabet and again mapped back onto the 3D template

used for the "design" category. Figure 31 shows the distributions of prediction accuracy.

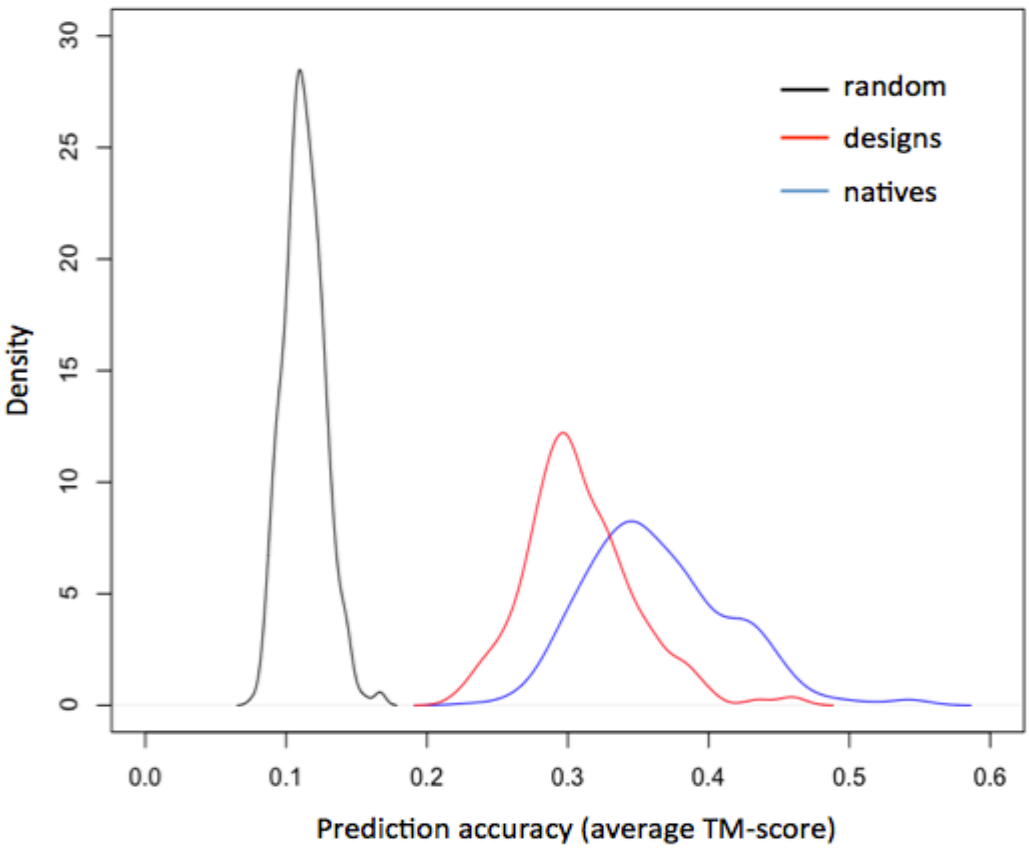

Figure 31. *Density plot showing the average predictability for different categories of sequence.* Each sequence was run through our accelerated prediction method, with 40 predicted models produced. Each of these models was then mapped onto the expected 3D structure and the average TM-score over all of these models was taken as the predictability of the sequence. The blue line represents native sequences mapped onto their known structure, and is the highest scoring. The red line indicates designed sequences mapped onto the desired template structure and the black line is random sequences mapped onto that same template structure.

As expected, the native sequences have the highest prediction accuracy (average TM-score) when mapped back onto their known 3D structures with a peak at around 0.35. The designed sequences are not too far behind in their distribution with a peak at 0.30 and the shape of the distributions look fairly similar. When the random sequence predictions are mapped back onto one of our templates, they give a very low score of 0.12 indicating that they are not very well tailored to the template (as expected). These

scores may seem generally low, but that is mainly due to the limitations of *ab initio* structure prediction[150,151]. We avoided using homology/comparative modeling in this case so as to not bias the native predictions with sequence-structure information and to offer an even level of comparison.

We also decided to look at native sequence recapitulation in our designs, as we thought this could offer an insight into whether we were approaching realistic designs. The logic behind this is that the closer a designed sequence is to the native, the closer the structure should also be. If our designs were far from the native sequence, it would not necessarily mean that they could not adopt our targeted structure. In fact, sequences as low as 15% identity can still adopt very similar 3D conformations[152,153]. However, increased native sequence recapitulation (especially in the core packing residues) is considered as another indicator of a more viable design[154]. Some design algorithms have a native sequence recapitulation of around 35 - 41% in the core, and 20 – 28% overall[155]. However, Baker et al suggest slightly higher values of 50% recapitulation in core residues and 33% overall are good reference values[123].

3CHY

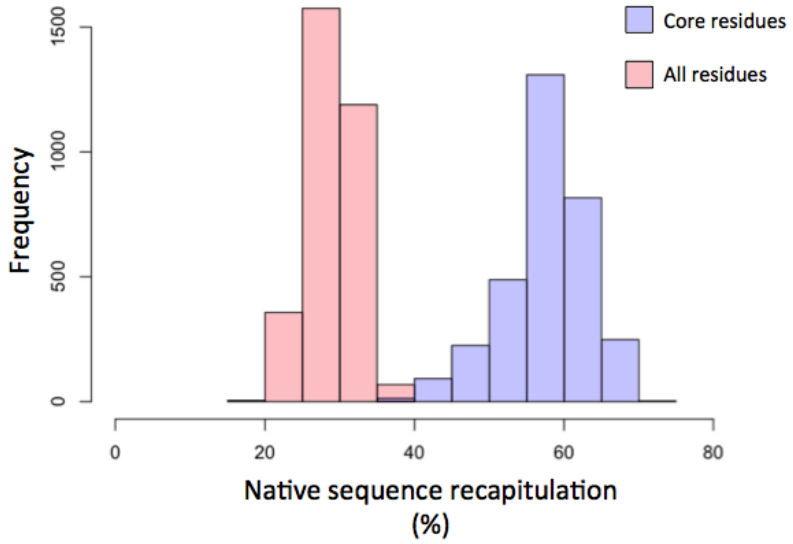

Figure 32. *Histogram of native sequence recapitulation for 3CHY redesigns.* Redesigns of the 3CHY protein appear to have very good native sequence recapitulation. ~60% of core residues maintain the same identity, with ~30% recapitulation overall.

Native sequence recapitulation for our 3CHY redesigns seems very acceptable, with around 60% of core residues being identical to the ones found in the native sequence (Figure 32). Around 30% overall sequence identity to the native is also acceptable and within the range we were aiming for.

1CC7

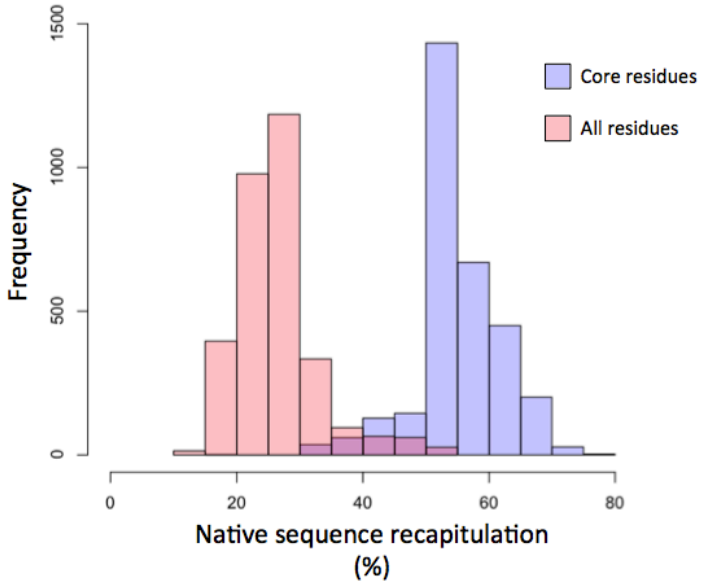

Figure 33. *Histogram of native sequence recapitulation for 1CC7 redesigns.* Redesigns of the 1CC7 protein appear to have very good native sequence recapitulation. ~55% of core residues maintain the same identity, with ~30% recapitulation overall.

The native sequence recapitulation for 1CC7 also seems to be good, with around 55% core identity to native and 30% overall (Figure 33). It is important to note that perhaps there are potentially more options for residues at the surface, as they can be less structurally integral than core residues. Since some surface residues may be involved with the function of the protein (e.g. binding surfaces) but not particularly critical in dictating the fold, these may vary more in our designs.

Based on the average predictability of our designs, coupled with the degree of native sequence recapitulation, we were encouraged that our design protocol has produced designs that we believe to have a degree of realism and viability.

4.3.3.2 Analysis

With 200 designs per bin (3,200 in total for each of the backbones), we obtained the following information;

1CC7

Effect parameter (set as "good"/+)	Estimate of effect coefficient	F-ratio
Intercept	0.39	
Buried divergence (bur)	0.018	1.42
Surface divergence (surf)	0.014	0.92
Secondary structure (ss)	-0.032	4.51
Rosetta energy (energy)	0.010	0.49
bur*surf	0	0
bur*ss	-0.010	1.00
surf*ss	-0.039	6.955
bur*energy	0.010	0.47
surf*energy	0	0
ss*energy	-0.039	6.71

Table 6. *Analysis of full factorial 1CC7 experiment, where prediction accuracy was the output.* 200 designs for each bin were predicted using Rosetta and then the predictive models were mapped back onto the target structure. After obtaining a score for each model, the data was analysed using previously outlined techniques (Sections 4.2.1 and 4.2.2). There are some significant effects when secondary structure prediction accuracy is at the "good" level, highlighted in green. An increased level of this factor leads to a negative effect on predictability, shown by a negative effect coefficient.

The data obtained from this experiment (Table 6) suggests that there could be a main effect from the *ss* variable, with possible interactions also present between *ss*energy* and *surf*ss*. The critical F-value for 1+3184 degrees of freedom is 3.84, meaning that they are also significant (highlighted in green). Surprisingly, the estimates for the magnitude of effect are all negative. This means that when the *ss* variable is set to the "good" setting, it has a negative impact on the predictability of a design. The negative impact of having a "good" secondary structure score is also seen in every interaction where this level is set (although it is not significant in the *bur*ss* interaction). Each significant effect estimate seems to only subtract ~10%

of the intercept value but when considered together, these effects could sum to a sizable amount.

3CHY

Effect parameter (set as "good"/+1)	Estimate of effect coefficient	F-ratio
Intercept	0.37	
Buried divergence (bur)	0.015	0.51
Surface divergence (surf)	0.030	2.74
Secondary structure (ss)	-0.035	3.94
Rosetta energy (energy)	0.035	3.20
bur*surf	0.025	1.39
bur*ss	0.011	0.26
surf*ss	-0.015	0.51
bur*energy	0.016	0.57
surf*energy	0.015	0.55
ss*energy	-0.048	7.59

Table 7. *Analysis of 3CHY computational results.* A similar pattern to the 1CC7 computational results is seen. The only significant effects (green) are when secondary structure prediction accuracy is high. This leads to a negative effect coefficient, meaning a higher level of this factor corresponds to a poorer structure prediction.

A similar trend is seen in the 3CHY designs (Table 7). When the *ss* variable is set to a "good" level, we see a significant main effect that contributes negatively to the predictability and a significant interaction between *ss*energy*.

A "good" setting for the *ss* variable means that there is a high degree of correlation between the prediction of secondary structure by PsiPred (on a single sequence) and the position definitions given by DSSP in the known 3D model. Intuitively, we assumed that having amino acids that intrinsically adopt the defined secondary structures for each residue position would be beneficial. The more residues that are found in the preferred secondary

structure, the more easily that structure would form and this could lead to a more stable design. However, this appears to not be the case and attempting to increase the secondary structure score for a design can have a negative impact on predictability. A lot of information is missing from the single sequence PsiPred secondary structure output, including local environmental knowledge. The 3D context of each residue is hugely important in producing stable protein folds and over-optimization of a 1D sequence for a given set of secondary structures maybe ignores this aspect too much. PsiPred not only uses intrinsic properties of amino acids, but also Bayesian inference to deduce the conditional probability of the amino acid forming a structure given that its immediate neighbours in the sequence have already adopted that structure. However, the greater context of each residue in 3D space is not represented at all. Having a large number of residues that intrinsically form a desired structure does not necessarily mean that sequence will have the required long-range interactions needed to form a compacted protein. Designs with a high *ss* score may have less predictability because they are less adept at forming the required contacts in 3D space. It has been suggested that the conservation of sequence based on secondary structure propensities without optimized tertiary interactions to stabilize the global fold will not constitute a native-like fold[156]. Our results seem to take this a step further and say that optimizing too far towards secondary structure propensity can have a negative impact on the viability of a design.

4.4 Discussion

We have attempted to use the Rosetta Design protocol in a large-scale test to investigate what makes a more successful protein design. Given past successes reported when using the Rosetta Design protocol, we assumed that valuable information could be obtained from screening a large number of native redesigns to look for correlations in factors that we thought to be important. Factorial analysis can be a very powerful tool for investigating large numbers of factors, and fractional factorials can help to reduce the time and cost of these screens. However, the sparse information obtained when attempting these types of investigations in the field of protein design meant that we could see no trends in what would constitute a "good" redesign at the experimental level.

The designs we put forward during this investigation and attempted to synthesize looked realistic by our measures, and to the design protocol used to make them. They seemed energetically favorable, with some having a high secondary structure prediction accuracy as well as low sequence divergence. This experiment was still useful because now we have a "negative" set of designs to probe for information on protein design. The PDB is a depositary for successful protein structures but the sequences that do not fold correctly are not present because they cannot conduct their function and so are not conserved by evolution. Now we have a set of proteins that look realistic to a number of measures, but have not adopted the correct 3D conformation. If we look for more differences between this large data set of unsuccessful designs and successful ones found in nature, we may be able to glean some more indications of what is needed to push our designs towards being better. This aspect will be elaborated more in Chapter 6.

By using predictability as a proxy for viable designs, we attempted to expand our experiment beyond what would have been possible in the given time frame if we were only looking at experimental validation. The

computational analysis of our design sets indicated that there could be a potential negative contribution present when optimizing towards idealized secondary structures. This has been touched upon by other groups in the field, and may be a feature to take notice of when designing sequences in future studies.

Our investigations included predictability as a proxy measurement for the viability of a design. However, we did not attempt to optimize towards a higher score. By using a protocol to increase the overall computational predictability of a design, we could hope that this would lead to an increase in real-world stability and foldability. The subject of the next chapter will address our attempts to produce higher TM-scoring and more consistent sequences for our given folds using a genetic algorithm made for that purpose.

Acknowledgements

The DoE approach to experimentation was planned with the help of Michael 'Sid' Sadowski.

5. A genetic algorithm for protein design

5.1 Introduction

In the previous chapter, we attempted to investigate a number of factors and the impact they had on producing viable sequence redesigns for known backbones. Unfortunately, the level of information obtained from experimental testing was not ideal, with data too sparse to glean any particularly useful information. After discovering this lack of information, we used computational predictability as a proxy for judging viable sequence designs in the hope that this would allow us to expand our sample size to the point where interesting information could be seen. However, the only relationship found through this was between increased secondary structure prediction accuracy and a lower computational predictability. So far in our studies, we have not made an attempt to directly optimise towards a greater predictability. Since predictability is thought to closely mirror the real-world foldedness of a protein[77,79,149], it would seem reasonable to optimise towards this as much as possible. To this end, we created a genetic algorithm that starts with randomly generated, low-predictability sequences and attempts to head towards increased predictability using a variety of mutational techniques. We used the same 3CHY and 1CC7 target structures as in the factorial design experiments, and attempted to produce highly predictable sequences *de novo*. If successful native redesigns were found using this methodology, we could then move on to designing sequences for our novel folds.

5.2 Methods

5.2.1 Basic genetic algorithm (GA)

To begin, we created a very basic genetic algorithm that mimics real-life evolution (Figure 34). The input for the protocol is a set of 20 randomly generated "parent" sequences. Two of these sequences are chosen to produce two "children" sequences via 2-point recombination at random positions. This can also result in 1-point recombination if the end of a sequence is chosen as a break. Using the set of 20 parent sequences, 100 children per generation are created in this manner. After all children have been produced, every location within a sequence can spontaneously "mutate" to any other amino acid (except Cys) with a given probability. As a measure of fitness for each sequence, 40 models are made using the *ab initio* prediction method and then mapped onto the target structure before being TM-scored. Again, this is rather a small number of models to make, but we felt this to be a nice number to gain some insight into potential structures. Minimizing the number of predictive models needed is especially important for this method, as it is very computationally expensive. For a generation of 100 children sequences, it takes between 1 – 2 hours on 100 nodes for our setup. This means that for a set of 250 generations, it can take anywhere between 10 – 20 days for the protocol to run.

The 15 child sequences with the highest average TM-score are ranked as the "most fit" and are put through as the parents for the next generation. Along with the highest scoring children, the top 5 parents from the previous generation are also maintained as seeds for the next generation. This cycle of selection continues until the average TM-score appears to converge and maintain a constant level. In this way, we optimise towards producing designed sequences with a high degree of predictability.

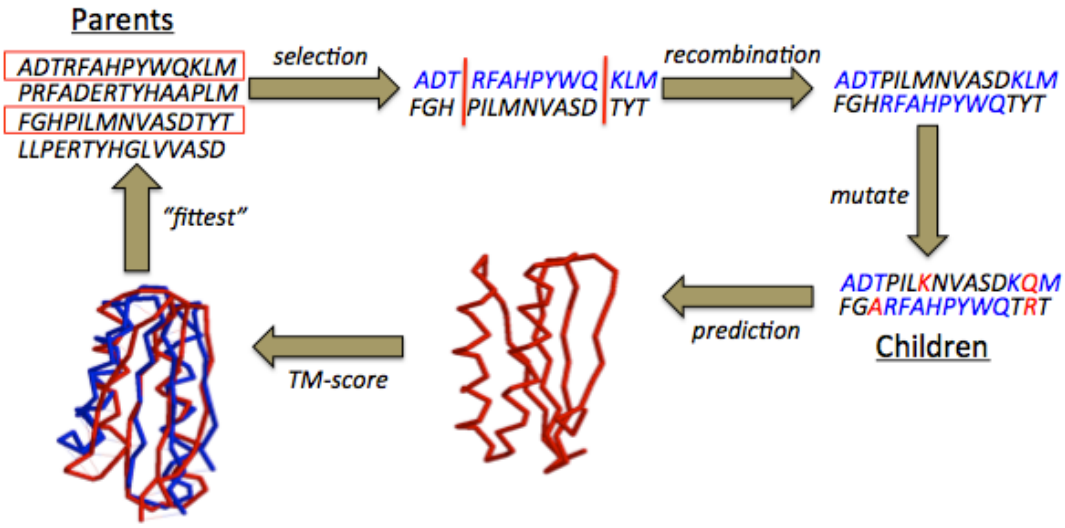

Figure 34. *Schematic of the genetic algorithm protocol*. The process starts with 20 randomly generated parent sequences. Two of these are chosen, break points are generated and recombination between these locations results in 2 children sequences. This is repeated until 100 children are produced. Each residue in the child sequence then has a given percentage chance to mutate into any other amino acid (apart from Cys). 40 *ab inito* predictions are then made for each child sequence and the average TM-score of these to the target structure is taken as "fitness". The fittest 15 children are taken through as the parental seeds for the next generation, along with the 5 top scoring parents from the previous round. This process continues until the average TM-score converges and no longer improves between generations.

5.2.2 Rosetta GA

Rather than solely relying on chance mutations to improve the predictability of a design, we also decided to make a version of the genetic algorithm that possessed a more "intelligent" mutation operator. The same overall methodology remained the same, with the only difference being that instead of a random change at the mutated positions, RosettaDesign was used to select an amino acid suitable for that location (again excluding Cys). In this way, we hoped that our designs would be better directed and converge to a better predictability in a shorter amount of time. The sequence to be mutated was first threaded onto the target backbone, the whole structure was relaxed and sidechain positions were then optimised. Then, rotamers for all residues were tested at mutable locations and the amino

acid/rotamer combination that gave the lowest energy state was accepted at this position.

5.2.3 Dynamic GA

Another feature we thought could increase the rate of convergence was to use dynamic mutation rates when selecting which positions to mutate. In our basic protocol, each residue has the same probability to be chosen for change at all times. However, if we already have part of a well-formed structure then we do not want to necessarily make any changes within those parts that predict well. By keeping well-predicted substructures constant and mutating residues that are not in the correct positions, we may be able to produce better sequence designs. SAP (Structural Alignment Program)[157,158] is a structural comparison program that can offer an indication on how similar local structures are in two models. Using the local similarity score obtained from this program, we can weight our mutation rate at each location based on how well that area is predicted. We can assign a lower mutation rate to regions that are well predicted, and a higher mutation rate to the locations that are not predicted very well in the hopes that area will assume the correct structure more quickly (Figure 35).

Being able to adjust the mutation rate dynamically based on the previous predictions would be quite hard to implement in our GA if we kept the recombination step in our protocol. Breaking two different sequences and knitting then back together is unlikely to maintain the predictability of individual substructures, meaning that our adjusted mutation rates would not be relevant for the next generation. For this reason, we modified our protocol to not include the recombination step. A baseline mutation rate was set to begin with, and the structures of the parent sequences were predicted. Mutation rates were then adjusted based on the local similarity score given by SAP. Areas of low local structural similarity, indicated by a SAP score below 1 for that position, were given a mutation rate 4 times the original baseline mutation rate. Regions with high similarity, and therefore

good predictability, were given a reduced mutation rate of ¼ the value of the baseline mutation rate. Mutations then occurred, using RosettaDesign for residue selection, according to the given probabilities (without any recombination step) to produce children sequences. Structures for these were then predicted and the 15 highest scoring were taken as parents of the next generation, along with 5 parents from the previous seeds, with the mutation rates adjusted accordingly for each location.

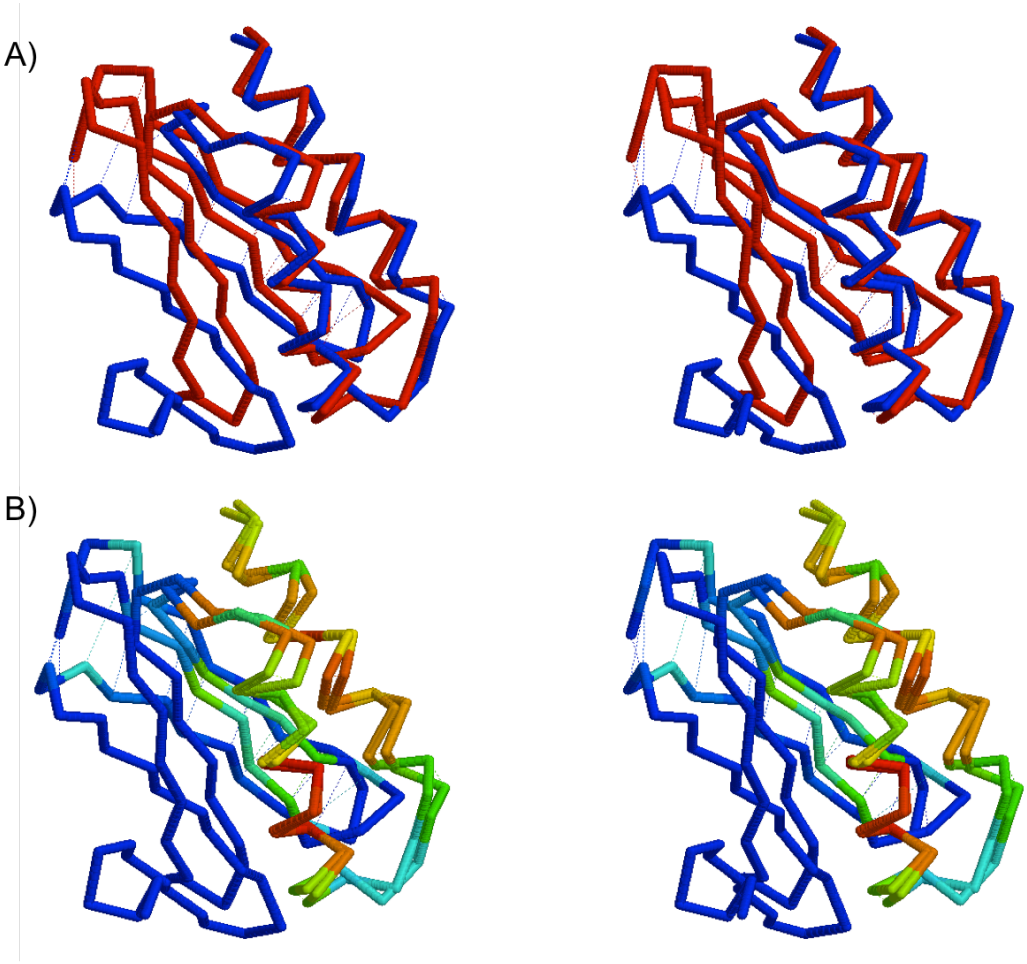

Figure 35. *Stereo images illustrating local structure similarity scores produced by SAP*. A) The target model is shown as the red chain, whereas the predicted model is shown as the blue chain. The helices are well predicted, with a high degree of correct local structure present. However, the strands for the predicted model are quite far from the target. B) Coloured by the local similarity score produced by SAP, blue parts of the structures indicate regions that have low similarity (score <1) and therefore will possess a higher mutation rate for those locations in the next generation. The areas that are not blue have a score greater than 1 and so will have a severely reduced mutation rate in the next iteration of design in the hopes that structure will be conserved.

