# Reengineering Bacterial Toxins into Virus-Like Particles

**James Finnian Ross**

**Submitted in accordance with requirements for the degree of Doctor of Philosophy**

**The University of Leeds**
**Faculty of Biological Sciences and the**
**School of Chemistry**

**October 2013**

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

“Study hard what interests you the most in the most undisciplined, irreverent and original manner possible.” — Richard P. Feynman

# Acknowledgements

I would first like to thank my primary supervisor, Assoc. Prof. Bruce Turnbull. Bruce has supplied unparalleled levels of enthusiasm, thought and emotional and mental support over the last three years. Additionally Bruce has steered me out of the woods and way up the mountain, away from the bear!

I would like to thank my computational supervisor, Assoc. Prof. Emanuele Paci. Emanuele has provided me with many, many, many answers to my naïve scripting problems. Additionally Emanuele helped me try to think more like a physicist and appreciate nearly all problems from an entropic perspective.

A big thank you to my co-supervisor, Dr Mike Webb. Mike has been on hand to answer all molecular biology and biochemical questions I have had, which he has done so with immediacy and vigour.

Additionally I would like to thank the members of my supervisors' labs, I've gone for an alphabetical order to keep things friendly: Matthew Batchelor, Tom Branson, Heather Cox, Dr Martin Fascione, James Gowdy, Dr Jeff Hollins, Kristian Hollingsworth, Kat Horner, Darren Machin, Dr Richard Malham, Dr Tom McAllister, Diana Monteiro, Phil Morrison, Kostas Papachristos, Chadamas Pinitsoontorn, Gael Radou, Dan Williamson, Dr Zu Thur Yew, Briony York.

Technical and/or emotional support was also provided by Dr Arwen Pearson, Dr Roman Tuma, Dr Andy Wilson, Martin Fuller, Dr Peiyi Wang, Dr Neil Ranson, Prof. Mike McPherson, Dr Chris West, Dr Simon Connell, Dr Anders Aufderhorst-Roberts, the whole of the Astbury Society, the Leeds Wellcome Trust kids (esp. Lucy Chappell, Thodoris Karamanos, Alex Borodavka and Briony York (again)).

Of course I would not be here now without my parents, literally, who have supported me emotionally (and financially!) throughout the whole of my further education stint. My brother and his beautiful family. My friends from Burton on Trent.

Last and certainly not least, The Wellcome Trust, who have funded my research and my lifestyle for the duration of the PhD and have provided me with opportunity to pursue my own research agenda.

# Abstract

The re-design and controlled self-assembly of natural systems into non-natural functional products is a quickly developing area of Synthetic Biology. Specifically, the manipulation of existing, and the introduction of new protein-protein interactions will allow great advances in bionanotechnology. In nature, protein-protein assemblies mediate many cellular processes and exhibit complex and efficient functions. It is thus rational to assume human-guided biomolecular assemblies could embody equally complex functionality designed to address current human needs.

Here we present the design and preparation of a Virus-Like Particle (VLP) engineered from the cholera toxin B-subunit (CTB). This was achieved via the *de novo* design of a protein-protein interface between CTB subunits consisting of coiled-coil C-terminal extensions and modification to the CTB surface. A combination of computational methods was used to suggest mutations which should reduce the ΔΔG of interaction across the interface. CTB is a natural homopentamer with inbuilt cell targeting and endocytic triggering mechanism. Future applications for the VLP could include use as a drug delivery vehicle to transport protected therapeutic agents to targeted cell types.

Through our investigations it became apparent that the CTB-VLP structures behaved in a similar manner to naturally occurring virus coat proteins, which suggests the successful biomimicry of these complex systems. This study provides a basis for the development of further VLPs from other homomultimeric proteins, especially further classes of homopentameric bacterial toxins.

# Abbreviations

| | |
|---|---|
| AFM | Atomic force microscopy |
| CIP | Calf-intestinal phosphatase |
| CT | Cholera toxin |
| CTA1 | Cholera toxin A-subunit 1 |
| CTA2 | Cholera toxin A-subunit 2 |
| CTB | Cholera toxin B-pentamer |
| DeCrE | A negatively charged coiled-coil motif |
| DeCrK | A positively charge coiled-coil motif |
| DLS | Dynamic light scattering |
| DNA | Deoxyribonucleic acid |
| dNTP | Deoxyribonucleotide triphosphate |
| DSC | Differential scanning calorimetry |
| DSF | Differential scanning fluorimetry |
| E. coli | Escherichia coli |
| ESI-MS | Electrospray ionisation-mass spectrometry |
| HPLC | High Performance Liquid Chromatography |
| JR-coil2 | A putative trimeric designed coiled-coil |
| JR-coil7 | A putative dimeric designed coiled-coil |
| LB | Lysogeny broth |
| LTB | Heat-labile toxin B-pentamer |
| MBP | Maltose-binding protein |
| MBP-$AB_5$ | MBP-CTA2 fusion, with CTB assemble around CTA2 |
| MS | Mass spectrometry |
| PCR | Polymerase chain reaction |
| PFU | DNA polymerase (from Pyrococcus furiosus) |
| PPI | Protein-protein interaction |

| | |
|---|---|
| PWO | DNA polymerase (from Pyrococcus woesei) |
| RNA | Ribonucleic acid |
| SDS | Sodium dodecyl sulfate |
| SDS-PAGE | Sodium dodecyl sulfate-polyacrylamide gel electrophoresis |
| SEC | Size exclusion chromatography |
| STNV | Satellite Tobacco Necrosis Virus |
| SV40 | Simian Virus 40 |
| SynTri | A homotrimeric coiled-coil |
| TEM | Transmission electron microscopy |
| TEMED | Tetramethylethylenediamine |
| TEV | Tobacco etch virus |
| Tris | Tris(hydroxymethyl)aminomethane |
| VLP | Virus-Like Particle |
| VP1 | Polyomavirus capsid protein |
| v/v | Volume per volume |
| w/v | Weight per volume |

# Contents

# Chapter 1
# Introduction

# 1.1 Overview

In this project we aim to create a novel virus-like particle (VLP). Synthetic biology is a field which aims to design from scratch, or redesign existing natural, molecules and systems to create desirable functions for human needs. The VLP we aim to construct is designed from a modified natural non-capsid protein to which we have introduced the ability to form enclosed capsid-like structures. The VLP will be constructed through the addition of scaffolding extensions followed by the redesign of the putative protein-protein interaction interface. This redesign process will use a number of computational protein-protein interface design techniques to elucidate mutations which increase the propensity for association. Some natural viruses are also known to assemble through the use of scaffolding molecules and extensions and we hope to emulate the process of viral capsid assembly with our designed protein capsid. The subunit for our construction is cholera toxin B-subunit (CTB) which naturally triggers its own endocytosis. Thus a VLP constructed from CTB should mediate endocytosis of the constructed capsid. This would make the CTB-VLP useful as a delivery vehicle for the intra-cellular targeting of therapeutic agents.

# 1.2 Synthetic Biology

## 1.2.1 Overview

Synthetic Biology is an emerging bioscience which applies biological molecular theory to non-natural systems in order to create useful technologies. This objective is achieved through constructing synthetic gene circuits, by redesigning proteins, building novel DNA-based structures[1] or creating new forms of life[2]. Synthetic Biology represents the most recent genre smudging paradigm and draws interest from a broad range of scientific disciplines including mathematics, physics, chemistry, biology, engineering, computing, electronics and materials science, due to its capacity to have impact on future work in these fields. The current concept of synthetic biology is seen by many as a form of biological engineering[3], mixed with investigative science. In 1988 Richard Feynman left his final black board scribblings for us including the famous quote “What I cannot create, I do not understand”. Hopefully as we experiment in synthetic biology and learn how to mimic biological mechanisms we will enhance our understanding of the physical processes which nature exploits in order to create them in the first place[4]. Applying this new knowledge to

the *de novo* engineering of biology will further teach us the limits of natural biological components. Then, through the redesign or *de novo* construction of biological systems we will overcome these evolutionary constraints.

### 1.2.2 Historical perspective

In the 1980s synthetic biology was generally described as the production of synthetic forms of life via artificial gene circuits, then in 2004 as the unnatural assembly of natural polymers[5], seen by many as a form of biological engineering [3]. More recently The Royal Society of Engineering described Synthetic Biology's aim as "design and engineer biologically based parts, novel devices and systems as well as redesigning existing, natural biological systems."

### 1.2.3 Standardisation

Synthetic Biology has made dramatic propositions about fundamentally changing biology by providing a platform for standardisation and hence manipulation and design of biological parts. This standardisation is the true power of synthetic biology, and though we are not at that stage yet, when it comes, it should allow click and drag biological design and snap-together biological functionality. A deciding factor in this movement will occur when the cost of mail order DNA synthesis out competes the cost of the current, more labourious molecular biology technologies for designing DNA to the required specification. A standardised approach to strains, plasmids, expression systems etc. would allow greater productivity, easier sharing and higher compatibility between laboratories[3]. In 1996 Eric S. Lander of MIT stated in 'The New Genomics: Global Views of Biology', "One can envision a growing handbook of reusable modular components, much as architects of electronic circuits use today that would propel studies ranging from basic research to applied gene therapy"[6]. The Registry of Standard Biological Parts[7] was founded at MIT in 2003 to apply that very idea, through collaboration with the BioBricks Foundation[8] and the International Genetically Engineered Machines (iGEM) competition[9]. They aim to create standardised biological parts that can be used for a given purpose in almost all circumstances, i.e., snapping any promoter to any gene in any plasmid to transform any chassis (organism). In creating this kind of standardisation they hope to attract non-biologists in order to bring interdisciplinary groups into the field. As biology is highly complex, synthetic biology will

have to undergo more standardisation than mechanical and electrical engineering to achieve an appropriate degree of standardisation[3]. However, several biological data standards already exist, such as the International Nucleotide Sequence Database Collaboration[10], microarray data[11], the Protein Data Bank[12], enzyme nomenclature[13], systems biology models[14] and restriction endonuclease activities[15], which is at least a move in the correct direction.

### 1.2.4 Synthetic Genomics

Synthetic biology is a broad subject with many facets, however, the field is currently dominated by the sub-field of synthetic genomics. Much of this work involves the standardisation of the genetic code, tuning promoters and ribosome binding sites for tailored and timely protein expression[16]. Genes from one, or in fact many different organisms, are re-assembled into a specific host cell to create new genetic circuitry that can then carry out a unique function not previously available[17]. This area of synthetic biology, while extremely powerful, is limited to the molecular machinery currently available through biological evolution. The design of novel nucleotides[18] or redesign of ribosomes to accept four-digit codons[19] are examples of attempts to remove this limitation by re-engineering native functions or by introducing *de novo* functionality[20]. A further step forward is the prediction and production of nascent biomolecules from scratch[21]. This step is yet to be practically applicable as standard, nevertheless, the re-engineering of protein domains and motifs is tractable[22]. The redesign or enhancement of enzymatic function to increase catalysis or alter substrate specificity is a field well developed long before the arrival synthetic biology[23,24]. However, the scaffolding of biomolecules to create larger, more complex structures and the control of addressable surfaces is still in its infancy. Controlling the proximity and orientation of biomolecular assemblies will enhance the repertoire of components available for future synthetic biology applications both *in vivo* and *in vitro*. This area of synthetic biology I will refer to as Synthetic Proteomics.

### 1.2.5 Synthetic Proteomics

The rational design of proteins as synthetic components is currently approached in a number of ways, from focusing on rationally searching homologue database matches to required functions[25], to molecular dynamic simulations allowing *ab initio* protein design.

Other groups are building up from small peptides to larger proteins, in the speculation that protein motif scaffolds (tectons) will be instrumental in *de novo* protein design and understanding[26]. Currently there is a heavy reliance upon top-down methods such as homology modelling, site directed mutagenesis and directed evolution. These approaches have produced good results, such as advances in DNA sequencing[27], PCR amplification[28], addition of unnatural amino acids into protein sequences[29], manufacturing of pharmaceutical compounds[3], and the classic, laundry detergents, but have produce very few examples of truly novel, functional proteins or protein-protein interactions.

Nature evolves structure and function from those molecules already in existence, which may not produce proteins with maximum turn-over or optimal functionality. Thus there must be structural solutions that will lead to optimised kinetics and diverse functions that can be developed outside of nature's evolutionary context[25]. This is particularly pertinent as the computational power required to search homology databases of the future will need to be immense, due to the sheer quantities of data. Truly novel design (atomistic) requires high computation levels beyond the processing power of today's machines. However, with ever increasing computational power, fully parameterised atomistic simulation will one day become a reality. Whatever the method, computational biology will play a major role. Today it is widely believed that rational design of future synthetic biological systems will be achieved via further advances in computational biology[3].

The design of tectons and novel proteins, built to purpose, will allow the construction of simple synthetic macromolecules which, in turn, will lead to multifunctional protein assemblies and finally to advanced complex synthetic cells and tissues. Bromley, Woolfson and co-workers *et al.* present this concept in an overview of what they term 'synthetic-biology space'[30] (Figure 1-1).

### 1.2.6 Synthetic Assemblies

Self-assembly is an important phenomenon for constructing nano-scale biological structures and maintaining this property in designed system components will be essential to synthetic biology's success[31]. By controlling association between biological molecules it will be possible to build fibers and filaments, precisely designed surfaces and macromolecular assemblies with diverse functionality.

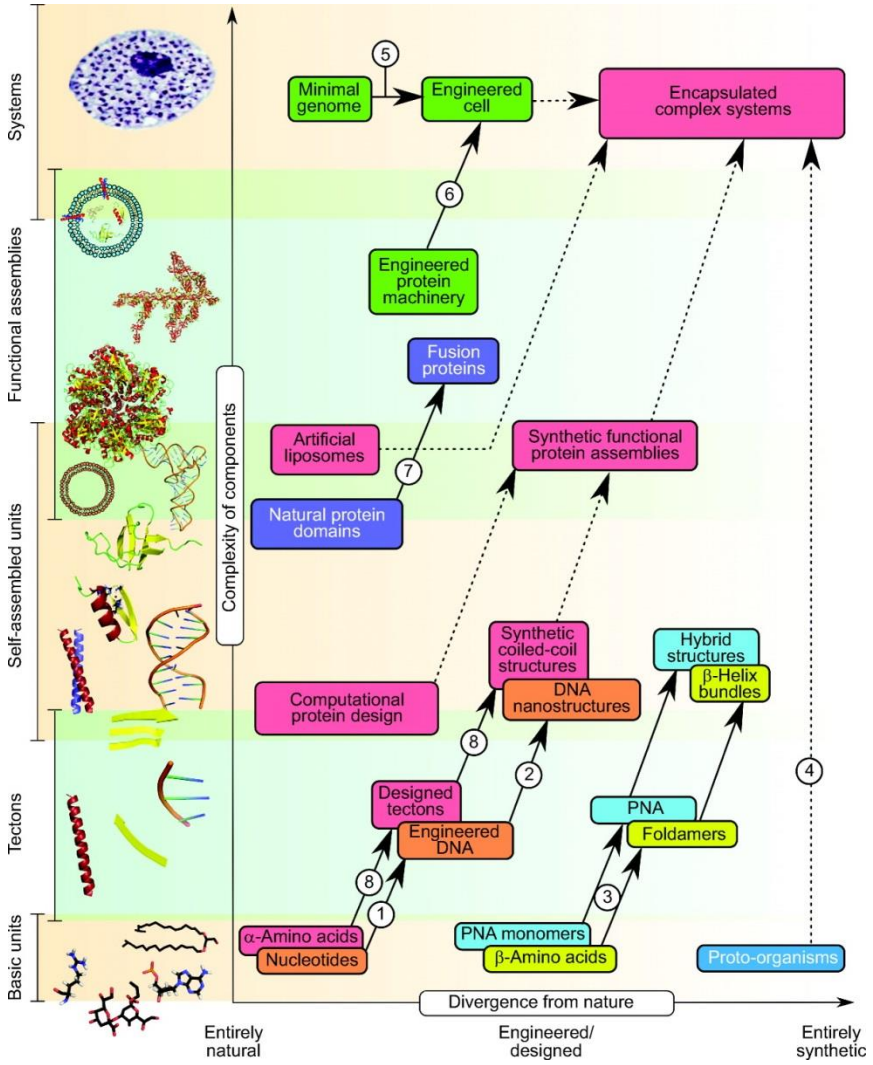



**Figure 1-1:** Synthetic-biology space. A description of the steps required to increase complexity and divergence from nature in the construction of synthetic components which will lead to fully synthetic cell-like systems. Starting from natural building blocks such as DNA or peptides, non-natural proteins and DNA structures can be constructed. Additionally non-natural building blocks can also be used such as PNA, peptide nucleic acids, and foldamers, synthetic polymers which reliably fold into higher order structures, in order to create more complicated components for further synthetic biology uses. Reprinted with permission from[30]. Copyright 2014 American Chemical Society.

The ability to precisely position biofunctional macromolecules on an addressable surface with nanoscale accuracy is an important step toward the goal of creating nanotechnological devices. Fu *et al.* from the Shanghai Institute of Applied Physics have demonstrated the non-covalent assembly of DNA origami sheets into tubes which could be addressed with two proteins that form part of an enzymatic cascade. The group showed that proximal anchoring of the proteins increased the overall reaction rates and when the enzymes were enclosed within tubes that the reaction rates were further increased [32], Figure 1-2. This work demonstrated that simple bioreactors can greatly increase serial enzymatic reactions through diffusional restriction of the reaction intermediates.

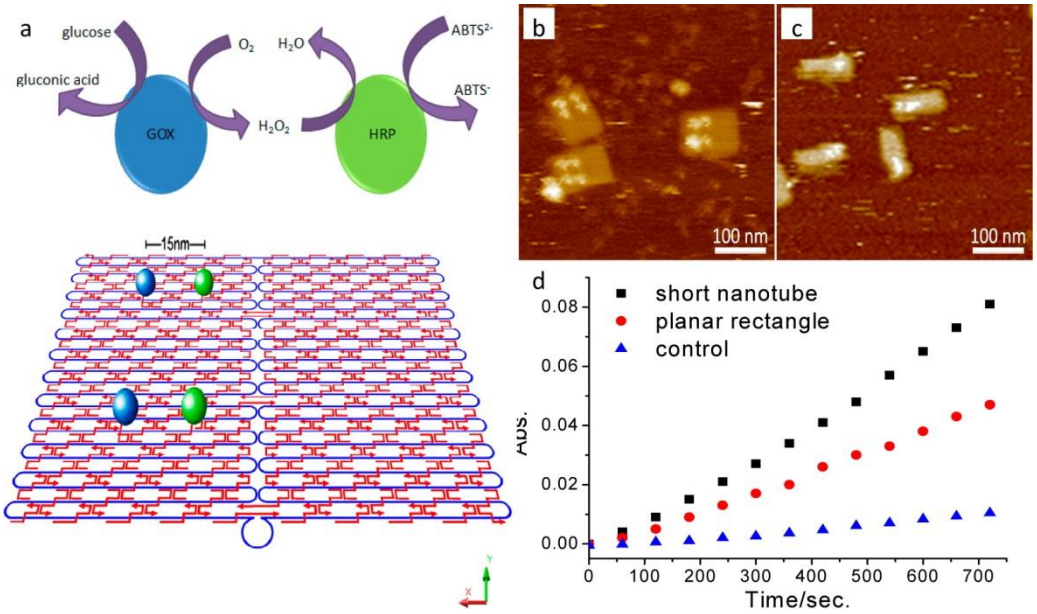



**Figure 1-2:** Addressable Nanoscale Bioreactors. **a)** GOx forms peroxide upon glucose reduction, HRP utilises this peroxide to convert $ABTS^{2-}$ to the chromophore $ABTS^{\cdot-}$. A schematic of the enzymes positioned on the DNA origami substrate. **b)** AFM of the DNA origami sheets with the enzymes associated. **c)** DNA origami tubes with the enzymes presented on the interior. **d)** The change in enzyme kinetics from enzymes free in solution, addressed to origami sheets or confined to the interior of the origami tubes. Reprinted with permission from [32]. Copyright 2014 American Chemical Society.

### 1.2.7 Current Government, Academic and Public Awareness

In recent years the UK government has expressed increasing levels of interest in synthetic biology. In 2008 a postnote was released from the Office of Science and Technology[33] to inform the public and government on the pros and cons of synthetic biology. More recently Research Councils UK have released a Synthetic Biology Roadmap[34] which details how to advance the UK's position as a world leader in synthetic biology. Furthermore, the government has announced a £600M investment in science, including synthetic biology as one of the principal beneficiaries through the creation of six centres for excellence in synthetic biology research.

The academic community has also become increasingly interested in synthetic biology with the creation of networks in synthetic biology between UK universities (almost £900K to establish communication and networking[35], the iGEM competition[9], as well as a host of recent UK-based international conferences such as a recent Harden Conference[36], a joint

Biochemical Society and Protein Society conference on protein engineering[37] and this year's BioBricks Foundation Synthetic Biology 6.0 conference[38].

# 1.3 Protein-Protein Interactions

With the wealth of structural information available on protein-protein interactions (PPI) and current theories on PPI requirements converging, the design of protein interfaces is now becoming a realistic vision which is bound to have an impact on future synthetic biology systems.

## 1.3.1 Overview

Protein-protein interactions (PPI) mediate many cellular processes from organisation of cellular scaffolds and intracellular signalling cascades to initiation of transcription and the immune response. The mapping of these protein interactions will provide a wealth of information on biochemical and metabolic networks[39]. The specific structure of a protein's interaction site, *e.g.*, the position of the van der Waals interactions[40] as well as polar contacts and salt bridges[41] along with more general features such as size, shape and intrinsic flexibility determines the complementarity with which proteins can bind to form stable complexes[42-44]. Methods for designing these interactions into *de novo* protein-protein interfaces will allow the construction of larger assemblies to be realised.

## 1.3.2 Homomeric vs. Heteromeric Interactions

PPIs can be divided into two classes: homomeric and heteromeric. Though this classification seems obvious and potentially arbitrary, the reality is that these two types of interaction differ greatly. Homomeric protein complexes often associate to form "permanent" interactions through extensive van der Waals forces mediated by sterically complimentary hydrophobic interaction sites[45]. These specialised and long-lasting interactions typically have a solvent-excluded surface area (SESA) of between 1200-4660 Å$^2$[46]. Conversely, heteromeric complexes more often associate dynamically[47] and, as the surfaces are at times solvent exposed, they cannot contain too many hydrophobic residues. As these surfaces typically interact through complementary hydrogen bonding, dynamic protein interfaces often resemble non-interface surfaces[41]. Heteromeric interfaces often have a SESA of between 600-2000 Å$^2$ [46]. A SESA of 600 Å$^2$ is thought to be a reasonable

minimum to prevent rapid invasion of the bulk solvent into an interaction site formed from a single 'hot spot'[48].

### 1.3.2.1 Hot Spots

Most PPIs are dominated by only a small percentage of the interaction site residues, termed 'hot-spots'. Hot-Spots are motifs of the interface that contribute a large portion of the binding energy involved in the interaction. Typically they are defined as individual residues which are capable of contributing a $\Delta\Delta G_{binding}$ (change in free energy of binding) of -2 kcal/mol or more to the PPI[49]. It has been found that the main hot spot residues identified are tryptophan (21%), arginine (13.3%) and tyrosine (12.3%)[50], which can contribute up to -4 kcal/mol to the $\Delta\Delta G_{binding}$[49]. These three residues are, of course, large residues and their significant improvements in $\Delta\Delta G_{binding}$ compared to small residues like alanine could arise from their larger interaction surface area. Alternatively, these residues are more structurally diverse than the likes of leucine, serine and valine and they could thus perform a more diverse range of roles at the interface; for example tyrosine can form π-π stacking interactions, hydrophobic interactions and hydrogen bonds. So, is the larger $\Delta\Delta G_{binding}$ for these residues solely a consequence of their size, or is it the result of their specific structural features? The latter argument is supported by tyrosine being three times more likely to be a hot spot residue than phenylalanine, despite reasonably similar structures[48].

### 1.3.2.2 Solvent Effects

Protein-protein interfaces generally contain a mixture of hydrophobic and polar residues[51]. Therefore, PPIs are often characterised by a favourable energy gain in burying the hydrophobic parts of the interface from the bulk solvent[52], as well as a compromise between polar residues interacting with solvent or forming hydrogen bonds with the binding partner[53]. Over larger (wider) interfaces, residues that promote solvation must be present to maintain solubility. The displacement of water within and around the binding site in to the bulk solvent often provides, in general terms, an increase in disorder of the system even when the increase in order of the new complex is taken into account. This is achieved, in part, by the destruction of hydration shells, i.e., ordered layers of water, which extend further into the bulk solvent from unbound PPI than for standard solvated protein surfaces.

### 1.3.3 Experimental Characterisation of PPIs

Experimental validation of PPIs can be assessed by a number of techniques. Examples of well used methods to assess PPIs are the yeast two-hybrid system[54], immuno-precipitation and affinity pull-down, as well as mass-spectroscopy[55,56]. However these methods are not capable of delivering information based on the binding mode of the complex. It is widely accepted that accurate characterisation of PPIs requires the use of several experimental techniques in a combinatorial manner[57,58]. Techniques which can provide medium to high resolution structures for PPIs are X-ray crystallography, NMR spectroscopy, electron microscopy and cryo-electron tomography. Lower resolution techniques such as FRET, analytical ultracentrifugation, ITC and SAXS[59] can also deliver important information on PPI formation. However, which experimental techniques can be successfully used is dependent on the type of PPI. For example, assessing homodimeric PPIs eliminates two-hybrid, Immuno-precipitation and pull-down as viable choices. However, given a highly stable quaternary structure, simple methods such as SDS-PAGE or SEC, or quick and cheap techniques such as Differential Scanning Fluorimetry may be all that is necessary to deliver the required information.