5.2.4 Setting a mutation rate

Setting the mutation rates correctly for our GAs is key to their performance. Too low a mutation rate can mean that the method will take a long time to converge, as only small changes are explored in each generation. Setting the rate too high can mean that we do not get the convergence we want, as too many positions change each generation and not enough consistency is maintained. Therefore, we thought it prudent to run a few different GAs with different mutation rates in order to choose the optimal one. We ran both the basic GA and the RosettaDesign GA with mutation rates of 1%, 2.5% and 5% to see which led to the best convergence rate. For both protocols, a 2.5% change to mutate each residue appeared to provide the best convergence after 150 generations (Figure 36). After an initial 50 generations, the 2.5% mutation rate had a maximum average TM-score of 0.35, as opposed to the other two rates that were around 0.30 TM-score. The lower mutation rate of 1% appears to still be improving just before the run was completed, but the rate of convergence is much lower than the one seen with a mutation rate of 2.5%. By the end of 150 generations, a 2.5% mutation rate gives a maximum average TM-score of ~0.45 within a generation and the 1% rate is still at ~0.35. The 5% rate is interesting, as it initially appears to be comparable to the 1%. However, it seems to reach a peak of ~0.35 and drop off afterwards. This could be due to the reason stated earlier, that we do not maintain much consistency with a higher mutation rate and so achieve lower scores by changing too many residues at a time. Based on this information, we used the 2.5% mutation rate as a basis for our GA methods. A similar pattern was seen for both the basic and RosettaDesign GAs.

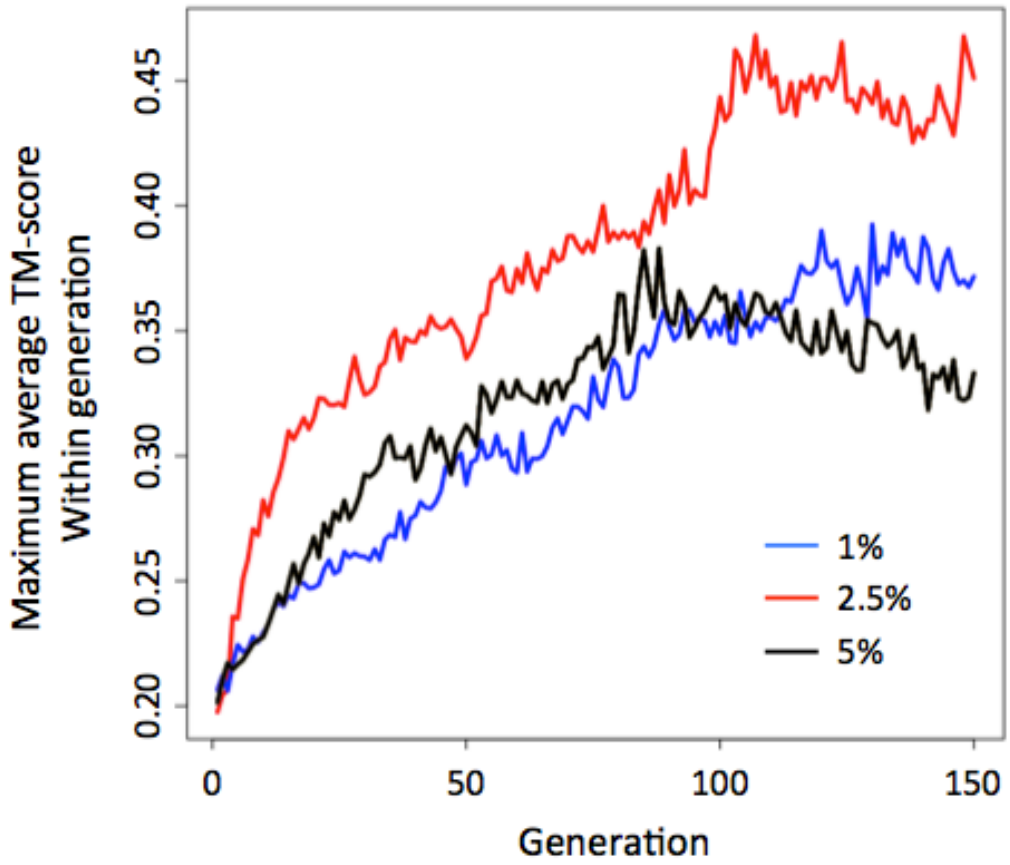

Figure 36. *The effect of different mutation rates in the basic genetic algorithm*. Different mutation rates were tested to investigate which had the best convergence. All rates seem to offer some improvement to the maximum average TM-score within a generation, but a 2.5% chance seems to offer the highest score and in the lowest number of generations.

5.2.5 Experimental testing

We ran each of the GA variants for both the 1CC7 and 3CHY backbones to try and produce realistic designs for the already known structures. From each of these runs, we chose the 8 sequences that had the highest average TM-score to the target model and took them through to experimental testing. We followed the expression protocols previously outlined (Section 2.2), and conducted experiments without solubility tag fusions. Checking the soluble fraction of cell lysates, we were able to determine which designs to take forward to SEC. SEC was first performed under the normal conditions,

followed by a second round using the aggregation prevention buffer outlined in the main methods.

5.3 Results

5.3.1 Random genetic algorithm

To begin with, we ran a basic genetic algorithm that had a random mutation operator where a location was mutated to a different residue with a 2.5% probability. There were no selection criteria for what types of residue should be at given locations, just a random substitution if that location was chosen. This means that we are heavily relying on the *ab initio* prediction step to judge if a new sequence is more likely to adopt the correct structure.

The starting set of 20 randomly generated sequences has a low maximum average TM-score, with the genetic algorithm starting at ~0.15 TM-score to the target structure (Figure 37). The lower bound for native protein predictability is around 0.30, with the higher bound being 0.45. The basic GA reaches the lower bound after 30 generations, with 0.45 TM-score being reached after a total of 100 generations. At this point, a plateau is reached and the designs appear to stop improving in their predictability. As a standard for well-designed proteins, we also used Top7 as a comparison. Top7 is a previously *de novo* designed protein by the Baker group that is very stable, has a high predictability and is robust to a significant number of mutations[77,78,159,160]. We ran the Top7 sequence through our *ab initio* prediction protocol to produce 40 models, and then compared each of these to the known structure. The average TM-score between model and known structure for Top7 was 0.50, higher than our GA managed to produce and also better than all native proteins. As an aside, Top7 is known for it's atomic-level prediction accuracy that means it should be predicted almost perfectly with a TM-score of 1.0 when compared to the actual structure. However, this is when using standard prediction protocols where 10,000 models are produced and the lowest energy cluster is chosen as the most

viable structure. Since producing this many models would take a very large amount of time, we used a drastically reduced set of 40 models to screen proteins with the rationale that more stable proteins will adopt structures close to the target even at this level. This means that the accuracy of prediction will be severely hindered, but can still offer a good means to assess a large number of sequences in a much shorter amount of time.

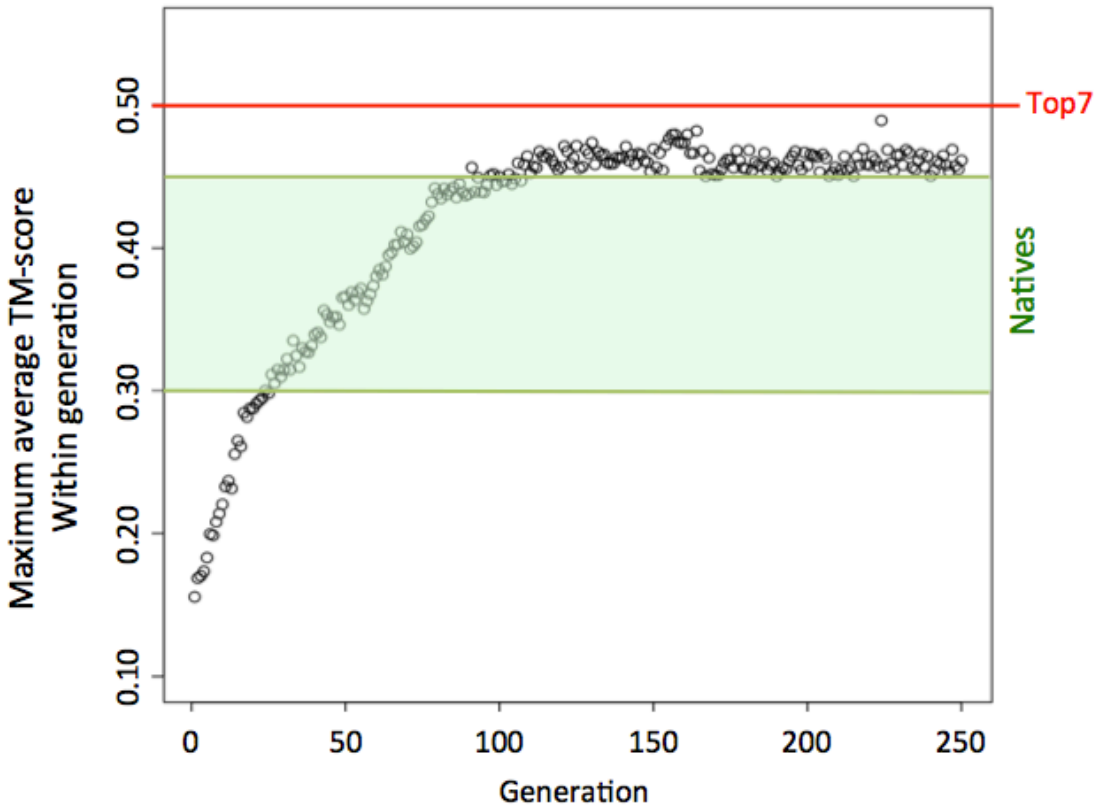

Figure 37. *Maximum average TM-score increases with generations in our basic genetic algorithm.* Randomly generated sequences produce a low TM-score to the target model, but selecting the "fittest" sequences to take through to seed the next generation quickly results in an increase in scoring. By 30 generations, sequences are comparable to the lowest-scoring native sequences mapped onto the respective known structures. By 100 generations, designed sequences are at the same level of predictability as the highest scoring natives and appear to plateau here for the next 150 generations. However, the very stable Top7 protein has a higher predictability than both the native proteins and all of our designed sequences.

5.3.2 RosettaDesign GA

Using a random mutation operator seemed to work quite well, producing sequences that have a level of predictability comparable to the highest

scoring native sequences. In an attempt to improve the process, we used a more "intelligent" mutation operator in the form of the RosettaDesign protocol. In this situation, sequences are threaded onto the target structure and then positions marked for mutation are assigned a new residue by using rotamer libraries combined with a potential energy function. It was our hope that using a directed method for residue selection would lead to an increased rate of TM-score improvement and a final TM-score plateau that was higher than using a random amino acid selection.

We ran our RosettaDesign GA and both of these hypotheses proved to be true (Figure 38). Again, randomly generated sequences produced a low maximum TM-score as a starting point, but scoring quickly improved as more generations were made. After 30 generations, scores were already within the range of "native" scores, a slight improvement in the rate of increase when compared to the basic GA. The native scores were quickly surpassed, and by 125 generations we had reached the level of predictability that Top7 possesses. After this, our sequences plateaued at a higher score of 0.54 for the remaining generations. These sequences are predicted very well, and we would therefore expect them to be very stable when we experimentally test them.

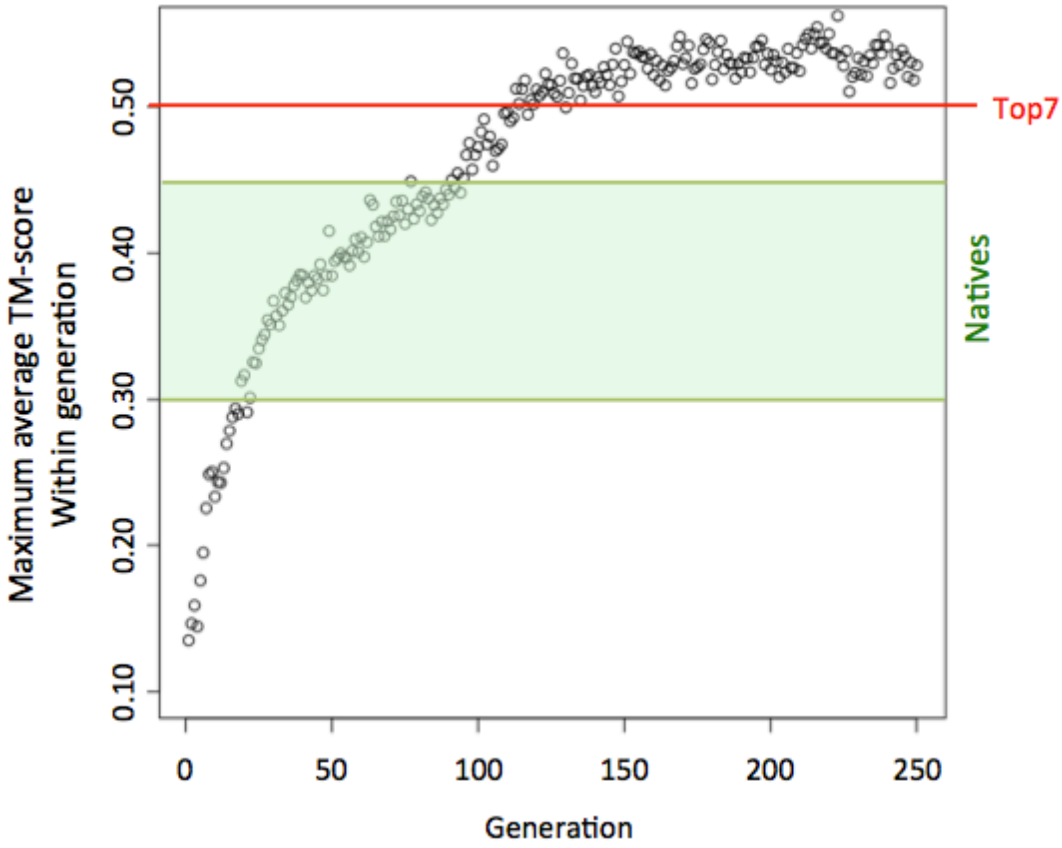

Figure 38. *Using a RosettaDesign mutation operator results in a quicker TM-score increase and reaches a plateau at a higher level*. The low starting scores given by random sequences are quickly improved upon, with a TM-score around 0.34 after 50 generations. This is in the middle of the distribution of native scores, and so we would expect our sequences at this point to be relatively realistic. After 125 generations, our sequences have surpassed even the highest scoring native sequences and are comparable with Top7. In the following 125 generations, designed sequences score slightly higher than Top7 at ~0.53.

5.3.3 Dynamic GA

In order to increase the rate of improvement for designed proteins in subsequent generations, we attempted to use a method that changed the mutation rates at each location based upon how well that part of the structure was predicted previously. Regions that were predicted incorrectly had an increased mutation rate, whereas regions that were predicted correctly when compared to the target structure had a reduced mutation rate. We hoped that by doing this, well-designed regions would be maintained through generations but structural features needing

improvement would explore more sequence space in order to find a suitable fit.

Using this method appeared to offer a dramatic increase in the rate of TM-score improvement within the first few generations (Figure 39), with designed sequences reaching native-like predictability in only 4 generations. A steady increase in the maximum average TM-score is then seen as the GA progresses, until a higher predictability than native sequences is obtained after ~50 generations. By 100 generations, designs had a TM-score similar to Top7. After this, scores increased slightly until a plateau was reached at about 0.53.

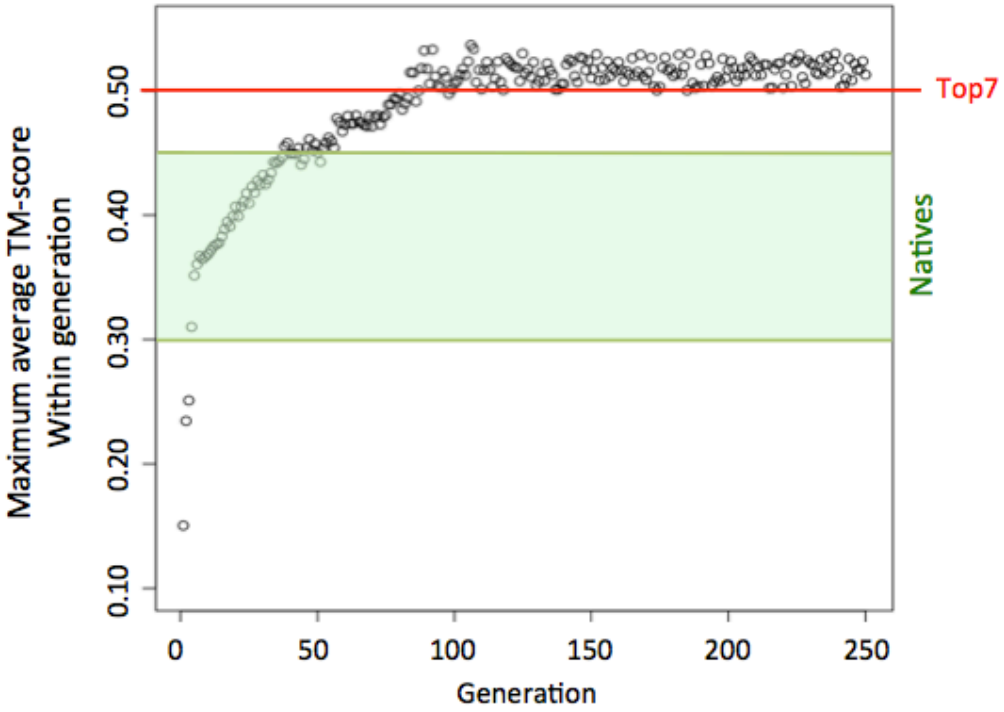

Figure 39. *Using a dynamic mutation rate scaled to the local predictability in that region, along with a RosettaDesign mutation operator, results in a quick TM-score increase until a plateau is reached*. After a low initial starting point from randomly generated sequences, TM-score quickly improves so that native-like predictability is attained within 4 generations. Native predictability is surpassed after 50 generations, with Top7-like scores being obtained by 100 generations. Much like in the design GA, a plateau is reached at ~0.53 TM-score.

The large increase in score for the first few generations is impressive, but not too surprising. Because our process starts with randomly generated sequences, they are not likely to have any similarity to the target structure. This means that a considerable number of positions will have a high mutation rate assigned to them, leading to redesign by the Rosetta protocol. Having a very high mutation rate at numerous locations with is essentially the same as attempting to redesign the majority of the structure, which we have already performed during the factorial design experiments. When the majority of positions are open for a complete redesign by Rosetta, the average TM-score of the resulting proteins is around 0.30 (Figure 31) and we also see this in the initial generations of our dynamic GA. When some well-predicted regions of structure begin to form in the designed sequences and mutation rates are lowered again, the normal GA process seems to take over and optimise towards increasing the TM-score until a plateau is reached.

5.3.4 Sequence recapitulation

As mentioned previously, we can assume that prediction accuracy and native sequence recapitulation are good proxies for viable protein designs. We are already optimising towards a better predictability with our GAs, but it would also be interesting to investigate how native sequence recapitulation changes as more generations are made.

When using a random mutation operator, where any residue can be replaced with any other residue, there doesn't seem to be any increase in native sequence recapitulation with generation progression (Figure 40). Although average TM-score increases as the method progresses, sequence identity to the native remains below 10%. One of the reasons for this could be that residues with similar properties to native are being placed in the locations to be mutated, giving a similar effect on predictability but without having exact identity.

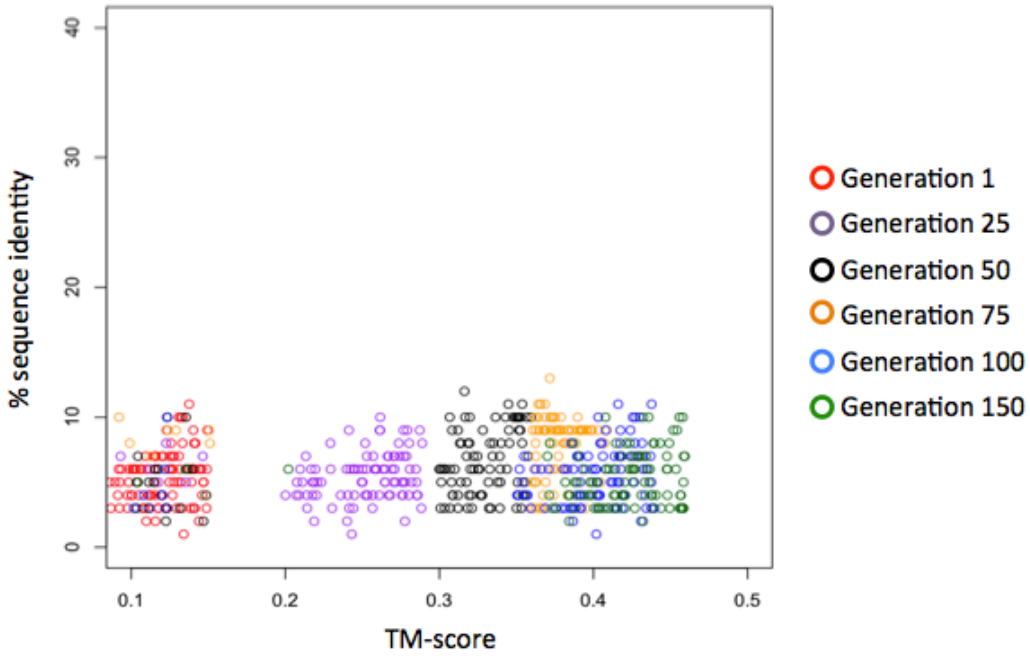

Figure 40. *Native sequence recapitulation for each designed sequence in a generation when using a random mutation operator.* Although the average TM-score of designed sequences increases between generations, native sequence recapitulation appears to remain below 10%.

When using a RosettaDesign mutation operator to make residue selections based on energy minimization and structural information, native sequence recapitulation can be seen to increase with more generations (Figure 41). TM-score increases as the method progresses and it appears that sequence identity also follows the same trend. After 150 generations, a peak of 0.5 TM-score is reached and there is a 40% sequence identity to the native at this point. Between 150 and 250 generations, when the TM-score plateau is reached, native sequence recapitulation also appeared to remain constant at ~40% (not shown). By the end of our design process, we obtain a slightly higher total sequence recapitulation than the 33% suggested by Baker in previous work as being indicative of a viable design[81,123]. There was no difference in the pattern when using either the RosettaDesign GA or the dynamic GA, possibly because they both used the same mutation operators to make residue selections.

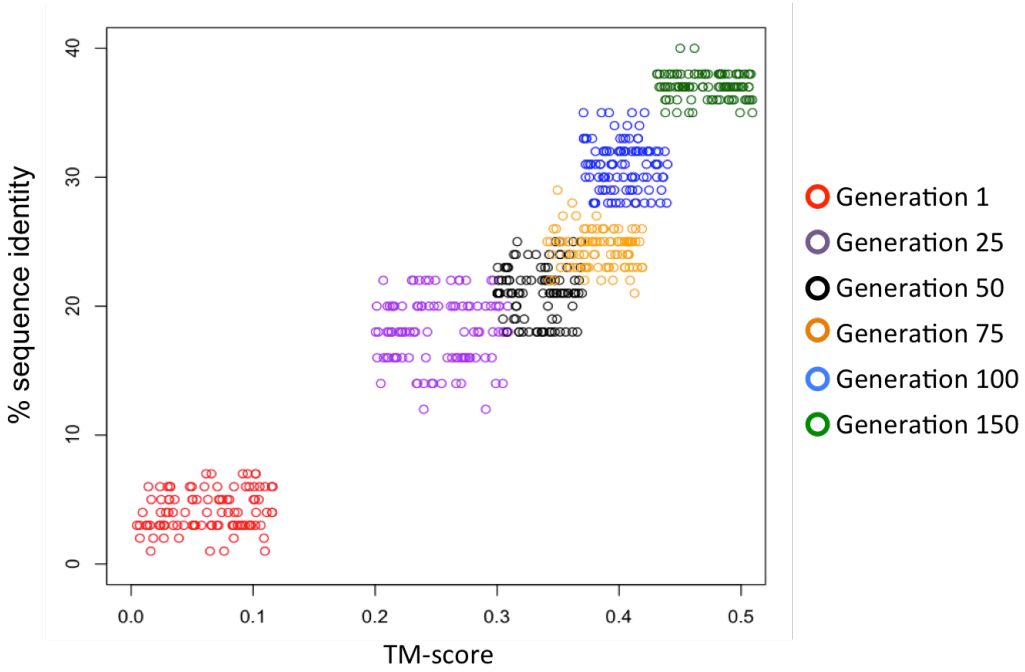

Figure 41. *Native sequence recapitulation for each designed sequence in a generation when using the RosettaDesign mutation operator.* As TM-score increases with generations, so too does the level of native sequence recapitulation. By the end of 150 generations, TM-score and sequence identity are both higher than previously published protocols.

5.3.5 Comparative modelling predictions

For each of the designs we chose to put forward into experimental testing, we first decided to use the comparative modelling method of structure prediction to see how well the sequences would perform. We are already using an *ab initio* method of prediction to direct our designs in their predictability, but we still thought it prudent to also use a more robust prediction method alongside it. There are some problems with *ab initio* methods, and using the comparative modelling as an additional step is a good check to see if we are drifting too far from real structures.

For the random mutation operator GAs, comparative modelling produced structures that were very close to the targets we were aiming for (Figure 42A). The global fold is good, with all secondary structure elements in the correct places. However, there is still some shifting in the frames of the helices that makes the atomistic detail of our designs not quite the same as the target. Overall, even using a random mutation operator in our GA allows

us to produce semi-realistic designs when it comes to comparative modelling prediction outputs with a TM-score of between 0.67 - 0.76 to the target structure.

Using the RosettaDesign mutation operator in our GA also appeared to produce designs that were very close to the native structure (Figure 42B). With a TM-score of 0.83 - 0.91 of prediction model to target structure, these designs are very high scoring when using comparative modelling. There are some minor differences in some loop regions, but the majority of backbone atoms are predicted to be in the correct place and the shifts seen when using the random mutation operator are not present. Again, there was no real difference seen when using the RosettaDesign or dynamic GA.

Based on the encouraging results from both *ab initio* and comparative modelling prediction methods, we continued with our experiments to test our designs in the lab.

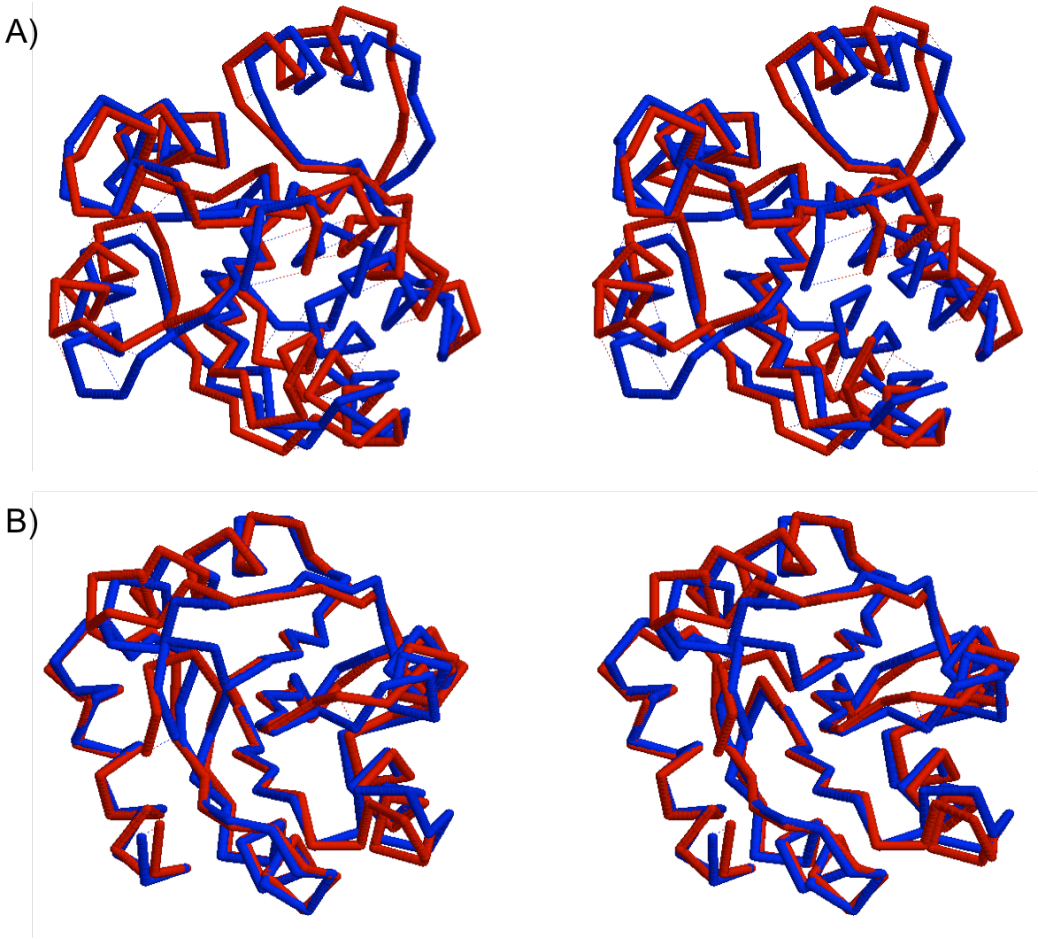

Figure 42. *Stereo images demonstrating the comparative modelling results of designs produced using the GA method.* A) Comparative modelling of designs made using the GA with a random mutation operator seems to generate models that are quite close to the target structure (TM-score = 0.76). There is some frame shifts in the helices, so atomistic detail is missing, but the global fold is correctly predicted and secondary structures are generally in the right place. B) Using a RosettaDesign mutation operator leads to a predicted model that almost has atomistic detail (TM-score = 0.91). There are some slight differences in the locations of some loops, but backbone atoms are almost all in the same place as in the target model.

5.3.6 Experimental testing

For each of our genetic algorithm protocols (random, RosettaDesign and dynamic), we chose the 8 sequences for each backbone that had the highest *ab initio* TM-score to the target structure and experimentally tested these for solubility (Table 8).

Backbone	GA variant		
	Basic	Design	Dynamic
1CC7	0	2	2
3CHY	0	1	1

Table 8. *Raw results for solubility of designs in each GA category*. 8 different sequences were expressed in *E. Coli* and only a very small number of these were soluble. No designs produced by using the "basic" GA were soluble, and only a very small number were soluble from the "Design" and "Dynamic" GAs.

Unfortunately, only a total of 6 designs (out of 48 tested) were soluble from this experiment and these were all from using the RosettaDesign mutation operator. When these 6 designs were taken forward to the size exclusion chromatography step for further purification and characterisation, they eluted in the void zone (similar to what was seen in Figure 30). This indicates that again, the designs we have produced exist as soluble aggregates rather than globular proteins and are not likely to adopt the correct structure.

5.4 Discussion

Our experiments so far have used structure prediction as a proxy for viable designs when varying a number of other factors but have not focused on directly improving predictability. Using a genetic algorithm approach, we attempted to optimise the TM-score between predicted structure of a sequence and the target model. As a mutation operator, we used either a random method of residue selection or a more directed selection with the RosettaDesign forcefield. We began all our genetic algorithm variations with random sequences, so TM-scores start off low but then steadily increase with each generation until a plateau is reached and no more improvement is seen. Interestingly, regardless of the GA process used, TM-score improvement is complete by 150 generations.

The basic version of our GA was able to produce sequences that were comparable to the best-scoring native sequences, but not quite as good as the very stable Top7 protein. Changing the mutation operator and using a dynamic mutation rate improved upon this early success, to quickly surpass natives and even Top7 scores. Each of the GAs also generated sequences that looked very good when using a comparative modelling approach to predict the structure. Predicted models had a TM-score between 0.67 - 0.91 when compared to the target structure, which is a very good score and suggests that the structures have a high degree of similarity[112]. The templates picked our by the comparative modelling protocol to predict our sequences were all mainly 3CHY-like proteins, which is encouraging as it means we are getting close to the native sequence and structures.

The predictability of sequences produced by the basic GA was comparable with that of the best native sequences, but native sequence recapitulation remained at a low level and it is hard to judge exactly what this may mean. The same fold topology may be adopted by sequences with very little exact identity[152,153]. However, using native sequence recapitulation as a proxy for viable designs can be invaluable in helping to screen a multitude of different sequences as the entire process of computational design and structural characterisation is very involved and time-consuming. We know that the native sequence for that given structure is already a suitable one, given that it already exists in nature. The closer we can get to this should be a good measure of how well our design process is doing. When using the RosettaDesign mutation operator in our GAs, we can see an increase in recapitulation as the generations progress. This means that it is possible to get closer to the native sequence using the design method and maybe we would expect these designs to be performing better than when just using the basic GA. We give the RosettaDesign GA some context as to where residues are located in the structure when making a selection, whereas the basic GA randomly attempts to fit a different residue in. This means that the RosettaDesign protocol uses structural information unavailable to the basic

version to guide the amino acid selection process at a residue position, and this may lead to improved recapitulation through better understanding of the steric environment at that location.

Since both the *ab initio* and comparative modelling predictions appeared to be very good, we would have perhaps expected our sequences to perform better under experimental testing. For the RosettaDesign and dynamic GAs, we also had a high level of native sequence recapitulation. Unfortunately, very few of our designed sequences were soluble and none of them seemed to possess a fully compacted tertiary structure that could pass into the pores of a correctly sized size exclusion column. This suggests predictability and native-like sequences are not the only criteria that need to be met in order to produce viable designs. There could be other factors that we have not taken into account when producing our designs, and our understanding of the process could do with improvement.

There is also a lot of room for improvement in the GA method itself. We chose to parameterise it by using values that we thought were reasonable (e.g. making 40 predicted models to score for fitness, or scaling the mutation rate in the dynamic GA by a factor of 4). Although the values that we chose appeared to allow the GA to function properly and produce realistic designs, there is still a lot of optimization that may improve overall performance. Typically, these types of parameters are investigated with multiple runs of a GA and the best performing are chosen to be the final values. However, as mentioned previously, a typical GA run can take anywhere between 10 and 20 days on 100 computer nodes. This is a huge amount of time and rather than spend an excessive amount of time performing multiple runs for parameter optimization when we were unsure the methodology would even work, we selected what we thought were reasonable parameters and pressed forward to synthesizing proteins as soon as possible. In the end, we did produce proteins that looked realistic to all of our computational measures but met with no experimental success. Perhaps then it was an informed choice to leave the complete optimization of the GA for a later

date, when we have more of an indication about what comprises a truly viable design.

In all of our experiments so far, we have produced protein designs that look realistic to a number of measures and are predicted very well by state-of-the-art prediction programs. Unfortunately, we have met with little success as few have been soluble and none have been folded. On the positive side, we now have a large number of viable-looking designs that we thought would adopt correct structures but do not. As mentioned previously, there are only 'good' sequences in nature as the ones that do not fit will be discarded by evolution. We now have a set of 480 total designs that we have experimentally tested but do not possess the correct tertiary structure. If we compare some parameters of this set to a similar number of native proteins (that will adopt the correct structure), we may be able to obtain some insight into what the differences are. This will be the subject of the next chapter.

Acknowledgements

All computational and experimental work was conducted myself.

6. Machine learning for improved design

6.1 Introduction

In our studies so far, we have attempted to make viable protein designs through a number of different protocols. We have used established sequence design protocols to investigate the effect of various factors on solubility and foldedness for both novel and native structures. We also proposed a new method for design, using a genetic algorithm to optimise towards producing sequences that are predicted to closely resemble the target structure. Using our assorted techniques, we have produced a total of 480 sequence designs that look realistic to our eyes and to numerous computational measures. Unfortunately, none of the designs synthesized so far appeared to possess a true defined globular structure. In an attempt to glean more information from these failed attempts at protein design, we compared them to a set of real proteins of known structure. By taking equivalent subsets of proteins from each group and defining attributes for each member, we could apply machine learning techniques to discern a difference between the two groups. Upon finding dissimilarities, we attempted to direct our design process to a previously untested area of attribute space. Once designs fitting these parameters were produced, we again looked at native sequence recapitulation and predictability but this time added molecular dynamics simulations as a further test of stability. Proteins were then tested experimentally, without the use of a solubility tag.

6.2 Methods

6.2.1 J48 decision tree

To enable us to determine some decision rules and better direct our design process to areas that could be interesting, we created a J48 decision tree using the Weka data analytics program[161]. This method is a Java-implemented version of the C4.5 algorithm[162] that generates decision trees which can be used for classification. For a training set (*S*) of classified samples, each sample (x) has a number of attributes (*y*) (x_{1i}, x_{2i}, x_{3i}…x_{yi}). For each iteration of the C4.5 algorithm, it iterates through each unused attribute of set *S* and calculates the information gain (entropy) of that attribute. The attribute with the largest information gain is chosen to split the set by that attribute to produce a subset of data. For example, if we have a simple data set;

Attribute 1	Attribute 2	Class
A	True	1
A	True	1
B	True	1
A	False	2
B	False	2

Table 9. *Example training set for machine learning techniques.* If a set of data has two different "Attributes" and a resultant "Class", machine learning techniques can be used to attempt to find commonalities that exist within a class.

Then the total entropy for the set (*S*) is;

$$Info(S) = -\sum_{i=1}^{k}((freq(C_i, S)/|S|).ln(freq(C_i, S)/|S|))$$

where $freq(C_i,S)$ is the number of the set that belong to class C_i (out of k possible classes) and $|S|$ is the number of samples in set *S*.

$$\text{Info}(S) = -(3/5)\ln(3/5) - (2/5)\ln(2/5)$$
$$= 0.6730$$

To find the entropy of attribute *y*, we must first partition the set in accordance with *n* outcomes of the attribute test. On this partitioned set (*T*), the following equation can be applied;

$$Info_y(T) = \sum_{i=1}^{n}((|T_i|/T).info(T_i))$$

Entropy of attribute 1;

$$\text{Info}_{y1}(T) = 3/5\,(-2/3\ln(2/3)) - 1/3\ln(1/3))$$
$$+ 2/5\,(-1/2\ln(1/2) - 1/2\ln(1/2))$$
$$= 0.6592$$

Entropy of attribute 2;

$$\text{Info}_{y2}(T) = 3/5\,(-3/3\ln(3/3) - 0/3\ln(0/3))$$
$$+ 2/5\,(-0/2\ln(0/2) - 2/2\ln(2/2))$$
$$= 0$$

From this, we can estimate the information gain from each attribute (*y*) by;

$$Gain(Y) = Info(S) - Info_y(T)$$

Attribute 1 information gain = 0.6730 – 0.6592 = 0.0138

Attribute 2 information gain = 0.6730 – 0 = 0.6730

Because the information gain is much higher for attribute 2, we would choose this as our first splitting node. This is fairly obvious from the data, but is a good illustration of how the method works.

Once a node has been formed, the process recurs for each subset. The process can stop when every element in the subset belongs to the same class. This branch is then turned into a leaf of the tree and labelled with the class that belongs here. The method can also stop when there is no significant information gain for attributes, but there are still mixed classes. In this case, the branch is turned into a leaf and labelled with the most common class. Once the tree is completed, the C4.5 restrospectively prunes the tree, discarding one or more subtrees and replacing them with leaves that simplify the decision tree. This is useful because large decision trees can be difficult to understand if each node has a specific context established by the outcomes of antecedent nodes.

To define tests at each node of the tree, the C4.5 can use one of two algorithms. There is a "standard" test for use on a discrete attribute with one outcome and branch for each possible value of the attribute being tested. However, if an attribute (*Y*) has continuous numeric values, a binary test with the outcomes $Y \leq Z$ or $Y \geq Z$ can be used with a threshold value (*Z*).

To define the threshold Z, the samples in the set are first sorted by the values of the attribute being considered (e.g. v_1, v_2, v_3 … v_m). Any threshold value lying between vi and vi+1 will have the same effect of dividing cases into those whose value of attribute lies in {v_1, v_2 … v_i} and those whose value is in {v_{i+1}, v_{i+2} … v_m}. Thus, there are only *m-1* splits, all of which can be examined systematically to obtain an optimal split. Although it is usual in other methods to take the midpoint of (v_i, v_{i+1}) as the threshold, the C4.5 algorithm chooses the lower value (v_i).

For example, if we have a set;

	Attribute 1	Class
	10	1
Z_1		
	20	1
Z_2		
	30	1
Z_3		
	40	2

Total entropy of set = - 3/4 ln (3/4) – 1/4 ln (1/4) = 0.5623

Entropy of each threshold value;

Z_1 = 1/4 (- 1/1 ln (1/1) – 0/1 ln (0/1))
+ 3/4 (- 2/3 ln (2/3) – 1/3 ln (1/3))
= 0.4774

Z_2 = 2/4 (- 2/2 ln (2/2) – 0/1 ln (0/1))
+ 2/4 (- 1/2 ln (1/2) – 1/2 ln (1/2))
= 0.3466

Z_3 = 3/4 (- 3/3 ln (3/3) – 0/3 ln (0/3))
+ 1/4 (- 0/1 ln (0/1) – 1/1 ln (1/1))
= 0

Information gain for each threshold;

Z_1 = 0.5623 – 0.4774 = 0.0849
Z_2 = 0.5623 – 0.3466 = 0.2157
Z_3 = 0.5623 – 0 = 0.5623

From this data, we can see that threshold Z_3 gives the most information gain and is therefore the most optimal split for attribute that defines class group. Again, this is obvious from the data but serves as an illustrative example.

This technique works the same way for much larger and complicated datasets.

We used this method to try and discover some new information, using 480 native proteins as one class ("viable") and our unfolded designs as another ("not viable"). We kept the same factors as used previously in the factorial experiment (surface divergence, buried divergence, secondary structure prediction accuracy and Rosetta energy score) but this time kept each as continuous variables instead of having "good" and "bad" categories. To this, we also added *ab inito* and comparative modelling accuracies for each sequence-structure relationship. Although we have already investigated these variables in some way, keeping the continuous variables should allow a more in-depth look at how they affect the viability of designs. If they didn't offer up any information, more attributes could be added to the investigation to gain more insight. 480 proteins of comparable size to our designs (70 - 130 residues) were taken and analysed as the "viable" set of sequences.

10-fold cross validation was used to verify our J48 method. This means that the data is partitioned into 10 equal subsets, with a single subsample retained as the validation data for testing the model produced by the 9 other training sets. The cross validation is then repeated 9 more times, with each of the subsets used as the validation set once. The performance of the 10 classifiers is then averaged and this is given as the overall model.

6.2.2 A genetic algorithm to direct design

After finding information indicating a difference between our designs and native proteins, we would be able to assign potential new designs into the correct category of "viable" or "not viable". This would be useful, but we still may waste time making a multitude of designs that do not possess the correct level of attributes to be viable. To increase the number of designs that fell into the right category, we used a version of our genetic algorithm

that had a modified objective function. Instead of optimising solely towards predictability, a point is given for each attribute level that is within the range of the "native" sequence scores. Optimisation of the process is then just towards having the most collective points. We used this modified fitness function in both the RosettaDesign (Section 5.2.2) and dynamic (Section 5.2.3) versions of the genetic algorithm to produce native-looking designs for the 3CHY backbone, based on the information we obtained in the previous analysis.

6.2.3 Experimental testing

When realistic designs had been produced, we chose 12 sequences at random that scored the maximum points in the fitness function of the genetic algorithm. We then expressed these in *E. Coli* and screened for solubility and foldedness (protocols in Section 2.2).

6.2.4 Computational testing

At the same time as we were experimentally testing our designed sequences, we also tested our designs computationally. This involved structure prediction methods, both *ab initio* and comparative modelling, as used previously, but molecular dynamics (MD) simulations were also added. By employing MD, we hoped to achieve some insight into the stability and behaviour of the designed sequences that we were testing.

For each designed structure, we began with an equilibration stage that started at a temperature of 10 K and ramped up to 300 K over 20,000 steps. During this time, harmonic restraints were applied to the alpha-carbons with a spring constant of 10 kcal/mol. After equilibration had been completed, we had a production stage of MD, with constraints removed, for 20 ns. We checked the progression of our structure both visually, to see if there were any movements of secondary structures, and looked at total root-mean square deviation (RMSD) over the whole structure to see how

much global change there had been. We performed these experiments using the AMBER[163] (99SB-ILDN[164]) force field.

6.3 Results

6.3.1 Decision tree

In order to try and compare our group of unfolded designs to folded native structures, we applied a J48 machine learning algorithm to construct a decision tree that attempts to categorise different classes based on their defined attributes. The 7 attributes we chose to investigate were buried and surface composition divergences, secondary structure prediction accuracy, Rosetta energy score, *ab initio* prediction scores, and comparative modelling scores. For each of the 480 proteins in each group, the levels of these attributes were calculated and then analysed using the Weka data analytics program.

The decision tree produced by the J48 algorithm indicated that there were some criteria that seemingly led to either "viable" or "not viable" sequences (Figure 43). There are some leaves with mixed classes, such as when buried divergence is less than or equal to 490 and secondary structure prediction accuracy is above 71, which means that we have explored this area already with designed sequences that were not folded. These types of leaves can potentially be separated into more subtrees, but this would involve adding more and more specific criteria for each instance rather than finding a general trend for each class. One leaf stands out from the rest as unexplored by our design process, with a large number of viable native sequences (173) but only 2 designed sequences. It would be interesting to explore this area more, as it appears that we have not fully covered the full range of native sequence properties. This leaf has two criteria that seem intuitive, buried divergence is kept below 490 and Rosetta score is also quite negative. However, secondary structure prediction accuracy being below 71 is something that we would not have expected. It would be safe to assume

that having residues possessing a higher propensity of adopting the correct secondary structures would be a beneficial step in creating viable protein designs. This does not appear to be the case, with sequences scoring higher than 71 classed as "not viable". This is very similar to the pattern we saw in the factorial design experiments in Chapter 4. Also of interest is the lack of prediction scores in the decision making process. Neither *ab initio* nor comparative modelling scores are found as nodes for splitting the two classes, indicating that there is very low information gain in these two attributes. As we have already seen in previous chapters, our designs predict about the same as native sequence-structure relationships and so it is no surprise then that these two attributes do not offer any usefulness to segregate the groups. Surface divergence also only makes an appearance towards the bottom of the tree, and results in quite mixed leaves. This suggests that it may be less important than the preceding nodes, but may still offer some information when looking at more specific cases.

As stated previously, one branch of the tree appears to stand out as densely populated with viable sequences with very few designed sequences present in that region. This could offer a new place to look for viable sequence designs for our proteins, and so we took the criteria outlined by the decision tree and used them as the fitness function for our genetic algorithm.