### 1.3.4 Coiled-Coils

Coiled-coils are potentially the simplest of all protein-protein interaction domains. They are ubiquitous amongst prokaryotic and eukaryotic cells and are also found in some viruses. Consequently, they have received much attention and are well characterised in the literature[60]. Coiled-coils are formed by the association of two or more α-helices which may associate in parallel, antiparallel or mixed orientations. The total length can vary from 14 amino acids to many hundreds of amino acids. Shorter coiled-coils often form oligomerisation motifs of larger proteins, while longer coiled-coils have structural roles in the cytoskeleton and in motor proteins. Crick first proposed their existence in 1953[61] and the first crystal structure was obtained by Phillips in 1986[62].

The repeating unit of a coiled-coil is a ‘heptad’, i.e., a sequence of seven amino acids. The helices of the coiled-coil twist around the central axis of the oligomer in a left handed manner, this superhelical distortion lowers their pitch to 3.5 residues per helical turn. Thus the heptad can repeat every two turns of the helix. This heptad repeat has the nomenclature `abcdefg`. Depending on the oligomeric state of the coiled-coil, the

residues in the heptad repeat take on different functions (Figure 1-3)[60]. In dimeric, trimeric and most tetrameric interactions, the a and d' residues form the hydrophobic core, supported by electrostatic interactions between e and g' residues.

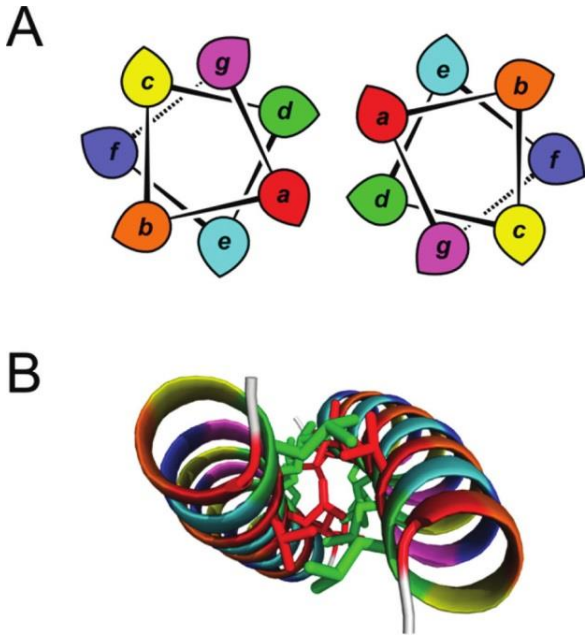



**Figure 1-3:** The structure of a dimeric coiled-coil. **a)** The helical wheel representing the coiled-coil heptad repeat. a and d positions associate to form the hydrophobic core, while e and g position support the structure with electrostatic interactions. **b)** A view down the central axis showing the hydrophobic core. Reprinted with permission from[60]. Copyright 2014 American Chemical Society.

In dimeric and trimeric coiled-coils, hydrophobic residues dominate at positions a and d which creates a hydrophobic interaction surface between the two coils. Smaller hydrophobic residues such as valine often determine the dimeric oligomerisation site and bulkier residues such as isoleucine often prescribe trimeric coiled-coils. These hydrophobic residues interact in a so-called 'knobs-into-holes' packing, where side-chains (knobs) of one helix pack into gaps between side-chains (holes) on the binding partner. This interaction provides the majority of the coiled-coil binding affinity. Positions e and g run alongside the hydrophobic core and often form salt-bridges between subunits which help to determine the specificity by which binding occurs. The remaining positions b, c, and f are usually residues with a high helical propensity and also help to maintain the solubility of the coils by interacting with the solvent.

Many natural coiled-coil sequences contain irregularities that deviate from the basic function of the residues in the heptad repeat listed above. These can include the incorporation of electrostatically repulsive residues or residues with poor helical propensities. Although these deviations likely perform a biological function, such as recognition etc., they are of limited use for synthetic biology applications. However as the structural rules for coiled-coils are relatively simple compared to other protein-protein interfaces, , these motifs can be designed for specific uses. The heptad repeat denotes the function of the amino acid, therefore each residue can be purposefully chosen to contribute to the design. The first redesign of a coiled-coils, by Hodges[63] produced an 86 residue disulphide-linked parallel dimer.

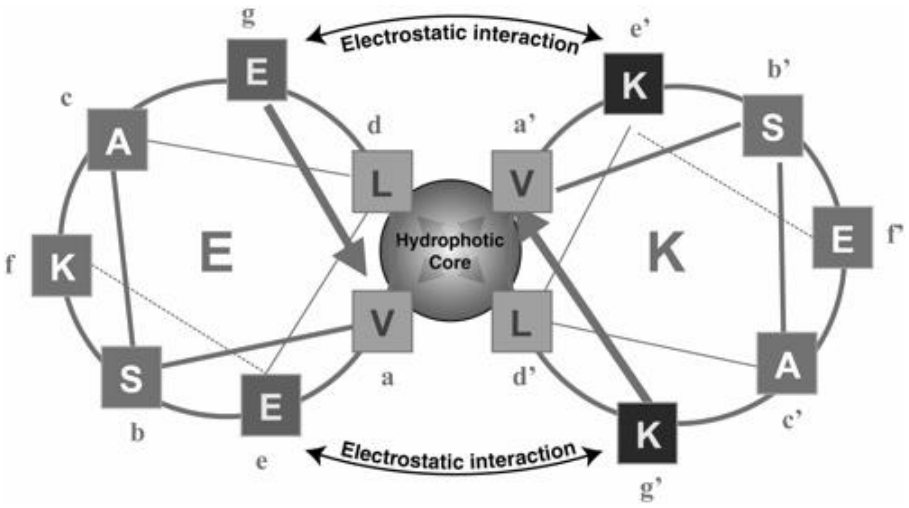



**Figure 1-4:** The de Crescenzo coilE (left) and coilK (right) which heterodimerise. The hydrophobic core residues consist of valine and leucine. The electrostatic interaction help prevent homodimerisation through electrostatic repulsion. Reprinted with permission from[64]. Copyright 2014 American Chemical Society.

In 2003 de Crescenzo et al., in the O'Connor-McCourt group at the Canadian National Research Council, designed a number of heterodimeric parallel coiled-coils with varying lengths of the canonical heptad repeats VSALEKE and VSALKEK[64] (Figure 1-4). One coil (coilE) had glutamate residues at the e and g positions, while the other coil (coilK) contained lysines at the same positions. The coiled-coil constructed with five heptads had a $K_d$ of 63 pM. They measured the effect of heptad length on the association of the dimer and found that smaller coils associated via more complex mechanisms involving conformational rearrangements.

More recently Burton *et al.*, in the Woolfson group at the University of Bristol, have recently demonstrated the incorporation of reactive thiol groups into a *de novo* designed coiled-coil scaffold and shown these groups to be solvent accessible through their reaction with iodoacetamide[65] (Figure 1-5). This incorporation of reactive groups to truly novel scaffolds is a good example of our current ability to engineer proteins at this level. As this *de novo* protein assembly was created from scratch, the function of each residue is known with precision, which leads these kinds of structures to be amenable to further design.

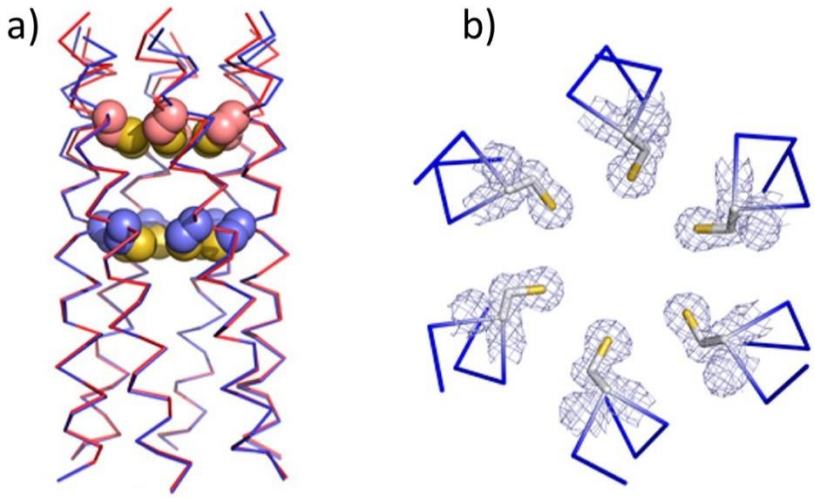



**Figure 1-5:** Incorporation of reactive thiol groups into a channel of a de novo protein scaffold. **a)** An overlay of two hexameric coiled-coils with a ring of cysteines incorporated at different heptad repeats. **b)** a cross-sectional view of a ring of cysteine residues. Reprinted with permission from[65]. Copyright 2014 American Chemical Society.

## 1.3.5 Symmetrical Protein Assemblies

The production of larger symmetrical protein cages can be exemplified with four examples. Firstly Padilla *et al*., in the Yeates group at the University of California, Los Angeles, constructed large symmetrical assemblies via protein fusions including a tetrahedron formed from the homodimeric M1 matrix protein of the influenza virus and the homotrimeric bromoperoxidase, via the incorporation of 9 residue helical linker[66] (Figure 1-6a/b). The Noble group, at the University of Oxford, constructed 2D arrays with the streptavidin homodimer and the a homotetrameric ALAD-streptag fusion[67], (Figure 1-6c/d). The resulting structures are an impressive feat of construction which rely on the correct choice of linker between the fusion proteins to hold the two protein domains at specific angles. In each case, they rely on the use of native binding interfaces provided by nature.

In a third example, the Woolfson group at the University of Bristol, described the construction of protein cages from coiled-coils[68] (Figure 1-7). These self-assembled cage-like particles (SAGEs) were built in an analogous fashion to the protein cage from the Yeates group, in this case, employing the use of homotrimer and heterodimer building blocks that were linked by disulfide bonds. However, these coiled-coils were designed *ab initio* to interact with each other which presents another level of complexity in the design process. Molecular dynamic simulations were used to show the propensity for the 2D coiled-coil array to form curvature which would lead to enclosed particles.

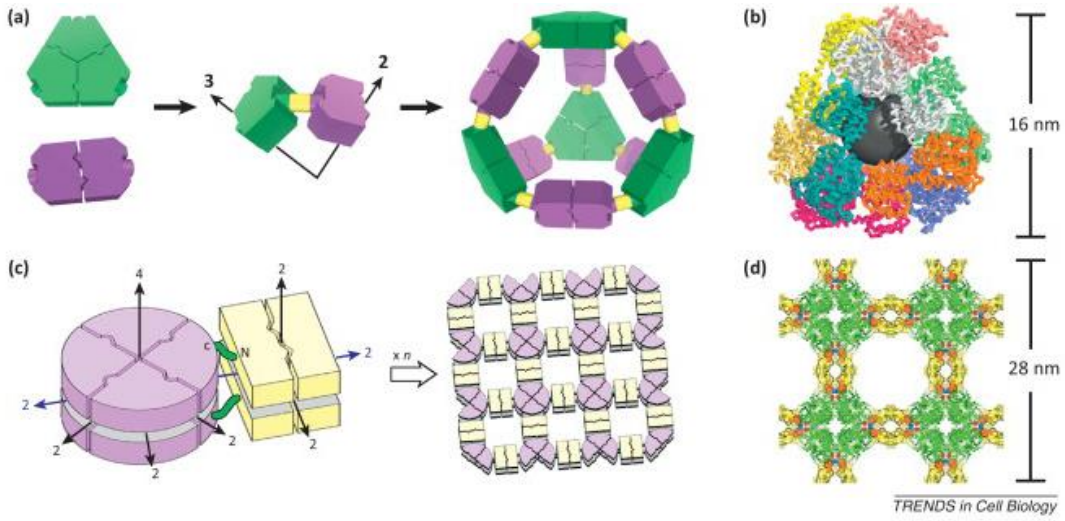



**Figure 1-6:** Oligomeric fusion strategies which bring about 2D and 3D assemblies. **a)** A semi-ridging linker region (yellow) is used to prescribe an enclosed structure of a dimer-trimer fusion. **b)** The crystal structure of the tetrahedral particle. **c)** Fusion proteins of dimer-tetramers which bind along the same axis of symmetry form 2D arrays. **d)** The crystal structure of the 2D array. Entire figure reprinted from[69], with permission from Elsevier. **a and b)** from[66]. Reprinted with permission from AAAS. **c and d)** Reprinted by permission from Macmillan Publishers Ltd: Nature nanotechnology[67,70], copyright 2010 and 2011.

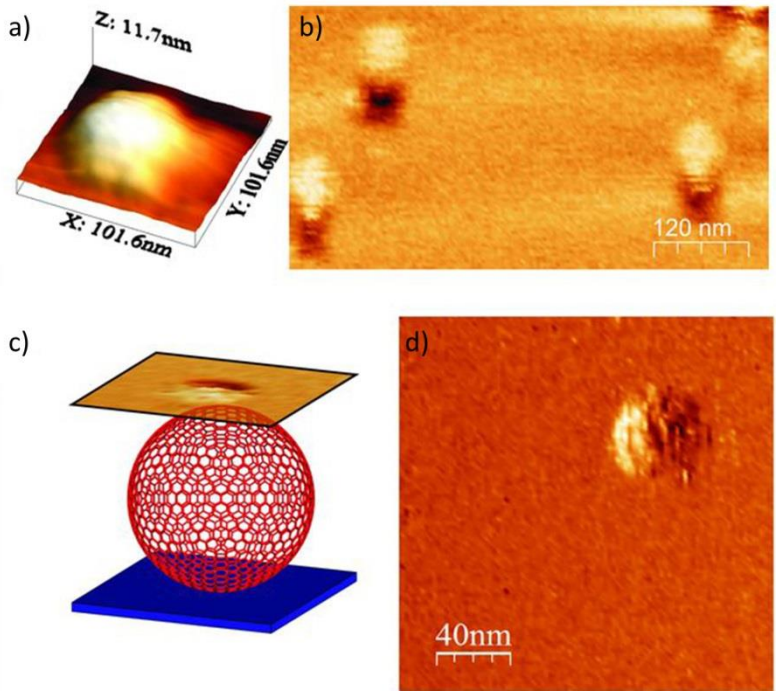



**Figure 1-7:** SAGE coiled-coils forming nano-scale protein cages visualised by scanning probe microscopy. **a)** AFM 3D projection showing the height and relative diameter of the collapsed protein cage. **b)** Liquid AFM of protein cages. **c)** schematic of the non-contact LMFM scanning regime. **d)** magnified image of the protein cage, revealing hexagonal structure. From[68]. Reprinted with permission from AAAS.

Finally, the Baker group, at the University of Washington, have investigated the construction of larger synthetic assemblies in the form of symmetrical protein cages[71]. They used symmetrical docking screens of large protein libraries, followed by redesign of the new protein-protein interface with the program Rosetta, to produce both tetrahedrons and octahedrons. It is worth mentioning that the proteins used to construct the octahedron naturally form protein cages of a larger dimension, thus this protein cage was made by re-engineering of a PPI. However, the protein which form the tetrahedron, galactoside *O*-acetyltransferase, is not thought to naturally form any higher order complexes. TEM and crystal structures of the tetrahedral T3-10 protein cage are shown in Figure 1-8. This work presents the meticulous engineering of a *de novo* PPI, demonstrating our fundamental ability to construct macromolecular assemblies without any need for covalent modification or scaffolding of an interaction. However a screening process was used to search for proteins which matched a 'designability' criterion to allow the ease of redesigning the interface. Thus, the proteins which form the assemblies were not chosen for their native functions; rather, they were selected because they were expected to be amenable to redesign into protein assemblies.

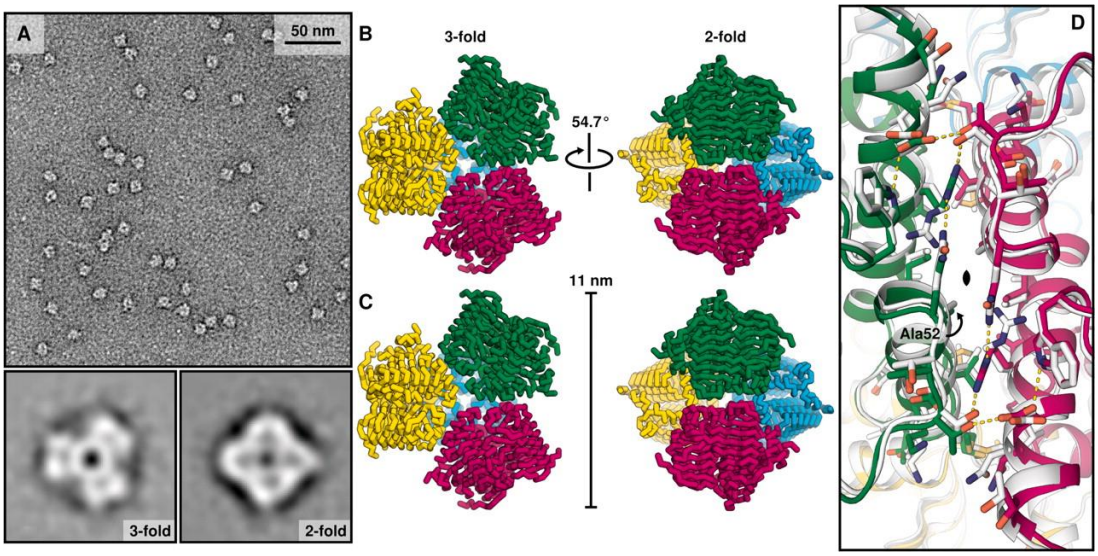



**Figure 1-8:** *de novo* assembly of a tetrahedral particle T3-10. **a)** A representative TEM micrograph with averaged structures below. **b) and c)** backbone crystal structures of the T3-08-T3-10 and the T3-10 assemblies, respectively. **d)** A comparison of the crystal structure (green and magenta) to the computational prediction (white) from which the designs were constructed. From[71]. Reprinted with permission from AAAS.

The ability to control the assembly of functional biomolecules with atomic precision will allow future technologies to be realised, from simple targeted drug delivery vehicles to more complicated assemblies that could be used for biomolecule-based computational devices.

# 1.4 Biomolecular Simulations

## 1.4.1 Overview

Richard Feynman proclaimed in his 1963 'The Feynman lectures on physics, vol-1' that 'everything that living things do can be understood in terms of the jiggling's and wiggling's of atoms', and through biomolecular simulation we have a means of visualising these movements and extracting useful information from them. Biomolecular simulation is a technique based on the numerical solution of the equations of motion over time, for a model that represents a system of molecules. Models differ in the degrees of freedom assumed (electrons, nuclei, group of atoms (amino acids), entire proteins, etc.) and the interaction between these degrees of freedom. Force-fields which accurately represent interactions between atoms in proteins, nucleic acids, lipids, water and small molecules of biological and pharmaceutical importance have been developed in the past 30 years. These force-fields have been incorporated into a number of software packages, such as AMBER and CHARMM, which have been built for the evolving needs of the modern biologist. This type of software allows the study of protein folding, dynamics and changes in conformation[72-74], protein association with small molecules and proteins[74-76] and assessment of the free energies of ligand binding[77].

### 1.4.1.1 Simplified models

As the size of the system of study increases, the use of such detailed models and force-fields dramatically limits the timescales that can be investigated, up to a point where the necessary computational processing power is reached[78,79]. When studying the association of a large number of proteins, fully flexible atomistic models are thus not viable. In order to extend modelling beyond the limits of atomistic simulations, a number of coarse-grained approaches have been proposed wherein biomolecules are represented by simple spheres, cylinders or ellipsoids, or sometimes as rigid atomistic models. The all-atom approach and the course grain approach can be combined to allow the simple modelling of parts of the system which are less informative, while rigorously simulating the part of the system which is more informative[80].

In all cases (atomistic and course grained) the environment around the target, often the solvent, must also be represented numerically. In a further attempt to reduce complexity, the effect of the environment is often represented by an average calculation of its influence on the protein, rather than calculated atom by atom (explicit). For these reasons, biomolecular simulations cannot routinely give precisely correct numerical predictions of a system[81]. Therefore, it is still essential to combine simulation with experiment to allow simulation to steer the experimentation and experiment to optimise the computation[82].

#### 1.4.1.2 Binding orientations and assessment.

With regard to PPIs, protein-protein docking techniques have become fashionable as of late. Such docking programs include Z-dock[83], Hex[84] and Rosetta-Dock[85]. These programs find the most energy-minimised configuration between two proteins, thus predicting their binding conformation. The use of simple linear interaction energy methods should also not be over looked. For example the FoldX package contains a simple energy function and a side-chain rotamer library. It is used to measure the free energy change of unfolding, compared to a hypothetical model of the same sequence. Recently Christensen and Kepp[86] have used FoldX in conjunction with MD simulation to successfully estimate the stability changes upon mutation of a fungal laccase.

### 1.4.2 Computational Protein-Protein Interface Design

The generation of specific protein-protein interactions with designed affinity, specificity and kinetics is a long term goal of computational design. Protein interaction affinities can be influenced by a number of factors such as temperature, pH and the ionic strength of the solvent. Interaction sites can occur on any surface of the protein including unstructured loop regions. This binding affinity can be described as a function of the free energy difference between the bound complex and the components of the complex, which includes both inter-subunit and intra-subunit energy changes as well as entropic contributions and 'external factors' such as solvent effects. This variable menagerie is partly responsible for the lagging of PPI design compared to computational design advances in enzymatic functions[87] and protein folds[21].

Many residues in proteins are critical for their structural viability and biological function. Identification of mutable positions which maintain tertiary and quaternary structure and function but allow a degree of surface association manipulation is an essential first step in designing protein interfaces. A number of initial strategies can be employed to identify designable positions in a protein sequence, such as protein sequence analysis to identify evolutionarily conserved residues, mutational differences between strains of a given species, as well as between structural homologues. Structurally important residues can often be determined in such searches. Furthermore, these conserved residues have been shown to overlap with hot spot residues[88]. A more rigorous test for key residues is alanine scanning mutagenesis[89] which can be conducted *in silico* by simulation using a range of commercial modelling packages [48,90]. The use of alanine scans has confirmed that the majority of the protein interface energy gain is contributed by just a few important hot spot residues[91].

Schreiber and Fleishman[92] have suggested a number of strategies which can lead to successful protein interface design which are highlighted in Figure 1-9. Here we will focus on the *de novo* design strategies. Docking protein scaffolds to find binding partners with the most complementary shapes and then optimising the amino acid residues at the interface was described above through Baker's design and assembly of octahedral and tetrahedral protein cages [71].

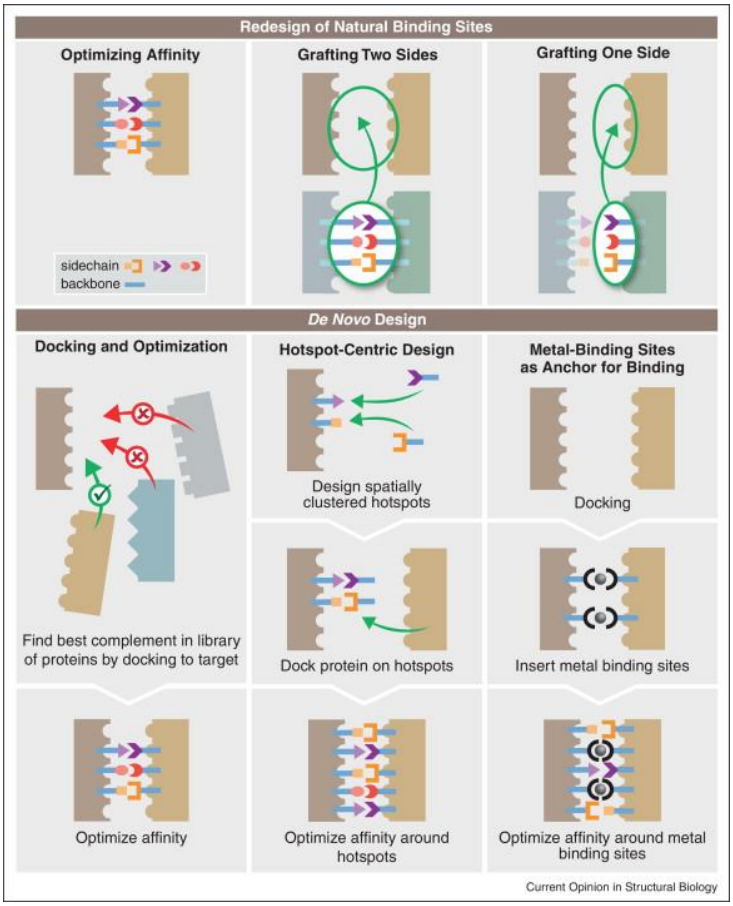



**Figure 1-9:** A schematic description of possible protein-protein design approaches. Redesign can be accomplished through the grafting of interacting residues from a known PPI on to a new putative interface. d*e novo* design can be achieved through the docking of scaffolds, hot-spot design, which may include a level of grafting, or through metal, or ligand, binding site design, where each strategy is followed by optimisation of the binding site. Reprinted from[92]. Copyright 2013, with permission from Elsevier.

A second strategy for de novo interface design is to dock amino acid side chains to the target protein surface in a similar way to *in silico* fragment screening, and then link them together using a suitable protein scaffold. This strategy is illustrated by the work of Fleishman *et al.* from the Baker lab at the University of Washington. They designed a binding protein to the conserved stem region of the 1912 H1N1 influenza hemagglutinin protein[22]. After docking amino acid residues to the surface of the hemagglutinin protein, the possible positions of the Cα rotamers were calculated. Finally these Cα ensembles were screened against a library of proteins to find scaffolds with the correct Cα positions. Once these hotspot residues and scaffolds had been deduced, Rosetta was used to suggest packing residues which would hold the hotspot residues in the correct orientations and contribute additionally to the interface. Figure 1-10 shows the crystal structure of a designed protein HB36 in association with the haemagglutinin protein. An additional mutant HB80 was shown to inhibit the fusogenic conformational change required for influenza infectivity at low pH.