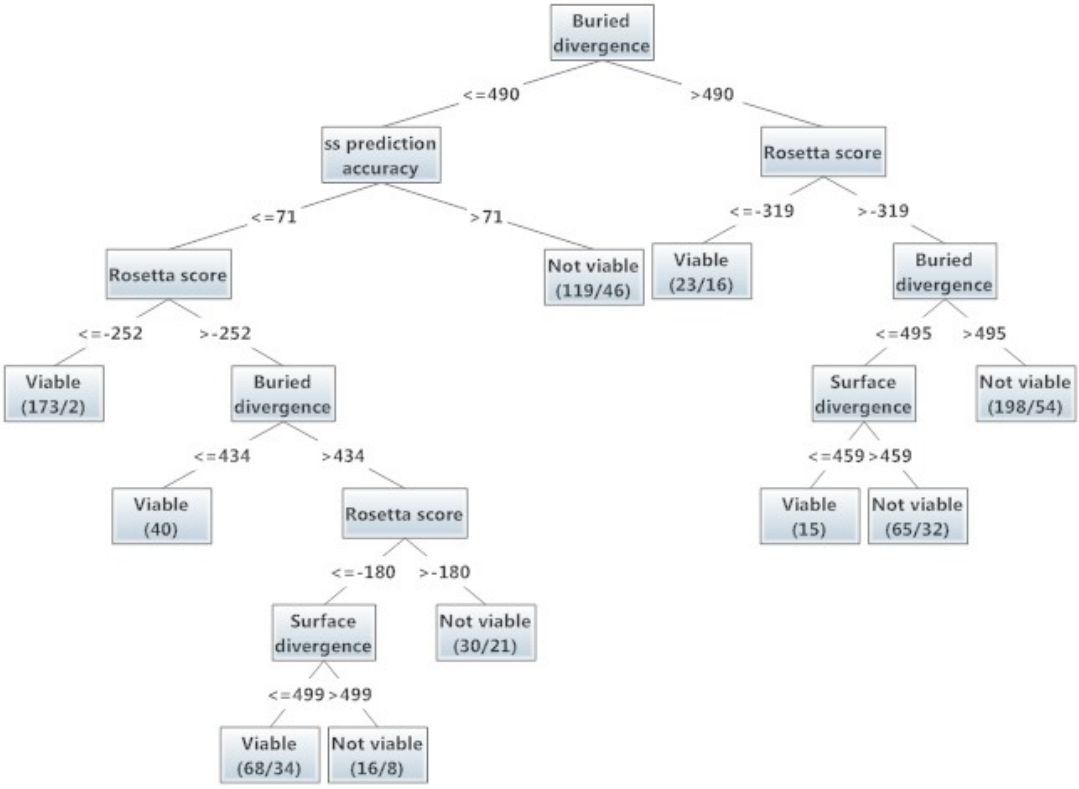

Figure 43. *J48 decision tree showing the various criteria for classifying native "viable" sequences and our "not viable" designs*. There are numerous leaves that result in mixed classes, indicating that we have covered this area of attribute space with our previous designs. The mixed leaves may be split into smaller subtrees, but the rules governing each of these get very specific for each individual sequence. However, there is a leaf at the far left of the tree that is highly enriched in "viable" sequences. With only two designed sequences present in this area, it could be interesting to make more sequences that fit into this category to see if they are folded. The criteria for this leaf are a low buried compositional divergence and a low Rosetta energy score, which are expected. The other criterion is to have a secondary structure prediction accuracy of less than 71, which is slightly counterintuitive. Both *ab initio* and comparative modelling scores do not factor into the decision making process and so there is low information gain between the classes in this respect (i.e. the groups of scores are too similar)

6.3.2 Sequence design outputs

Using machine learning to classify "viable" and "not viable" sequences, we gained some insight into which attribute regions we had explored and which we had not. One of the main discoveries was that there is a large group of "viable" designs with a certain criteria, and we have only produced two designs that fit into this branch. To direct our designs towards this space, we took the decision rules for this branch and used them in our RosettaDesign genetic algorithm fitness function. For each rule that was met, the sequence scored a point. The rules were that the sequence must have a buried composition divergence below 490, a secondary structure prediction score below 71 and a Rosetta energy score below -252. By selecting the sequences scoring the most points within a generation as parents for the next generation, we hoped to produce designs that explored an area we have not explored.

Again, starting the GA with random sequences initially provides very low scores, with the best scoring sequence having only one target rule met (Figure 44). After a few generations, the best sequence tends to score an increased 2 points. By 100 generations, the maximum of 3 points is reached and all of the decision criteria are fulfilled by some designs. Although there does seem to be a slight trend towards an increasing points score, the trajectory of designs appears to contain a lot more noise than previously seen. This is most likely because we are working with a point-scoring binary system. Even small changes may affect the levels of the continuous attributes being optimised towards. If our designs are close to the boundary of the decision rule, changing just a few residues could knock the scoring down for that sequence.

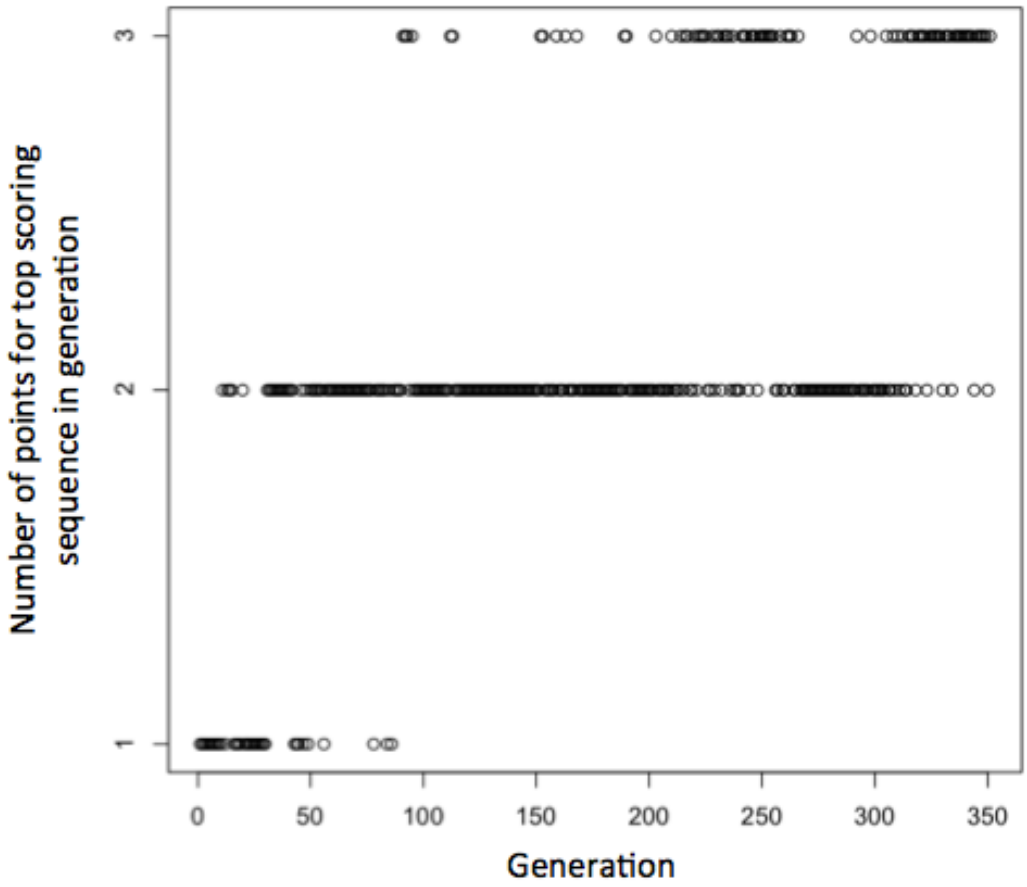

Figure 44. *A genetic algorithm directed towards designing proteins in a specific region of attribute space.* The rules that resulted in a large number of viable sequences in the leaf of the decision tree were used in a points-scoring system for the fitness function of a genetic algorithm. Each time a rule was met by a sequence, a point is scored and the highest scoring sequences are taken through as parents for the next generation. This method started off with a set of random sequences that scored very low, the maximum scoring sequence only meeting one rule. Soon, two rules were met and the sequences started scoring more highly. After 100 generations, sequences meeting all the decision rules were produced. The process appears to be less efficient than our previous GAs, but in this case we are dealing with binary decision boundaries and it is easier to disrupt a previously good score with fewer changes.

Along with a general trend of increasing score with generation times, we can also see that there is an increase in the average points scored for all sequences in a generation (Figure 45). Initially, average score for a sequence in the first generation is hovering around 1. This is to be expected, as they are just randomly generated sequences with no design performed on them yet. The first 100 generations still have a low score, indicating that the

sequences scoring 2 points in this time (in the previous graph) are quite rare. After 100 generations, the average score for sequences in a generation begins increasing, and still appears to do so even after 350 generations. This suggests that we are indeed optimising towards a better score with increasing generations, but at a relatively slow pace.

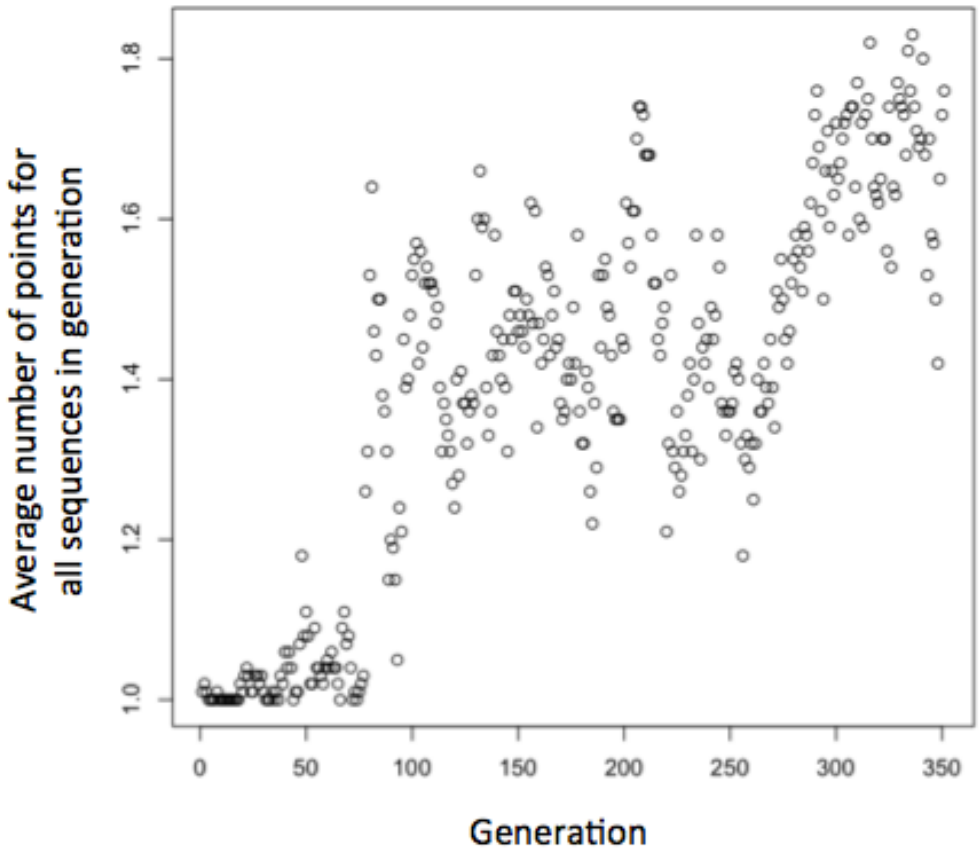

Figure 45. *Average score for each sequence increases with generations.* After starting off with every sequence having a low score for the first 100 generations, the GA appears to result in a generally increasing average score as more generations are produced.

We also thought it a good idea to check how this design method, with the modified fitness function, performed with regards to native sequence recapitulation. As before, the closer we are to reproducing the native sequence (especially core residues), then the better we can assume our designs to be. As generations progressed, the maximum level of native sequence recapitulation increased both in the core and overall (Figure 46). The first 100 generations saw the greatest rate of increase, with both core and total recapitulation reaching much higher levels. At this point, the total

recapitulation appears to reach a plateau and remain around 25-30% for the next 250 generations. Core recapitulation appears to steady at about 40% between generations 100-200, before starting to increase again from 200-350. At the end of 350 generations, the core recapitulation is at ~55% and still looks to be on a slight upward trajectory. Overall, there is a promising similarity to native sequence both in the core and in total.

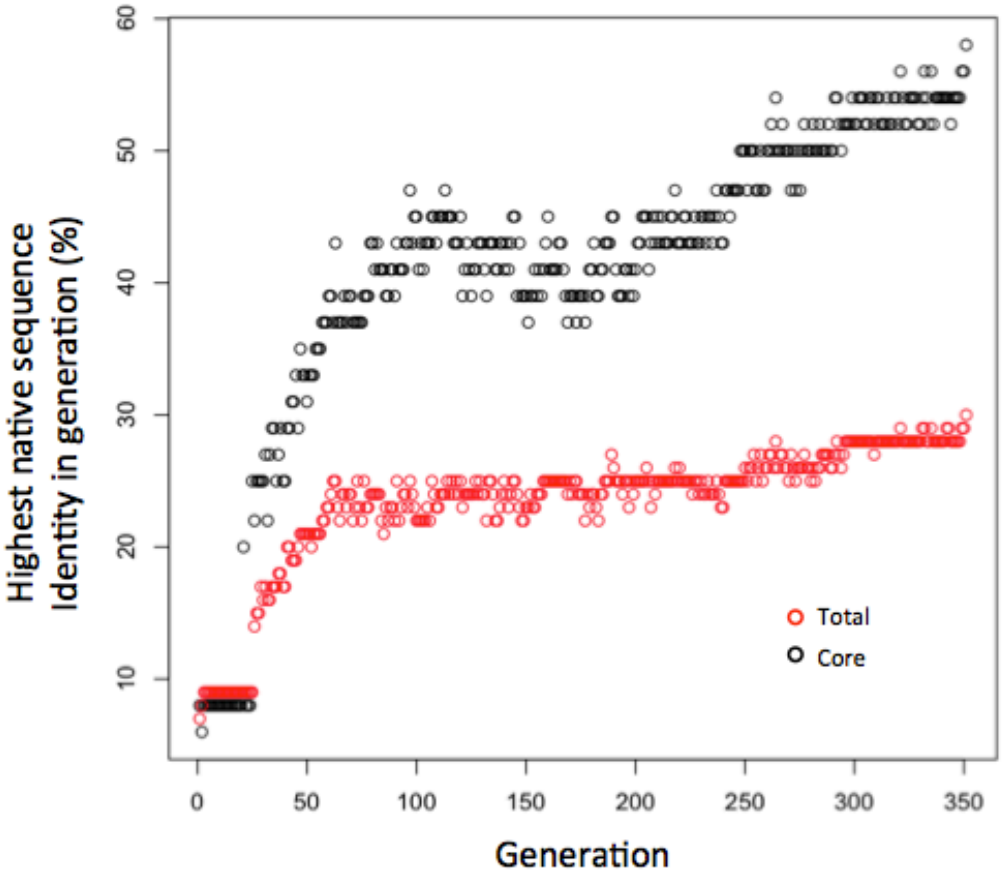

Figure 46. *Maximum sequence recapitulation increases with generations*. Initially there is hardly any similarity between the native sequence and our randomly generated ones. However, in the first 100 generations, a large increase is seen in both core and overall similarity. This then plateaus for total residues, and remains at ~25-30% for the remainder. Core similarity also seems to plateau between generations 100-200 at around 40%, but starts increasing again after 200 generations. By the end of the run, it has reached ~55% and still appears to be increasing.

Our aim was to produce designs that fell into three specific selection criteria outlined by the decision tree analysis, and our process has appeared to do that. From the selection of sequences scoring the maximum number of points, we chose 12 to take forward to both experimental and

computational testing to see what information could be obtained. The 12 we chose were also taken from generations after 250 iterations. We thought this to reasonable, as they also scored highly on native sequence recapitulation.

6.3.3 Structure prediction

We also thought it useful to still check how each sequence fared when the tertiary structure was predicted by both *ab initio* and comparative modelling methods. When *ab initio* prediction was used, TM-scores to the target model were between 0.40 - 0.52 for all of the designed sequences. This is similar to the level that we have seen previously with some of our GA methods. The comparative modelling scored higher than this, with designs having a TM-score between 0.63 - 0.72 (Figure 47). The level of accuracy for these prediction is again very high, with the global fold being recognised but still with minor differences in the exact atomistic details

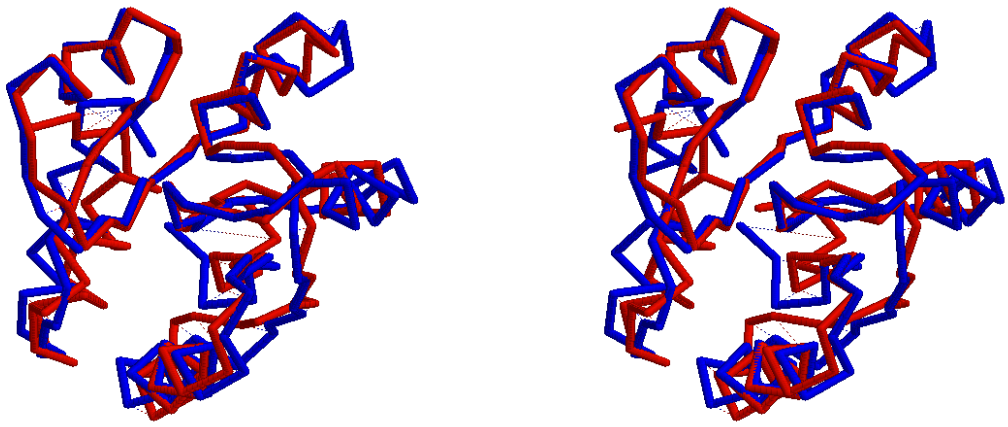

Figure 47. *An example of the prediction accuracy of one of the best designs produced by this method*. The global fold is predicted well through comparative modelling, but there are some shifts in the exact atomistic details. This results in a TM-score to the target template of 0.72.

6.3.4 Molecular Dynamics (MD)

We also performed MD simulations on some of our sequence designs to attempt to understand how stable they would be under normal conditions and to potentially pinpoint any possible problems in the structure. Each designed sequence taken from the GA was threaded onto the 3CHY backbone, and then the structure was relaxed before energy minimization was performed with Rosetta. These threaded structures were then equilibrated with the water molecules, and then simulated for 20 ns as a production stage. During these 20 ns, RMSD from the original structure was taken as a measure of stability. If a large RMSD occurred after a short time, we could assume that the protein would not be very stable. If RMSD remained relatively constant, then we could assume that the structure is quite stable.

As a control, we simulated the native 3CHY sequence and structure for 20 ns under the same conditions as we were using for our designs. This would give us a baseline that we could compare the behaviour of our designs to. The native 3CHY protein seemed very stable under our simulated conditions, with the RMSD from the original structure never reaching above 0.2 Å (Figure 48). The movement of the structure appears to remain constant throughout the length of the simulation, with not much variability. Most of the designs tested unfortunately seemed to have a much higher RMSD than the native. Designs 1, 2, and 3 all seemed to have a large starting RMSD and appear to have increases that suggest some unfolding of the structure. For Design 1, this happens very early, with an increase of 0.2 Å after only 5 ns. Designs 2 and 3 also have an increase at around 10 - 12 ns, but this is of significantly less magnitude than Design 1. Design 4 has a high starting RMSD, but the structure appears to remain around this level for the duration of the simulation. This implies that although our starting structure may not be perfect to begin with (a large ask for any design process) the structure adopted in the simulation is relatively stable and there is no unfolding event.

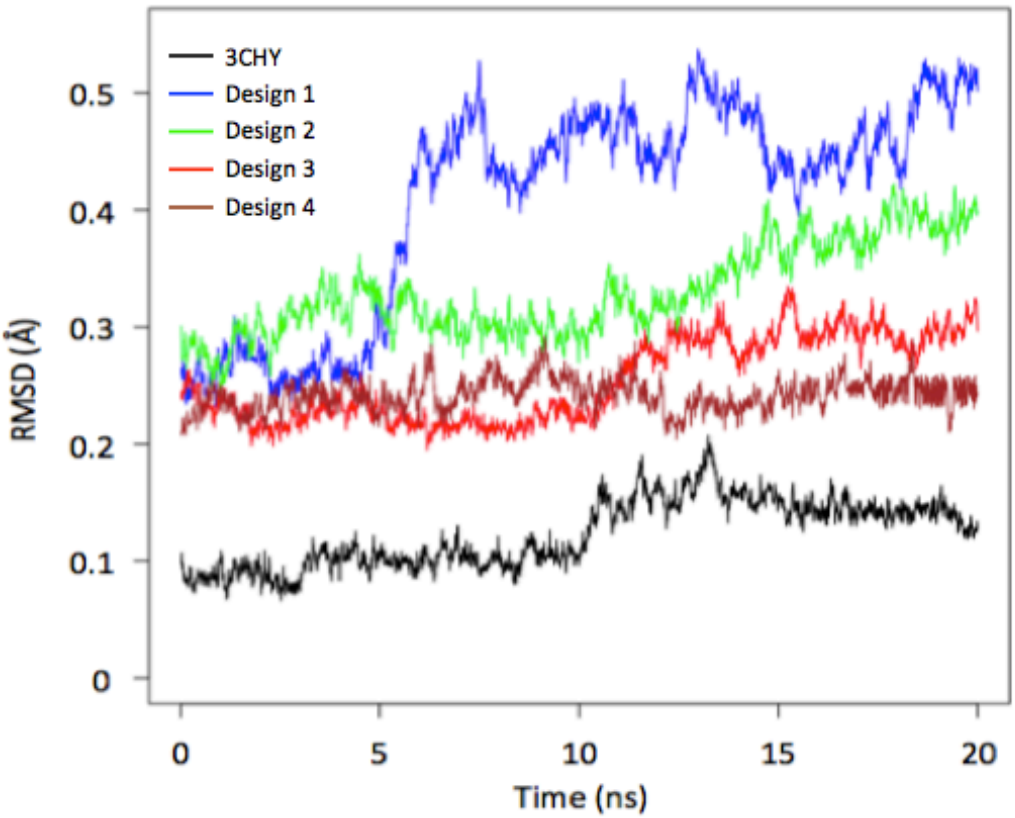

Figure 48. *RMSD from starting structure over 20ns of molecular dynamics simulations for the native 3CHY protein and four of our sequence designs*. The 3CHY protein has a low RMSD from starting structure, indicating that it is stable under our MD simulation. All of our designs seem to have a large RMSD from the starting structure, and most of them have large increases that indicate an unfolding event. For Design 1, this happens at 5 ns and for Designs 2 and 3, this happens between 10 – 12 ns. Although Design 4 has a quite a high RMSD from the starting structure, it doesn't seem to have an unfolding event suggesting that although the starting structure may not be perfect, the structure it adopts under simulated conditions is relatively stable.

The differences between the starting and ending structures for the native 3CHY were hardly noticeable upon visual inspection, with slight differences only visible in some loop regions (Figure 49A). For the designs that show an increase in RMSD during the simulation, there does appear to be an unfolding of one of the terminal helices (Figure 49B). The helix breaks into two, with one part drifting to a position orthogonal to the other. This causes a large instability in the protein, as now the hydrophobic core is exposed and could lead to unfolding. One of our designs doesn't follow this trend and

although RMSD from the starting structure is high, there doesn't seem to be any unfolding and the structure is relatively stable (Figure 49C).

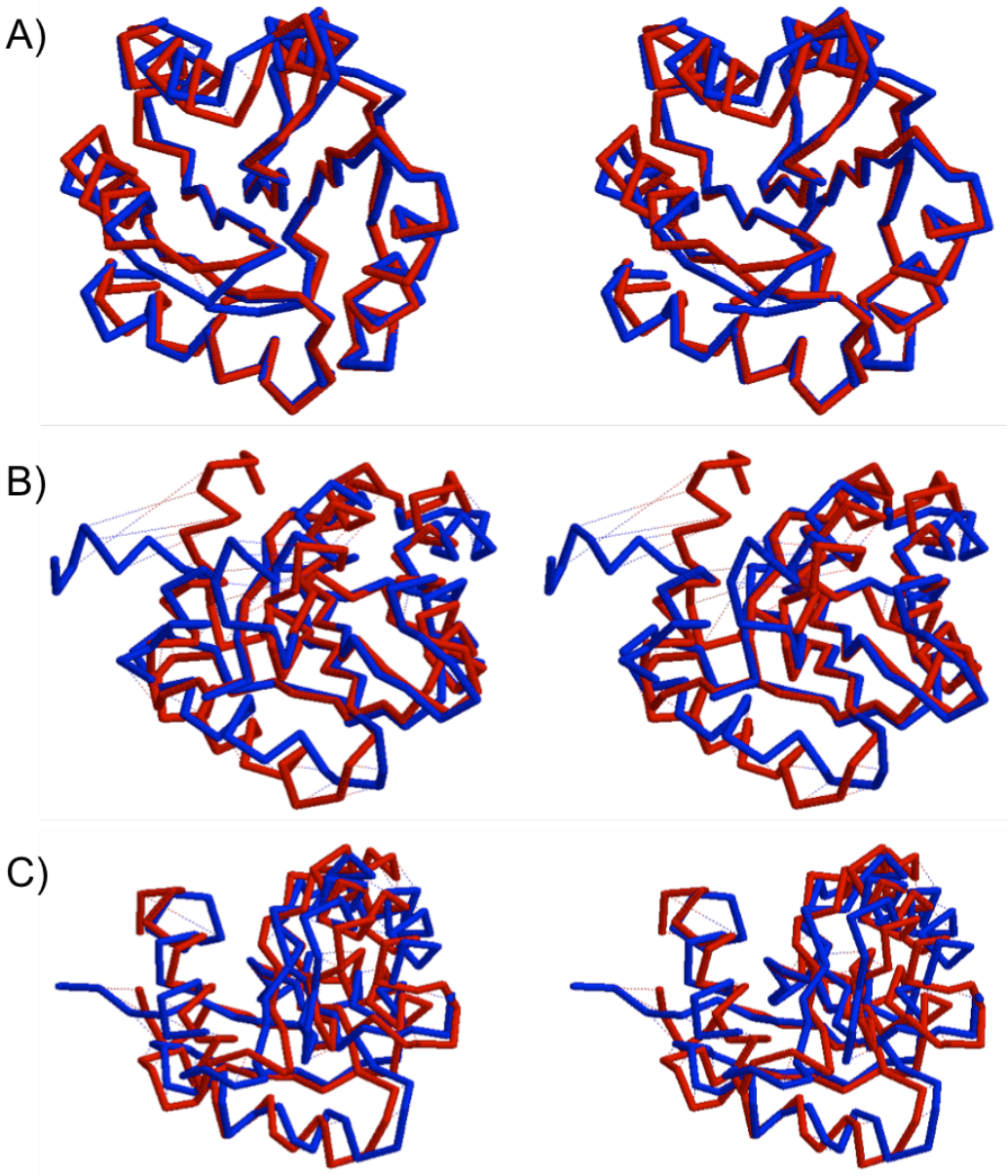

Figure 49. *Stereo-images showing overlap between start (red) and end (blue) structures for native 3CHY and our designs*. A) The native 3CHY shows very little difference in structure after the full simulation. Minor differences are seen between two structures, mainly in the loop regions. B) For most of our designs, there is an unfolding event where the terminal helix breaks and half adopts a position orthogonal to the rest. This leads to instability and can cause unfolding due to the hydrophobic core now being exposed. C) For Design 4, a large difference is seen between the two structures but the global fold is retained. There are no unfolding events, suggesting that even though there is a high RMSD, the structure adopted is a relatively stable one.

We only subjected 4 of the 12 designs to MD simulations, as they were run in parallel with the experimental testing. In the time it took to produce our proteins in the lab and screen then, we only managed a third of the simulations we intended to do.

6.3.5 Experimental testing

The 12 designs we chose from the design method were also taken through to experimental testing in the lab. They were expressed as outlined previously, without the use of a solubility tag and sequence verified externally.

Unfortunately, all of the designs tested did not appear to be soluble. After expression in BL21 cells, cells were lysed and the whole/soluble fractions ran on an SDS-PAGE gel. All of the 12 designs were expressed in the cells and were present as a band of the appropriate size, but were not seen in the soluble fraction (Figure 50).

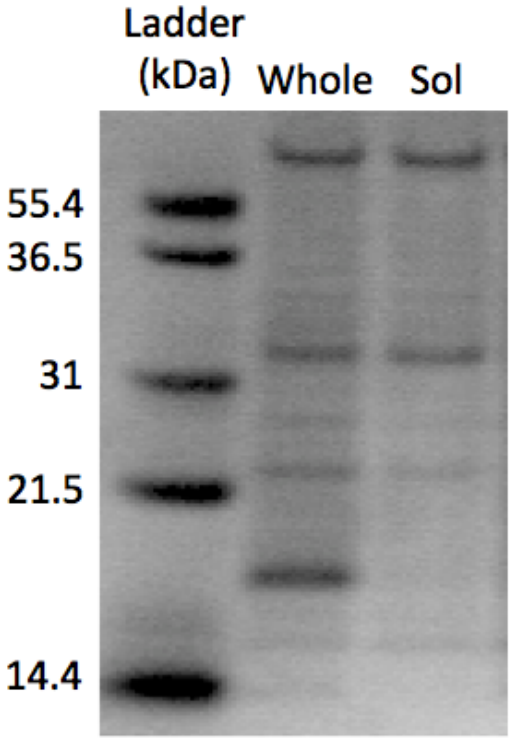

Figure 50. *Example of the level of solubility shown by our designs.* Our designed sequences were present as a band at appropriate size (~17kDa) in the whole cell fraction, indicating that they were being expressed properly. However, no designs were present in the soluble fraction after centrifugation. These results suggest that the designs aggregate and the resulting precipitate is removed in the pellet upon centrifugation.

The fact that none of the designed sequences were soluble suggests that they are not properly folded, so aggregate and precipitate out of solution.

6.4 Discussion

In this chapter, we aimed to find some potential differences between "viable" real sequences and "not viable" designed sequences that we have made previously. We used a J48 machine learning algorithm to investigate 6 different attributes and produce a decision tree to categorise the two classes. Of the various attributes used, both methods of structure prediction (*ab initio* and comparative modelling) did not seem to offer any information gain between classes. This does not mean that predictability is not a necessity for good designs, more that both of the categories were equal in their overall levels of these two attributes. Surface composition divergence also appeared to only have some information in specific instances and did not seem to be a major decision node between classes. This is slightly intuitive, as divergence from a model surface could be less important for sequences. Surface residues are not usually as vitally important as core residues for dictating the fold of a protein and so are more amenable to variation. The three main factors for making distinctions between classes were buried composition divergence, Rosetta energy score and secondary structure prediction accuracy. A particular leaf of the tree appeared to be unexplored by our design processes and so we attempted to direct a genetic algorithm to this area to see if we could produce more viable sequences. This leaf had a low buried sequence divergence and a low Rosetta energy score, which is to be expected. The cores of proteins are usually very well conserved and mostly hydrophobic, meaning that more divergence from a standard model could severely disrupt the fold. Rosetta energy is a relative measure of the stability of the protein so a more negative score correlates with a more stable fold. Unexpectedly, the leaf with a high proportion of viable sequences has a low secondary structure prediction accuracy. We have seen this trend before in the computational analysis of the factorial experiments, where a higher secondary structure prediction accuracy

correlates with a worse predictability score. In this analysis, a similar pattern is seen where a high secondary structure prediction accuracy (>71) branches away from the "viable" designs and produces a leaf containing mainly "not viable" classes. Other groups have suggested that secondary structure propensities are not sufficient for viable protein structures, as the local 3D environment of each residue position is not considered[156]. We have seen that optimising secondary structures had a negative effect on structures in the factorial analysis (Chapter 4), and a similar effect is seen here.

Perhaps one criticism of our investigations in this chapter is that our "not viable" set contains quite a high proportion of the same fold-types, whereas the "viable" natives potentially cover a larger area of fold space. This is true, but the aim of this experiment was to gain some insight into the general rules that could possibly govern viable protein sequences and can be applied in a greater context. After the experiments, we did check to see how the process would have changed if we would have just used 3CHY-like folds for comparisons and the results were strikingly similar (Figure 51). 48 homologues of the 3CHY fold were chosen as the "viable" set and 48 of our re-designs for the 3CHY structure were the "not viable" group. The first two nodes splitting the groups are secondary structure prediction accuracy and buried compositional divergence, both following the same trend of lower levels leading to a leaf of "viable" sequences. Rosetta energy is missing as a decision node, whereas it was present in the process previously. Again, surface divergence seems to potentially have some information but only a small part in the decision making method. Of the new designs we made and then took forward to experimental investigation in this chapter (shown in red), 9 fell into the leaf containing all "viable" sequences. The remaining 3 sequences had slightly higher secondary structure prediction accuracy than 65 and so fell into a slightly different branch of the tree.

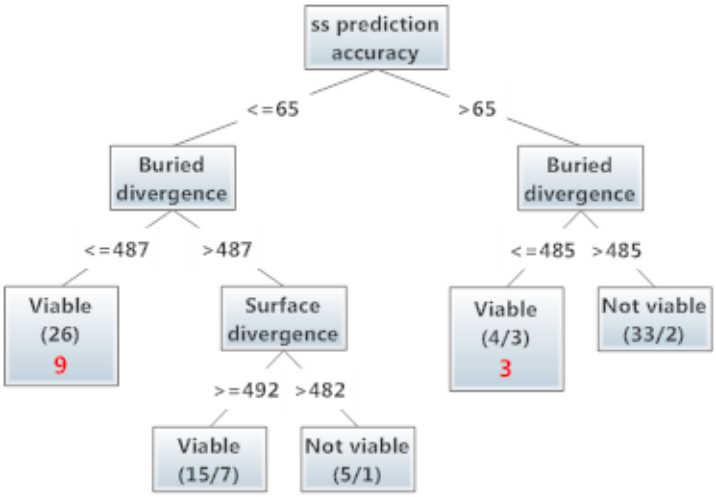

Figure 51. *Smaller, more simple decision tree made by using 48 native 3CHY homologues as "viable" and 48 3CHY sequence redesigns as "not viable".* The pattern seems similar to the general trend in the larger decision tree. Secondary structure prediction accuracy and buried divergence seem to have a large impact on where sequences fall. Rosetta energy is missing from this tree, indicating that there is no information gain between the groups, as are prediction accuracies. There is a leaf containing mainly "viable" sequences present on the far left, and the majority of the sequences we made during this experiment still fall into this category (shown in red)

The molecular dynamics simulations offered up some interesting results. Out of the 4 designed sequences tested, 3 of them suggested that the structures were not stable and had some unfolding at a particular helix each time. This is information that hasn't been captured anywhere else. Structure prediction methods failed to see any problems with this helix and indicated that all secondary structures should be well formed. The other design that was tested had quite a bit of RMSD from the starting structure but the structure that it adopted seemed stable, with no unfolding present. However, although this structure performed well in every aspect, it was still not soluble under lab conditions. We have started looking at MD simulations quite late in the project, and due to time constraints we have not had a chance to investigate this aspect fully. The results gained from just 4 simulations hint that there could be a lot more to look at, and given more time it could be valuable to pursue this further.

During this experiment, we were still having problems producing soluble proteins for further investigation. If our designs are falling at the first hurdle and precipitating when screened for solubility, there is not much else afterwards that we can examine. To overcome this problem, a solubility tag could be used again when expressing our designs in the bacterial cells. Solubility tags are useful for increasing solubility for proteins that can be difficult to produce under normal conditions. The use of tags was dropped in the previous experiments because they gave a high level of false-positive results, leading to a lot of wasted time chasing up designs in large screening processes. Now that our experimental testing sets are reduced in number, we could explore our potential designs in more depth. Adding solubility tags or screening different expression conditions to try and increase the amount of information gained from each experiment is much more tractable if we have smaller sets of proteins to test. In the large-scale batches containing 48+ proteins, screening would have been very time consuming for potentially little gain but now these types of investigations would be easier. Unfortunately, due to time constraints on the project, we had to end our investigations here without pursuing more data from proteins under different conditions.

Acknowledgements

Thanks to Enrico Spiga for help with the molecular dynamics simulations.

7. Design of a Leucine-Rich Repeat and production by peptide synthesis

7.1 Introduction

In the past 50 years, advances in chemical peptide synthesis have meant that its use is now becoming more commonplace in biological research and drug development[165]. Besides being able to produce peptides found in biological systems, chemical peptide synthesis allows a high degree of freedom in the types of molecules produced. At a time when synthetic biology is realizing the potential for introducing non-canonical amino acids into a system[166–168], peptide chemistry offers a means to introduce new modifications much more easily than modification of the genetic code[169]. In order to begin to understand the potential benefits of applying these methods to synthetic biology and protein design, we attempted to make a single 28-residue Leucine-Rich Repeat (LRR) segment that we had designed.

LRRs, like other repeat proteins, are composed of highly regular consecutive structural units that stack to form elongated domains[170]. The structural motif of LRRs is highly conserved across the whole family, principally containing a β-strand followed by a helix lying in an antiparallel fashion[171] (Figure 52A). These individual units, ranging from 20 – 30 residues in length, stack together to form an elongated domain with a continuous hydrophobic core and large interaction surfaces[172] (Figure 52B). The large interaction surfaces provided by repeat proteins has led nature to adopt them as scaffolds to support a wide range of protein-protein interactions[173]. In particular, the innate immune system relies heavily on LRRs are the primary method for target recognition[174].

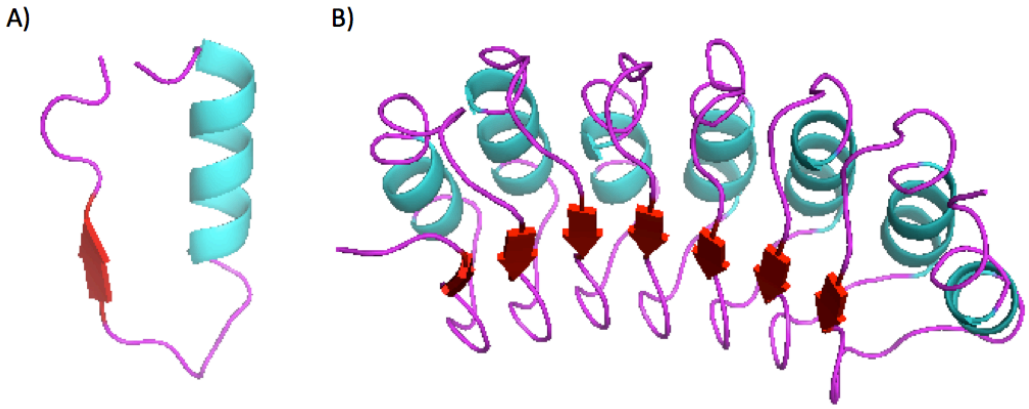

Figure 52. *Leucine-Rich Repeat (LRR) protein structures.* A) An individual unit of an LRR protein is somewhere between 20-30 residues and contains a β-strand, followed by a helix linked in an antiparallel fashion. B) These individual units link together to form an extended domain with a hydrophobic core and large binding surfaces.

Because of their intrinsic modularity, high stability and potential for interface engineering, repeat proteins have attracted much attention as alternative scaffolds to antibodies[175,176] and as building blocks of protein nanomaterials[177,178].

There have already been multiple efforts to design repeat proteins with the main emphasis on creating consensus repeat proteins[179–181], varying the number of repeat modules[182], or fusing naturally occurring repeat proteins together[183]. So far, nobody has managed to produce a single LRR repeat unit that can be linked together to form a stable protein without the use of terminal capping repeats. It is our aim in this chapter to condense the LRR sequences of the ribonuclease inhibitor (RI) protein into a single repeat, use Rosetta to design the remaining non-consensus residues, and then produce the resulting peptide using chemical synthesis. There has already been some success redesigning the RI protein using consensus sequences, but they rely on having two different repeats linked together in an alternating manner[171]. We hope to improve on this published data by taking it a step further and condensing to a single repeat.

Because of the modularity of repeat proteins, we see chemical peptide synthesis as an ideal route to produce the single repeats before attempting to link them together.

7.2 Methods

7.2.1 Design of a single LRR unit

The ribonuclease inhibitor (RI) protein is an LRR protein, which has 16 individual 28-mer repeat units stacked together to form a horseshoe-shaped domain (Figure 53). Each unit varies in exact sequence, but striking similarities exist between each one.

Figure 53. *The ribonuclease inhibitor protein structure.* 16 individual LRR units are linked together to form a horseshoe domain. The sequences for each repeat are slightly different, but there is a high degree of similarity between all of them.

We began by decomposing all individual unit repeat sequences into the most conserved residues, along with maintaining all positions where Leu was present (Figure 54). Upon reaching a general sequence across all RI repeats, we then compared this sequence to the alternating repeats that had previously been successful[171]. Any residues in consensus were kept in our sequence, with the remaining positions being redesigned using Rosetta. 3D models were constructed for the design stages by taking one of the RI LRRs and copying the structure to form a 7-mer repeat. Onto this 7-mer scaffold, the sequence was threaded onto each repeat. Identical positions for each repeat were designed simultaneously so whenever a residue changed in one repeat, the equivalent residue changed in each of the other repeats.

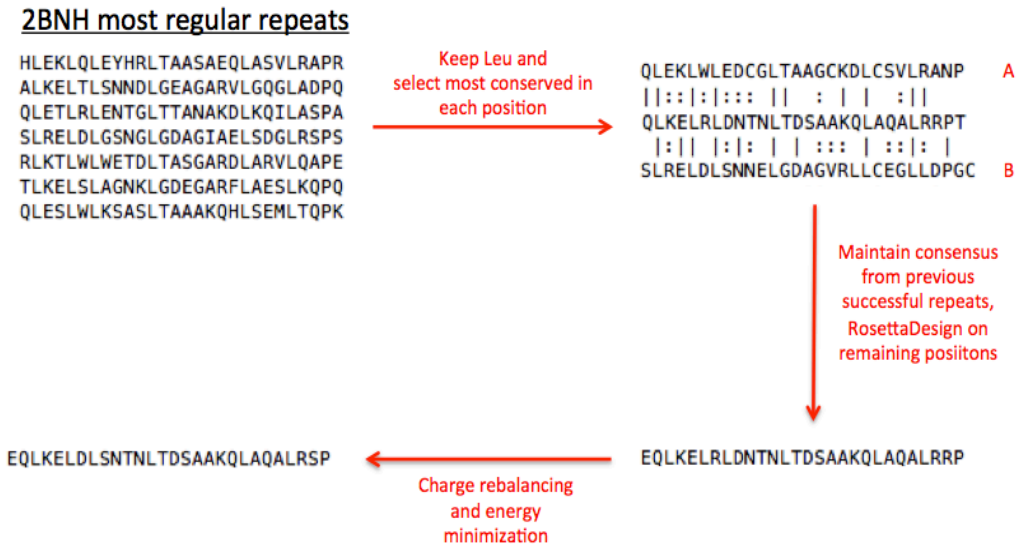

Figure 54. *Design workflow for the single LRR repeat.* We began by decomposing the most regular repeats from the ribonuclease inhibitor protein into a consensus sequence. This was then compared to the previously designed "A" and "B" units[171], and the residues in agreement were maintained. Positions not in agreement were left open to redesign by Rosetta. The resulting sequence underwent compositional analysis, following which charges were rebalanced and energy minimization performed to produce the final sequence for production in the lab.

The same compositional analysis used previously was performed on the resulting sequence, where amino acids are grouped based on their properties (Positive = Y,W,H,R,K,N,Q; Negative = E,D,S,T,G,P,C; Hydrophobic = A,V,I,L,M,F). The total composition of our sequence was compared against a set of 1000 native proteins under 100 residues, and we found that our LRR was on the border of realistic sequences (Figure 55). To address this, we performed charge rebalancing by modifying previous non-consensus positions. The final sequence then underwent energy minimization through Rosetta, before being taken forward to synthesis in the lab.

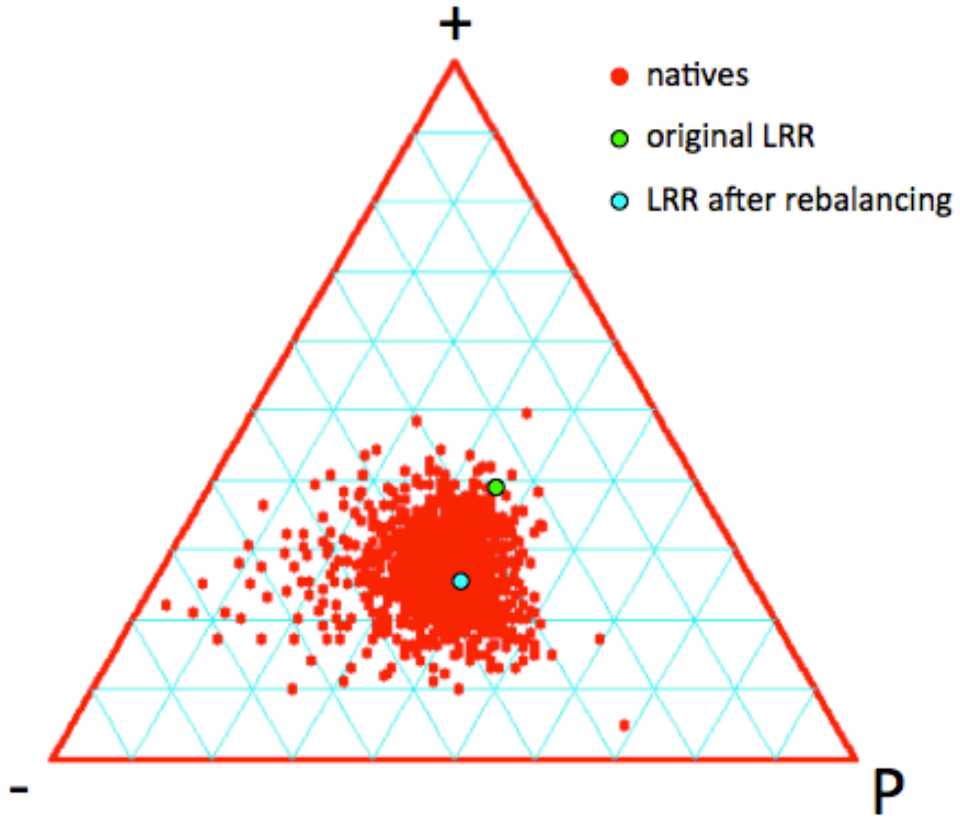

Figure 55. *Compositional analysis of LRR designs compared to native sequences.* Total compositional analysis revealed that our original LRR design (green) was on the boundary of native-like sequences (red). We modified some non-conserved residues in order to rebalance the charges, and the resulting sequence was much closer to realistic sequences (blue). Classification of residues; "P" = hydrophobic, "+" = positive, "-" = negative.

7.2.2 Solid phase peptide synthesis (SPPS)

Solid phase peptide synthesis was first introduced by Merrifield in the early 1960s[184], and numerous developments since then have pushed the method to the forefront of synthetic peptide production[165]. The C-terminus of a peptide is linked to a solid polymeric support via a linker, and synthesis proceeds through the sequential addition of amino acids in a C- to N-terminal manner. Amino acids have protected sidechains, to prevent side-reactions, along with initially protected α–amine groups. For each new cycle of residue addition, the protection of the N-terminal amine is removed and a new fully protected residue is coupled. When the full sequence has been synthesized in this progressive manner, the full peptide is cleaved from the resin and can be fully deprotected[185] (Figure 56). We chose to use an

automated Fmoc/tBu protocol that has previously been established and for a greater understanding of the exact chemistry, the relevant literature can be consulted[165,185,186].