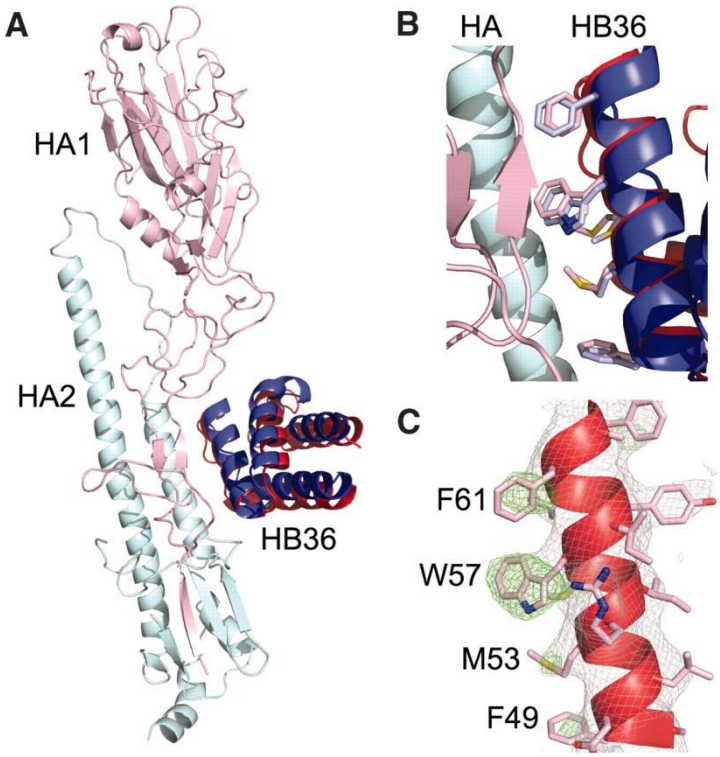



**Figure 1-10** The HB36 designed protein binding to the stem region of hemagglutinin. **a)** A superposition of the crystal structure of HB36 (red) binding to the two subunits of hemagglutinin (pink and cyan) with the computational model (blue). **b)** A close up of the designed interaction site. **c)** An unbiased electron density map from the X-ray data. From[22]. Reprinted with permission from AAAS.

A third strategy for de novo interface design is to incorporate metal-mediated interactions, which may then be augmented by other interactions between amino acids. The Kuhlman group at the University of North Carolina recently attempted to control protein-protein interactions through zinc mediated association[93]; they succeeded while inadvertently creating a protein capable of catalysing carboxyester and phosphodiester hydrolysis[94]. They designed a homodimer, wherein each monomer was a helix-turn-helix. Each monomer was designed to coordinate two zinc atoms using two histidine residues. As zinc has a preference for tetrahedral coordination, a dimeric protein complex should form with the additional histidines on the binding partner coordinating the free sites on the zinc (Figure 1-11). The authors used the Rosetta program to search the pdb for other zinc binding scaffolds and to attempt to graft these onto selected regions of the dimer.

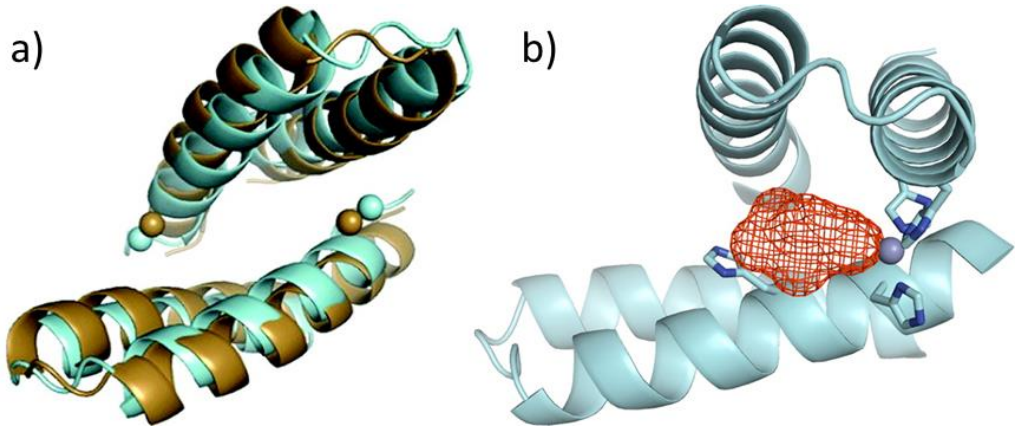



**Figure 1-11:** The MIDI-zinc homodimer, interacts primarily through zinc mediated association. **a)** the crystal structure (blue) and the computational mode (brown) with an Cα rmsd of 1.4 Å. **b)** One of the zinc coordination site is not occupied by a histidine which allows the zinc to mediate carboxyester and phosphodiester hydrolysis. Reprinted with permission from[94]. Copyright 2012 American Chemical Society.

The discovery and implementation of new design protocols is occurring regularly. Established force-fields such as CHARMM and AMBER are regularly updated, however they do not incorporate some of the more interesting features present in packages like Rosetta. Rosetta however does not perform the more conventional MD simulation of the former packages. Additional programs perform a number of specific and important measurements however, each of these programs is written in different code and are not regularly compatible.

# 1.5 Virus Capsid Assembly

## 1.5.1 Overview

The regulation of capsid assembly is paramount in the construction of infectious virus particles. Often the virus genome provides the common structural denominator for viral capsid protein assembly. However under particular conditions the virus coat proteins themselves can spontaneously assemble into virus-like particles. How do natural viruses prevent the formation of these non-infectious particles? This is achieved through the 'process of assembly' mediated through the 'regulation of assembly'. The process of assembly is similar to that of many other protein-protein interactions and is based around the physicochemistry of the interaction site and the environment in which assembly occurs. The regulation of assembly however is more biological in nature, in that the assembly of the particle must be spatially and temporally regulated to ensure packaging of the genome and the correct assembly of the coat. Allostery is thought to play a key regulatory role through effector molecules and viral or host factors. Often these allosteric regulators make a conformational alteration in the coat protein which leads to ordered assembly.

## 1.5.2 Assembly

From a simplistic perspective a capsid protein should associate through Brownian motion, in a geometrically precise orientation and bind with high affinity to form stable particles. However, if the binding affinity is very high, then full virus particles would rarely occur as most of the subunits would become associated in partially constructed intermediates leading to kinetic traps[95]. The use of master equations [96,97] and simulation[98,99] has revealed that a general requirement for efficient and precise capsid assembly is the interactions between the coat proteins must be weak. The low affinity ensures that intermediate states may disassemble almost as readily as they assemble which minimises aberrant capsid growth[100] and formation of kinetic traps (Figure 1-12). For example the average association energy between two hepatitis B virus capsid subunits is approximately -4 kcal/mol[101]. However, the multivalency of the full assembly is strong enough to prevent the dissociation seen with the part assemblies. This weak association has an additional positive effect on the assembly process: nucleation events are less frequent than the association of additional subunits to a growing particle. The association between two subunits is mediated by one interface, however, given a trimer, each subunit make two

interactions which mutually support each other. Any further subunits joining the growing particle from this point make at least two further interactions and support the global stability of the particle. Furthermore, as addition is more likely than nucleation, this ensures that a reservoir of subunits is available for each nucleation event to rapidly progress to a full capsid.

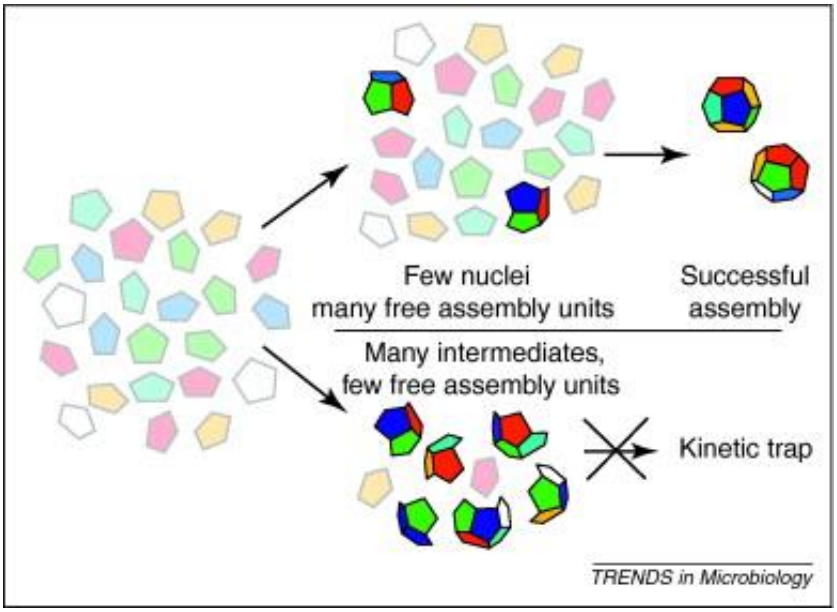



**Figure 1-12:** Efficient virus assembly and kinetic traps. Given few nucleation events full size capsid formation is achieved with pool of capsid protein. However, if nucleation occurs too frequently aberrant part capsids become the dominating species. Reprinted from[95]. Copyright 2011 with permission from Elsevier.

## 1.5.3 Viral Capsid Scaffolding

Scaffolding proteins are found in many viruses, for example the bacteriophage P22 contains a thermodynamically and kinetically contributive scaffold protein[102,103]. Scaffolding proteins can help mediate morphological changes required for the maturation steps that lead to the fully assembled virion[104]. These scaffolding effects can also be mediated through chain extensions of the virus protein coat, for example by donor strand exchange as seen with simian virus 40[105], or with coiled-coils, as seen with satellite tobacco necrosis virus (STNV)[106] (Figure 1-13). In the case of simian virus 40, these donor strands take up relatively different positions dependent on the quasi-symmetry of the particle's geometry. The STNV coiled-coils are just one and half heptads in length but act to help assemble the coat proteins by providing a stage for RNA binding.

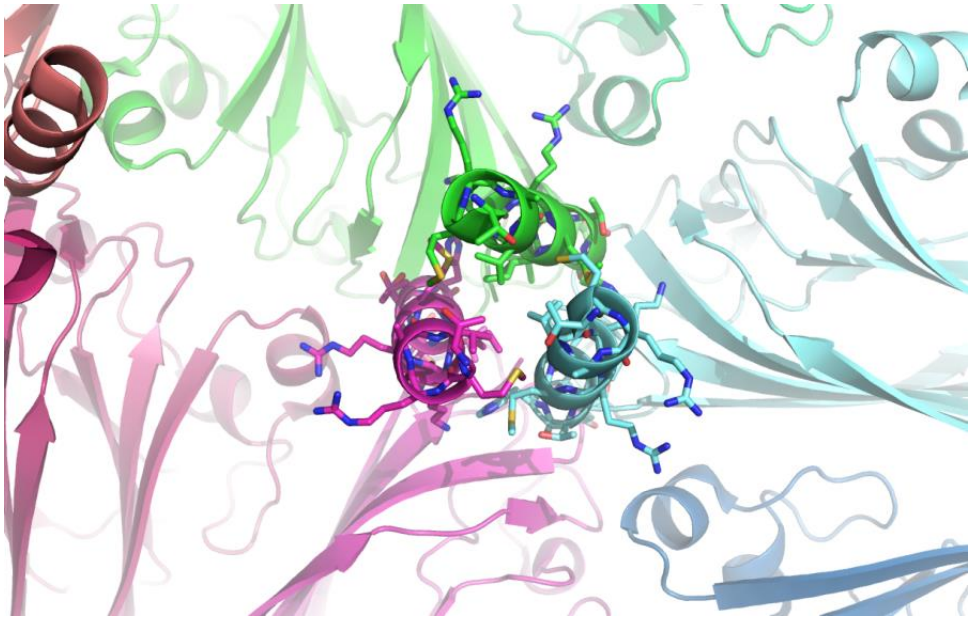


**Figure 1-13:** *N*-terminal scaffolding coiled-coil at the 3-fold axis of the satellite tobacco necrosis virus. From 1A34.pdb[106]

## 1.5.4 Pentagonal Tiling

Many virus particles associate to form structures with icosahedral symmetry such as the T=1, T=3, T=4 and T=7 structures. Each of these structures, barring the T=1 dodecahedron, requires a mixture of both pentagons and hexagons to complete the particle. However, the protein chosen for this project is just a pentagon. Does this restrict the size of the capsid to that of the dodecahedron? The short answer is no. Nature provides examples of viruses having the T=7d geometry in which five pentagons associate around a central pentagon to creates a 'super pentagon' which can then go on to form a 'super dodecahedron'. This structure is exemplified by the families of *papillomaviridae* and *polyomaviridae*. In theory, an intermediate structure comprising an octahedron of trimers of pentamers[107] (Figure 1-14a) is also accessible. The caveat of these T=7d assemblies, however, is the subunit interactions do not occur in a uniform manner. Figure 1-14b shows the tiling arrangement in the T=7d capsid, as seen for the simian virus 40. The grey pentamer sits are the centre of the 5-fold axis of symmetry and interacts with five pentamers. However, each of those five pentamers interacts with a total of six pentamers with different orientations of interactions. This unusual assembly arrangement is mediated by the scaffolding extensions which contribute the majority of the binding surface area. Therefore if we are to design a larger capsid from pentagons, we must follow suit.

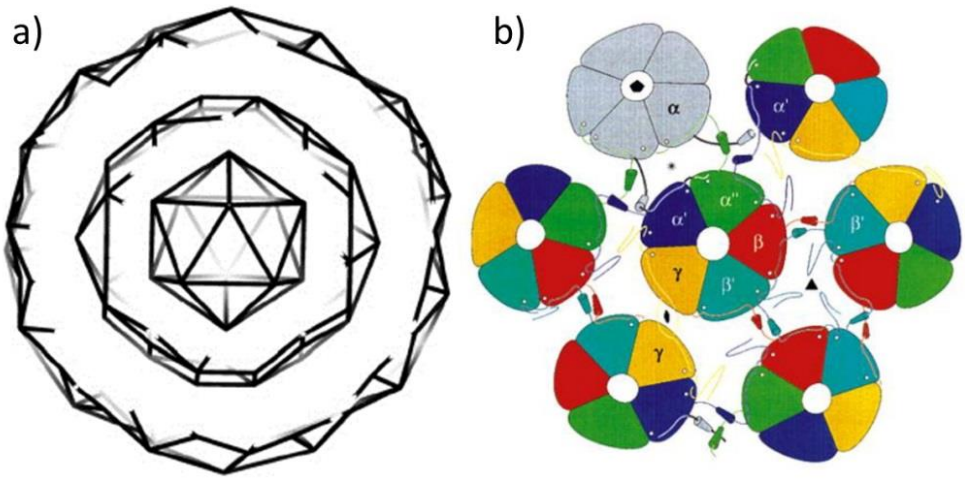



**Figure 1-14:** Particle options for pentagonal tiling capsids. **a)** A dodecahedron pentamers, an octahedron of trimers of pentamers and super dodecahedron. Reprinted from[107]. Copyright 2008, with permission from Elsevier. **b)** Pentagon tiling arrangement for the super dodecahedron. Reprinted from[105]. Copyright 1996, with permission from Elsevier.

## 1.5.5 Structural polymorphism

The capsids of almost all spherical viruses have icosahedral symmetry. Assembly of capsids without the viral genome often leads to misassembly steps which result in aberrant capsid structures[98]. As discussed, dependent on the type of virus, capsid proteins can often bind in different orientations to the same subunit. This feature, combined with Brownian fluctuations, can give rise to non-target structures. This point can be exemplified again with the SV40 virus: under non-physiological conditions, SV40 has been shown to assemble rapidly[108]. This phenomenon leads to a number of kinetic assembly traps producing particles of 22 nm icosahedra, spherical particles of 32-35 nm and even tubular structures[109].

The Handa group at the Tokyo Institute of Technology showed that SV40 assembled into a range of aberrant capsid structures under non-physiological conditions, 1 M NaCl and 2 mM $CaCl_2$ (pH 7.2) at room temperature[108] (Figure 1-15). This assembly is thought to be due to the high NaCl concentration which inhibits disassembly of the VP1 proteins. Furthermore, they found no evidence of any assembly occurring in the same conditions but using 150 mM NaCl.

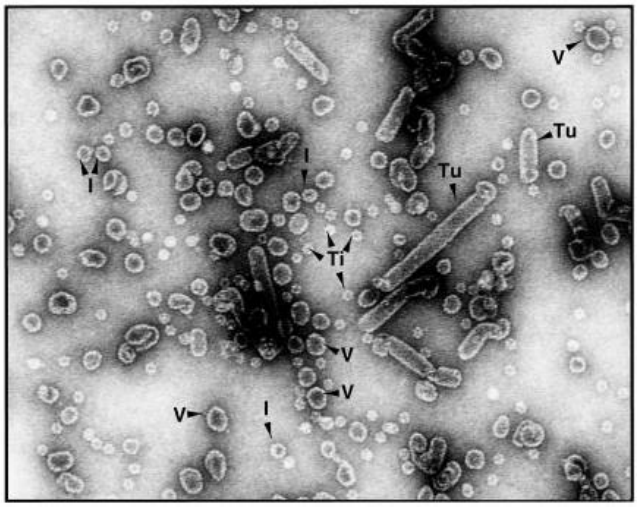



**Figure 1-15:** SV40 assembly in 1 M NaCl, 2 mM $CaCl_2$ (pH. 7.2) at room temperature. Tu are tubular structures, Ti are tiny particles, V are VLP, I are intermediate particles. Republished with permission of the Society for General Microbiology, from [108]; permission conveyed through Copyright Clearance Center, Inc.

The Zlotnick group at the Hebrew University-Hadassah Medical School, have recently shown that incubation of SV40 VP1 with differing sizes of DNA can bring about different capsid sizes[110]. This experiment highlights the role that scaffolding molecules play in virus capsid assembly. The group assembled VP1 on both ssDNA at 3200 nucleotide and dsDNA 2.4 kbp and 5.2 kbp (Figure 1-16). Perhaps unsurprisingly they found larger particles assembled around scaffolds of larger DNA.

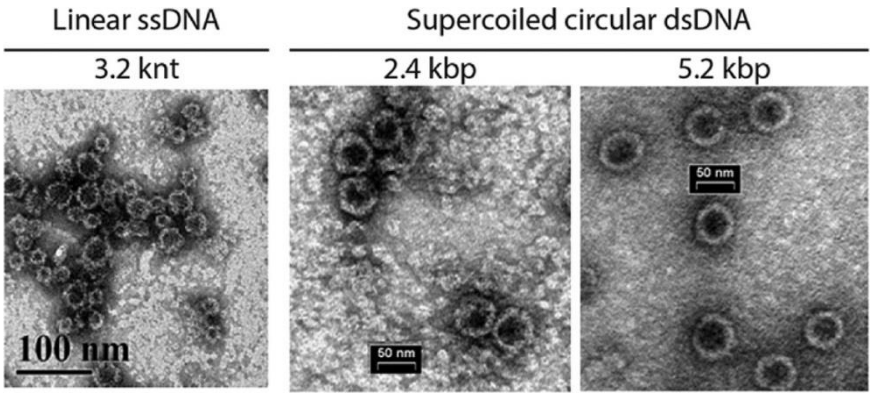



**Figure 1-16**: SV40 VP1 assembled around different sizes of DNA scaffolds, as denoted. Reprinted with permission from[110]. Copyright 2013 American Chemical Society.

Finally, Nguyen and Brooks, from the University of Michigan, used simulations to demonstrate a number of T=7 capsid polymorphisms that can theoretically occur[111]. They show that from T=1-7 that the aberrant structures produced are larger than the target structure, however from T=7-19, the aberrant structures created are often smaller than the target structure. Interestingly, T=7 capsids lie within both groups and thus should be able to form aberrant capsids that are either larger or smaller than the expected size.

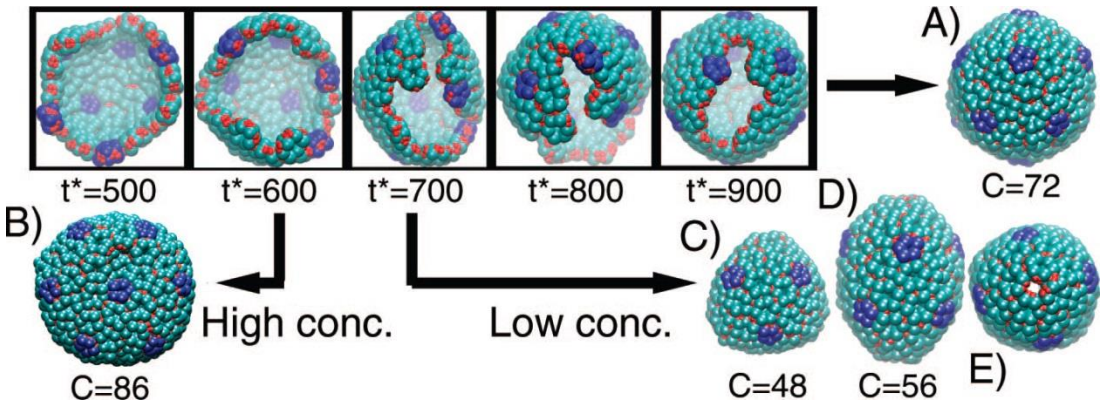



**Figure 1-17:** Snapshots of the growth of T=7 capsids over time at a concentration of 43.5 μM and 308 K. a) the target T=7 capsid. **b)** At high concentrations a larger capsid can form containing 14 extra coat proteins. **c-e)** At lower concentrations, smaller aberrant capsids can form. Structure **c)** is sometimes referred to as the 'intermediate' structure containing octahedral symmetry. Reprinted with permission from[111]. Copyright 2008 American Chemical Society.

# 1.6 Cholera Toxin

## 1.6.1 Overview

The starting point for the project design is the cholera toxin (CT) (Figure 1-18a), which is an $AB_5$ protein toxin of around 85 kDa naturally expressed by *Vibrio cholerae*. CT and the heat-labile enterotoxin (LT), from enterotoxigenic *E. coli*, cause cholera and 'travellers' diarrhoea', respectively, which induce severe diarrhoea and vomiting, leading to a loss of up to 30 litres of fluid a day[112]. Both are rife within the developing world and cause over a billion cases of diarrhoea annually[112,113], and together with, the shiga toxin and pertussis toxin, form the $AB_5$ protein toxin family which are responsible for over a million deaths annually[114].

## 1.6.2 Structure

CT and LT share 83% sequence identity and consist of an extremely stable[115], homopentameric doughnut-shaped B-subunit (Figure 1-18c,d) with a toxic A-subunit tethered through its central hole (Figure 1-18b)[116,117]. The toxic component of the A-subunit (CTA1) is a globular domain which is linked to the B-subunit pentamer via an α-helical rod (CTA2) that is a non-toxic proteolytic fragment of the A-subunit[118]. The B-subunit protomer (the monomeric unit) consists of a large α-helix (central, within the pentamer), and two adjoining anti-parallel β-sheets. These sheets twist in such a way as to make parallel β-sheet interactions in the centre of the molecule; however, they are almost perpendicular to each other at the edges of the protomer.

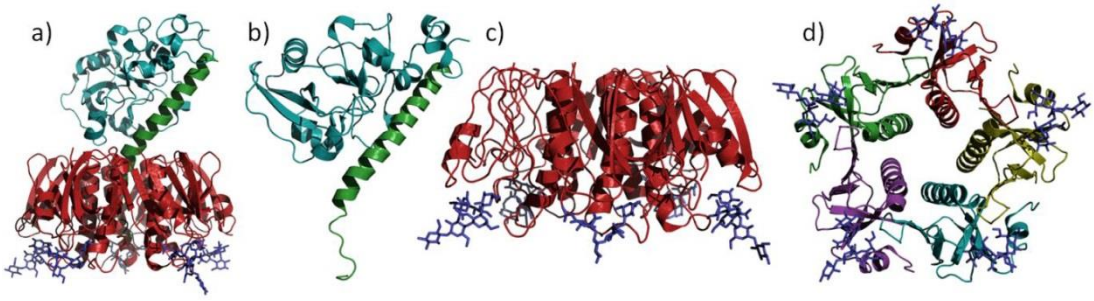



**Figure 1-18:** Structures of cholera toxin (composite of PDB 1S5E and 3CHB), however all $AB_5$ toxins have this same generic structure. **a)** The $AB_5$ holotoxin, A-subunit in teal and green, B-subunits in red and GM1 ganglioside in dark blue. **b)** The A-subunit, A1 in teal and A2 in green. **c)** The B-subunit pentamer, in red with GM1 in dark blue. **d)** A top down view of the B-subunit pentamer revealing its five-fold symmetry, pentagonal shape and GM1 binding sites, in dark blue.

Two β-sheets from adjoining protomers bind in an anti-parallel fashion to form a virtually unbroken β-sheet core in a ring around the pentamer. Each protomer contributes residues to a carbohydrate binding site located distally to the A-subunit binding site and situated on the seam of the dimeric interaction of the protomers. This binding site recognises the GM1 ganglioside, (Galβ3-GalNAcβ4(Neu5Acα3)Galβ4GlcβCeramide (Figure 1-19), which is found on the surface of intestinal epithelium cells[119]. The five binding sites on the toxin[120] have one of the highest affinities known for protein-carbohydrate interactions ($K_d$ = 43 nM)[121]. The porcine variant of LTB (pLTB) is also know to bind neolacto-(Galβ4GlcNAcβ-)terminated glycoconjugates and the human variant (hLTB) is known to bind blood group A type 2-(GalNAcα3(Fucα2)Galβ4GlcNAcβ-)terminated glycoconjugates[122,123].