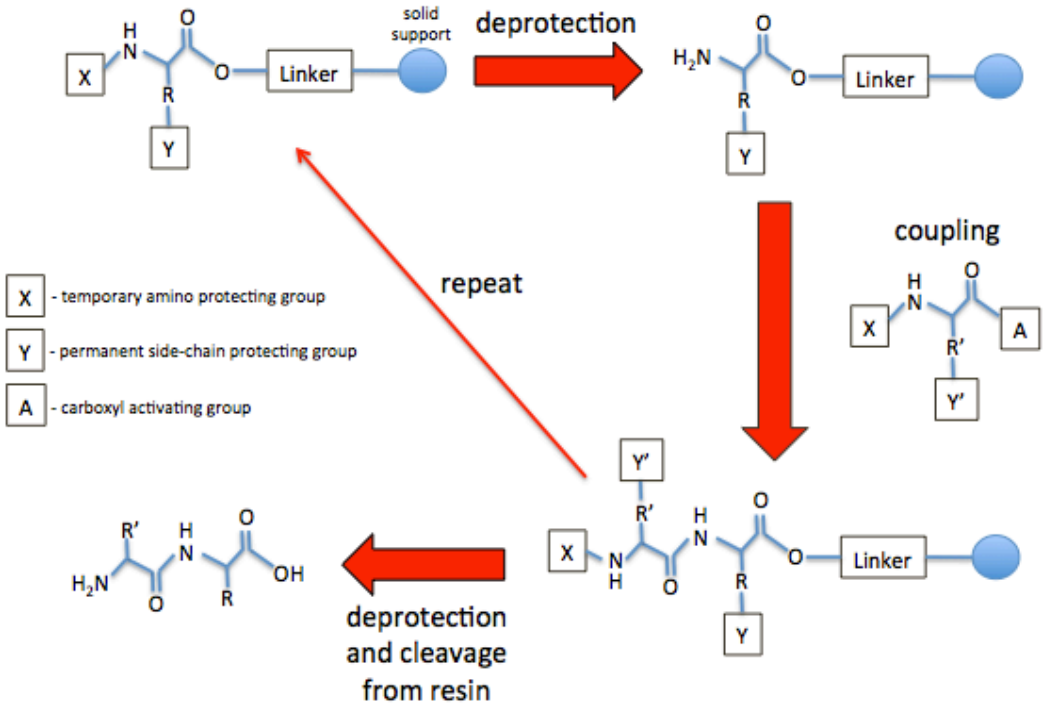

Figure 56. *Basic methodology of solid-phase peptide synthesis.* The first residue is coupled to a polymeric support via a linker region. Residues begin by having two different protecting groups for the α–amine and the sidechain. The N-terminal amine is deprotected, allowing addition of a second fully protected residue and the peptide grows in size. Rounds of deprotection and coupling occur until the full peptide segment is synthesized, after which the peptide is cleaved from the resin and sidechains are deprotected.

Even with protecting groups on the residues, problems can still occur with the synthesis of a full peptide. In a perfect scenario, the peptide and polymer chains would be fully solvated as extended chains (Figure 57A). However, multiple problems can occur that result in an incomplete synthesis. Among the possibilities for bad scenarios: the peptide chains could be intramolecularly aggregated while the polymer is solvated (Figure 57B), the peptide chains could be well solvated whilst the polymer backbone is poorly solvated (Figure 57C), or the peptide chains could be intermolecularly aggregated but the polymer solvated (Figure 57D). All of these cases are less mobile and accessible than the optimal solution and so

reduce reaction rates or lead to truncation of the peptide. The intrinsic properties of the peptide sequence itself govern which scenario will take place for a given "difficult peptide".

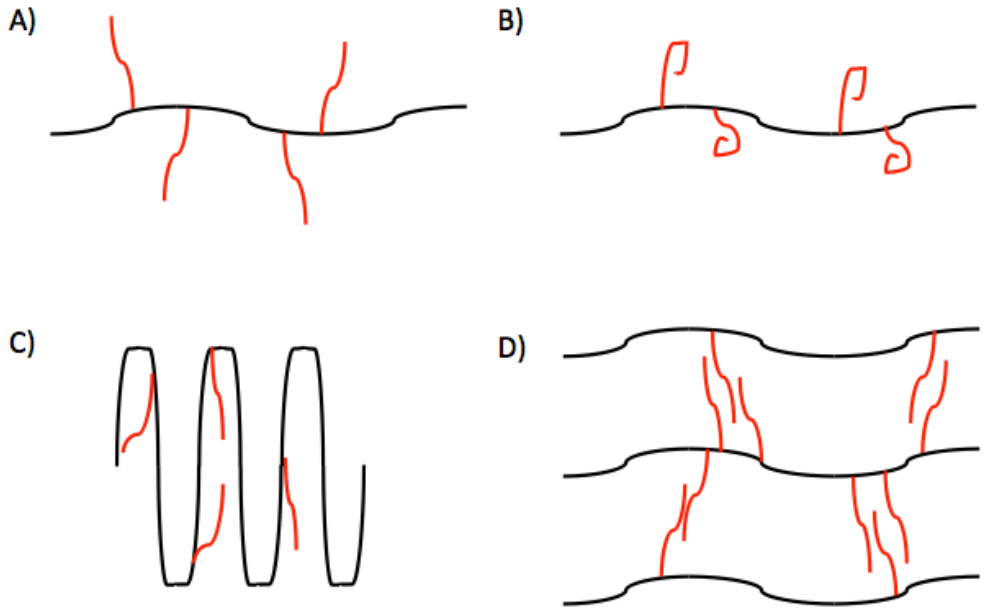

Figure 57. *Potential scenarios during peptide synthesis.* A) The perfect scenario, when peptide (red) and polymer (black) chains are fully solvated. B) Peptide chains are intra-molecularly aggregated, polymer backbone is solvated. C) Peptide chains are solvated, polmer backbone poorly solvated. D) Peptide chains are inter-molecularly aggregated, polymer is solvated. All scenarios except for "A" are less accessible and mobile, resulting in lower reaction rates or incomplete synthesis.

The major contributor to aggregation of "difficult peptides" is inter-chain backbone hydrogen bonding between growing peptide chains (Figure 57D)[187]. Backbone protection techniques have arisen to combat this type of aggregation, removing potential hydrogen bonding by temporary protection of the secondary amide bond[188]. The introduction of proline into aggregating sequences has been shown to prevent inter-chain associations by removing backbone hydrogen bonds[189]. This has led to commercially available 'pseudoprolines' (Figure 58A), which allow the synthesis of previously intractable peptides[190]. These can be introduced with great convenience as dipeptide building blocks, but are limited to positions containing X-Ser or X-Thr[191]. Another form of backbone protection is through addition of a Hmsb auxiliary (2-hydroxy-4-methoxy-5-methylsulfinyl benzl, Figure 58B). This again offers temporary protection of

the secondary amide bond to reduce association of peptides during synthesis, but has less specific circumstances for use compared to pseudoprolines[186].

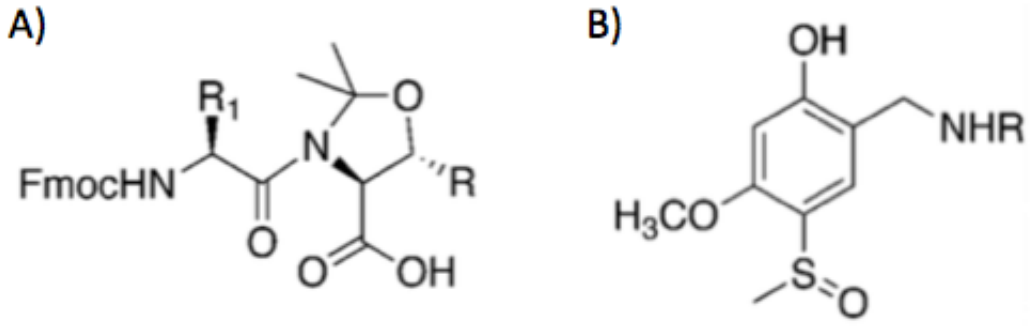

Figure 58. *The structures of two auxiliaries used in backbone protection methods.* A) Pseudoprolines can be used in some limited circumstances. R = H, Ser; R = CH_3, Thr. B) 2-hydroxy-4-methoxy-5-methylsufinyl benzyl (Hmsb) can be used in a wider variety of situations.

The efficiency of synthesis can be assessed by analytical reverse-phased high-pressure liquid chromatography (RP-HPLC) on a reversed-phase column, followed by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (MS).

7.2.3 Peptide synthesis (automated protocol)

Standardized automated protocols were used for Fmoc/tBu peptide synthesis[186]. Peptides were prepared using HCTU/DIEA (CS Bio 336 automated synthesizer) activation for Fmoc/tBu chemistry. A Rink amide resin was used as the solid support for the synthesis (100 mg, 72 μmol). Fmoc deprotection was performed using 20% piperidine in DMF in two stages with an initial 3 min followed by 7 min deprotection. All coupling reactions were performed with 10-fold excess Fmoc-Amino acid-OH/HCTU/DIEA for 30 min. The backbone protecting group was introduced in two steps. The salicylaldehyde dissolved in DMF (0.01 M) was added to peptide-resin for two 30 min cycles, followed by DMF wash cycle. This gave a strong yellow colour to the resin indicating imine formation. Reduction with $NaBH_4$ dissolved in DMF (0.1 M, filtered w/ PVDF 0.2 μm) was then

performed by two 15-min-cycles followed by thorough DMF wash. The next amino acid following insertion of the backbone protecting group was coupled to the alkylated amino group, as previously, and followed by a 10 min DCM wash, 1 hour shaking in DCM, 10 min DMF wash, and 30 min shaking in DMF.

All Fmoc amino acids and pseudoprolines used in these experiments were sourced from *Merck Millipore, UK*.

7.2.4 Peptide characterization (HPLC and MS)

Peptides were prepared as outlined in the previous section. After synthesis, cleavage was performed by addition of TFA/TMSBr/EDT (10:1:0.25) for 1 hour. Following peptide cleavage, the TFA-cleavage cocktail was filtered, filtrate sparged (nitrogen) and peptides were precipitated with Et_2O (Na-dried, 4 ^oC) before being freeze-dried. Peptides were purified by semi-preparative HPLC on a RP-C18 column (22 x 250 mm, Vydac) using linear gradients of CH_3CN in 0.1 % TFA/H_2O with a flow rate of 10 $mL.min^{-1}$. HPLC gradients were prepared using solvent A (0.1% TFA in H_2O) and solvent B (90 % CH_3CN in 0.1 % TFA). Detection was performed at 214 nm. Peptides were characterized by MALDI-TOF MS on a BRUKER microflex (ion positive linear and reflector mode) using CHCA matrix (10 $mg.mL^{-1}$ in CH_3CN/H_2O/TFA, 50:50:0.1).

7.2.5 Circular Dichroism (CD)

CD measurements were performed using a Jasco J-715 spectropolarimeter fitted with a cell holder thermostatted by a PTC 348-WI Peltier unit. Far-UV spectra were recorded at 20 ^oC over a range of λ 195 – 235 nm, using 1 mm fused silica cuvettes (Hellma) with peptide concentrations of 100/250/500 μM diluted in 10 mM potassium phosphate buffer. Spectra were typically recorded with 0.2 nm resolution and baseline corrected by subtraction of the appropriate buffer spectrum. The results were reported as mean residue ellipticity, calculated using the Jasco Spectra Analysis Tool.

Thermal denaturation experiments were performed at a fixed wavelength of 221 nm in a 2 mm fused silica cuvette (Hellma) with a 250 μM protein sample, again diluted in 10 mM potassium phosphate. The temperature ranged from 2 – 70 ^{o}C and was controlled by a Jasco PTC-348WI Peltier system. Data were converted to mean residue ellipticity, again using Jasco Spectral Analysis Tool.

CD specta prediction was obtained using the DichroCalc web tool, which can give an estimate of the CD spectra expected from a given PDB file. For the in-depth methodology, see relevant literature[192].

7.2.6 Fragment condensation

For large peptide chains, fully protected peptide segments can be coupled together to make a chain larger than can be produced by SPPS. This is known as convergent solid-phase peptide synthesis (CSPPS). Firstly, the fully protected segments are synthesized by normal SPPS. After cleavage from the resin, the segments are kept fully protected (apart from the C-terminal where the linker used to be) to avoid any off-target reactions in the subsequent coupling stages[193]. The fully protected segments are isolated and characterised, usually by RP-HPLC and MALDI-MS to ensure they have a high purity for the coupling stages. Segment condensation then begins with a single peptide chain bound to a coupling resin, to which further segments are added using a number of amide bond forming methods[194]. Solubility of the protected peptides plays a vital role in enabling characterisation through RP-HPLC, along with having a major impact on how efficient the coupling reaction will be. Diffusion of reactants into the support matrix is essential, meaning solubility of a fragment is key but unfortunately not consistently predictable for any designed peptide[195].

When we have produced our peptide and characterised a single repeat, it is our intention to link these individual units together using CSPPS methods and produce a full repeat protein.

7.3 Results

7.3.1 LRR synthesis

To check if our design was one of these "difficult peptides", we first attempted to synthesise the full sequence with traditional Fmoc SPPS methodologies and tBu protecting groups[185]. After synthesis, we cleaved the peptide from the resin and deprotected the sidechains using TFA. We then characterised purity by HPLC (Figure 59A) before analysing fractions with MALDI-MS (Figures 59B, 59C).

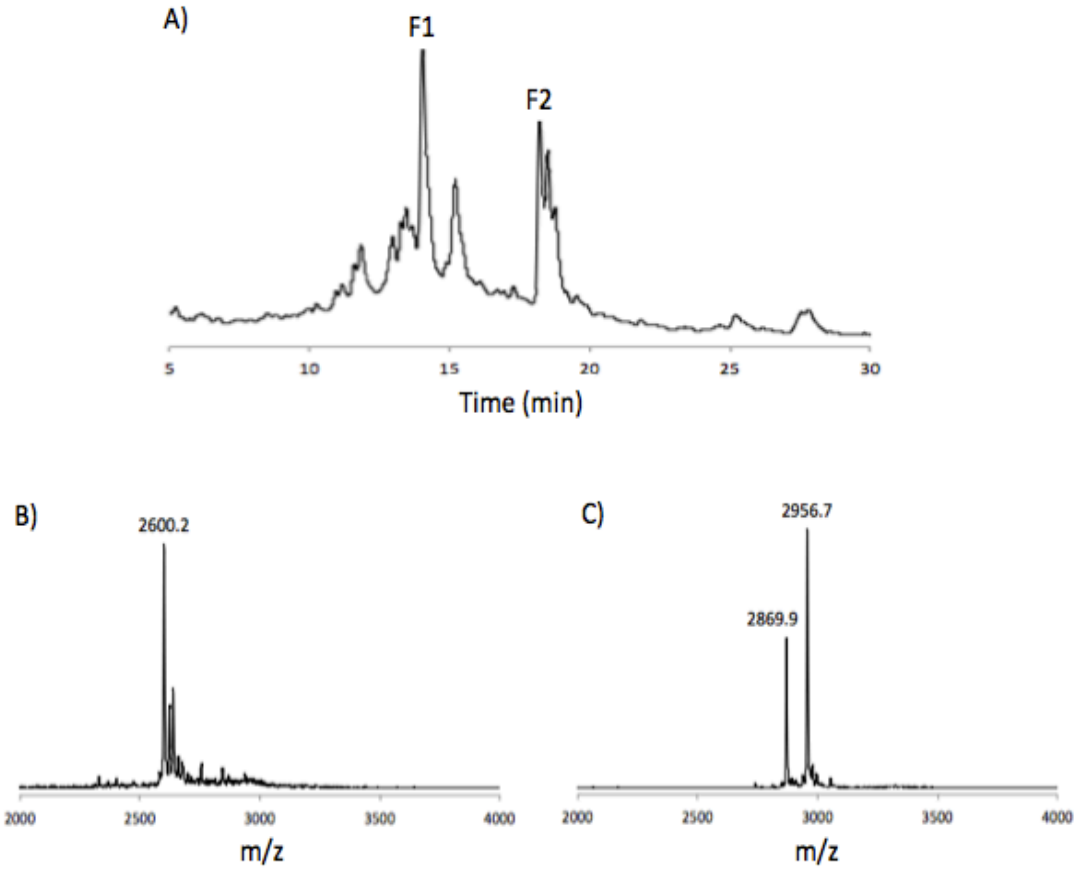

Figure 59. *Synthesis of LRR using standard Fmoc/tBu solid phase peptide synthesis techniques*. A) Analytical HPLC traces of the crude product indicated that there were two main species of peptide present in the sample. Mass spectrometry of F1 (B) and F2 (C) indicated that neither fraction contained the peptide we were attempting to make. The calculated mass of the desired peptide was 3054.6 m/z in the protonated form, and 3077.6 in the presence of Na^{2+}.

Using standard SPPS techniques did not appear to be sufficient to produce our designed LRR, with analytical HPLC and MS suggesting that incomplete synthesis was a problem. In an attempt to combat this, we used a Hmsb auxiliary towards the N-terminal end of the peptide (Figure 60A). The crude product contained 4 four different species of peptide (Figure 60B). Calculated mass of the peptide with H^+ is 3054.6 m/z, with Na^+ is 3077.6 m/z. With the backbone protection group still present, masses change to 3252.6m/z [H^+] and 3274.6 m/z [Na^+]. MS analysis showed that the peptide by itself was present (Figure 60F), along with a species where the backbone protecting group had not been cleaved off (Figure 60E). However, there were other incomplete species present (Figure 60C + 60D) and the crude was not very pure.

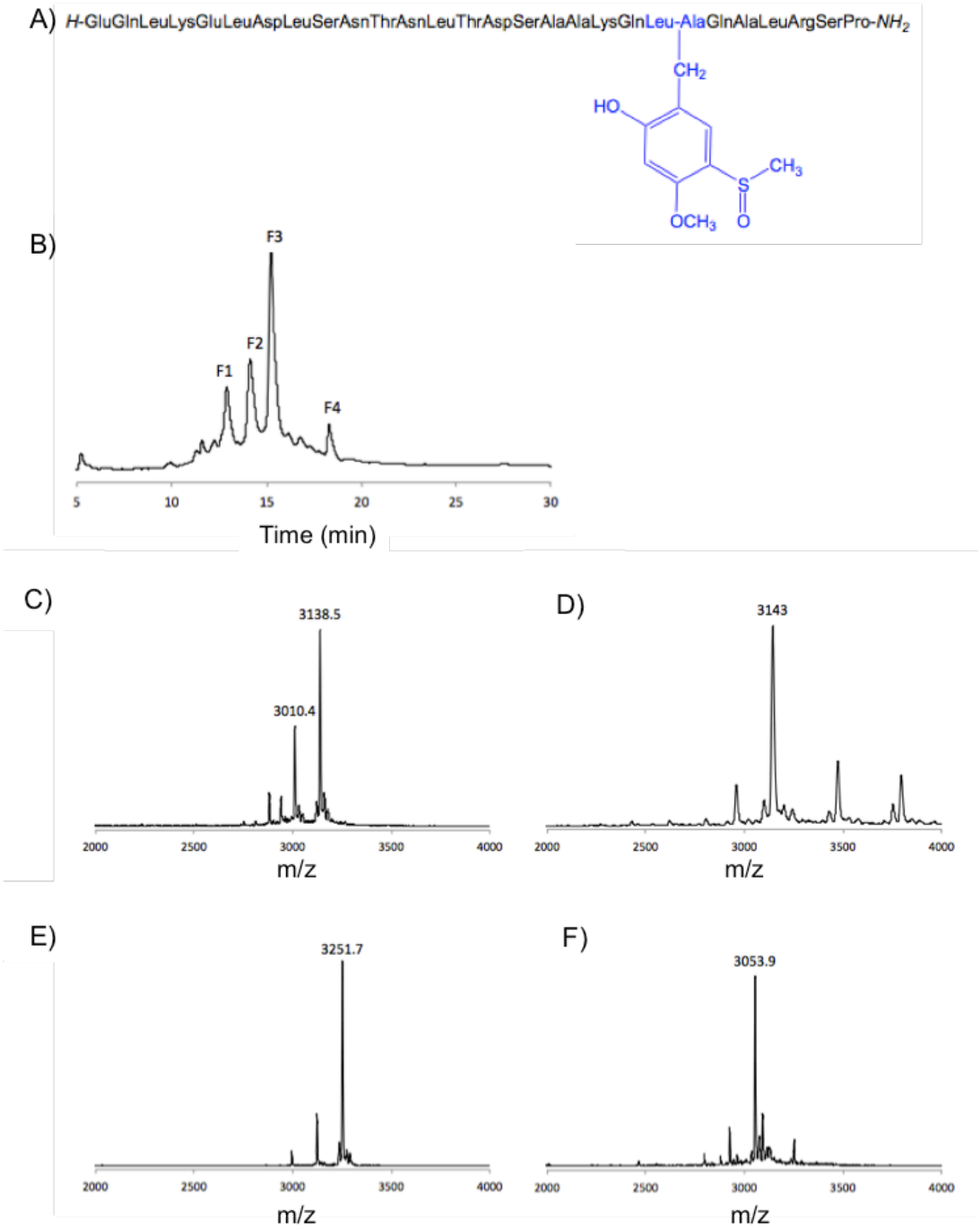

Figure 60. *Synthesis of LRR using a single Hmsb backbone protection group.* A) Insertion site for the backbone protecting group. B) Analytical HPLC of crude product shows four species of peptide. The peptide (F) was present in peak F1, and also with the Hmsb group still present in F2 (E). However, the crude wasn't pure and there were other species present in peaks F3 and F4 (C + D). The calculated mass of the desired peptide was 3054.6 m/z [H^+], and 3077.6 m/z [Na^{2+}]. With the backbone protection group still present, mass was calculated at 3252.6 m/z [H^+], 3274.6 m/z [Na^{2+}]. Peptide was cleaved from the resin with TFA/TMSBr/EDT (10:1:0.25)

There were improvements in the synthesis when using a single Hmsb backbone protection group, but there is still a lot of room for improvement to increase purity. To offer further backbone protection, and hopefully a more pure synthesis, we added two pseudoproline residues to our already Hmsb-protected peptide (Figure 61A). By adding these pseudoprolines, we protect a large area of the peptide from forming hydrogen bonding and therefore synthesis is more pure. Under milder cleavage conditions, two species are seen by HPLC (Figure 61B) and pure segment is seen under slightly different cleavage conditions (Figure 61C). Subsequent MS of each fraction shows that F1 contains the peptide with the Hmsb group (Figure 61D) and F2 is the peptide by itself (Figure 61E).

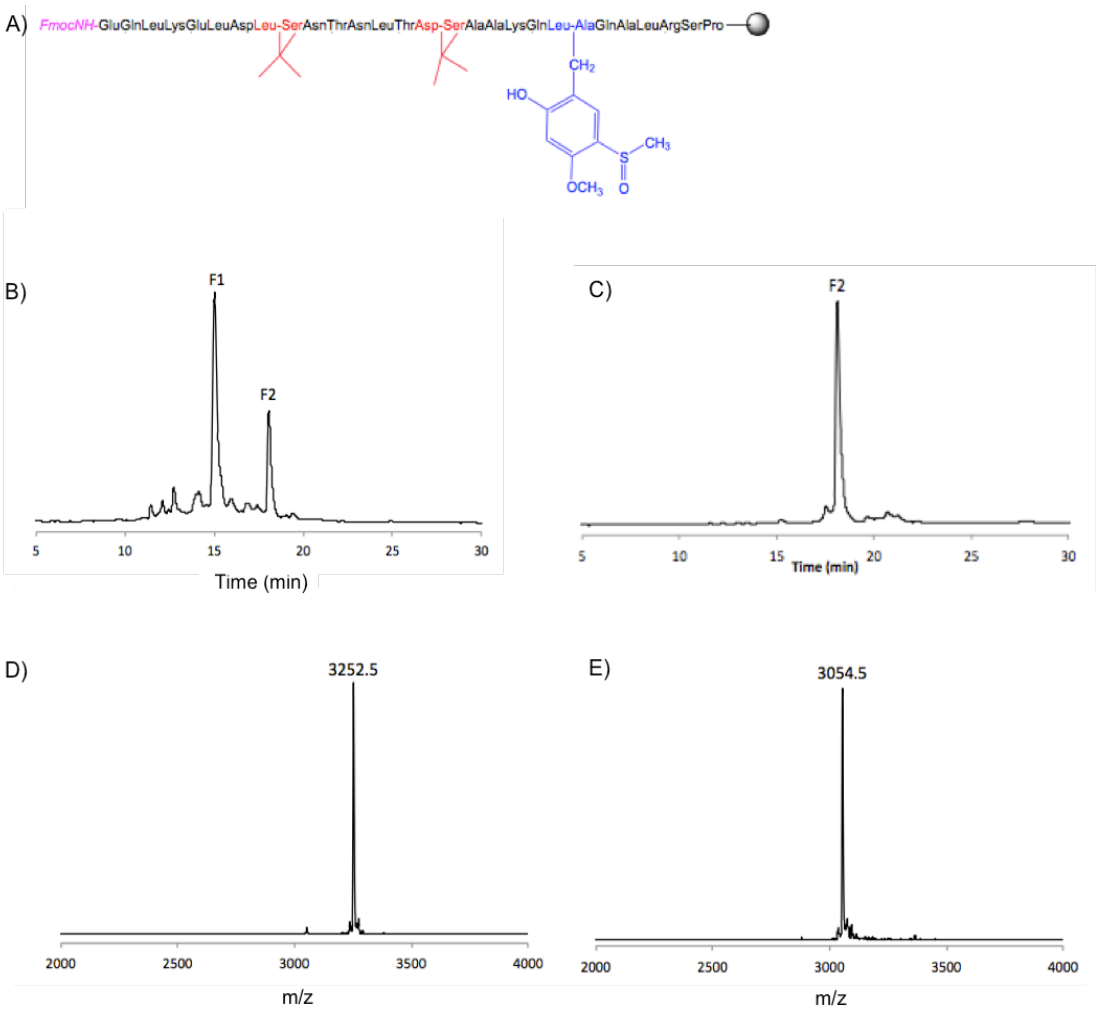

Figure 61. *Synthesis of LRR using pseudoprolines and Hmsb backbone protection groups*. A) Locations of backbone protection groups (red = pseudoproline, blue = Hmsb). B) Analytical HPLC of crude product shows two species of peptide. Peptide was cleaved with TFA/TES/H_20 (10:5:0.5) C) HPLC of cleavage under different conditions shows a pure species. TFA/TMSBr/TA/EDT (10:1:0.5:0.25) D) F1 contains the peptide with the Hmsb protecting group. E) F2 contains the unprotected peptide. Calculated mass of the desired peptide was 3054.6 m/z [H^+], and 3077.6 m/z [Na^{2+}]. With the backbone protection group still present, mass was calculated at 3252.6 m/z [H^+], 3274.6 m/z [Na^{2+}]

7.3.2 Structural information from a single repeat

Now that we had made a pure sample of our peptide, we characterised any potential secondary structure of the monomer by circular dichroism (CD).

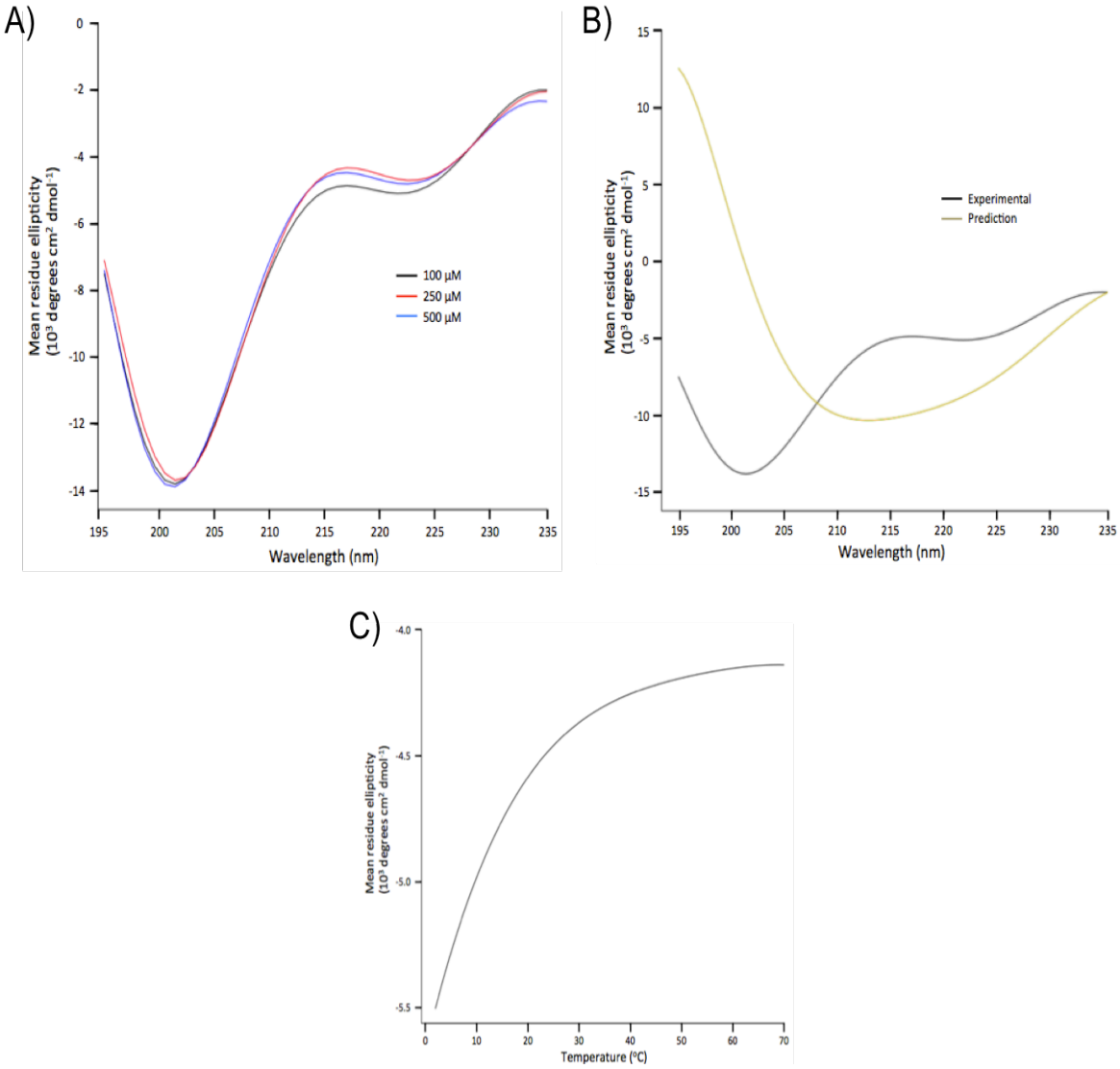

Figure 62. *Circular dichroism analysis of single LRR repeat.* A) CD spectra obtained at various concentrations of peptide suggests that the single repeat LRR could possess some ordered secondary structure. CD scans are the same, independent of concentration, indicating that individual units are not associating together. B) When compared to the predicted CD spectra of a single unit, the signal (especially around ~222 nm) is much less prominent than it should be for a fully folded monomer. This implies that there is less secondary structure content than expected. C) A thermal scan of a 250 µM solution at 221 nm did indicate that there was some secondary structure present, which was lost upon heating.

CD spectra obtained for the single LRR unit suggests that it does possess some secondary structure content (Figure 62A). Signal from random coils are seen in the 195 - 210 nm range, and the mean residue ellipticity in this region of our spectra indicates that this type of conformation is present in our synthesized peptide. However, there is also some signal present around 222 nm. This region is characteristic of helical content in a protein, hinting that there could be some secondary structure formation. This is further supported by a thermal scan of the 250 μM sample (Figure 62C). Signal at the 221 nm wavelength is lost upon heating of the peptide, meaning that the secondary structure content is reduced as higher temperatures are reached. We also used DichroCalc[192] to predict the anticipated spectrum of our peptide, based on the PDB of the designed monomer (Figure 62B). Judging from the spectra prediction, if the monomer possessed the exact correct structure then we would expect to see a stronger signal in the 210 – 225 nm region. However, our design protocol aimed to produce an extended LRR with multiple units linked together. When the single unit is in solution, hydrophobic regions will be exposed to solvent that would otherwise be packed into the core of the extended protein. Having exposed hydrophobes will reduce the formation of accurate secondary and tertiary structure, but the effect of this should be minimised when repeats are linked together and there is a tightly packed core. That we see some helical content at ~222 nm is therefore still encouraging and we hope that by linking more units together, we will discover the exact desired conformation.

Another interesting feature is that the CD spectra obtained at various concentrations are all very similar (Figure 62A). This suggests that the spectra obtained are independent of concentration and so there is no association of individual units. We may have hoped for some associating behaviour that would be characteristic of self-assembling peptides, but that does not appear to be the case and our peptide remains as a monomer even if concentration is increased. Again, this is not completely discouraging as we designed the individual unit to be part of a larger protein. It is our hope

that by linking repeats together via convergent peptide synthesis, we can still produce an extended LRR protein.

7.3.3 Producing a fully protected peptide segment

As stated previously, performing convergent solid-phase peptide synthesis to link fragments of peptides together requires the fragments to be fully protected and also highly soluble. We therefore attempted to construct a segment using solid-phase synthesis and then cleave only the linker region, leaving the sidechain and N-terminal amine protecting groups intact. Using the same Hmsb and pseudoproline locations as in the previous experiments, we synthesised a new batch of LRR single repeat peptide but cleaved using only TFE/DCM (1:1). This meant that the C-terminal linker was cleaved to release the peptide from the resin, but all other protecting groups were maintained (Figure 63A). The fully protected peptide appeared to be highly soluble in water, meaning it could also be characterised by RP-HPLC. Analytical RP-HPLC traces of the crude product showed it to be quite pure (Figure 63B), and collecting the fractions containing only our peptide increased the purity for use in coupling experiments (Figure 63C). MS of the pure product showed a peak at the correct mass (Figure 63D), further confirming the purity of our peptide.

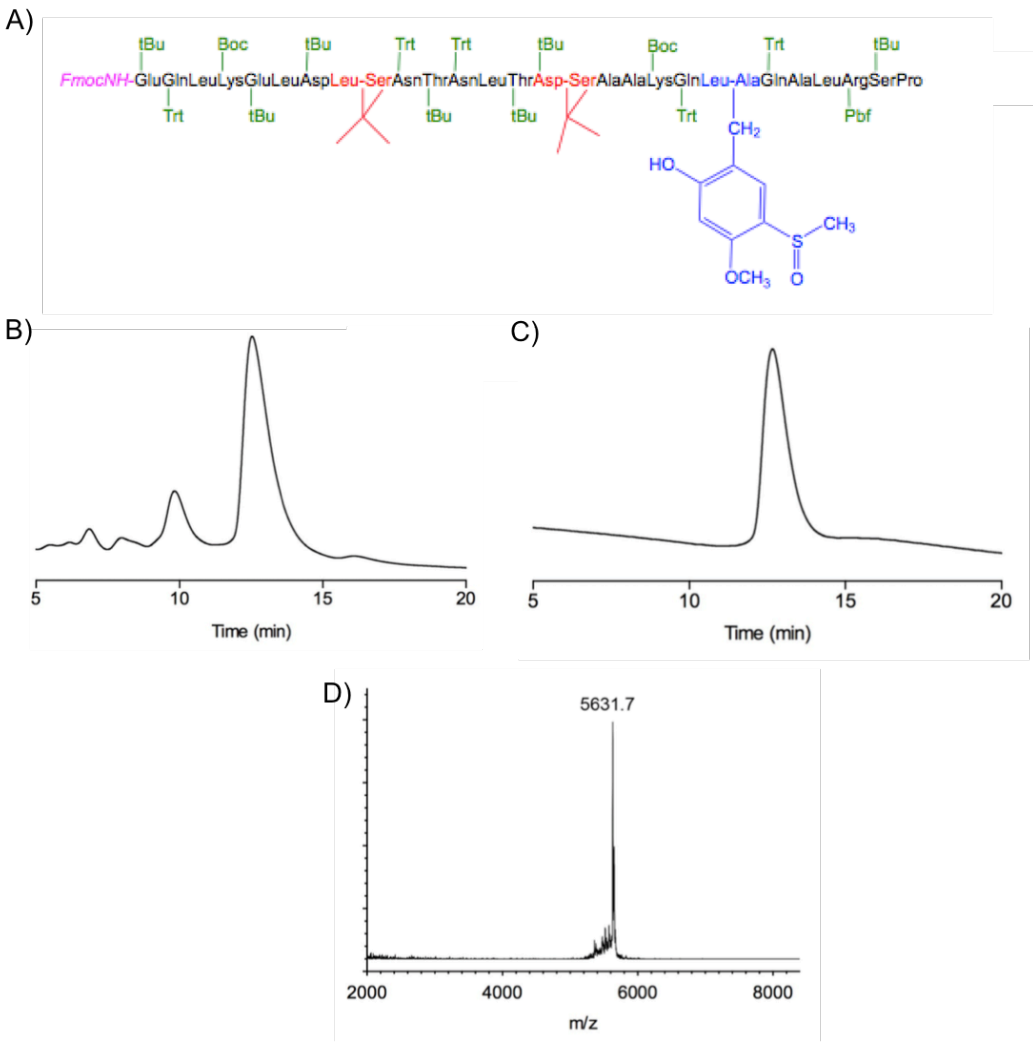

Figure 63. *Synthesis of the fully protected LRR single repeat unit.* A) The fully protected peptide was made with two pseudoprolines and one Hmsb backbone protection groups. B) RP-HPLC of the crude product suggested a very clean and efficient reaction. C) Collection of fractions containing our target peptide only allowed a much purer sample to be obtained for use in coupling. D) MS also confirmed the purity of our peptide, with a peak observed at the correct mass. Calculated mass of fully protected peptide = 5630.9 m/z [Na^{2+}].

Now that we have produced a fully protected single repeat unit of our LRR, we can begin to look at ways to start coupling the segments and hopefully produce an extended repeat protein. Currently, we are looking at which particular methods and resins can be used for this purpose and hope to produce and characterise a 7-mer, before extending the protein even further.

7.4 Discussion

LRRs are very interesting targets for protein design, as they are composed of repeating units of relatively simple complexity. Their modularity makes them attractive candidates for production through peptide chemistry, and the options to add non-canonical residues or functional groups could expand functions beyond what is currently seen in nature. We attempted to design a single repeat LRR that could be able to be linked together to produce the extended repeat protein.

We have found a way, using pseudoprolines and Hmsb backbone protection groups, to produce a pure peptide with high efficiency. Early results for the design produced by this method are encouraging, with the single repeat unit containing some secondary structure that is lost upon a thermal scan. We have also produced a large quantity (~1 g) of fully protected peptide that is very soluble in water. The next stage of experimentation is to start developing an efficient coupling setup by finding a suitable resin to use as the solid state matrix, finding out the optimal loading ratio of initial peptide, and defining which method of amide bond formation to use.

Very recently, while our experiments were ongoing, there were a series of papers in quick succession that addressed the same topic as we were investigating[149,196]. Starting from backbones created *de novo*, simultaneous sequence design was performed on all repeats in the structures using the Rosetta program and sequence constraints derived from multiple sequence alignments. They produced a large number of designs for various repeat proteins, including LRRs, which had a high success rate when tested experimentally (80% soluble, 40% folded). Another paper also elaborated on the control of repeat protein curvature, allowing the shape of binding surface to be designed[197]. This involved producing self-compatible building blocks that were extended into repeats, with these linked together by 'junction modules' to produce a curvature with atomic level accuracy to the design.

The generalised protocol outlined by Baker for producing repeat protein designs seems very similar to ours, using consensus design coupled with Rosetta. However, they bring in full multiple sequence alignments where we have just used consensus design for our specific protein. One potential downfall of our protocol may be that peptide synthesis is still very time-consuming. Although automated synthesisers have made it much less hands-on to produce peptides, 100 mg of a single segment for us takes about a week. Baker *et al* used more traditional methods of recombinant protein expression of synthetic genes and may have been able to screen much more designs in the same amount of time by doing this. However, as mentioned before, chemical synthesis does allow much greater control over the protein produced (with regards to ease of difficult functional groups being added) and this could be an advantage for us by using this method in the future.

Acknowledgements

Computational design was conducted with Willie Taylor, and peptide synthesis was aided by Abubaker 'Bak' El Sayed .

8. Conclusions

The field of protein design is one with huge potential for a variety of different areas. There have been many recent advances that have started pushing design to the forefront of modern technology, with other groups reporting success with published methods.

In total, we have made ~500 novel protein sequences using methods based on protocols that have been successful for other people and yet we have met with very limited success. Only a handful of our proteins throughout this thesis appeared to be soluble, and we have been unable to produce any fully folded structures. The NMR experiments in Chapter 3 could warrant further investigation, as it is possible that these proteins may exist as molten globules. If we perform some further experiments (such as ANS binding or CD spectrometry) and determine if this is the case, we could potentially modify some of these sequences to try and find a completely folded structure.

More molecular dynamics (MD) experiments could also be useful for future design experiments. By all other computational measures, including tertiary structure prediction, the designs we produced looked very similar to native proteins. The MD experiments in Chapter 6 contradicted all of the other computational measures within that chapter, and seemed to be the only correct indicator of the actual outcome. MD simulations have a much more complex energy function than structure prediction, but this comes at the cost of hugely increased computational time. However, since structure prediction techniques do not appear to offer a useful indicator of success and MD simulations do, it might be worth investing that extra time in the hope of creating better designs. Because we were so focused on creating fully folded proteins, we have not attempted to glean as much information as we potentially could have. CD spectra for unfolded designs could still offer some information about how close we are to making better protein sequences. If more secondary structure content is present in some designs than others, then these designs can be seen to be heading in the right direction and so we could try to optimize towards this. We may be able to

do this in an incremental fashion, improving structural content with each design round, rather than going straight for completely folded designs.

We can also speculate on some broad conclusions about the experiments we have performed. Although the targets we have tried to design sequences for may appear simplistic and rather small, they are still more complex than some of the previous successes in protein design so far. As outlined at the start of this thesis, large success has been had designing *de novo* proteins that have a high degree of symmetry or easily definable parameters (e.g. helical bundles)[73,76]. Perhaps the backbones that we have been trying to design sequences for are more difficult than we would have anticipated. We chose the ferrodoxin and Rossman folds as the basis for most of our design studies because the sequence space around these are highly populated in nature, leading us to assume that the structure itself is quite robust. However, finding a *de novo* sequence to fold into the correct backbone has proven immensely challenging. In future studies, it may be worthwhile attempting to concentrate on finding a defined protocol for more simplistic folds and expanding on any potential success we may have.

We have predominately used the Rosetta program for our designs, and this was because of the apparent success of other groups using simple protocols to produce very good results. In our hands, this has not been the case and perhaps it would be prudent to start investigating other design methodologies. Numerous different attempts have been made at creating design protocols[81,198–202], and concentrating on the just one may have restricted our efforts. From our experiments, it is looking more and more likely that Rosetta is not ready to be used for generalised design problems and instead should be treated as a program that may have some success in bespoke instances, with a lot of optimisation needed to make it work for a particular problem. We have applied the standard Rosetta protocols, outlined by the developers of the program, and have had no success with tackling our design problems. Perhaps with more focus on an individual protein using targeted mutations and the manual curation of designs, we

could have produced a fully folded protein. We could also make some changes to the weighting functions of the Rosetta program itself, and see if that could make any different to the success of our experiments. There is a large Rosetta community online (called Rosetta Commons[203]) aimed at allowing people to discuss the methods that they have implemented and the outcomes of those experiments. There is a wealth of information contained on the forum, but each instance appears to use slightly different weighting parameters or methodologies for each instance. Again, this is indicative of Rosetta not being a generalised method but more of a tool that needs a lot of optimization to work for a particular problem. Rather than following the guidelines outlined by the developers of the program, perhaps we could have more success by attempting to modify Rosetta in some way. This could be one of the directions the lab could take in the future.

Rosetta is one of the foremost design programs available today, yet it is still very difficult to use successfully and appears not to be applicable to general design problems. This is disappointing, but past successes should still be taken into account. There have been successful designs by members of other groups in the field but, if anything, our research highlights just how far the technology needs to progress before we can get consistent results from using this specific program. Perhaps our ambitions were too high, aiming for a broad design protocol and uncovering some rules for successful design, but the information we produced is still valuable. Despite encouraging computational results throughout our studies, experimental success was sparse, at best, and the gap between *in silico* and *in vitro* understanding is still some way from being bridged.

At the end of the thesis, we investigated the potential role peptide synthesis has in the synthetic biology field by producing a single-repeat LRR that can be linked together to form an extended protein. An individual unit was chemically synthesised and showed potentially encouraging results from CD analysis, suggesting that there may be some secondary structure forming already. Experiments are still on-going to define which coupling conditions

will be best for creating an extended repeat protein and we are hopeful of making an extended LRR. Although peptide synthesis does appear to offer more control over a given protein, we were "beaten" to the LRR re-design aspect of the project by groups that had used a more traditional *in vitro* expression approach. Traditional biological methods are still much higher throughput than peptide synthesis methods, and it may have been useful to stick to these when testing our LRR experimentally.

9. References

1. Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. SCOP: A structural classification of proteins database. *Nucleic Acids Research* **27,** 254–256 (1999).
2. Orengo, C. *et al.* CATH - a hierarchic classification of protein domain structures. *Structure* 1093–1109 (1997). doi:10.1016/S0969-2126(97)00260-8
3. Dunker, a K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC genomics* **9 Suppl 2,** S1 (2008).
4. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (New York, N.Y.)* **181,** 223–230 (1973).
5. Epstein, C. J. & Anfinsen, C. B. The reversible reduction of disulfide bonds in trypsin and ribonuclease coupled to carboxymethyl cellulose. *The Journal of biological chemistry* **237,** 2175–2179 (1962).
6. Levinthal, C. Are there pathways for protein folding? *Journal Chemical Physics* **65,** 44–45 (1968).
7. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **84,** 4240 (1962).
8. Dinner, a R., Sali, a, Smith, L. J., Dobson, C. M. & Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences* **25,** 331–9 (2000).
9. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475,** 324–332 (2011).
10. Ciryam, P., Tartaglia, G., Morimoto, R. I., Dobson, C. M. & Vendruscolo, M. Widespread Aggregation and Neurodegenerative Diseases Are Associated with Supersaturated Proteins. *Cell Reports* **5,** 781–790 (2013).
11. Pace, C. N., Treviño, S., Prabhakaran, E. & Scholtz, J. M. Protein structure, stability and solubility in water and other solvents. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **359,** 1225–1234; discussion 1234–1235 (2004).
12. Lawrence, M. S., Phillips, K. J. & Liu, D. R. Supercharging proteins can impart unusual resilience. *Journal of the American Chemical Society* **129,** 10110–10112 (2007).
13. Poso, D., Sessions, R. B., Lorch, M. & Clarke, a. R. Progressive stabilization of intermediate and transition states in protein folding reactions by introducing surface hydrophobic residues. *Journal of Biological Chemistry* **275,** 35723–35726 (2000).
14. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M.

Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424,** 805–808 (2003).

15. Dantas, G. *et al.* High-resolution Structural and Thermodynamic Analysis of Extreme Stabilization of Human Procarboxypeptidase by Computational Protein Design. *Journal of Molecular Biology* **366,** 1209–1221 (2007).

16. Dobson, C. M. Protein folding and misfolding. *Nature* **426,** 884–890 (2003).

17. Kopito, R. R. Aggresomes, inclusion bodies and protein aggregation. *Trends in Cell Biology* **10,** 524–530 (2000).

18. Sørensen, H. P. & Mortensen, K. K. Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. *Microbial cell factories* **4,** 1 (2005).

19. Nallamsetty, S. & Waugh, D. S. A generic protocol for the expression and purification of recombinant proteins in Escherichia coli using a combinatorial His6-maltose binding protein fusion tag. *Nature protocols* **2,** 383–391 (2007).

20. Magnan, C. N., Randall, A. & Baldi, P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics (Oxford, England)* **25,** 2200–7 (2009).

21. Sormanni, P., Aprile, F. a & Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology* **427,** 478–490 (2015).

22. Hirose, S. & Noguchi, T. Espresso: A system for estimating protein expression and solubility in protein expression systems. *Proteomics* **13,** 1444–1456 (2013).

23. Hollup, S. M., Sadowski, M. I., Jonassen, I. & Taylor, W. R. Exploring the limits of fold discrimination by structural alignment: A large scale benchmark using decoys of known fold. *Computational Biology and Chemistry* **35,** 174–188 (2011).