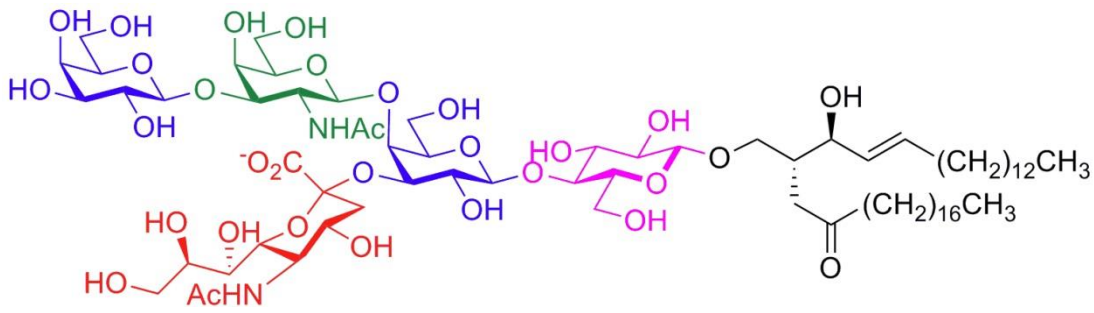



**Figure 1-19:** Structure of the GM1 Ganglioside: GM1 is a pentasaccharide containing two galactose (blue), one N-acetylgalatosamine (green), one glucose (pink) and one sialic acid (red) residue. The ceramide lipid forms the linker attaching the oligosaccharide to the cell surface[124]. Image obtained from Dr Bruce Turnbull, University of Leeds.

### 1.6.3 Expression and Secretion

The CT gene cassette, ctxAB, is found on a virulence plasmid naturally occurring in *Vibrio cholera* as well as the CTXϕ phage of *V. cholera*, which is responsible for horizontal gene transfer of the cassette[125]. The genes for the A-subunit and B-subunit are found to overlap in a polycistronic gene where the Shine Dalgarno sequence for the B-subunit is located within the 3' district of the CTA gene[126]. CTB has a propeptide leader sequence which targets CTB expression to the periplasm. In the periplasm the leader sequence is cleaved and a disulfide bond is formed between C9 and C86. Once correctly folded, the CTB protomers associate around the CTA2 C-terminal domain to form the holotoxin which is secreted. The secretion of CTB involves the type 2 secretion system (T2SS), a sophisticated piston driven mechanism seen in Figure 1-20[127]. The T2SS is a macromolecular complex which spans both the inner and outer membrane of the periplasm and is capable of secreting fully folded proteins. The T2SS contains a dodecameric hollow shaft which loads the CT holotoxin; a pseudopilus is then used to purge the complex.

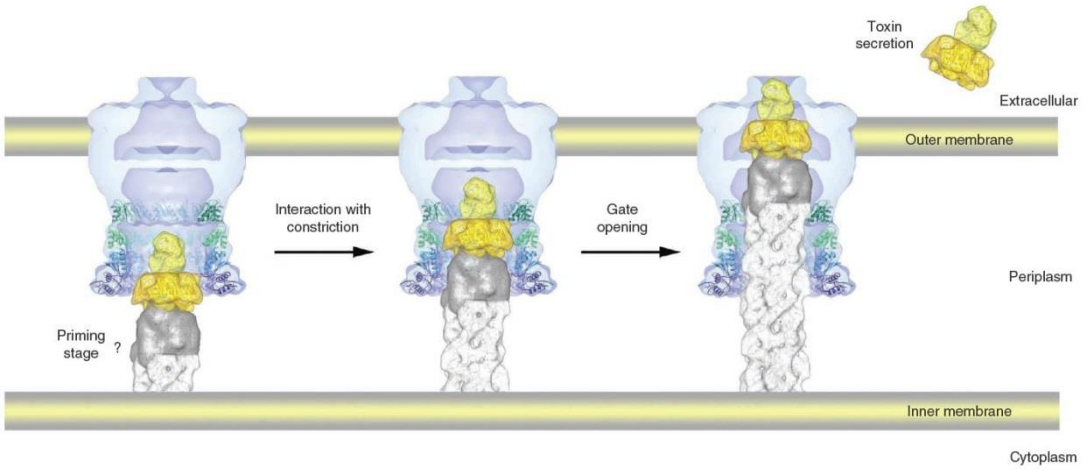



**Figure 1-20:** Structure of the Type 2 Secretion System: Proposed mode of action of the T2SS, CT (yellow) binds to the pseudopilus tip (grey) while the secretin (blue) is closed or directly to secretin. Pseudopilus tip extension pushes the CT free from the complex, whereby CT dissociates from the pseudopilus tip. Reprinted by permission from Macmillan Publishers Ltd: Nat. Struct. Mol. Biol[127] , copyright 2010.

## 1.6.4 Function

Once secreted from the pathogen the peptide backbone between A1 and A2 chains is cleaved by host proteases in the case of LT, or secreted proteases in the case of CT, leaving only a disulfide bond between the A1- and A2-subunits[125]. Binding to GM1 at the cell surface stimulates receptor-mediated endocytosis of the AB5 toxin[128]. The toxins which have a KDEL or QDEL sequence at the C-terminus of CTA2, follow the retrograde translocation pathway through the Golgi to the Endoplasmic Reticulum (Figure 1-21). There, protein disulfide isomerise is thought to break the remaining disulfide bond between CTA1 and CTA2[118,129] and help shuttle CTA1 into the cytosol via the sec61p channel.

Once CTA1 enters the cytosol it ADP-ribosylates adenylate cyclase, thus mediating constitutive activation of the enzyme. The resulting build-up in cAMP leads to an expulsion of salts and movement of water out of the cells into the intestinal lumen[130]. CT and LT thereby cause life-threatening diarrhoea. In the absence of the A-subunit, the B-subunit pentamer is not toxic; however, it continues to trigger receptor-mediated endocytosis and gains access to cell's interior.

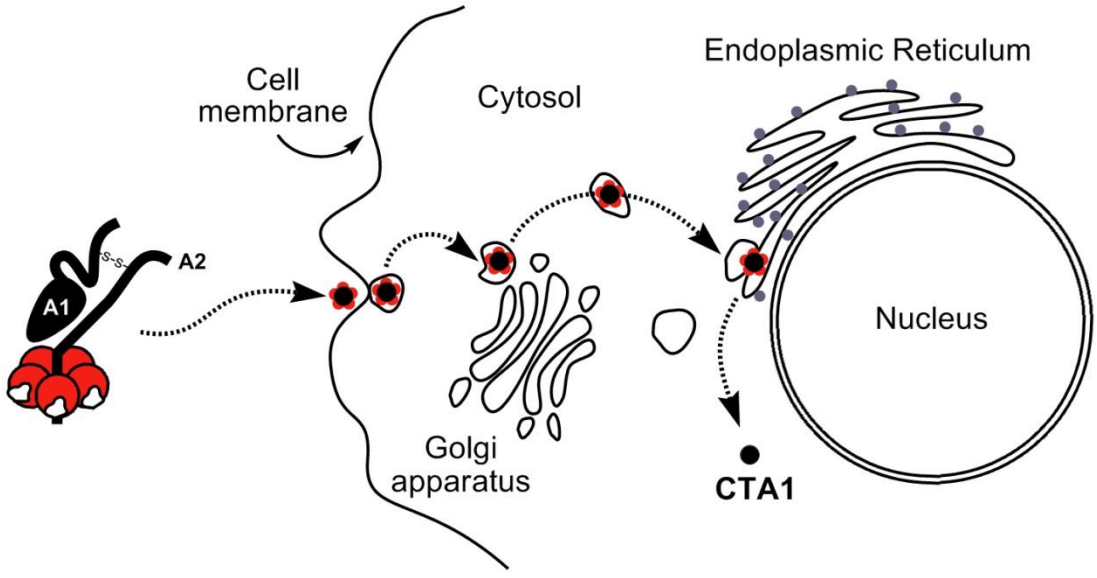



**Figure 1-21** The CTB subunit binds to cell surface GM1 gangliosides which trigger the endocytosis of the holotoxin. The C-terminal residues of the CTA subunit target the vesicle into the retrograde translocation pathway, through the Golgi to the endoplasmic reticulum. Here the disulfide bond between CTA1 and CTA2 is cleaved and CTA1 is shuttled into the cytoplasm where it binds to adenylate cyclase, thus mediating its constitutive activation. This leads to a build-up in cAMP and transport of salts out the cell, these ions are followed osmotically by water[124]. Image obtained from Dr Bruce Turnbull, University of Leeds.

## 1.6.5 CTB as a Target for Bionanotechnology

CT, CTB, LT and LTB are regularly expressed using *Vibrio sp. 60*, which is a non-pathogenic relative of *V. cholera*[131]. Co-expression of CTA and CTB has been found to give up to five-fold increase in overall CTB production[132]. It is thought that the presence of the A-subunit facilitates the B-subunit association, reducing degradation (by reducing free protomer) and thereby increasing the relative expression of the B-subunit. As the protein is naturally secreted, the B-subunit pentamer or holotoxin is isolated from the media supernatant by filtration, affinity purification or precipitation. The benefit of this procedure is a relative lack of contaminating proteins; however, this advantage is countered by the substantial dilution of the protein in the media. Some groups report purification from the periplasm[133], but no groups have reported expression from the cytoplasm.

Due to the abilities of these proteins to induce endocytosis and their high stability, they are uniquely positioned to make good oral carriers for transport of products around the body[134]. Furthermore, this high structural stability suggests that these proteins could harbour more mutations compared to the average protein without compromising its stability, making these proteins attractive design targets. The three most similar B-subunits of the $AB_5$ family are CTB, LTB (83% sequence identity to CTB El Tor biotype 3)[135], and CFXB, a cholera-like toxin found in *Citrobacter freundii* (73% sequence identity to the CTB El Tor biotype 3)[136]. Each of these proteins exist as several different strains, harbouring many different amino acid variations at the positions of the protomer structure. The full sequence of the CTB classical Biotype and some homologues are shown in Figure 1-22.

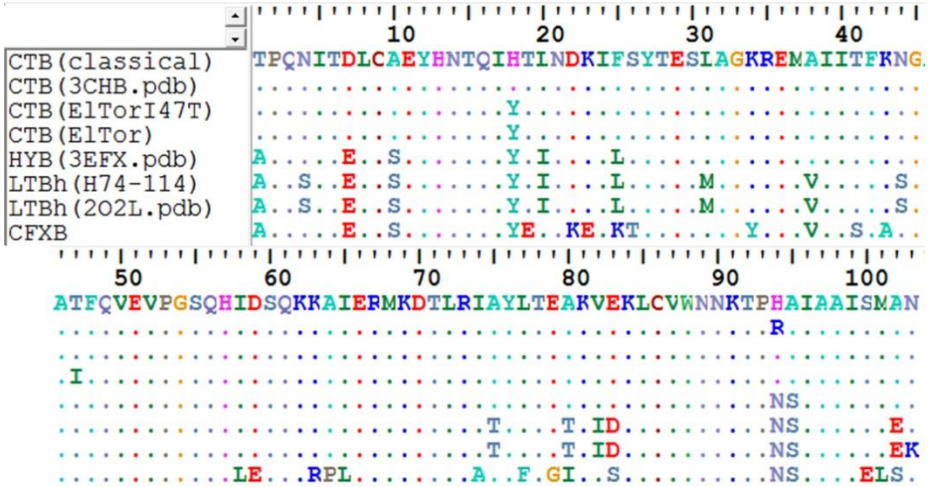



**Figure 1-22:** Cholera toxin B-subunit (CTB), heat labile enterotoxin (LTBh) and cholera-like toxin found in *Citrobacter freundii* (CFXB) of various strains. Full sequence for CTB classical biotype is shown in colours corresponding to each residues biophysical characteristic. Identical residues displayed as dots. Note subtle differences within species and larger changes between species. Figure produced with Bioedit.

Mutational studies have been carried out on CTB and LTB since the 1990s mapping the expression, transport, assembly, secretion and virulence conferred by the change[137-140]. Particular interest has been paid to the GM1 binding site as work has concentrated on developing therapeutic inhibitors to prevent holotoxin endocytosis, thus the positions 33, 34, 35 and 88 have been thoroughly mapped. By combining the information from naturally occurring variants and human mutated constructs, a map of the proteins' mutational space can be fashioned which shows the levels of malleability currently known to maintain features such as pentamerisation and holotoxin formation, etc. shown in Figure 1-23.

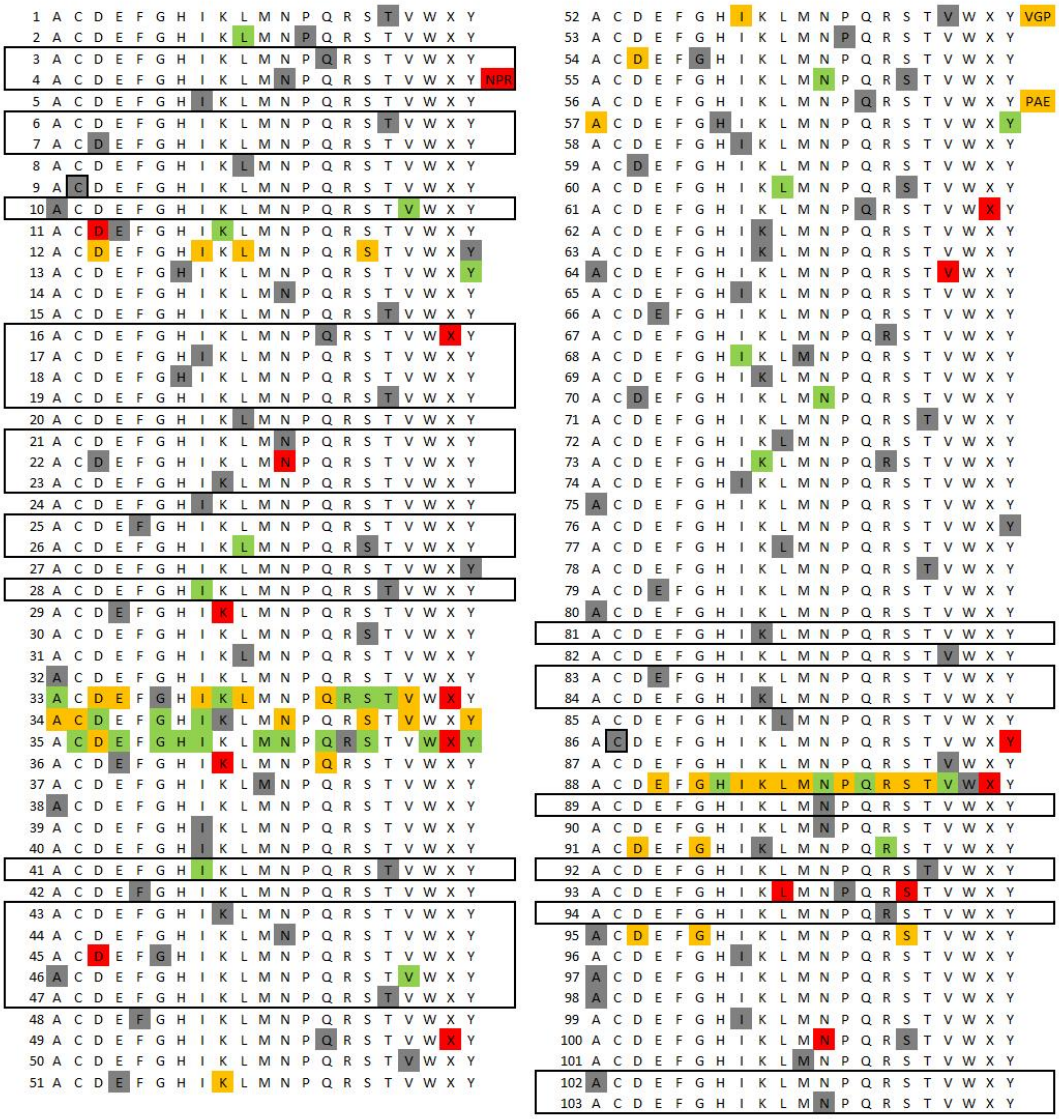


**Figure 1-23:** Mutational Space Map: The map shows the residue position down the left-hand side, and lists all possible amino acids adjacent to the position number (X = stop codon). The CTB El Tor biotype 3 residues are highlighted in dark grey. Light grey positions indicate deviations from this CTB sequence made by close sequence homologues (see figure 1.6). Coloured residues represent mutants described in the literature as able to secrete as holotoxin (green); able to form pentamers but not the holotoxin (orange); and as not leading to expression (red). The NPR, VGP and PAE are C-terminal extensions.[137-140]

# 1.7 Project Summary

In this project, we aimed to construct a novel virus-like protein capsid to test the ability of current computational techniques to estimate changes in free energy of a PPI upon mutation as well as to create a useful therapeutic delivery device. The principles used in its design and validation will help lay the foundations for further rational protein design routines that are generically applicable. As a secondary aim, the VLP constructed could be useful as a drug delivery vehicle or nano-particle scaffold for further augmentation.

Cholera Toxin B-subunit (CTB) is a pentameric, pentagonal protein and, by its fivefold symmetry, can be used as a building block for a dodecahedron containing 12 pentamers or larger structures which comply with 3D pentagon tiling theory. We aim to exploit the CTB property which mediates endocytosis at epithelial cell surfaces in our capsid structure. CTB tapers naturally at an angle which would provide a large interaction surface for a capsid-like assembly. Utilising this natural tapering to create a curved surface would also present the CTB GM1 ganglioside binding site on the outside of the particle, which should allow triggering of endocytosis, as found in the wild-type protein.

In order to achieve this assembly process our CTB subunits were modified with C-terminal coiled-coil 'scaffolding' extensions (Figure 1-24) to bring about self-association. An *in silico* symmetrical docking protocol was developed to assess the most likely interaction face for our assembled subunits, based on which a list of interacting residues was constructed. These residues were used for an *in silico* saturation mutagenesis experiment to identify which mutations would contribute to binding of the subunits. Hits from the computational process where expressed *in vitro* with the coiled-coil extensions and characterised for their ability to form higher order structures by size exclusion chromatography, dynamic light scattering and transmission electron microscopy.

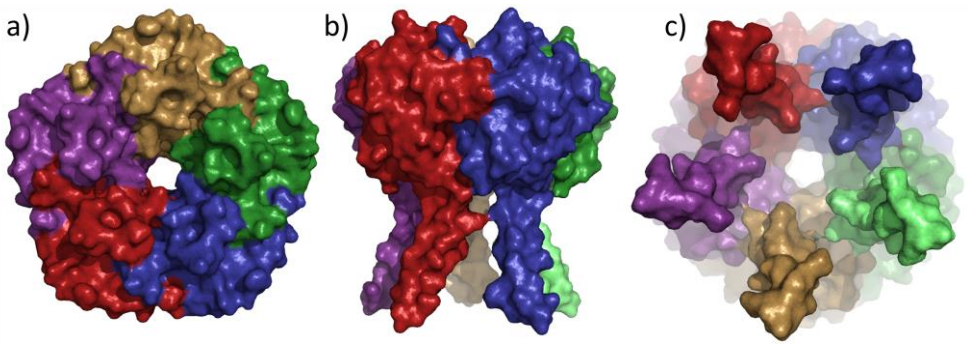



**Figure 1-24:** Modified CTB with *C*-terminal coiled-coils (3-heptad). **a)** Top-down view of CTB. **b)** Side-on view of CTB revealing coiled-coils. **c)** Bottom-up view of CTB. Figure produced from modified pdb 3CHB.

# Chapter 2

# Materials and Methods

# 2.1 Experimental

## 2.1.1 Materials

### 2.1.1.1 Instrumentation

PCR reactions were conducted using a Techne Thermocycler TC-512. Glassware, plasticware and media were sterilised in either a Prestige Medical bench-top autoclave or a LTE Touchclave-R autoclave. Sterile environments were mediated with a Thermo Electron Corporation SAFE 2010 Class II laminar flow cabinet or a bench-top bunsen. Incubation of mini-cultures, starter cultures and transformations were performed with a Stuart Orbital Incubator, petri dishes and digestions were incubated in a Binder BD23 incubator and scale-up cultures were incubated in a Kuhner Climo-Shaker ISF1-X. A Bio-Rad Geldoc XR was used for the UV visualisation of ethidium bromide-stained agarose gels and stained protein gels under white light. Centrifugation of volumes up to 1.5 mL was conducted with a Heraeus Biofuge *pico*, volumes up to 50 mL with a Heraeus Multifuge 3 S-R and volumes up to 500 mL in a Beckman Coulter Avanti J-30I. SDS-PAGE was performed in a Bio-Rad Mini Protean Apparatus. Mass spectrometry was conducted with a Bruker HCT Ultra MS System with an Agilent 1200 Series Auto Sampler. DSF experiments were performed with a Bio-Rad CFX96 Real-Time PCR Detection System. SEC was conducted on a GE healthcare ÄKTA FPLC system. DLS was performed with a Malvern Zetasizer Nano ZS. TEM was conducted on a Tecnai G2 Spirit electron microscope.

### 2.1.1.2 General Materials

General reagents were purchased from Sigma Aldrich, Fisher Scientific, VWR International, Roche Diagnostics, Generon, Merck Chemicals Ltd and MP Biomedicals Europe. DNA Ladders were obtained from New England Biolabs and protein molecular weight markers were purchased from Fisher Scientific. Protein gels were stained with Instant Blue (Triple Red Limited). Antibiotic stock concentrations of 1000× were prepared from 100 mg/mL ampicillin supplied by Fisher Scientific. Pwo Polymerase, stored at 3 U/µL, was prepared by Dr Louise Kime. All other DNA modifying enzymes were purchased from New England Biolabs. The gel stain Sypro Orange (5000×) was obtained from Sigma Aldrich and used at a working concentration 10× in DSF experiments. Ni Superflow Plus Resin and Ni-NTA resin was purchased from Generon and Qiagen respectively. α-Lactose Agarose resin was supplied by Sigma-Aldrich, amylose-agarose from New England Biolabs and a Superose 6

10/300 GL SEC column was obtained from GE Healthcare. Sequencing reactions were carried out by GATC Biotech via ABI 3730xl Sanger sequencing.

### 2.1.1.3 Buffers

**Pfu Buffer (10×):** 200mM Tris-HCl (pH 8.8 at 25 °C), 100 mM $(NH_4)_2SO_4$, 100 mM KCl, 1% (*v/v*) Triton X-100, 1 mg/ml BSA and 20 mM $MgSO_4$.
**TAE Buffer (50×):** 2 M Tris base, 5.71% (*v/v*) acetic acid, 0.5 M EDTA.
**SDS-PAGE Running Buffer (5×):** 125 mM Tris-HCl, 960 mM Glycine, 0.5% (*w/v*) SDS
**SDS-PAGE Sample Buffer (2×):** 100 mM Tris-HCl, 0.1% (*w/v*) Bromophenol Blue, 200 mM DTT, 20% (*v/v*) Glycerol, 4% (*w/v*) SDS
**Nickel Suspension Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl, 20 mM Imidazole
**Nickel Elution Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl, 500 mM Imidazole
**Lactose Suspension Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl
**Lactose Elution Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl, 300 mM Lactose
**Maltose Suspension Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl
**Maltose Elution Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl, 10 mM Maltose
**SEC Buffer:** 100 mM Tris-HCl (pH 7.4 at 25 °C), 100 mM NaCl
**EB Buffer:** 10 mM Tris-HCl, pH 8.5

### 2.1.1.4 Nucleotide Synthesis

Oligonucleotides from 18 to 95 nt were purchased from Integrated DNA Technologies. These included oligonucleotides for priming PCR reactions against DNA templates, for sequencing or for the *de novo* construction of the CTB gene, and its extensions, by assembly PCR.

### 2.1.1.5 Heat Shock Competent Cell Strains

Heat shock competent *E. coli* strains XL10 and TG1 were obtained from Stratagene and Lucigen respectively. Incompetent DE3 expression stains BL21 Star*, BL21 Gold, C41 and C43 were provided by Dr Thomas Edwards, University of Leeds, which all contained the pRARE2 plasmid. The pRARE2 plasmid encodes the enhanced production of tRNA for a variety of underrepresented codons in *E. coli,* levelling the codon bias.

## 2.1.2 General DNA Manipulation

### 2.1.2.1 Starter Culture or Mini culture

Mini-cultures were prepared by adding 5 mL of sterilised LB media to a 50 mL falcon tube containing 100 µg/mL ampicillin. Either a single colony was picked from a petri dish, a glycerol stock scraping was added or 5µL of liquid culture was added to the falcon tube followed by agitation. The culture was incubated for approximately 16 hours at 37 °C and 200 rpm.

### 2.1.2.2 Isolation of Plasmid DNA

Plasmid DNA was obtained from cells isolated from overnight mini-cultures pelleted via centrifugation for 10 minutes at 6000*g*. DNA extraction was conducted by alkaline lysis using the Qiagen plasmid mini-preparation kit eluting with 50 µL of EB buffer.

### 2.1.2.3 Polymerase Chain Reaction Amplification

General PCR amplifications were performed with a final concentration of 1× Pfu Buffer, 0.6 U *Pwo* Polymerase, 0.8 mM dNTP mix, approximately 10 ng of template DNA, 1 µM priming oligonucleotides and made to a final volume of 50 µL with water. This mixture was constructed adding the polymerase last, and mixed well, just before initiation of the thermal cycling. Table 2-1 describes a general PCR program. As standard after each PCR, the mixture was purified using the Qiagen PCR clean-up kit.

| Cycle | Process | Time (s) | Temperature (°C) |
|---|---|---|---|
| 1 | Initial denaturation | 180 | 95 |
| 30 | denaturation | 30 | 95 |
| - | Annealing | 30 | Primer $T_m$ -2 |
| - | Elongation | 60 per 500 bp | 70 |
| 1 | Final Elongation | 2.5× Elongation | 70 |
| 1 | Hold | - | 4 |

**Table 2-1**: Standard PCR reaction. This table shows the generalisations made to the method which most PCR reactions require.

**Assembly PCR –** Also known as Polymerase Cycling Assembly, allows the construction of a large section of dsDNA from short overlapping ssDNA molecules. This technique was used to construct a subset of mutant CTB genes. As shown in Figure 2-1a, the gene is routinely constructed from up to eight internal segments in the presence of an excess of flanking primers. PCR amplification leads to the production of the full length gene. In some cases a second amplification is conducted using the product of the first PCR and the flanking primers to obtain large quantities of the full length gene. The exact same method used for the wild-type gene construction was used to construct the CTB mutants, however the wild-type parts were substituted for mutant parts. This technique led to a nomenclature based on the parts used to construct each gene. CTB was constructed from six parts, whose wild-type parts were labelled `A`, the wild-type can be written `AAAAAA`, however, the mutant exemplified in the Figure 2-1b, which contains mutant parts, can be written as `IQVAGD`.