24. Taylor, W. R. A 'periodic table' for protein structures. *Nature* **416,** 657–660 (2002).

25. Sadowski, M. I. & Taylor, W. R. Protein structures, folds and fold spaces. *Journal of physics. Condensed matter : an Institute of Physics journal* **22,** 033103 (2010).

26. Taylor, W. R., Chelliah, V., Hollup, S. M., MacDonald, J. T. & Jonassen, I. Probing the 'dark matter' of protein fold space. *Structure (London, England : 1993)* **17,** 1244–52 (2009).

27. RCSB PDB Homepage. (2015).

28. Grant, A., Lee, D. & Orengo, C. Progress towards mapping the universe of protein folds. *Genome biology* **5,** 107 (2004).

29. Zhang, Y. & Skolnick, J. The protein structure prediction problem

could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 1029–1034 (2005).

30. Crooks, G. E., Wolfe, J. & Brenner, S. E. Measurements of protein sequence-structure correlations. *Proteins: Structure, Function and Genetics* **57,** 804–810 (2004).

31. Sadowski, M. I. & Jones, D. T. The sequence-structure relationship and protein function prediction. *Current Opinion in Structural Biology* **19,** 357–362 (2009).

32. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function and Bioinformatics* **82,** 1–6 (2014).

33. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annual review of biochemistry* **77,** 363–382 (2008).

34. Taylor, W. R. Evolutionary transitions in protein fold space. *Current Opinion in Structural Biology* **17,** 354–361 (2007).

35. Wang, Z. X. How many fold types of protein are there in nature? *Proteins: Structure, Function and Genetics* **26,** 186–191 (1996).

36. Coulson, A. F. W. & Moult, J. A unifold, mesofold, and superfold model of protein fold use. *Proteins: Structure, Function and Genetics* **46,** 61–71 (2002).

37. Govindarajan, S., Recabarren, R. & Goldstein, R. a. Estimating the total number of protein folds. *Proteins* **35,** 408–414 (1999).

38. Levitt, M. Growth of novel protein structural data. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 3183–3188 (2007).

39. Rohl, C. a., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology* **383,** 66–93 (2004).

40. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12,** 7–8 (2014).

41. Lesk, A. M. in *Introduction to Bioinformatics (3rd edition)* 333–357 (Oxford University Press, 2008).

42. Eddy, S. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).

43. Altschul, S. *et al.* Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic acids Res* **25,** 3389–3402 (1997).

44. Wu, S. & Zhang, Y. in *Bioinformatics: Tools and Applications* (eds. Edwards, D., Stajich, J. & Hansen, D.) 225–242 (2009).

45. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).

46. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology* **405,** 607–618 (2011).

47. Thompson, J. Rosetta 3.4: Documentation comparative modeling of protein structures. (2010). at <https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/d5/d4e/comparative_modeling.html>

48. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: What the R osetta protein modeling suite can do for you. *Biochemistry* **49,** 2987–2998 (2010).

49. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annual review of biochemistry* **77,** 363–82 (2008).

50. Leaver-fay, A. *et al.* R OSETTA 3 : An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Cancer Research* **487,** 545–574 (2011).

51. Kim, D. E. Documentation for ab initio protein structure modeling application (AbinitioRelax). (2010). at <https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/d0/dd9/abinitio.html>

52. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry* **49,** 2987–2998 (2010).

53. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology* **268,** 209–225 (1997).

54. Lazaridis, T. & Karplus, M. Effective energy function for proteins in solution. *Proteins* **35,** 133–52 (1999).

55. Morozov, A. V, Kortemme, T., Tsemekhman, K. & Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6946–6951 (2004).

56. Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology* **326,** 1239–1259 (2003).

57. Raval, A., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function and Bioinformatics* **80,** 2071–2079 (2012).

58. Zhou, H. & Skolnick, J. Ab initio protein structure prediction using chunk-TASSER. *Biophysical journal* **93,** 1510–1518 (2007).

59. Xu, D. & Zhang, Y. Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. *Scientific reports* **3,** 1895 (2013).

60. Bystroff, C. & Shao, Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics (Oxford, England)* **18 Suppl 1,** S54–S61 (2002).

61. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science (New York, N.Y.)* **253,** 164–170 (1991).

62. Gutte, B. A synthetic 70-amino acid residue analog of ribonuclease S-protein with enzymic activity. *J. Biol. Chem.* **250,** 889–904 (1975).

63. Gutte, B., Daumigen, M. & Wittschieber, E. Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids. *Nature* **281,** 650–655 (1979).

64. Moser, R., Thomas, R. M. & Gutte, B. An artificial crystalline DDT-binding polypeptide. *FEBS Letters* **157,** 247–251 (1983).

65. Eisenberg, D. *et al.* The design, synthesis, and crystallization of an alpha-helical peptide. *Proteins* **1,** 16–22 (1986).

66. Ho, S. P. & DeGrado, W. F. Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein. *Journal of the American Chemical Society* **109,** 6751–6758 (1987).

67. Regan, L. & DeGrado, W. Characterization of a helical protein designed from first principles. *Science* **241,** 976–978 (1988).

68. Dahiyat, B. I. & Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **278,** 82–87 (1997).

69. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science (New York, N.Y.)* **282,** 1462–1467 (1998).

70. Cohen, C. & Parry, D. A. Alpha-helical coiled coils: more facts and better predictions. *Science (New York, N.Y.)* **263,** 488–489 (1994).

71. Offer, G., Hicks, M. R. & Woolfson, D. N. Generalized Crick Equations for Modeling Noncanonical Coiled Coils. *Journal of Structural Biology* **137,** 41–53 (2002).

72. Grigoryan, G. & Degrado, W. F. Probing designability via a generalized model of helical bundle geometry. *Journal of molecular biology* **405,** 1079–100 (2011).

73. Wood, C. W. *et al.* CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics (Oxford, England)* **30,** 1–7 (2014).

74. Grigoryan, G. *et al.* Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science (New York, N.Y.)* **332,** 1071–1076 (2011).

75. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346,** 481–485 (2014).

76. Thomson, A. R. *et al.* Computational design of water-soluble alpha-helical barrels. *Science* **346,** 485–488 (2014).

77. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)* **302,** 1364–8 (2003).

78. Boschek, C. B. *et al.* Engineering an ultra-stable affinity reagent based on top 7. *Protein Engineering, Design and Selection* **22,** 325–332 (2009).

79. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–7 (2012).

80. Saraf, M. C. *et al.* IPRO: an iterative computational protein library redesign and optimization procedure. *Biophysical journal* **90,** 4167–4180 (2006).

81. Mitra, P., Shultis, D. & Zhang, Y. EvoDesign: de novo protein design based on structural and evolutionary profiles. *Nucleic acids research* 1–8 (2013). doi:10.1093/nar/gkt384

82. Gainza, P. *et al. OSPREY: protein design with ensembles, flexibility, and provable algorithms. Methods in enzymology* **523,** (Elsevier Inc., 2013).

83. Eisenbeis, S. *et al.* Potential of fragment recombination for rational design of proteins. *Journal of the American Chemical Society* **134,** 4019–22 (2012).

84. Höcker, B. Engineering chimaeric proteins from fold fragments: 'hopeful monsters' in protein design. *Biochemical Society Transactions* **41,** 1137–1140 (2013).

85. Soding, J. & Lupas, A. N. 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,** 837–846 (2003).

86. Grishin, N. V. Fold change in evolution of protein structures. *Journal of structural biology* **134,** 167–185 (2001).

87. Ostermeier, M., Shim, J. H. & Benkovic, S. J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature biotechnology* **17,** 1205–1209 (1999).

88. Otey, C. R. *et al.* Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. *PLoS Biology* **4,** e112 (2006).

89. Bharat, T. A. M., Eisenbeis, S., Zeth, K. & Höcker, B. 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,** 9942–7 (2008).

90. Moss, A. J., Sharma, S. & Brindle, N. P. J. Rational design and protein engineering of growth factors for regenerative medicine and tissue engineering. *Biochemical Society transactions* **37,** 717–721 (2009).

91. Strauch, E., Fleishman, S. J. & Baker, D. Computational design of a pH-sensitive IgG binding protein. 1–6 (2013). doi:10.1073/pnas.1313605111

92. Martí, S. *et al.* Computational design of biological catalysts. *Chemical Society reviews* **37,** 2634–2643 (2008).

93. Liu, Y. *et al.* De Novo -Designed Enzymes as Small-Molecule-Regulated Fluorescence Imaging Tags and Fluorescent Reporters. 1–4 (2015).

94. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19,** 596–604 (2009).

95. Kiss, G., Celebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational enzyme design. *Angewandte Chemie (International ed. in English)* **52,** 5700–25 (2013).

96. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science (New York, N.Y.)* **319,** 1387–1391 (2008).

97. Pederson, G. T. *et al.* Computational design of an enzyme catalyst for a stereoselective biomolecular Diels-Alder reaction. *Science* **329,** 309–313 (2010).

98. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453,** 190–195 (2008).

99. Critical Assessment of protein Structure Prediction. at <www.predictioncenter.org/index.cgi>

100. Liu, Y. Rosetta 3.5: Documentation for the fixed backbone design application, 'fixbb'. (2008). at <https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/d4/d68/fixbb.html>

101. Sievers, S. a *et al.* Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* **475,** 96–100 (2011).

102. Fleishman, S. J., Whitehead, T. a, Ekiert, D. C., Dreyfus, C. & Jacob, E. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332,** 816–821 (2012).

103. King, N. P. *et al.* Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **336,** 1171–1174 (2012).

104. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510,** 103–8 (2014).

105. Khare, S. D. *et al.* Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nature Chemical Biology* **8,** 294–300 (2012).

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

107. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **34,** 827–828 (1978).

108. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research* **31,** 3370–3374 (2003).

109. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics* **57,** 702–710 (2004).

110. Carugo, O. & Pongor, S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein science : a publication of the Protein Society* **10,** 1470–1473 (2001).

111. Betancourt, M. R. & Skolnick, J. Universal similarity measure for comparing protein structures. *Biopolymers* **59,** 305–309 (2001).

112. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics (Oxford, England)* **26,** 889–95 (2010).

113. Fraternali, F. & Cavallo, L. Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome. *Nucleic acids research* **30,** 2950–2960 (2002).

114. Gao, X. Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Research* **31,** 143e–143 (2003).

115. Studier, F. Protein production by auto-induction in high-density shaking cultures. *Protein expression and purification* **41,** 207–234 (2005).

116. Bondos, S. E. & Bicknell, A. Detection and prevention of protein aggregation before, during, and after purification. *Analytical Biochemistry* **316,** 223–231 (2003).

117. MacDonald, J. T., Maksimiak, K., Sadowski, M. I. & Taylor, W. R. De novo backbone scaffolds for protein design. *Proteins* **78,** 1311–25 (2010).

118. Taylor, W. R. Defining linear segments in protein structure. *Journal of molecular biology* **310,** 1135–1150 (2001).

119. Taylor, W. R. Protein structure comparison using bipartite graph matching and its application to protein structure classification. *Molecular & cellular proteomics : MCP* **1,** 334–339 (2002).

120. Taylor, W. R. Searching for the ideal forms of proteins. *Biochemical Society transactions* **28,** 264–269 (2000).

121. Taylor, W. R. *et al.* Prediction of protein structure from ideal forms. *Proteins* **70,** 1610–9 (2008).

122. Pandini, A., Fornili, A. & Kleinjung, J. Structural alphabets derived from attractors in conformational space. *BMC bioinformatics* **11,** 97 (2010).

123. Saunders, C. T. & Baker, D. Recapitulation of protein family divergence using flexible backbone protein design. *Journal of Molecular Biology* **346,** 631–644 (2005).

124. Huang, P. S. *et al.* Rosettaremodel: A generalized framework for flexible backbone protein design. *PLoS ONE* **6,** (2011).

125. Tyka, M. Rosetta 3.4: Commands for the relax application. (2010). at <https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/d6/d41/relax_commands.html>

126. DeLuca, S., Dorr, B. & Meiler, J. Design of native-like proteins through an exposure-dependent environment potential. *Biochemistry* **50,** 8521–8 (2011).

127. Nivón, L. G., Bjelic, S., King, C. & Baker, D. Automating human intuition for protein design. *Proteins* 1–9 (2013). doi:10.1002/prot.24463

128. Correia, B. E. *et al.* Proof of principle for epitope-focused vaccine design. *Nature* **507,** 201–6 (2014).

129. Laskowski, R. a., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26,** 283–291 (1993).

130. Havranek, J. Rosetta 3.4: Documentation for RosettaVIP Void Indentification and Packing application. (2012). at <https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/d5/d7a/vip.html>

131. Taylor, W. R. Residual colours: a proposal for aminochromography. *Protein engineering* **10,** 743–746 (1997).

132. Huth, J. R. *et al.* Design of an expression system for detecting folded protein domains and mapping macromolecular interactions by NMR. *Protein science : a publication of the Protein Society* **6,** 2359–2364 (1997).

133. Cheng, Y. & Patel, D. J. An efficient system for small protein expression and refolding. *Biochemical and Biophysical Research Communications* **317,** 401–405 (2004).

134. Zhou, P. & Wagner, G. Overcoming the solubility limit with solubility-enhancement tags: Successful applications in biomolecular NMR studies. *Journal of Biomolecular NMR* **46,** 23–31 (2010).

135. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of biomolecular NMR* **6,** 277–293 (1995).

136. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy:

Development of a software pipeline. *Proteins: Structure, Function and Genetics* **59,** 687–696 (2005).

137. Hirose, S. ESPRESSO: EStimation of PRotein ExpreSsion and SOlubility. (2015). at <mbs.cbrc.jp/ESPRESSO/>

138. Hirose, S. *et al.* Statistical analysis of features associated with protein expression/solubility in an in vivo Escherichia coli expression system and a wheat germ cell-free expression system. *Journal of Biochemistry* **150,** 73–81 (2011).

139. Baldwin, R. L. & Rose, G. D. Molten globules, entropy-driven conformational change and protein folding. *Current opinion in structural biology* **23,** 4–10 (2013).

140. Redfield, C. Using nuclear magnetic resonance spectroscopy to study molten globule states of proteins. *Methods (San Diego, Calif.)* **34,** 121–32 (2004).

141. Semisotnov, G. V *et al.* Study of the 'molten globule' intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers* **31,** 119–28 (1991).

142. Sachdev, D. & Chirgwin, J. M. Order of fusions between bacterial and mammalian proteins can determine solubility in Escherichia coli. *Biochemical and biophysical research communications* **244,** 933–937 (1998).

143. Donnelly, M. I. *et al.* An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein Expression and Purification* **47,** 446–454 (2006).

144. Nominé, Y. *et al.* Formation of soluble inclusion bodies by hpv e6 oncoprotein fused to maltose-binding protein. *Protein expression and purification* **23,** 22–32 (2001).

145. Ollikainen, N., Smith, C. a, Fraser, J. S. & Kortemme, T. *Flexible backbone sampling methods to model and design protein alternative conformations. Methods in enzymology* **523,** (Elsevier Inc., 2013).

146. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

147. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. 195–202 (1999). doi:10.1006/jmbi.1999.3091

148. Kullback, S. & Leibler, R. a. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22,** 79–86 (1951).

149. Parmeggiani, F. *et al.* A general computational approach for repeat protein design. *Journal of Molecular Biology* **427,** 563–575 (2014).

150. Zhang, Y. Protein structure prediction: when is it useful? *Current Opinion in Structural Biology* **19,** 145–155 (2009).

151. Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Current Opinion in Structural Biology* **14,** 70–75 (2004).

152. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology* **233,** 123–138 (1993).

153. Dokholyan, N. V, Shakhnovich, B. & Shakhnovich, E. I. Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 14132–6 (2002).

154. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 10383–8 (2000).

155. Mitra, P. *et al.* An Evolution-Based Approach to De Novo Protein Design and Case Study on Mycobacterium tuberculosis. *PLoS Comput Biol* **9,** e1003298 (2013).

156. Bazzoli, a, Tettamanzi, A. & Zhang, Y. Computational Protein Design and Large-Scale Assessment by I- TASSER Structure Assembly Simulations. *Journal of molecular biology* **407,** 764–776 (2011).

157. Taylor, W. R. Protein structure comparison using iterated double dynamic programming. *Protein science : a publication of the Protein Society* **8,** 654–665 (1999).

158. Taylor, W. R. Protein structure comparison using SAP. *Methods in molecular biology (Clifton, N.J.)* **143,** 19–32 (2000).

159. Zhang, Z. & Chan, H. S. Native topology of the designed protein Top7 is not conducive to cooperative folding. *Biophysical journal* **96,** L25–7 (2009).

160. Yadahalli, S. & Gosavi, S. Designing cooperativity into the designed protein Top 7. *Proteins: Structure, Function and Bioinformatics* **82,** 364–374 (2014).

161. Hall, M. *et al.* The WEKA data mining software. *ACM SIGKDD Explorations* **11,** 10–18 (2009).

162. Quinlan, J. R. *C4.5: Programs for Machine Learning*. (Morgan Kaufmann Publishers Inc., 1993).

163. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc* **117,** 5179–5197 (1995).

164. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function and Bioinformatics* **78,** 1950–1958 (2010).

165. Chandrudu, S., Simerska, P. & Toth, I. Chemical methods for peptide and protein production. *Molecules* **18,** 4373–4388 (2013).

166. Acevedo-Rocha, C. G. *et al.* Non-canonical amino acids as a useful synthetic biological tool for lipase-catalysed reactions in hostile

environments. *Catalysis Science & Technology* **3,** 1198 (2013).

167. Johnson, J. A., Lu, Y. Y., Van Deventer, J. A. & Tirrell, D. A. Residue-specific incorporation of non-canonical amino acids into proteins: Recent developments and applications. *Current Opinion in Chemical Biology* **14,** 774–780 (2010).

168. Offenbacher, A. R., Pagba, C. V, Polander, B. C., Brahmachari, U. & Barry, B. a. First site-specific incorporation of a noncanonical amino Acid into the photosynthetic oxygen-evolving complex. *ACS chemical biology* **9,** 891–6 (2014).

169. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annual review of biochemistry* **79,** 413–444 (2010).

170. Kobe, B. & Kajava, A. V. When protein folding is simplified to protein coiling: The continuum of solenoid protein structures. *Trends in Biochemical Sciences* **25,** 509–515 (2000).

171. Stumpp, M. T., Forrer, P., Binz, H. K. & Plückthun, A. Designing repeat proteins: Modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *Journal of Molecular Biology* **332,** 471–487 (2003).

172. Kobe, B. & Deisenhofer, J. Crystallization and preliminary X-ray analysis of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Journal of molecular biology* **231,** 137–140 (1993).

173. Kobe, B. & Kajava, a V. The leucine-rich repeat as a protein recognition motif. *Current opinion in structural biology* **11,** 725–732 (2001).

174. Janeway, C. a & Medzhitov, R. Innate immune recognition. *Annual review of immunology* **20,** 197–216 (2002).

175. Skerra, A. Alternative non-antibody scaffolds for molecular recognition. *Current Opinion in Biotechnology* **18,** 295–304 (2007).

176. Javadi, Y. & Itzhaki, L. S. Tandem-repeat proteins: Regularity plus modularity equals design-ability. *Current Opinion in Structural Biology* **23,** 622–631 (2013).

177. Grove, T. Z., Regan, L. & Cortajarena, A. L. Nanostructured functional films from engineered repeat proteins. *Journal of the Royal Society, Interface / the Royal Society* **10,** 20130051 (2013).

178. Phillips, J. J., Millership, C. & Main, E. R. G. Fibrous nanostructures from the self-assembly of designed repeat protein modules. *Angewandte Chemie - International Edition* **51,** 13132–13135 (2012).

179. Parker, R., Mercedes-Camacho, A. & Grove, T. Z. Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Science* **23,** 790–800 (2014).

180. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A.

Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *Journal of Molecular Biology* **332,** 489–503 (2003).

181. Grove, T. Z., Cortajarena, A. L. & Regan, L. Ligand binding by repeat proteins: natural and designed. *Current Opinion in Structural Biology* **18,** 507–515 (2008).

182. Reichen, C., Madhurantakam, C., Plückthun, A. & Mittl, P. R. E. Crystal structures of designed armadillo repeat proteins: Implications of construct design and crystallization conditions on overall structure. *Protein Science* **23,** 1572–1583 (2014).

183. Kim, H. M. *et al.* Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell* **130,** 906–917 (2007).

184. Merrifield, R. B. Solid Phase Peptide Synthesis. I. The Synthesis of. *Journal of the American Chemical Society* **85,** 2149 (1963).

185. Chan, W. C. & White, P. D. *Fmoc Solid Phase Peptide Synthesis: A Practical Approach*. *The practical approach series 222* **222,** (Oxford University Press, 2000).

186. Abdel-Aal, A.-B. M., Papageorgiou, G., Quibell, M. & Offer, J. Automated synthesis of backbone protected peptides. *Chemical communications (Cambridge, England)* **50,** 8316–9 (2014).

187. Schnölzer, M., Alewood, P., Jones, A., Alewood, D. & Kent, S. B. H. In Situ Neutralization in Boc-chemistry Solid Phase Peptide Synthesis. *International Journal of Peptide Research and Therapeutics* **13,** 31–44 (2007).

188. Howe, J., Quibell, M. & Johnson, T. A new generation of reversible backbone-amide protection for the solid phase synthesis of difficult sequences. *Tetrahedron Letters* **41,** 3997–4001 (2000).

189. Hyde, C., Johnson, T., Owen, D., Quibell, M. & Sheppard, R. C. Some 'difficult sequences' made easy. A study of interchain association in solid-phase peptide synthesis. *International journal of peptide and protein research* **43,** 431–440 (1994).

190. Wöhr, T. *et al.* Pseudo-prolines as a solubilizing, structure-disrupting protection technique in peptide synthesis. *Journal of the American Chemical Society* **118,** 9218–9227 (1996).

191. Wöhr, T. & Mutter, M. Pseudo-prolines in peptide synthesis: Direct insertion of serine and threonine derived oxazolidines in dipeptides. *Tetrahedron Letters* **36,** 3847–3848 (1995).

192. Bulheller, B. M. & Hirst, J. D. DichroCalc--circular and linear dichroism online. *Bioinformatics* **25,** 539–540 (2009).

193. Lloyd-Williams, P., Albericio, F. & Giralt, E. Convergent solid-phase peptide synthesis. *Tetrahedron* **49,** 11065 – 11133 (1993).

194. Hudson, D. Methodological Implications of Simultaneous Solid-Phase

Peptide Synthesis. 1. Comparison of Different Coupling Procedures. *Journal of Organic Chemistry* **53,** 617–624 (1988).

195. Nyfeler, R. Peptide Synthesis via Fragment Condensation. *Methods Mol. Biol.* **35,** 303–316 (1994).

196. Rämisch, S., Weininger, U., Martinsson, J., Akke, M. & André, I. Computational design of a leucine-rich repeat protein with a predefined geometry. *Proceedings of the National Academy of Sciences* **111,** 17875–17880 (2014).

197. Park, K. *et al.* Control of repeat-protein curvature by computational protein design. *Nature Structural & Molecular Biology* **22,** 167–174 (2015).

198. Kulp, D. W. *et al.* Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). *Journal of computational chemistry* **33,** 1645–61 (2012).

199. Jones, D. T. De novo protein design using pairwise potentials and a genetic algorithm. *Protein science : a publication of the Protein Society* **3,** 567–74 (1994).

200. Negron, C. & Keating, A. E. *Multistate protein design using CLEVER and CLASSY. Methods in enzymology* **523,** (Elsevier Inc., 2013).

201. Reynolds, K. a, Russ, W. P., Socolich, M. & Ranganathan, R. *Evolution-based design of proteins. Methods in enzymology* **523,** (Elsevier Inc., 2013).

202. Heinzelman, P., Romero, P. a & Arnold, F. H. *Efficient sampling of SCHEMA chimera families to identify useful sequence elements. Methods in enzymology* **523,** (Elsevier Inc., 2013).

203. Rosetta Commons. (2015). at <https://www.rosettacommons.org/>

204. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature structural biology* **4,** 10–19 (1997).