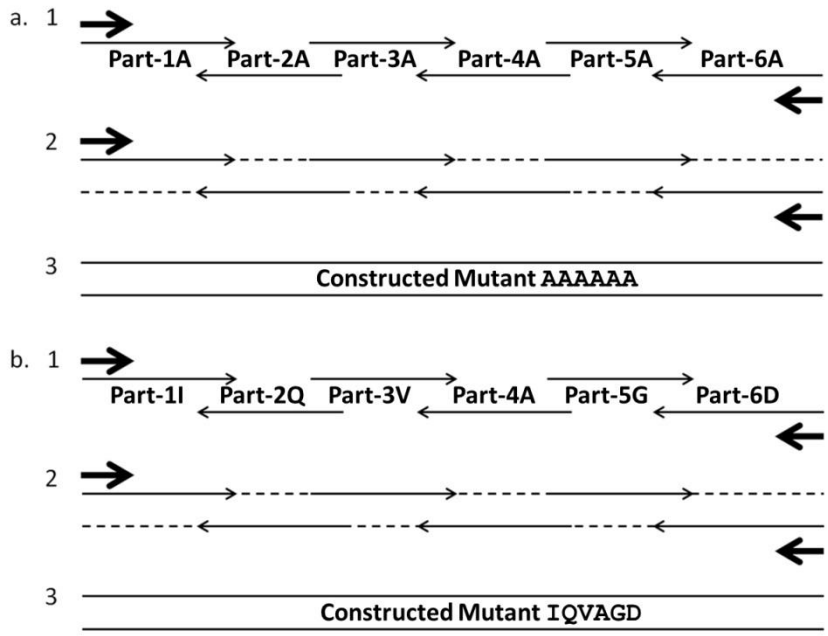



**Figure 2-1:** Assembly PCR. Short overlapping oligonucleotides, routinely altering between sense and anti-sense strands are assembled into one large section of DNA during a modified PCR program. Typically the concentrations of the parts are quite low, however, terminal primers are in excess, thick arrows, and these terminal primers dominate the amplification once the parts have bridged the gap between them. **a)** Construction of the wild-type gene by this technique. **b)** Construction of mutant genes. This technique allows multiple mutations to be made simultaneously and allows the construction of large potential libraries.

An alternative wild-type part 4B contains silent mutations to allow binding of a amplification primer was frequently used to aid subcloning. Example oligonucleotides for construction of the wild-type CTB gene are described in Table 2-2, mutant parts are available in the appendix.

| Part | Sequence |
|---|---|
| Part-1A-T | CTGTTCAGGCGCATG |
| Part-1A | CTGTTCAGGCGCATGCAACTCCTCAAAATATTACTGATTTGTGCGCAGAATACCACAACAC |
| Part-2A | CTCTTTTTCCCGCTAGCGATTCTGTATACGAAAAGATCTTATCATTTAGCGTATATATTTGTGTGTTGTGGTATTCTGCG |
| Part-3A | GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC |
| Part-4A | GCAATCCTCAGGGTATCCTTCATCCTTTCAATCGCTTTTTTTTGTGAATCTATATGTTGACTACCTGGTACCTCTACTTG |
| Part-5A | GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAA |
| Part-6A | CGAATTACCTGCAGGGAAAACTTAGTTTGCCATACTAATTGCGGCGATCGCATGAGGCGTTTTATTATTCCATACACATAA |
| Part-6A-T | CGAATTACCTGCAGGG |

**Table 2-2:** Example of the overlapping oligonucleotides required to assemble the full length CTB gene using the Assembly PCR method.

Assembly PCR was performed with a final concentration of 1× Pfu Buffer, 0.6 U Pwo Polymerase, 0.8 mM dNTP mix, 1 µM terminal priming oligonucleotides, 10 nM oligonucleotide parts (one each of parts 1-6 or 1-8) and made to a final volume of 50 µL with water. This mixture was prepared without polymerase, which was added shortly before initiation of the thermal cycling. The CTB Assembly PCR program can be seen in Table 2-3. As standard after each PCR, the mixture was purified using the Qiagen PCR clean-up kit.

| Cycle | Process | Time (s) | Temperature (°C) |
|---|---|---|---|
| 1 | Initial denaturation | 120 | 95 |
| 30 | denaturation | 30 | 95 |
| - | Annealing | 30 | 47 |
| - | Elongation | 60 | 70 |
| 1 | Final Elongation | 150 | 70 |
| 1 | Hold | - | 4 |

**Table 2-3:** Specific program for Assembly PCR of the CTB gene. This program can be used for both six part and eight part assemblies and also for the second round amplification containing flanking primers.

Assembly PCR was used to construct approximately half of the alanine scan mutants, the coiled-coil genes for MBP *C*-terminal extensions as well as all additional mutant CTB genes both with and without coiled-coils.

**Clonal Screening –** The presence of the correct insert was confirmed by PCR screening using primers Part-1A-T and part-4B-T. Part-4B-T is complimentary the sequence of Part-4B which is uniquely incorporated into assembly constructs.

**QuikChange Mutagenesis –** Routine site-directed mutagenesis of constructs was carried out using the Stratagene QuikChange protocol with a pair of complimentary primers containing the desired mutation. The PCR reactions generally described above were modified with a 14 min elongation time step. Following thermal cycling 20 U of DpnI was used to digest the methylated template DNA as described for the digestion protocols. QuikChange mutagenesis was used to construct approximately half of the alanine scan mutant genes.

### 2.1.2.4 Restriction Digest

**Basic Digest** – Restriction digests of PCR products and inserts were routinely carried out as follows. A 10 µL reaction mixture was prepared containing up to 1 µg of plasmid DNA, with 1× of the appropriate NEB buffer and 5-10 U of each restriction enzyme. Reactions were incubated at 37 °C for 1 h, after which samples were analysed 1% agarose. Analytical

double digests were conducted by repeating the above for each single enzyme and constructing a double digest containing 5-10 U of both enzymes. Digested samples were heat inactivated at 80 °C for 25 minutes or purified using the Qiagen PCR clean-up kit.

**Plasmid Digest for Ligation** – Restriction digestion, to generate linear vectors, was conducted as for PCR inserts, however the reaction mixture was scaled to allow 5-10 U of endonuclease per µg plasmid DNA and 5-10 U of Antarctic phosphatase per µg plasmid DNA. The digested samples were run on a 1% agarose gel. The gel was visualised briefly and the linearised plasmid band was excised with a clean scalpel and purified with the Qiagen Gel Extraction Kit.

### 2.1.2.5 Ligation

Ligation mixtures of 10 µL were constructed with 50 ng of linear plasmid vector, a 1:3 molar ratio of vector to insert, 1× Ligase Buffer, 20 U of T4 DNA Ligase and made up to 10 µL with $dH_2O$. These reactions were incubated at room temperature for 30 minutes before storage at 4°C prior to transformation.

### 2.1.2.6 Preparation of Heat shock competent cells

Chemically competent *E. coli* strains were grown and prepared as follows. 100 mL of LB media was inoculated with 100 µL of starter cultures of the required cell type using chloramphenicol for strains containing the pRARE2 plasmid. These cultures were incubated at 37 °C to a final $OD_{600}$ of 0.3 – 0.4. The cultures were immediately placed on ice. The cells were isolated by centrifugation in a pre-cooled rotor at 6000*g* and 4 °C for 10 min and the supernatant was discarded. The pellet was resuspended in 10 mL of cold, sterile 0.1 M $CaCl_2$ and incubated for 10 min on ice before isolating the cells again at 6000g and 4 °C for 10 mins. The pellets were resuspended in 4 mL of cold 0.1 M $CaCl_2$, 20% glycerol (*v/v*) and aliquoted into 100 µL fractions which were flash frozen in liquid $N_2$ and stored at -80 °C.

### 2.1.2.7 Transformations

5 µL of freshly thawed commercial chemically competent *E. coli* XL10 cells, or 20 µL of cells prepared as described above, were added to pre-cooled micro-centrifuge tubes and held on ice. 1 µL of plasmid DNA or 2 µL of ligation mix was added to the cells and mixed. The cells were incubated on ice for 10 min before being exposed to 42 °C for 45 s in a water bath. The samples were then returned to ice for a further 10 min. 990 µL of LB media was added to the samples and the culture was incubated at 37 °C for 1 h. 100 µL of the culture was applied to and spread on an LB-agar petri dishes of appropriate antibiotic resistance.

The remaining culture was centrifuged at 13000*g* for 4 min and decanted to leave 100 µL of media, the pellet was then resuspended and the cells were spread on LB-agar plates using sterile glass beads.

### 2.1.2.8 Glycerol Cell Stocks

Stock solutions of cells were prepared from mini-cultures, 500 µL of the culture was added to 500 µL of 80% glycerol, mixed and flash frozen in liquid $N_2$ before long term storage at -80 °C.

### 2.1.2.9 Agarose Gel (1%)

**Small gel –** 40 mL of TAE buffer was added to a 250 mL conical flask containing 400 mg of agarose which was then heated in a microwave until the agarose was fully dissolved. The molten agarose was decanted into a 50 mL falcon tube and made up to 40 mL with TAE buffer. This was held at 50 °C while the gel cast and comb were prepared. 1.5 µL of ethidium bromide was added to the agarose, which was then gently agitated and the agarose was poured into the cast. The agarose gel was fully set within 20 min. All DNA samples were prepared to give final loading buffer concentration of 1×. 10 µL of sample was loaded as default onto the gel. Gels where run at 100 V for 20 min as standard.

**Large gel** – 200 mL of TAE buffer was added to a 500 mL conical flask containing 2 g of agarose which was then heated in a microwave until the agarose was fully dissolved. This was held at 50 °C while the gel cast and comb were prepared. 7.5 µL of ethidium bromide was added directly to the gel cast and the agarose poured on top ensuring complete mixing of the ethidium bromide. The agarose gel was fully set within 40 min. All DNA samples were prepared to give final loading buffer concentration of 1×. 10 µL of sample was loaded as default onto the gel. Gels where run at 100 V for 40 min as standard.

### 2.1.2.10 Plasmid Gene Expression

Starter cultures were prepared from glycerol cell stocks. 3 mL of a starter culture was added per 1 L of autoclaved LB media containing 100 µg/mL of an appropriate antibiotic. Incubation was conducted at 37 °C and 200 rpm for 2½ hours. The optical density was then measured to reach 0.6-0.8 at λ600 nm. Induction was initiated with the addition of a final concentration of 300 µM IPTG. Incubation was then conducted at 30 °C and 200 rpm for 16 hours.

### 2.1.2.11 Fractionation

Overnight inductions were centrifuged at 10000*g* for 10 min, the pellet was discarded and the media supernatant was poured into a 2 L beaker per 1 L culture. 570 g of ammonium sulfate was added per 1 L initial culture. The ammonium sulfate was added slowly to stirred media and then incubated at room temperature for at least 2 hours while stirred. The resultant precipitated media was then centrifuged at 17500*g* for 30 min to collect the pelleted protein. The supernatant was discarded. This pellet was resuspended in 75 mL suspension buffer per 1 L of initial culture. These re-suspensions were transferred into 50 mL falcon tubes and centrifuged at 6000*g* for 5 minutes, the supernatant was passed through a 0.5 µm vacuum filter and then stored at 4 °C

### 2.1.2.12 Purification

Gravity drip columns were assembled with an appropriate affinity resin and allowed to form a bed. A syphon was assembled from a reservoir to the column. 5 column volumes of water was washed through the column, then 5 column volumes of suspension buffer. The precipitant fraction was then applied to the column and the flow through was collected. The column was then washed with 5 column volumes of suspension buffer. Elution was commenced with 5 column volumes of the appropriate elution buffer and fractions of 1-5 mL were collected, dependent on the experiment. This protocol was repeated with further samples or the column was washed with 5 column volumes of water, then 5 column volumes of 20% ethanol and stored in 20% ethanol at 4 °C. The fractions collected had EDTA free protease inhibitor cocktails added and were stored at 4 °C or incubated at room temperature as part of an assembly experiment.

### 2.1.2.13 SDS-PAGE (12%).

**Gel Preparation** - Clean glass slides were set up to cast the gel. A 15 mL, 12% resolving gel mix was constructed with 6.385 mL of $dH_2O$, 3.80 mL of 1.5M Tris-HCl pH8.8, 4.5 mL of 40% acrylamide, 150 µL of 20% SDS, 150 µL of 20% APS and 15µL of TEMED. This mixture was agitated and poured into the gel cast, a layer of isopropanol was applied and the gel was allowed to set for 10 min. Excess isopropanol was removed and the gel was thoroughly washed by applying a stream of water to the top of the glass slides. A 5 ml, 4% stacking gel mix was constructed with 3.325 mL of $dH_2O$, 0.945 mL of 0.5M Tris-HCl pH6.8, 0.625 mL of 40% Acrylamide, 50 µL of 20% SDS, 50 µL of 20% APS and 5 µL of TEMED. This mixture was agitated and applied to the gel cast, a comb was added and the gel was allowed to set for 20 min. Gels were used immediately.

**Gel Running** – A 12% SDS-PAGE gel was assembled in the running assembly. This assembly was filled with SDS-PAGE running buffer. The protein samples were mixed 1:1 with sample buffer and heated to 95 °C for 10 min. A protein molecular weight marker and each protein sample was applied to the gel. As standard SDS-PAGE gels were subjected to electrophoresis at 180 V for 47 min. Ran gels were saturated in Instant Blue for over one hour and destained overnight in water. CTB is known to run as a pentamer on SDS-PAGE gels if the sample, when mixed with SDS-PAGE sample buffer, is not boiled. For non-boiled CTB samples, mix with SDS-PAGE sample buffer at the last possible opportunity before running.

## 2.1.3 Plasmid Construction

A wide range of plasmid constructs for the expression of cholera toxin and its mutants were generated during this project. Plasmid pSAB2.1 was used for the production MBP-$AB_5$ complexes and plasmid pSAB2.2 was used for all other CTB mutants including those with *C*-terminal extensions. These two plasmids were derived from the plasmid pSAB2.0 which contains the pMAL-p5x backbone and contains a synthetic gene construct, bought from Genscript, named SAB2.0. The process is described below.

### 2.1.3.1 Plasmids

**pMAL-p5x and pMAL-c5x** – These plasmids were obtained from New England Biolabs and contain the MalE gene which expresses maltose binding protein, MBP, with a *C*-terminal polyasparagine linker region. A MCS follows this linker allowing splicing of an additional gene which will translate as an MBP fusion protein. A FactorX protease site sits just before the fusion, allowing cleavage of the product (Figure 2-2). pMal-p5x contains a periplasmic leading sequence which targets translation into the periplasmic space via the SEC pathway.

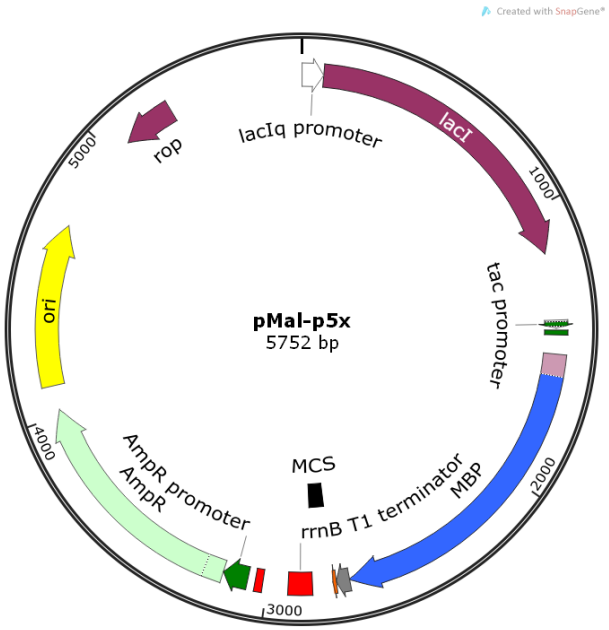



**Figure 2-2:** The MAL-p5x plasmid. This plasmid contains the ampicillin resistance gene, AmpR (pale green), the maltose binding protein (MBP) gene, MalE (blue) and is under the control of the lac repressor system (purple). A multiple cloning site sits downstream from the MalE gene, which allows the splicing of genes for expressing fusion proteins of MBP. This version of the pMAL plasmid translocated MBP expression to the periplasm. Figure produced with SnapGene.

### 2.1.3.2 SAB2.0 Gene Construct in pUC57

The SAB2.0 gene construct encoding CTA2 and CTB was purchased from Genscript in a pUC57 plasmid (Figure 2-3). This construct was specifically designed to be spliced into the pMal-p5x vector to create the plasmid pSAB2.0 encoding MBP-tagged CTA2 and CTB. This 981nt section of DNA contains, from 5' to 3', a TEV cleavage domain, the CTA2 gene, the rrnb termination sequence, an additional tac promoter, ribosome binding site and finally the CTB gene primed with the LTIIb periplasmic leader sequence. The CTB gene was designed to house a number of unique restriction sites throughout the gene to aid modification by restriction enzymes. This gene was also designed to allow facile derivatisation into two additional plasmids, pSAB2.1 and pSAB2.2

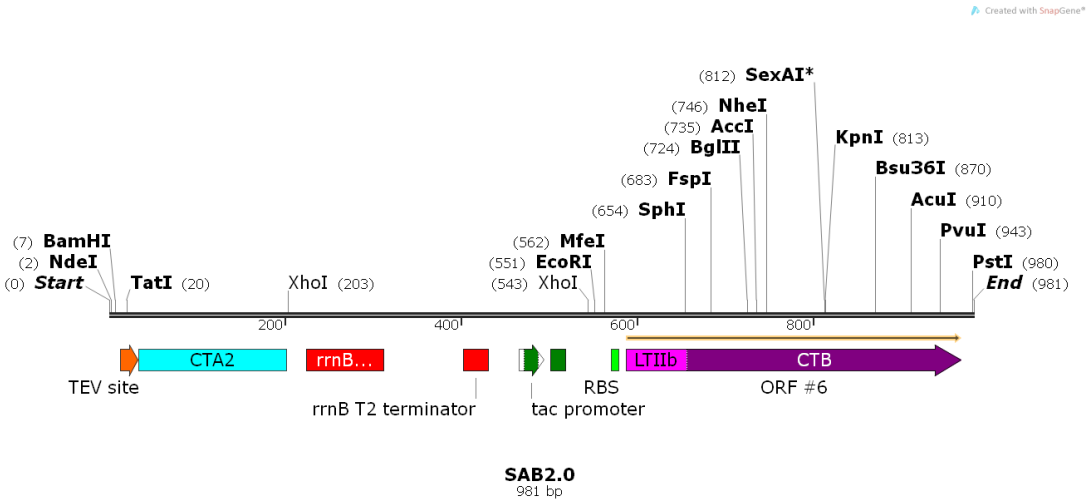



**Figure 2-3:** The SAB2.0 gene construct (digested from pUC57). This gene contains a TEV cleavage domain (orange), the CTA2 gene (light blue), the rrnb termination sequence (red), an additional tac promoter (dark green) and RBS (light green) and finally the CTB gene (deep purple) primed with the LTIIb periplasmic leader sequence (light purple). Figure produced with SnapGene.

### 2.1.3.3 pSAB2.0

pSAB2.0 should yield two products via individual mRNA, an MBP with a *C*-terminal fusion of the CTA2 protein and CTB. Both gene products are targeted to the periplasm to facilitate the assembly of CTB pentamers around the MBP-CTA2 fusion to create a MBP-$AB_5$ complex. pSAB2.0 was generated via subcloning of the BamHI-PstI fragment from pUC57-SAB2.0 into pMAL-p5x as shown in Figure 2-4.

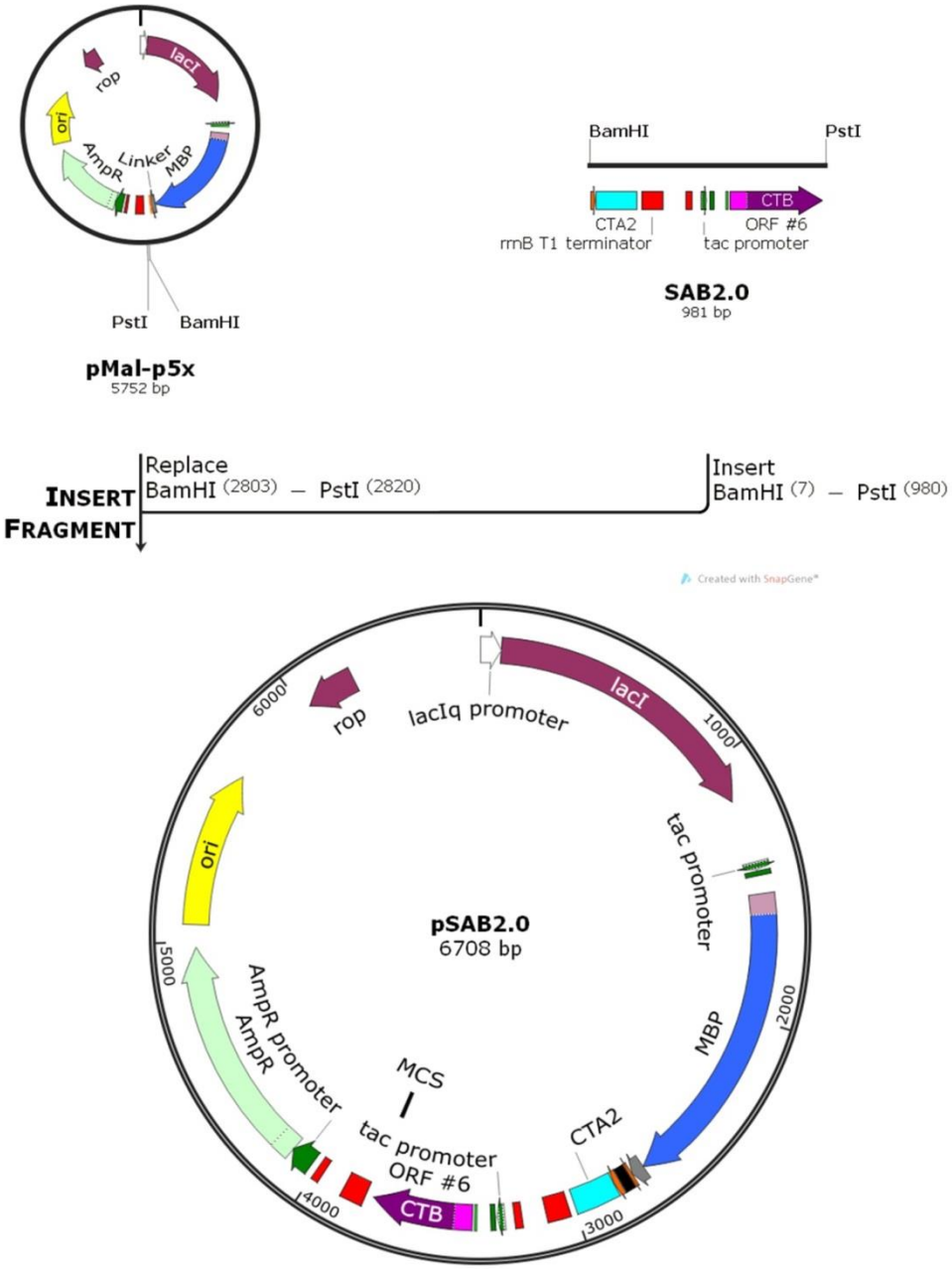



**Figure 2-4:** The pSAB2.0 plasmid. This plasmid is created from splicing the SAB2.0 gene construct into the pMAL-p5× vector after digesting both parts with BamHI and PstI. When induced the plasmid should express the MBP-CTA2 fusion protein (dark blue to light blue) containing a TEV cleavage site and CTB pentamers (purple), each on separate mRNA. Figure produced with SnapGene.

### 2.1.3.4 pSAB2.1

pSAB2.1 encodes a polycistronic mRNA for the expression of both MBP-CTA2 and for CTB, generated by the digestion of pSAB2.0 with XhoI followed by re-ligation (Figure 2-5). This digestion removes the additional terminator and promoter sequences between the MBP-CTA2 and the CTB genes.

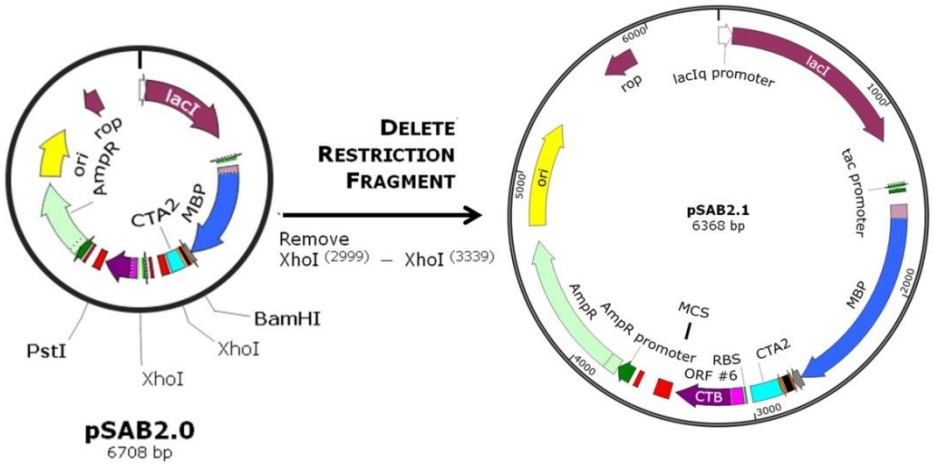



**Figure 2-5:** The pSAB2.1 plasmid. This plasmid is derived from the pSAB2.0 plasmid via digestion with XhoI. When induced this plasmid should promote transcription of a polycistronic mRNA containing the MBP-CTA2 (dark blue to light blue) fusion as well as CTB (purple). Figure produced with SnapGene.

### 2.1.3.5 pSAB2.2

Alternatively, pSAB2.2 encoding CTB alone was generated in a digest of pSAB2.0 with MfeI and re-ligation (Figure 2-6). This digestion removes the MBP gene (MalE), the CTA2 fusion protein as well as the additional terminator and promoter regions leaving the CTB gene downstream of the native tac promoter region of the pMAL-p5× plasmid. On expression, this plasmid yields a single mRNA containing the CTB gene.

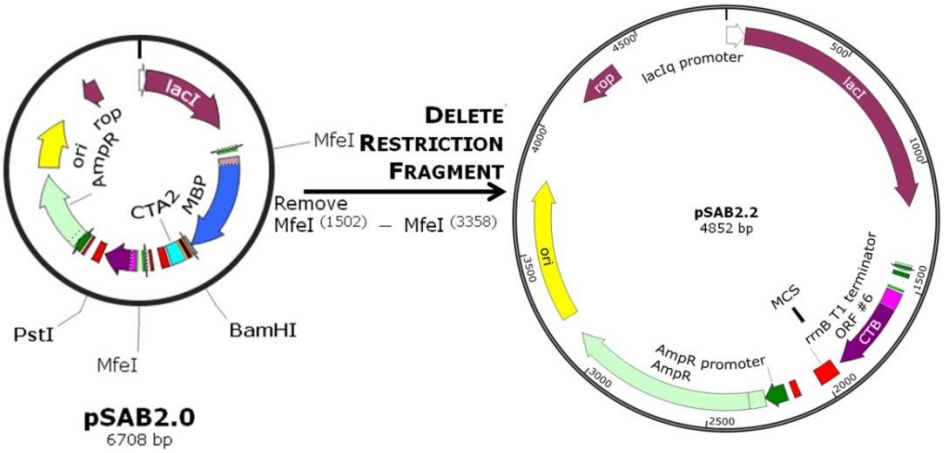



**Figure 2-6:** The pSAB2.2 plasmid. This plasmid is derived from the pSAB2.0 plasmid via digestion with MfeI. When induced this plasmid should promote transcription of a mRNA containing the CTB gene (purple). Figure produced with SnapGene.

### 2.1.3.6 pSAB2.2, Backbone Plasmid for CTB mutations

All site directed mutagenesis for the generation of alanine scan mutants was carried out on this plasmid. The mutants created through site directed mutagenesis were D7A, A10G, Q16A, I17A, T19A, N21A, K23A, F25A, S26A, T28A, G33E, T41A, K43A, A46G, K81A, W88K, N89A and H94A, mutagenesis primer pairs are available in the appendix. All

mutant CTB genes from assembly PCR, including the remaining alanine scan mutants and all interface design mutants, both with and without coiled-coil extensions, were spliced into this vector via SphI and PstI digests of the plasmid and PCR product, followed by ligation (Figure 2-7).

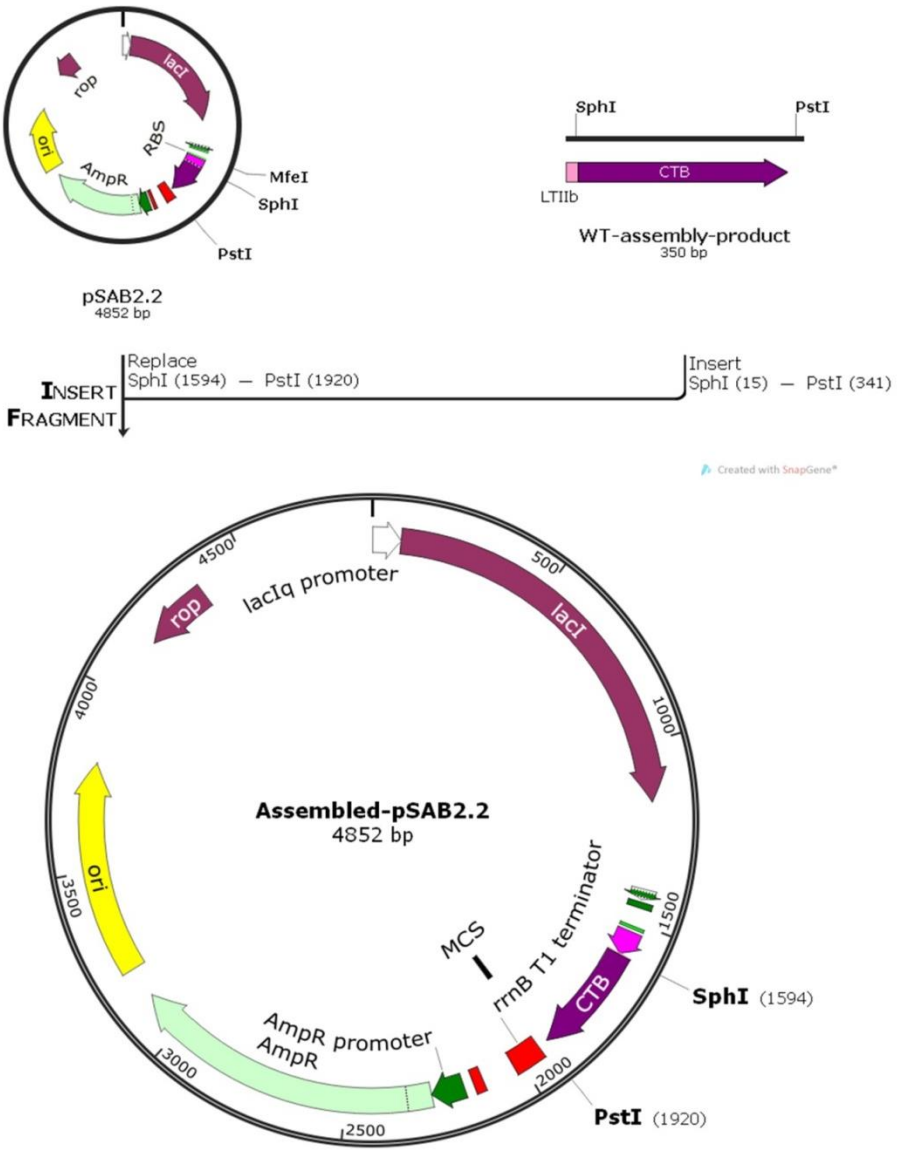



**Figure 2-7:** Introduction of mutant CTB genes into plasmid pSAB2.1. The pSAB2.1 plasmid and the assembly PCR product are digested with SphI and PstI. These are ligated together to form the pSAB2.1 plasmid containing the mutated gene. Figure produced with SnapGene.

For the addition of the *C*-terminal coiled-coils to CTB, eight parts were used instead of six. The general assembly PCR protocol used for assembling six parts was also used to assemble eight parts. The parts required for the extensions are listed in Table 2-4.

| Coiled-coil | Part-7 | Part-8 |
|---|---|---|
| coil-2 | A | A |
| coil-7 | E | F |
| DeCr-E | O | J |
| DeCr-K | P | K |
| SynTri | N | I |

**Table 2-4:** Assembly PCR parts required for the addition of coiled-coils to the CTB C-terminus

## 2.1.4 General Protein Manipulation

### 2.1.4.1 Plasmid Gene Expression

Starter cultures containing 100 µg/mL of the appropriate antibiotic in LB media were inoculated from glycerol cell stocks and incubated at 37 °C overnight. The starter cultures were used to inoculate 1 L of LB media, with 100 µg/mL of appropriate antibiotic, contained in 2 L conical flasks (3 mL of starter culture/L of media). The cultures were then incubated at 37 °C with shaking at 200 rpm until the optical density at 600 nm reached a value of 0.6-0.8 (typically 3 hours). Protein expression was induced by the addition of IPTG (final concentration 300 µM) and the cultures were incubated for 16 hours (overnight) at 30 °C and 300 rpm.

### 2.1.4.2 Fractionation

Cells were isolated by centrifugation at 10000*g* for 10 min and the cell pellet was discarded. Protein was isolated from the media supernatant by ammonium sulfate precipitation; ammonium sulfate (570 g per 1 L culture, 90% saturation at 0 °C) was added slowly to the stirred media before incubation with stirring for a further 2 h at room temperature. The precipitated protein was isolated via centrifugation at 17500*g* for 30 min and the supernatant was discarded. The pellet was resuspended in 75 mL suspension buffer per 1 L of initial culture, before clarification via centrifugation at 3500*g* for 5 min. The supernatant was filtered (0.5 µm vacuum filter) and stored, or immediately purified.

### 2.1.4.3 Purification

Protein was purified via affinity chromatography under gravity flow. The affinity resin (lactose-agarose, Ni-NTA agarose or amylose-agarose) was packed under gravity and equilibrated with five column volumes of water followed by five column volumes of

suspension buffer. The precipitant fraction was applied to the column. The column was washed with five further column volumes of suspension buffer. Protein was eluted using five column volumes of the appropriate elution buffer and collected in fractions of 1-5 mL. Following protein elution, the column was washed with five column volumes of water, followed by five column volumes of 20% ethanol before and storage in 20% ethanol at 4 °C. The protein containing fractions were pooled and had protease inhibitors added prior to storage at 4 °C.

### 2.1.4.4 SDS-PAGE

**Gel Preparation** – Protein identity and purity was routinely assessed by SDS-polyacrylamide electrophoresis using the tris-glycine buffer system. Stacking and resolving gels were prepared as shown in Table 2-5. This mixture was gently agitated and poured into the gel cast followed by a layer of isopropanol. After polymerisation the isopropanol was removed using a stream of water. Stacking gel mixture was then applied to the cast with a comb of choice and allowed to polymerise. Routinely 12 resolving gels were used.

| Reagent | Stacking | | Resolving | | | |
|---|---|---|---|---|---|---|
| | 5% | | 7% | 10% | 12% | 15% |
| H2O | 3.325 | | 8.260 | 7.135 | 6.385 | 5.260 |
| 0.5M Tris pH6.8 | 0.945 | | | | | |
| 1.5M Tris pH8.8 | | | 3.800 | 3.800 | 3.800 | 3.800 |
| 40% Acrylamide | 0.625 | | 2.625 | 3.750 | 4.500 | 5.625 |
| 10% SDS | 0.050 | | 0.150 | 0.150 | 0.150 | 0.150 |
| 10% APS | 0.050 | | 0.150 | 0.150 | 0.150 | 0.150 |
| TEMED | 0.005 | | 0.015 | 0.015 | 0.015 | 0.015 |
| total | 5.000 | | 15.000 | 15.000 | 15.000 | 15.000 |

**Table 2-5:** Preparation of SDS-PAGE gels. All values are in µL. Routinely 12% gels were made. TEMED was added last.

**Gel Running** – Protein samples were prepared by mixing 1:1 with SDS-PAGE sample buffer and heating to 95 °C for 10 min. SDS-PAGE gels were developed at using SDS-PAGE running buffer at 180 V for 47 min. After development, protein was visualised using Instant Blue gel stain. Monomers of CTB were obtained if the protein samples were boiled before loading. If the samples were prepared shortly before gel development and not boiled, pentamers of CTB were observed on the gel.

### 2.1.4.5 Assembly of Virus-Like Particles.

Assembly of VLP was conducted at a variety of CTB concentrations (often 2 mg/ml) via room temperature incubations over a variety of differing periods of time.

## 2.1.5 Biophysical Characterisation

### 2.1.5.1 Mass Spectrometry

#### 2.1.5.1.1 Overview

MS allows the measurement of the mass to charge ratio (m/z) for an ionised molecule. The samples in this project were ionised by electrospray ionization which creates quasi-molecular ions by the addition of protons to the sample. As proteins are large molecules there are often many possible positions for protons to bind. This causes a protein mass spectrum to give many charged states which can be deconvoluted to give the total mass.

#### 2.1.5.1.2 Method

Protein samples were prepared by diluting in water or by mixing 1:1 with 30% methanol, 1% formic acid, or alternatively were used neat. 10 µL samples were directly injected via the autosampling system and analysed by electrospray ionisation mass spectrometry. Data processing was conducted by ESI Compass 1.3 DataAnalysis V4.0 software (Bruker Daltonik).

### 2.1.5.2 Differential Scanning Fluorimetry

#### 2.1.5.2.1 Overview

Differential scanning Fluorimetry (DSF) was used to determine the melting temperature ($T_m$) of protein samples. In this assay the gel stain, Sypro Orange, is used as a fluorophore to detect the melting of the protein. Binding of the dye to hydrophobic portions of the protein causes a change in the dye's emission wavelength. Protein samples were heated at a rate of 0.125 °C/s in a 96 well real-time PCR machine. Each run takes approximately 17 min; making DSF highly suitable to high-throughput investigations. Differential scanning calorimetry, the gold standard for label free $T_m$ measurements, measures the change in heat upon unfolding; but cannot observe more entropically driven protein transitions such as disassembly of complexes and is not suited to high-throughput investigations. Correspondingly DSF cannot be used to observe the unfolding of proteins which lack a pronounced hydrophobic core, such as calmodulin.

#### 2.1.5.2.2 Method

**CTB Assay** – Assays mixtures were prepared to a total volume of 25 µL containing up to 22.5 µL purified CTB/mutant CTB and 2.5 µL 100× Sypro Orange (Sypro Orange is a gel stain supplied as a 5000× concentration, DSF uses a 10× final concentration). These assay mixtures were prepared directly in a 96 well polystyrene PCR plate and sealed with an optically transparent adhesive seal.

**Thermocycling Program** – Samples in the prepared plates were well mixed and any bubbles were removed by centrifugation before placing in the real-time PCR machine (Bio-Rad CFX96). The following program was run. Starting temperature of 25 °C, 75× cycles of 8 s, with 1 °C temperature increments per cycle. This takes the temperature from 25 °C to 100 °C over about 9 min. The dyes excitation wavelength is 488 nm and the emission is read at 570 nm.

#### 2.1.5.2.3 Analysis

The data from the DSF experiments was obtained as the total fluorescence emission over each 1 °C increment. The melting temperature ($T_m$) was calculated as the inflection point in this signal (Figure 2-8). The numerical differentiation of the data yields a simple visualisation of the $T_m$ as the peak maximum (Figure 2-8b). In Figure 2-8a the fluorescence of the most concentrated sample becomes saturated, in this instance the steepest point of the curve has already been reached maintaining the $T_m$ information. Those data which became saturated before the inflection point is reached were discarded.

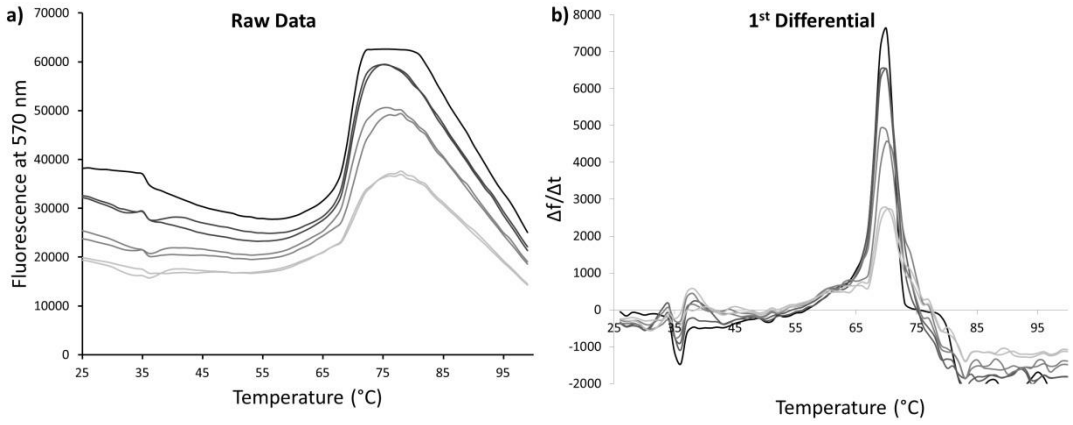



**Figure 2-8:** Data processing from DSF. **a)** The raw absorbance at 570 nm for each cycle. **b)** The derivative of the raw data, where the peak maxima represents the steepest incline of the curve from the raw data, revealing the protein melting temperature.

### 2.1.5.3 Size Exclusion Chromatography

#### 2.1.5.3.1 Overview

SEC was primarily used analytically in this project rather than as a method of purification. SEC separates proteins by their average hydrodynamic radius and can thus be used to determine the size of different species of proteins in complex mixtures. The components of the sample are separated by size and shape by as they flow through a bed of beads. The beads contain pores of differing sizes which either allow or exclude entry to particular parts of the beads interior. Samples with a cross-sectional area which permits entry to a given channel have a greater volume of diffusible solvent to access than those which are excluded from the channel and thus are retained for longer within the matrix; larger particles suffer exclusion from this greater diffusible volume and thus elute more rapidly. Particles which are too large to gain access to any of the pores in the beads all have the same volume to diffuse through and thus elute together in the void volume of the column.

#### 2.1.5.3.2 Method

Samples across a range of volumes and concentrations were applied to a Superose 6 10/300 column and isocratically eluted at 0.35 mL/min using SEC buffer. After 2 mL of elution, 300 µL fractions were collected in 96 well plates and their absorbance was measured at 280 nm. The plates were sealed and stored at 4 °C.

#### 2.1.5.3.3 Analysis

Peak sizes were quantified using Unicorn software (GE healthcare). Peak baselines were often manually adjusted before numerical integration was performed to yield the peak area. In the presentation of the data we often use the relative retention volume (rrv) rather than the retention volume in mL. The rrv was normalised about a peak from a protein ubiquitously purified with CTB preparations. Unlike mutant CTB peaks, this protein was present at consistent concentrations.

### 2.1.5.4 Dynamic Light Scattering

#### 2.1.5.4.1 Overview

DLS (or photon correlation spectroscopy) was used to calculate the hydrodynamic diameter of particles in a given solution. Particles in an aqueous solution diffuse in accordance with 3D Brownian motion and the scattering of monochromatic light (here 633 nm) through this liquid fluctuates in a time-dependent manner based on the particle's diffusion constant as governed by Brownian motion. This scattering undergoes constructive or deconstructive

interference from surrounding particles and this fluctuation in intensity yields information on the time-scale of diffusion of the particles. The non-randomness of the fluctuation can be quantified using the autocorrelation function which can then be used to determine the diffusion coefficient for the particles in solution. From this, the Stoke-Einstein equation can be applied to give the particle's hydrodynamic diameter.

$$d(H) = \frac{kT}{3\pi\eta D}$$

**Equation 1:** The Stokes-Einstein equation. ***d(H)*** is the hydrodynamic diameter, ***k*** is the Boltzmann constant, ***T*** is the temperature, ***η*** is the viscosity and ***D*** is the translational diffusion coefficient.

#### 2.1.5.4.2 Method

DLS experiments were conducted at 20 °C by measurement of the backscatter at 173° at a wavelength of 633 nm. Protein samples of 100 µL, and of at least 0.5 mg/mL were prepared after filtration through a 0.45 µm syringe filter. Each experimental reading was an average of ten independent 10 s measurements, three readings were taken per sample. Intensity values were transformed into volume and number percentages using the inbuilt function in the Malvern data processing software.

### 2.1.5.5 Transmission Electron Microscopy

#### 2.1.5.5.1 Overview

For transmission electron microscopy the sample of interest was applied to UV irradiated carbon coated grids; this UV irradiation applies a negative charge to the carbon making the surface hydrophilic. The samples were then washed with a negative stain, here uranyl acetate, to provide contrast. When visualised under the electron beam, the negative stain causes scattering of the electrons which are not detected on the CCD detector, thus areas heavy in stain appear dark. As proteins stand pronounced from the surface of the carbon, the stain only collects around the edge of the protein. Thus the protein scatters less electrons and appear light on a dark background.

#### 2.1.5.5.2 Method

Carbon TEM grids were UV irradiated for 30 minutes and left to rest for 10 minutes. 8 µL of protein sample was applied to the grid which was incubated for 30 s before the droplet was blotted off. The sample was left to air dry for 30 seconds before 8 µL of 1% uranyl acetate in $H_2O$ was added to the grid and blotted after 7 s. This sample was left to air dry for 30 seconds. Grids were inserted in to the G2 Spirit and visualised under a number of magnifications.

#### 2.1.5.5.3 Analysis

TEM micrographs were analysed with the 'Plot Profile' function of the program ImageJ (Figure 2-9). This function displays the average pixel intensity along a selected section of the image. The total width of the selection within the profile plot gives the total pixels of this new selection. By using this function with the scale bar a pixel per nanometre ratio is calculated with is then applied to measurements of particles within the micrograph.

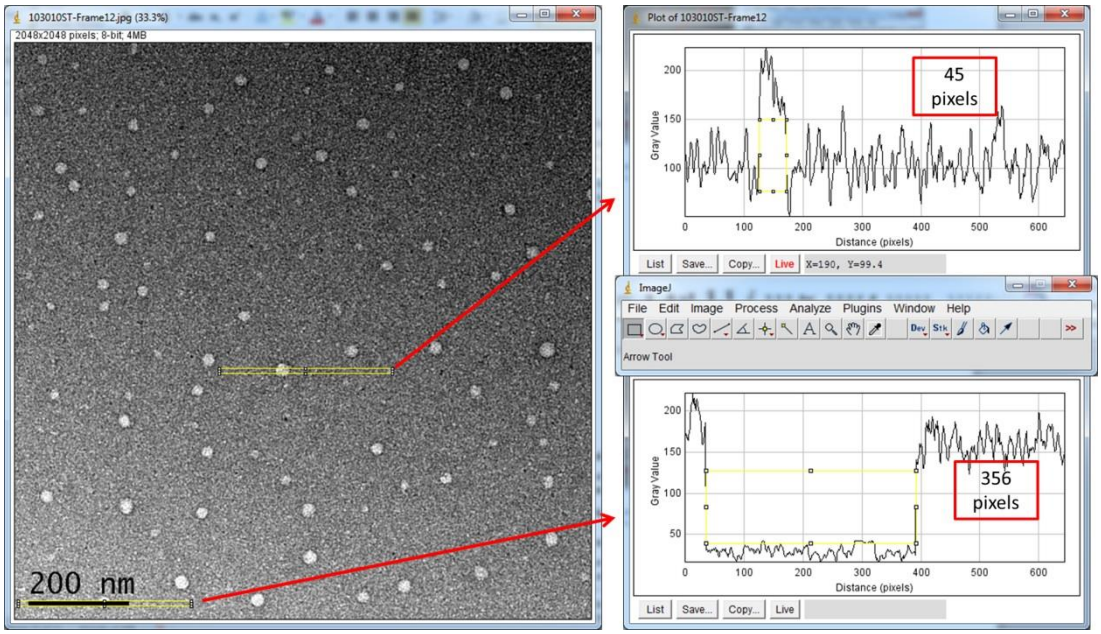



**Figure 2-9:** Analysis of micrographs by ImageJ. First the pixel per nm ration is determined, then this is applied to measurements of particles.

# 2.2 Computational

## 2.2.1 Materials

All computational work was carried out on a 64 bit Dell Precision WorkStation T7500 with 2 × 6 core Intel® Xeon® CPU X5690, 32 GB RAM, Nvidia GeForce GTX 680 and 4 TB MegaRAID SAS 1078 running Ubuntu 12.04.3

## 2.2.2 Methods

### 2.2.2.1 Molecular Dynamic Simulation

Molecular Dynamics (MD) is a simulation technique in which Newton's laws of motion are solved numerically for a set of atoms, interacting through a given potential under specific conditions over time. While in principle all the thermodynamic and kinetic properties of the system can be obtained from a trajectory (i.e., coordinates and velocities as a function of time) in practice this is only possible for simple, rapidly converging systems (e.g., small molecules in gas or liquid phase). Macromolecules of biological relevance cannot be fully characterised (e.g., folding and binding rates and equilibrium constants) by MD simulation. To reduce the complexity of all-atom simulations water can be considered "implicitly", i.e., its effect on the protein, such as screening charges, are included by modifying the protein-protein interaction potential. Coarse-graining of atomistic degrees of freedom is another strategy often used to reduce the complexity of the system. These simplifications allow long time course simulations to be run, which are required to sample large scale movements. However with simplification come inaccuracies. The best results still come from all-atom models in explicitly represented solvent under accurate electrostatic representations, this is however, infeasible for larger molecular systems, without the use of state-of-the-art machines[141].

#### 2.2.2.1.1 Basic Algorithms

The Chemistry at HARvard Molecular Mechanics package (CHARMM)[142] was used to conduct all molecular dynamic (MD) simulations. Through the numerical solution of Newton's equation of motion, $\boldsymbol{F_i} = m\boldsymbol{a_i} = -\nabla V(\boldsymbol{r_i})$, MD simulations calculate the trajectory of a set of atoms (i.e., coordinates, $\boldsymbol{r}_i$, and velocities, $\boldsymbol{v}_i$ for each atom, i) at each time $t$. Integration of the equations of motion can only be performed by discretising time. Many finite difference integrators are available; the Leap-Frog Verlet algorithm chosen here is accurate and stable with an integration timestep $\delta t$ of 2 fs when bonds to hydrogen

atoms are constrained. The Leap-Frog Verlet, unlike some Verlet schemes, require velocities to be calculated, for biomolecular simulations velocities are useful for calculating kinetic energies as well as for temperature control. However, as the name suggests, this algorithm leap-frogs position and velocity calculations, which are thus asynchronous.

$$1. \qquad \boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t\, \boldsymbol{v}\left(t+\frac{1}{2}\delta t\right)$$

$$2. \qquad \boldsymbol{v}\left(t+\frac{1}{2}\delta t\right) = \boldsymbol{v}\left(t-\frac{1}{2}\delta t\right) + \delta t\, \boldsymbol{a}(t)$$

$$3. \qquad \boldsymbol{v}(t) = \frac{1}{2}\left[\boldsymbol{v}\left(t+\frac{1}{2}\delta t\right) + \boldsymbol{v}\left(t-\frac{1}{2}\delta t\right)\right]$$

#### 2.2.2.1.2 The Empirical Energy Function

For a force, **F**, and accelerations, **a**, to be calculated on the atomic positions, **r**, using Newton's second law the potential energy function, $V(\boldsymbol{r})$, requires definition. The potential energy function is defined by the sum of the bonded and non-bonded terms.

$$4. \qquad V(\boldsymbol{r}) = E_{bonded} + E_{non-bonded}$$

The bonded term is described by the sum of bond stretching, angle bending and bond rotation (dihedrals and impropers) terms.

$$5. \qquad E_{bonded} = E_{bonds} + E_{angles} + E_{dihedral} + E_{impropers}$$

$$6. \qquad E_{bonded} = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedral} K_\Phi(1+\cos(n\Phi-\delta)) + \sum_{improper} K_\omega(\omega-\omega_0)^2$$

The $E_{bonds}$ term describes covalent bond stretching, where b is the bond length, $k_b$ is the bond force constant and $b_0$ is the equilibrium bond length. The $E_{angles}$ term describes the bending of angle θ between three bonded atoms, with the angle force constant $k_\theta$ and equilibrium bond angle $\theta_0$. The $E_{dihedral}$ term describes torsional steric clashes, where $k_\phi$ is the dihedral force constant, *n* is a coefficient of symmetry, ϕ is the dihedral angle and δ is the phase shift. The final $E_{improprers}$ term describes out of plane bending, or improper torsion which helps to maintain chirality, where $k_\omega$ is the force constant and ω and $\omega_0$ are the improper torsion and equilibrium improper torsion angles, respectively.

The non-bonded term of equation 4 includes van der Waals and electrostatic energy terms.

$$E_{non-bonded} = E_{van-der-Waals} + E_{electrostatics} \quad (7)$$

$$E_{non-bonded} = \sum_{non-bonded} \epsilon \left[ \left( \frac{A_{ij}}{r_{ij}} \right)^{12} - \left( \frac{B_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \quad (8)$$

The Lennard-Jones 6-12 potential was used to estimate the van der Waals interaction energy governing both short range repulsive electron-electron interaction forces and longer range attractive forces from dipole fluctuations which give rise to the classic energy minima. The constants $A_{ij}$ and $B_{ij}$ are determined by the chemical types of the atoms. The Coulomb potential determines the electrostatic interaction energy where ϵ is the effective dielectric function for a given medium between two atoms with $q_i$ and $q_j$ charges.

#### 2.2.2.1.3 Langevin Thermostat

Most experiments are carried out at a constant temperature, whereas the Verlet described above, based on Newton's laws of motion, represents a microcanonical ensemble where the number of particles (N), the volume (V) and total energy (E) are kept constant, not the temperature. All simulations run use Langevin dynamics to maintain the temperature (T), so called constant NVT ensemble rather than constant NVE with the addition of a friction coefficient $\gamma$. Two additional forces are used to supplement the force derived from the force-field, $F_{position}(t)$, which mimic the effect of a heat bath. These are $F_{friction}(t)$, which is directly proportional to the particle velocity, and the $F_{random}(t)$, which introduces solvent effects.

$$F(t) = F_{position}(t) + F_{friction}(t) + F_{random}(t) \quad (9)$$

$$F(t) = -\nabla V(\mathbf{r}(t)) - \gamma m \mathbf{v}(t) + R_{random}(t) \quad (10)$$

The $F_{random}(t)$ term is uncorrelated with the velocity of particles and the forces acting upon them, obeying a Guassian distribution about zero.

#### 2.2.2.1.4 The EEF1 Implicit Solvent Model

All atom simulations are time consuming due to simulation of the water molecules, MD performance is roughly proportional to the number of atoms in the system squared, thus removal of water atoms allows an improvement in efficiency of simulations. The Effective Energy Function EEF1 is an implicit solvent model[143] which reduces the solvent to a continuum representation incorporating its dielectric constant and surface tension. EEF1

assumes the total free-energy of solvation, $\Delta G^{solv}$, is given by the sum of the solvation energies of external chemical groups such as amino acid side chains or carbonyl carbons etc..

$$11. \qquad \Delta G^{solv} = \sum_i \Delta G_i^{solv}$$

$$12. \qquad \Delta G_i^{solv} = \Delta G_i^{ref} - \sum_{j \neq i} f_i(r_{ij}) V_j$$

The term $\Delta G_i^{solv}$ can be further broken down, equation 12, where $\Delta G_i^{ref}$ describes the solvation free energy of a fully solvent exposed chemical group, $r_{ij}$ is the distance between the two groups, *i* and *j*, $V_j$ gives the volume of the chemical group *j* in proximity with *i* and $f_i r_{ij}$ describes a Gaussian function that estimates the solvation free energy density around group *i*.

### 2.2.2.2 CHARMM MD runs

We used the CHARMM19 force-field with the EEF1 implicit solvent. Minimisations were conducted for 500 steps with the steepest descent to a local energy minima at 0 K. The system was then heated to 300 K over 100 ps and simulations were routinely run for 40 ps with 2 fs time steps. Coordinates and velocities were saved at intervals of 100 time-steps, 200 fs. The program VMD was used to extract PDB files from the trajectory. Scripts are available on the accompanying DVD.

### 2.2.2.3 FoldX

FoldX is an algorithm which was originally proposed to estimate the free energy of folding of a protein given its native conformation. FoldX contains a library of amino acid rotamers which it uses to sample the $\Delta G_{stability}$ (Equation 12) of different conformations and choose the most energy minimised protein structures. Therefore FoldX can be used to suggest conformations and mutations that increase the stability of a protein or a complex. These computed energies have been weighted to and validated from empirical protein engineering data. FoldX has been used to rapidly compute the free energy of proteins[144] and used to make quantitative assessments of the energetic impact of mutations on protein and complex stability[145]. There is an online FoldX webserver which can be used without downloading the software [146]. A benefit of the FoldX program is the ability to use a batch file which allows not only simultaneous running of a number of commands but also

for these commands to be used on a number of starting models, making the assessment of ensembles trivial.

$$13.\quad \Delta G_{stability} = W_{vdw}\Delta G_{vdw} + W_{solvH}\Delta G_{solvH} + W_{solvP}\Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{kon} + W_{mc}\, T\, \Delta S_{mc} + W_{sc}\, T\, \Delta S_{sc}$$

Where $\Delta G_{vdw}$ is the sum of all van der Waals interactions with respect to the same interactions against the solvent. $\Delta G_{solvH}$ and $\Delta G_{solvP}$ are changes upon unfolding of the solvation energy differences for apolar and polar groups respectively. $\Delta G_{hbond}$ is the intra-molecular hydrogen bond free energy difference compared to intra-molecular hydrogen bond formation to the solvent. $\Delta G_{wb}$ is an additional stabilising free energy for water molecules making more than one hydrogen bond to the protein that is not estimated with non-explicit solvent approximations. $\Delta G_{el}$ is the charged group electrostatic contribution. $\Delta S_{mc}$ accounts for the entropy penalty for the fixed backbone in the folded state. $\Delta S_{sc}$ accounts for the entropic penalty for fixing side-chains in a particular conformation. $\Delta G_{kon}$ accounts for the effect of electrostatic interaction on association constant $k_{on}$ for subunit binding energies in complexes.

The values for $\Delta G_{vdw}$, $\Delta G_{solvH}$ , $\Delta G_{solvP}$ and $\Delta G_{hbond}$ which are attributed to each atom are derived from experimental data, $\Delta S_{mc}$ and $\Delta S_{sc}$ are theoretical estimates and all weighting factors (W) are 1 except for the vdw weighting factor which is 0.33. The FoldX default conditions are 25 °C, pH 7, with a 50 mM ionic strength. The components of the program used mostly here are described below.

**Stability –** Measures the total $\Delta G_{stability}$ (equation 12) of unfolding of the molecule compared with a hypothetical unfolded protein of the same sequences and gives a breakdown of various energetic properties.

**AnalyseComplex –** Measures the $\Delta\Delta G_{interaction}$ between two molecules. The $\Delta G_{stability}$ of the whole complex is first calculated, then the $\Delta G_{stability}$ of the components of the complex. The difference in the free energy of unfolding between the entire complex and the sum of its components is given as the $\Delta\Delta G_{interaction}$ in kcal/mol (Equation 13)

$$14.\quad \Delta\Delta G_{interaction} = \Delta G_{stability_C} - \left(\Delta G_{stability_A} + \Delta G_{stability_B}\right)$$

Where *C* is the whole complex, *A* is one component of the complex and *B* is the rest of the complex.

**Alascan** – The FoldX alanine scan replaces each residue of a target structure sequentially with alanine and records the $\Delta G_{stability}$. This figure is then compared to the wild-type residue to assess the change in structural integrity given the alanine mutation. It should be noted that alanine is used to simulate the truncation of the residue's side-chain, thereby assessing the impact that particular residue has on the rest of the structure. The alanine scan applied by FoldX is very fast, and is achieved through the simple removal of the side chain. There is no attempt made by FoldX to refine the structure based on the removal of the side chain. This method was used on the initial wild-type structure of CTB and the results can be seen in Chapter 3.

**PositionScan –** Performs a saturation mutagenesis investigation on selected positions in the protein sequence. First the position is 'mutated' to the wild-type residue and then the surrounding residues sample rotamers to find the lowest $\Delta G_{stability}$, this is used as the wild-type reference. The selected position is then mutated to each of the 20 amino acids sequentially by the same process and the proteins unfolding energy is calculated and compared to the wild type reference. The difference is given as $\Delta\Delta G_{stability}$ in kcal/mol (Equation 14)

$$\Delta\Delta G_{stability} = \Delta G_{stability_{Mut}} - \Delta G_{stability_{WT}} \quad (15)$$

**BuildModel –** Given a template sequence and structure, FoldX builds new structures containing mutations present on additional sequences. During the construction process rotamers are sampled to find the lowest $\Delta G_{stability}$.

### 2.2.2.4 Pymol

Used to build, view and manipulate pdb files. It's "align", "sculpting" and "mutagenesis" functions were used extensively in rebuilding the mutant pentamers from monomers and applying new mutations.

# Chapter 3

# Assessing CTB Stability

# 3.1 Wild-Type CTB Expression and Characterisation

## 3.1.1 Overview

CT is expressed naturally in the Gram-negative bacterium *Vibrio cholerae.* Expression of CTB in the Turnbull/Webb lab was originally performed with the plasmid pATA13, obtained from Prof Tim Hirst at the Australia National University, for expression in *Vibrio* sp. 60, a marine fish pathogen commonly used for expression of CTB and LTB[147]. *Vibrio* sp. 60 does not express a CT homologue; however it does possess the secretory mechanisms required for secretion of the toxin. Whilst *Vibrio* sp.60 with the pATA13 plasmid expressed large quantities of CTB, the transformation of the *Vibrio* requires tripartite mating and the pATA13 plasmid was low copy number and quite large, 9kb, for site directed mutagenesis[148,149]. It was decided that transferring the gene into standardised laboratory strains of *E. coli* in commercially available plasmids would lead to a more efficient cloning methodology.

The structure of CTB (3CHB.pdb) contains an intra-monomer disulfide bond between the only two cysteine residues, C9 and C86. Additionally two of the three prolines, P53 and P92 occur in the *cis*-conformer opposed to the more common *trans*-conformer[117]. The proteins responsible for controlling these modifications, disulphide bond isomerase (DsbC) and peptidyl-prolyl cis-trans-isomerase (PPIase), function in the periplasmic space of *E. coli* [150,151]. Therefore, periplasmic targeting was deemed essential for our expression system. As a result of this decision, two fractionation regimes are possible, isolation of secreted protein from the culture media or recovery of the periplasmic fraction of cells, dependent on the secretion or retention of the protein, respectively.

CTB is known to bind to nickel NTA columns due to the pentavalency of surface-exposed histadine residues H13 and to a degree H94[152]. It has been shown that purification by this technique yields intact pentamers as the monomeric species lack the appropriate affinity. CTB can also be purified by lactose affinity chromatography columns, by way of the GM1-ganglioside binding site.

### 3.1.2 Expression Construct for *E. coli*

In order to provide a platform for reliable mutagenesis it was decided to move CTB expression from *Vibrio* sp.60 into *E. coli*. Transformation in *E. coli*, in comparison to *Vibrio* sp.60, is robust and efficient, also the plethora of available strains and collected knowledge make *E. coli* ideal for most routine protein expression.

pMal-p5x is a plasmid vector which allows the fusion of a protein of choice to maltose binding protein (MBP). MBP is frequently used to improve the solubility of recombinant proteins, however it is used here as a pseudo A-subunit as MBP has the same approximate size and globular nature as the toxic A-subunit. Additionally the MBP molecule provides an additional high specificity purification tag for isolation of the MBP-$AB_5$ complex. The multiple cloning site (MCS) is located to the C-proximal side of a poly-asparagine linker section and a Factor-X cleavage site. A 981nt section of DNA was designed (SAB2.0, obtained from Genscript) containing, from 5' to 3', a Tobacco Etch Virus protease (TEV) cleavage domain, the CTA2 gene (which links the A1-subunit to the B-subunit pentamer), the rrnb termination sequence, an additional tac promoter and ribosome binding site (RBS) and finally the CTB gene primed with the LTIIb periplasmic leader sequence[153]. This SAB2.0 section of DNA was cloned into pMAL p5x. It also contained a number of restriction sites to allow future manipulation of this parent plasmid, pSAB2.0, to give the truncated variations pSAB2.1 and pSAB2.2 (as described in the methods). The annotated sequences of each are available in the appendix (10.1.1 and 10.1.2). Expression of pSAB2.0 should yield two products. Firstly an MBP with a C-terminal fusion of the CTA2 protein, and secondly, CTB pentamers. Both gene products are targeted to the periplasm. With CTA2 being present in the periplasm at the same time as CTB is expressed into the periplasm, it was hoped that CTB pentamers would associate with the MBP-CTA2 fusion to create the non-toxic complex MBP-$AB_5$[154].

CTB expressed without the MBP complex can be affinity purified via two techniques. Firstly, by the pentameric nature of the protein and two surface histidines, CTB pentamers bind nickel NTA columns to a similar affinity as his-tagged proteins allowing nickel affinity purification. Furthermore, the native GM1 binding site which lies across the monomeric interfaces of the pentamer allow purification by lactose affinity.

### 3.1.2.1 Construction and Expression of pSAB2.0

pSAB2.0 was prepared by ligation of the synthetic gene construct SAB2.0 into pMal-p5x after digests of both parts with BamHI and PstI. This placed a TEV cleavage site directly after the FactorX cleavage site, included as 'Turbo TEV' was available in the laboratory. The plasmid was designed to express two mRNAs producing the MBP-CTA2 fusion and CTB, respectively. The rrnb terminator followed by the tac promoter was positioned between the MBP-CTA2 fusion and the CTB gene; however expression of this plasmid did not yield protein. It is possible that by duplicating the promoting and termination sequences in the plasmid, homologous recombination could have occurred[155].

### 3.1.2.2 Construction and Expression in pSAB2.1

Digestion of pSAB2.0 with XhoI allowed the removal of the additional terminator and promoter sequences between the MBP-CTA2 and the CTB genes, but maintained the second RBS. This construct was designed to produce a polycistronic mRNA containing both genes, which would allow the co-expression of an MBP-CTA2 fusion protein with CTB. Both proteins were directed to fold in the periplasm under control of the SEC translocon via the innate MalE periplasmic tag and the LTIIb periplasmic tag preceding the CTB gene[126]. Expression of this plasmid followed by SDS PAGE analysis of the culture medium confirmed the presence of both MBP-CTA2 and CTB. These proteins were extracted, purified and characterised.

Initial screens for optimal expression conditions, cell type and expression time were performed revealing qualitatively higher yields for this construct in C43 cells expressed at 30 °C for 16 hours. The cells were pelleted by centrifugation and the clarified media was passed over an amylose column. Elution with 10 mM maltose gave MBP and MBP-$AB_5$ complex (Figure 3-1). These elutions were further purified by nickel affinity chromatography to isolate pure MBP-$AB_5$ complex (Figure 3-1). The MBP-$AB_5$ complex migrates on SDS PAGE as the whole complex if the sample is not boiled prior to loading the gel, however if boiled for 10 minutes before running SDS PAGE, the proteins run as their constituent parts. After amylose purification, there were few contaminating species of other proteins; however there did appear to be at least two distinct sizes of MBP. This could be due to proteolysis of the CTA2 fusion from the MBP. CTB monomers can be seen in the boiled samples; however CTB pentamers, as seen in the positive control, run at an analogous size to MBP and are indistinguishable in non-boiled samples. Purification by amylose affinity chromatography retrieves only CTB pentamers which are associated with

the CTA2 fusion, thus in non-boiled samples most CTB pentamers run as the high weight MBP-$AB_5$ complex. In order remove the MBP molecules which are not bound to CTB, a nickel affinity column was run. CTB pentamers readily bind nickel columns without a his-tag, thus this technique was used to remove MBP molecules which were not associated with CTB. From the flow through it is noticeable that both sizes of MBP bands are collected suggesting that the ratio of MBP to CTB is not optimal for complex formation and the MBP-CTA2 fusion is in great excess. However after elution from nickel purification which is selective for CTB only, the primary band in the non-boiled sample is that of the MBP-$AB_5$ complex and two minor bands which correspond to MBP-CTA2 and CTB pentamer, respectively. Another important observation is the presence of a weak smudge in the nickel elution running from the MBP-$AB_5$ complex down to the MBP and CTB pentamer bands. This smudge is suggestive of disassembly during the course of SDS PAGE. This disassembly may be due the presence of SDS in the gel and buffers, the increase in temperature of the gel during the run, or a combination of both. In the boiled lane from the nickel elution there are only two bands seen, the MBP-CTA2 and CTB monomers.

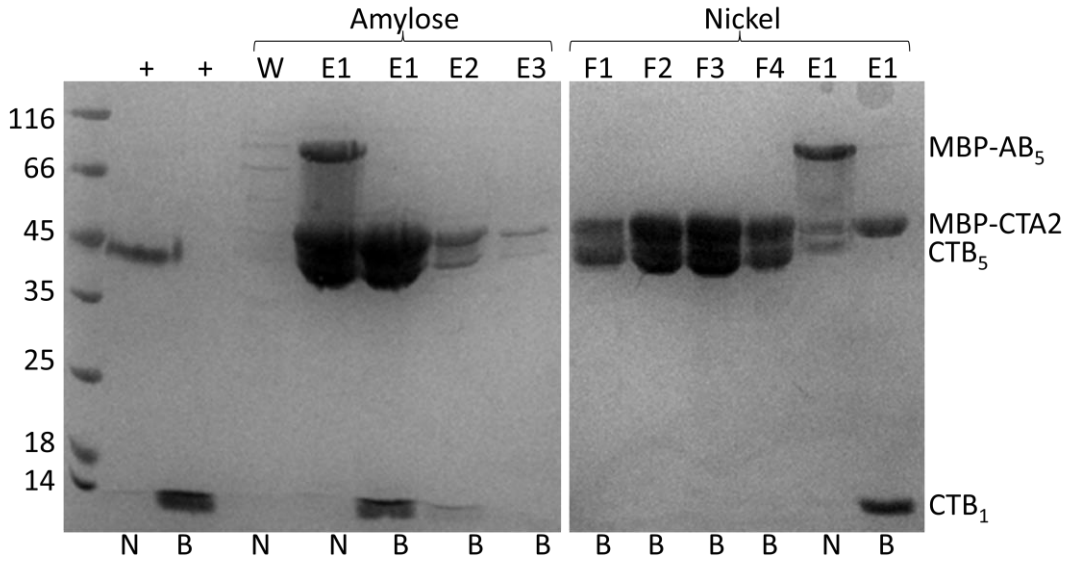



**Figure 3-1:** SDS-PAGE of the purifiedMBP-$AB_5$ complex stained with coomassie blue. The positive control (+) shows migration of WT-CTB both non-boiled (N) and boiled (B) prior to SDS PAGE. MBP was targeted for purification with an amylose column: the wash (W) shows the majority of contaminants were washed from the column; eluent E1 is shown non-boiled and boiled. The non-boiled lane shows the presence of a band at approximately 80-90 kDa which is not present in the boiled sample. The boiled sample shows a band running at the same size as the WT-CTB monomers. In both non-boiled and boiled samples two MBP variants are seen, presumably MBP and MBP-CTA2. These eluents were combined and CTB was then targeted for purification with a nickel column: the flow-through (F) from this purification reveals large amounts of MBP. The elutions from the nickel column (E) in the boiled lane show just two bands corresponding to MBP-CTA2 and CTB monomers, however in the non-boiled sample these can be seen complexed in the MBP-$AB_5$ complex, as well as products of 'on gel' complex disassembly at the sizes corresponding to MBP and CTB pentamers.

### 3.1.2.3 Expression in pSAB2.2

Digestion of plasmid pSAB2.0 with MfeI removed a section of DNA comprising everything from the MalE gene to the additional promoter and terminator regions. The resulting plasmid (pSAB2.2) encoded only the CTB gene with the LTIIb periplasmic targeting sequence, sited directly after the original pMal promoter and RBS sequence (plasmid map available in the appendix (10.1.1.7). During purification trials on pSAB2.1 it was found that application of the clarified media directly to nickel affinity columns caused stripping of the column by agents within the media. This presented a problem for the purification of CTB uncomplexed with MBP-CTA2. In order to overcome this problem, the protein component of the clarified media was precipitated with ammonium sulfate[156] over 2 hours at room temperature. Centrifugation yielded a protein precipitate which, when re-suspended in PBS and purified by nickel affinity chromatography, gave CTB pentamers. Further purification on a lactose affinity column, making use of the CTB GM1 binding site, removed impurities remaining after nickel purification. It is worth noting the faint high weight band in the non-boiled lanes highlighted by the red arrows. This is thought to be a dimeric interaction of two CTB pentamers as elucidated by tryptic MS in chapter 6.

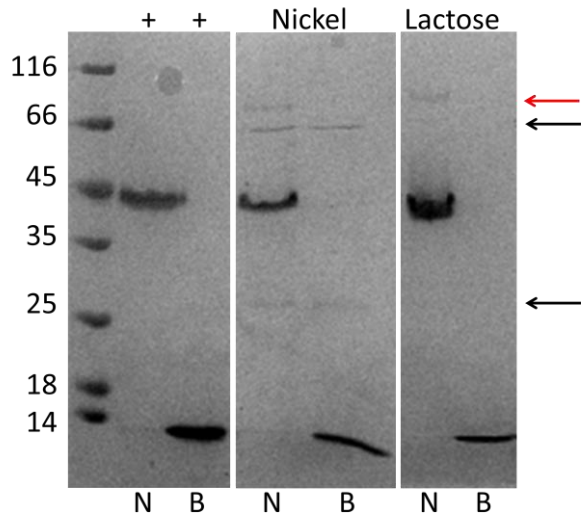



**Figure 3-2:** Purification of CTB. Elution of the ammonium sulfate precipitation fraction after binding to a nickel column reveals purified CTB in both the non-boiled (N) and boiled (B) lanes, which migrate in a similar manner as the control WT-CTB (+). When these elutions were applied to and eluted from a lactose affinity column, pure CTB was obtained. Other minor impurities are highlighted with black arrows and the dimer of CTB pentamers with red arrows

Despite the greater purity achieved though lactose purification, it was found that the process was rather inefficient, requiring three passes of the nickel eluent over the column to retrieve the same yields as from nickel affinity chromatography. This step was not routinely used in the purification of further mutants as the ratio of contaminants to CTB

was minimal. Towards the end of Chapter 6 a tryptic digest of these contaminants is performed.

## 3.1.3 Assessment of CTB Thermostability

In order to assess the stability of CTB and its mutants in a high throughput manner we turned to differential scanning fluorimetry (DSF). In this technique, 1 °C increments in temperature are made every 8 seconds over a 25 °C to 100 °C temperature range and the emission from a fluorophore is measured. Sypro Orange changes its emission wavelength upon binding hydrophobic moieties to 570 nm. Thus upon protein unfolding, and the resultant exposure of the hydrophobic core, an increase in the 570 nm emission is detected. Earlier experimentation by differential scanning calorimetry (DSC) identified CTB to have a $T_m$ of 76.5 °C at pH 7.5[157]. It is worth noting that DSC measures only the heat required to equilibrate the sample cell to the reference cell, and thus measures enthalpic changes but does not measure entropically based unfolding. However DSF detects the exposure of hydrophobic residues, thus a slight discrepancy between the techniques is expected.

### 3.1.3.1 Differential Scanning Fluorimetry of MBP, the MBP-$AB_5$ complex and CTB

The construct pSAB2.1 was used to express and purify the MBP-$AB_5$ complex. The protein analysed by DSF at a concentration of (0.3 mg/ml) 2.7 µM (Figure 3-3a). As the MBP-CTA2 and $CTB_5$ were likely to melt at different temperatures, two distinct melting events were expected. However, as seen in Figure 3-3a, only a melting peak at 58 °C is observed; this was attributed to the melting of MBP at 58 °C (discussed below). Unusually, a trough was observed with a minimum at 80 °C which is not often seen in DSF experiments. This experiment was conducted in parallel with CTB alone at a similar concentration of 5.2 µM (0.3 mg/ml) seen in Figure 3-3b. This trace did not contain the melting curve observed with the MBP-$AB_5$ complex but it did contain a small peak followed by a distinct trough as observed previously. Thus, the 58 °C melting peak is attributable to MBP and the trough, attributable to CTB. This trough is suggestive of a reduction in the fluorescence emission of Sypro Orange, which could arise from the dissociation of Sypro Orange from hydrophobic moieties. This observation is counter intuitive as the peak at 65 °C seems insufficient to signal full denaturation of the protein at the concentration assayed. In order to find an accurate CTB melting temperature a more rigorous DSF experiment was required.

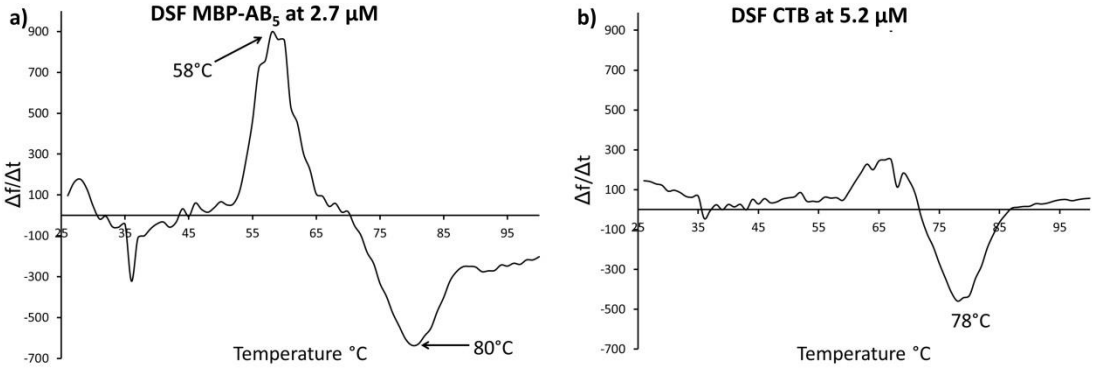



**Figure 3-3:** Differential Scanning Fluorimetry (DSF) of MBP-AB5 and CTB. **a)** shows the first differential of the 570 nm emission of the MBP-AB5 DSF analysis. The features of this trace are a peak at 58 °C and trough at 80 °C. Most proteins present a peak upon melting and thus the presence of the trough is unexpected. **b)** The trace for CTB alone; displays only a trough at 78 °C, suggesting that the 58 °C peak is attributable to MBP, and 78-80 °C trough to CTB. A small peak is also present at approximately 65 °C which appears to extend beyond the noise of the system.

A concentration gradient of MBP from 2 µM to 50 µM (0.2 to 5.5 mg/ml) was assayed by DSF to determine its melting temperature. MBP has been analysed by DSC and a melting temperature of 63 °C was determined at pH 7.4[158]. As seen in Figure 3-4a, a melting temperature of 62 °C is consistantly achieved across all concentrations. As the total protein content increases, an increase in total fluorescent emission is observed in line with DSF theory. This melting temperature is within 4 °C of the melting temperature attributed to MBP from the MBP-$AB_5$ complex DSF experiment (Figure 3-3a); the difference may be associated with a decrease in stability due to the addition of the CTA2 fusion.

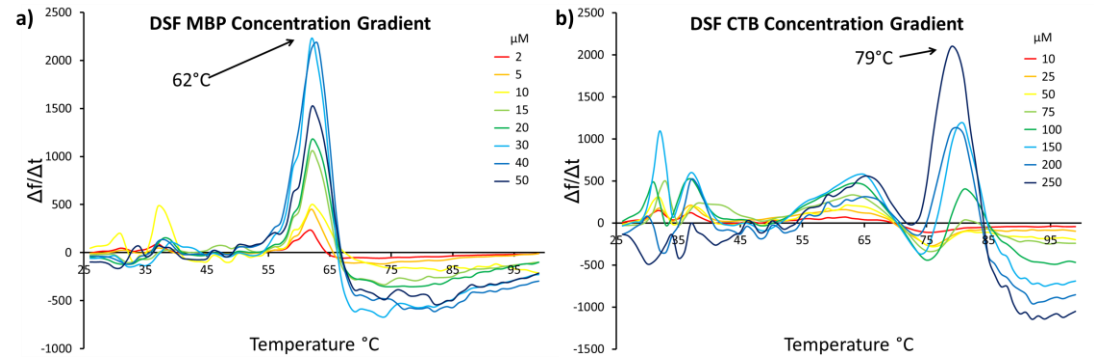



**Figure 3-4:** Further DSF of MBP and CTB. **a)** DSF was conducted on increasing concentrations of MBP from 2 to 50 µM. A $T_m$ of 62 °C was consistently achieved independent of concentration. **b)** DSF was conducted on increasing concentrations of CTB from 10 to 250 µM. A melting peak at 79 °C emerges, given increasing concentrations of CTB, from approximately 81 °C. This emerging peak also makes a left hand shift (reduction) in $T_m$ given increasing concentrations of CTB.

In order to address the CTB DSF trough issue seen in Figure 3-3b, a CTB concentration gradient was also assessed by DSF (Figure 3-4b). With an increase in concentration of CTB from 10 to 250 µM (0.6 to 14.5 mg/ml) a melting peak is gradually seen to protrude from the right hand side of the trough. As the peak intensity increases in a concentration dependent manner, there is a concomitant reduction in the apparent melting temperature. This observation is intriguing as it is unlikely that the melting temperature of the protein decreases with increasing concentration. A possible explanation for this phenomenon would be that the emergence of the melting peak is in competition with the unexplained trough; thus, as the intensity of the melting peak increases the trough minimum is eventually out-competed. An additional point of interest is the small peak at 65 °C which is barely noticeable above the noise of the original CTB DSF (Figure 3-3b); however, this peak is a consistent feature of the DSF data at all CTB concentrations. Finally, it is worth noting that the trough peak minima at 5.17 µM is 78 °C and peak maxima for CTB at 200 and 250 µM is 79 °C. I propose that by some obfuscated biophysical process, presumably based on the pentameric, doughnut shaped structure of CTB (its dissociation and subsequent association with Sypro Orange), that the trough at 5.17 µM represents the CTB melting temperature. However, all further CTB DSF melts were performed at concentrations which achieve standard melting curves.

# 3.2 Choosing Mutable Surface Residues

## 3.2.1 Overview

In order to increase the rate of particle assembly, deleterious side chains may need to be removed and replaced with side chains with no net repulsion to the new interface, or with residues which promote the association of the putative PPI. Creating this new PPI through mutations may allow the future removal the coiled coil scaffolding extensions. Before rationally designing surface mutations to CTB it was necessary to determine the effect of arbitrary surface mutations at appropriate positions on the CTB pentameric structure. A simple method for achieving this aim is to perform an ‘alanine scan’. During an alanine scan, each mutably desirable position is substituted for an alanine residue. The result is to truncate each side-chain back to Cβ or in the case of a glycine residue the addition of a Cβ. The removal of the specific interactions of the side-chain can help suggest the ‘reductive’ disruptive forces likely from mutating such a position. Of course this method does not allow prediction of any ‘additive’ disruptive forces incurred from introducing any other

side-chain. Figure 3-5 shows a selection of views of CTB with 30 chosen potential mutable positions highlighted in red. These positions can be seen to encircle the circumference of the protein. Additionally in Figure 3-5c green positions are highlighted for just one monomer. These positions were chosen as they fall on the proposed tapered interaction surface, yet leave the GM1 binding site free from mutation. Alanine scanning can be performed both computationally and experimentally. Here we utilise both of these approaches and use the alanine scan as an opportunity to compare the computational methods to experimental results.

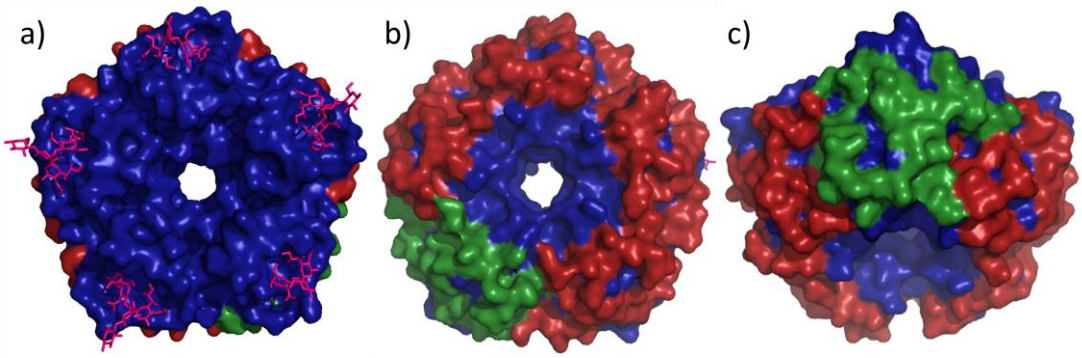



**Figure 3-5**: Highlighted alanine scan positions of CTB. Mutated residues are highlighted in red, for one monomer in green, non-mutated residues are highlighted in blue, GM1 oligosaccharides in pink. **a)** A top down view of CTB, GM1 binding on the front-side of CTB. **b)** A bottom up view of CTB, GM1 binding sites on the back-side of CTB. **c)** A side on view of CTB revealing the bottom face of CTB, mutable residues of a monomer. The tapering of the CTB molecule is easily discernible in this figures as in a) there are few visible mutable residues however in b) all residues are visible. Produced from pdb 3CHB.

## 3.2.2 Computational Alanine Scanning

Computational methods were used to deduce the change in free energy upon unfolding for the series of alanine mutants. The pdb 3CHB.pdb was used as the starting model for the computational analysis. The structure was subjected to a CHARMM minimisation and a 40 ps molecular dynamics simulation. A trajectory of this simulation was used to acquire 200 conformations of the CTB molecule. These 200 structures were fed into the FoldX program and the 'alanine scan' function was used to determine the $\Delta G_{stability}$ in kcal/mol, which measures the change in free energy between the folded structure and a hypothetical unfolded protein of the same sequence. This $\Delta G_{stability}$ was then compared to the wt-CTB energy value to give a $\Delta\Delta G_{stability}$ in kcal/mol, which predicts the change in $\Delta G_{stability}$ upon mutation. This process is discussed thoroughly in the methods section. Figure 3-6 shows

the wild-type residue position along with the FoldX-derived $\Delta\Delta G_{stability}$ upon substitution to alanine. These positions are also highlighted on one monomer of the CTB pentamer. Figure 3-6a displays a number of interesting values: I17A, G45A and T47A show relatively large destabilisations compared with the wild-type. I17, despite being on the surface of the protein, contributes Cδ to a hydrophobic patch shielded from the solvent by Cγ1 and 2, which could explain the large destabilisation upon removal of these atoms and exposure of this hydrophobic patch. G45A requires the addition of a side-chain Cβ atom, this would introduce a methyl group directly into the bulk solvent as well as create torsional constraints to the residue not experienced as glycine. T47 contributes its Cγ methyl group to an inter-subunit interface hydrophobic patch and uses the oxygen at Oγ2 to form inter-subunit water bridges. Some surprising values are the relatively weak destabilising effect of the E83A substitution, which removed an inter-subunit salt bridge. Figure 3-6b & c shows a 'chainbow' of the relative locations of the mutable positions on a monomer of the CTB structure. These two figures also demonstrate that these residues span the breadth of the interaction surface and are also confined to only that surface.

**a)**

| # | Residue | Position | Average ΔΔG |
|---|---|---|---|
| 1 | Q | 3 | 0.55 |
| 2 | N | 4 | 1.9 |
| 3 | T | 6 | -0.35 |
| 4 | D | 7 | 0.61 |
| 5 | A | 10 | - |
| 6 | Q | 16 | 0.52 |
| 7 | I | 17 | 2.38 |
| 8 | H | 18 | 0.42 |
| 9 | T | 19 | 0.99 |
| 10 | N | 21 | 0.58 |
| 11 | D | 22 | 0.52 |
| 12 | K | 23 | 0.29 |
| 13 | F | 25 | -0.01 |
| 14 | S | 26 | 0 |
| 15 | T | 28 | 0.95 |
| 16 | T | 41 | 0.87 |
| 17 | K | 43 | 0.57 |
| 18 | N | 44 | 0.77 |
| 19 | G | 45 | 3.03 |
| 20 | A | 46 | - |
| 21 | T | 47 | 1.78 |
| 22 | K | 81 | 0.64 |
| 23 | E | 83 | 0.69 |
| 24 | K | 84 | 0.56 |
| 25 | N | 89 | 0.22 |
| 26 | T | 92 | 0.8 |
| 27 | R | 94 | 0.19 |
| 28 | S | 100 | ? |
| 29 | A | 102 | - |
| 30 | N | 103 | -0.24 |

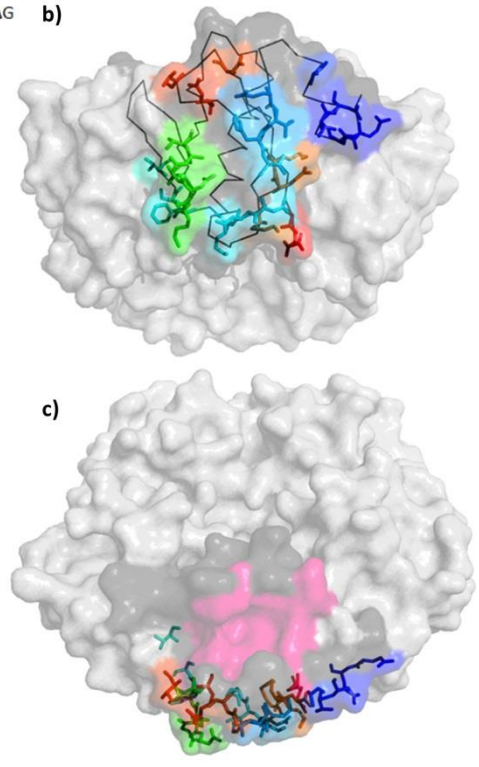



**Figure 3-6:** Computational alanine scan analysis. **a)** displays the residue substituted, the position on the sequence and the $\Delta\Delta G_{stability}$ in kcal/mol upon alanine substitution. The colours are comparable to those displayed in b) and c). **b)** shows that the residues selected span the breadth of the interaction site with residues matching the colours in a). **c)** shows an angle aligned to the proposed interface, which shows that the selected residues are clustered to the interface. Also present in the pink surface is the GM1 oligosaccharide binding pocket, which has be deliberately excluded from the list of potential mutations. Structures produced from pdb 3CHB.

## 3.2.3 Experimental Alanine Scanning

The alanine scan mutations were applied to the pSAB2.2 plasmid. Each mutant was created either by site-directed mutagenesis or by assembly PCR. Each of the 30 mutant constructs were expressed and the CTB was purified by nickel affinity and some by lactose affinity also as described earlier. SDS-PAGE was conducted on boiled and non-boiled samples to verify the pentameric stability and finally a DSF thermostability assay was conducted on concentrated samples to find the change in $T_m$ compared to the wild-type protein.

### 3.2.3.1 Expression and Purification of Alanine Scan Mutants

The majority of the alanine scan mutants expressed to produce viable pentamers as indicated by SDS PAGE of non-boiled samples (Figure 3-7). An interesting observation is the comparative gel retardation of D22A and E83A due to the removal of a negative charge and the counter gel acceleration of the K22A due to the removal of a positively charged residue. I47T is naturally occurring variant of the El Tor biotype of CTB, thus it expressed to a yield comparable to the wild-type (classical) CTB, in this figure this can be used to assess the resultant decrease in expression due to the alanine substitution. Two positions, I17A and G45A did not produce pentamers which endure SDS-PAGE. However, at the dye front of the I17A there is an indication that the monomer is present. As discussed previously, it is thought that nickel affinity purification is only able to purify pentamers of CTB as the monomers lack the multi-valency required for strong binding. Thus, the lack of pentamers on the gel hints to reduced overall pentameric stability for the mutant, rather than these actually consisting of monomers in solution. A band for monomers is absent from the G45A mutant, suggesting that this mutant form CTB pentamers. Reassuringly these two mutants, I17A and G45A, were highlighted to have large destabilising $\Delta\Delta G_{stability's}$ by the FoldX alanine scan. Where the native residue is an alanine, a substitution to glycine is performed.

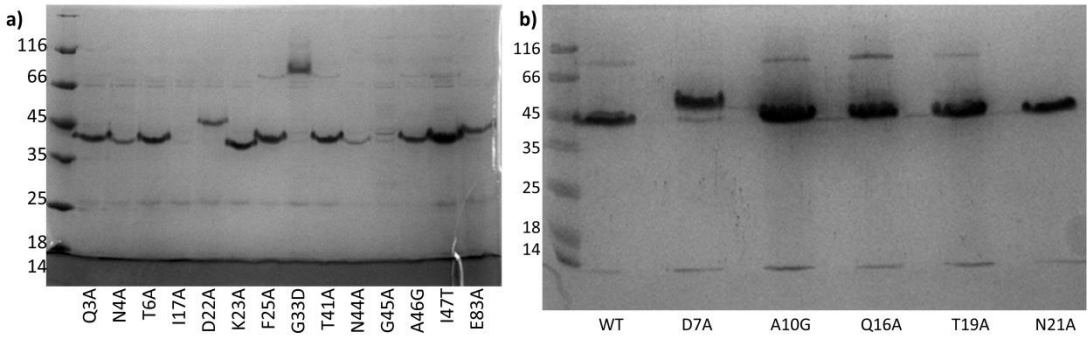



**Figure 3-7:** SDS-PAGE of the alanine scan mutants stained with coomassie blue. Each alanine scan mutant was run non-boiled on SDS-PAGE. Electrophoretic migration deviations were observed in mutants which remove either positively or negatively charged residues. **a)** shows a selection of mutants after nickel purification, various contaminants are observable. Of interest is G33D, not technically an alanine scan mutant, but a mutation which prevents GM1 binding at the associated site. This mutant appears to preferentially form the dimer of pentamers discussed earlier. Additionally of interest are I17A and G45A which do not seem to produce quantities of CTB pentamers discernible on the gel. **b)** Shows a selection of mutants purified by lactose affinity after the nickel affinity purification. The samples appear clean with the major band as pentamers, some monomers present in each sample and some mutants maintaining the dimer of pentamers, while some mutants lose this complex.

### 3.2.3.2 Thermostability Characterisation by DSF of Alanine Scan Mutants

Each protein was concentrated from a 15 ml elution to 1 ml. As the total expression of each protein was variable it was sometimes not possible to reach the standard protein concentration, thus protein concentrations submitted to DSF ranged from 1.2 mg/ml to 3.8 mg/ml. A concentration gradient of each mutant was analysed to assess any changes in the total signal ($\Delta f/\Delta t$) from the DSF output.. A selection of DSF traces are shown in Figure 3-8 as representations of the results obtained, all DSF curves are available in the appendix. As seen above in Figure 3-7a, an unboiled sample of `I17A` does not maintain its pentameric structure on SDS-PAGE, however, a monomer band was seen at the dye front. DSF of this mutant, Figure 3-8a, revealed a $T_m$ of 61 °C confirming the presence of protein and that it had been destabilised. However, DSF of the `G45A` mutant, Figure 3-8d, revealed no melting curve. This trace is useful to show the levels of underlying signal cause by contaminants remaining from the nickel affinity purification; at ~25 kDa and 66 kDa in the Figure 3-7a. These contaminants are not at concentrations required to compete with signals attributable to CTB. The DSF of mutant `F25A` seen in Figure 3-8c shows a left hand shift in $T_m$ as the concentration of CTB increases, which is in line with the DSF data shown for the wild-type CTB.

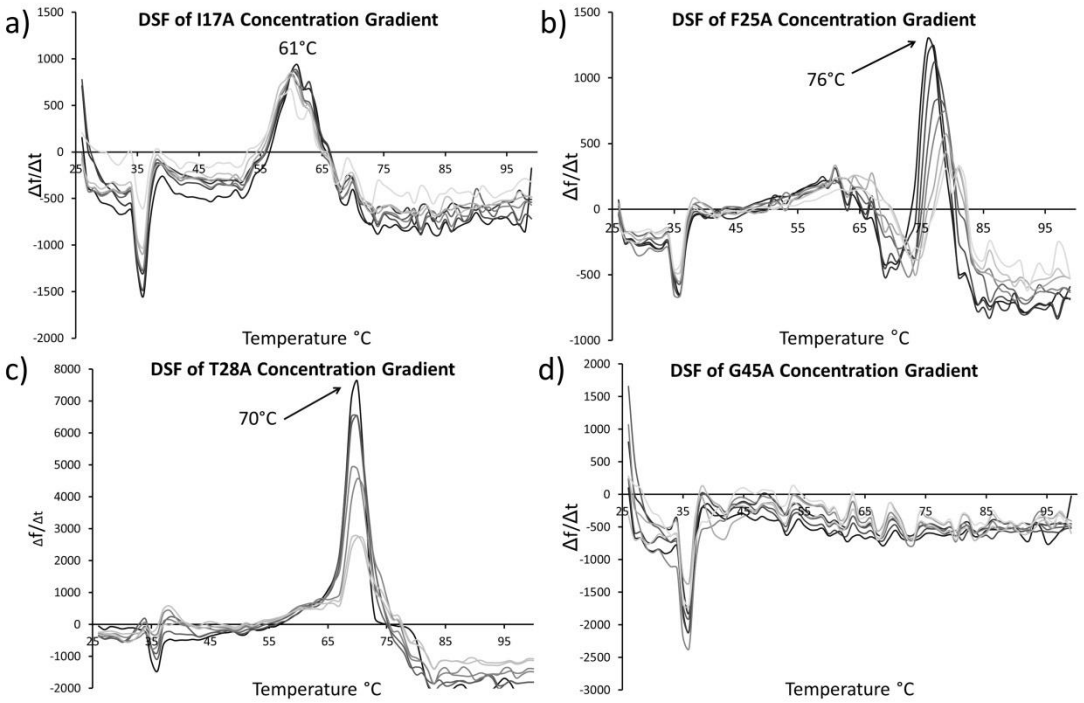



**Figure 3-8:** Example DSF traces of alanine scan mutants. **a)** shows DSF of I17A which has a relatively low $T_m$ and weak signal intensity resulting from a low level of protein expression. **b)** a DSF trace of F25A shows a $T_m$ that deceases with increasing concentration, as discussed in the main text this is presumably due to competition with the trough typical for low concentrations of CTB. **c)** shows a representative trace from the majority of the other DSF traces, in this case T28A. **d)** shows the DSF trace for G45A: no melting peak is observable suggesting a lack of a major protein species in the sample.

The majority of the DSF traces do not show this left hand movement of the $T_m$ with increasing concentrations. Some traces show an initial lowering of the $T_m$ with increasing protein concentrations, , however, the $T_m$ then becomes constant at higher protein concentrations. Additionally, this phenomenon seems to be independent of a concentration threshold as both fixed $T_m$ and shifting $T_m$ are observed among preparations of both, comparatively, high and low protein concentrations. Of interest, the locations of mutations which do cause this left hand shift are clustered to a specific location on the CTB pentamer shown in purple in Figure 3-9. This observation shall not be expanded on here, but perhaps points to the occurrence of an interesting kinetic event in that area of the protein.

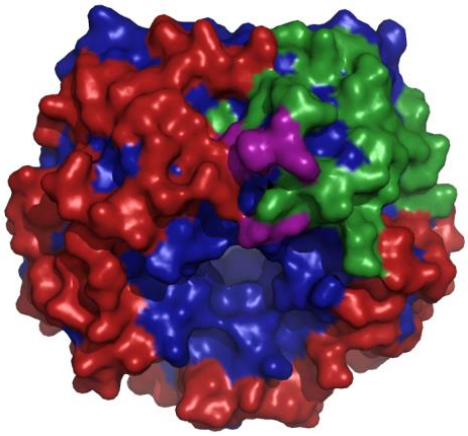


**Figure 3-9:** Location of mutations which cause a reduction in $T_m$ with increasing concentrations, by DSF. Alanine scan mutations are highlighted in red against the blue of the un-mutated protein. Green residues highlight the mutable resides present in a single monomer. The purple residues show the location of residues which present a reduction in $T_m$ by DSF upon increasing protein concentration. Produced from pdb 3CHB.

Figure 3-10 shows the combined DSF data across the alanine scan carried out in replicates of eight. As can be seen, the majority of the alanine scan mutants do not perturb the $T_m$ of CTB by more than 5 °C. However the yield of some mutants allowed no reliable $T_m$ could be determined, e.g. `T41A` (low signal), and `G45A`. Others mutants showed concerning levels of thermostability such `N4A`, `I17A` and `K84A`, which suggests these positions may not be suitable for mutation. This screen makes suggestions as to the structural significance of each residue and will inform future mutational propositions.

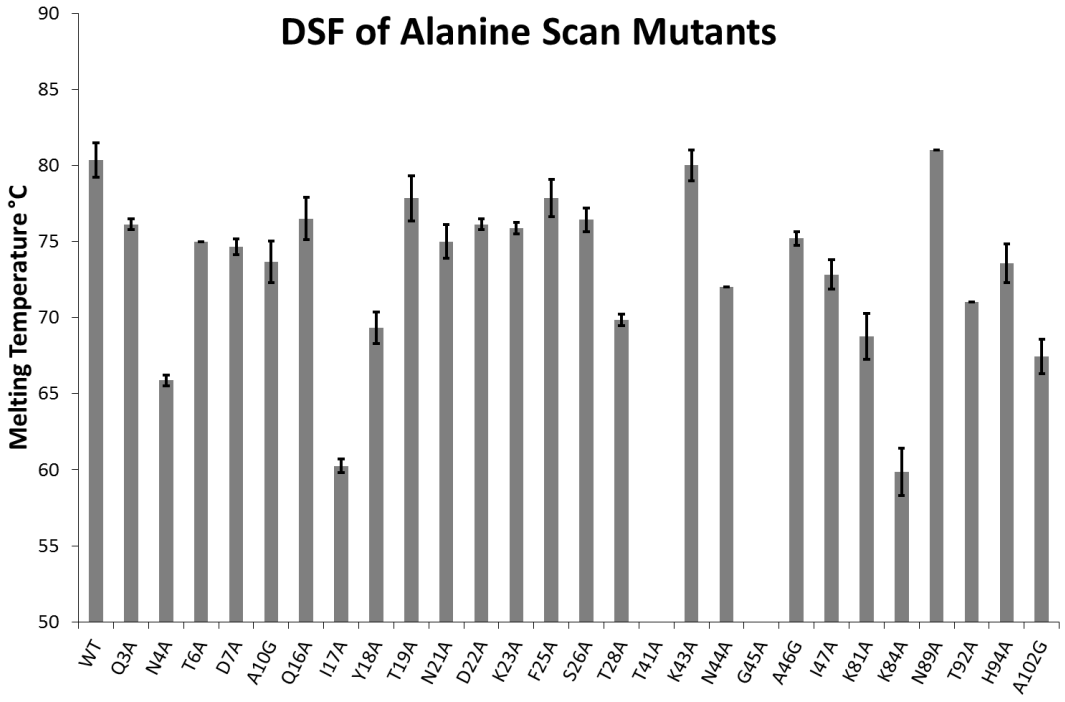



**Figure 3-10:** The total DSF derived $T_m$ across alanine scan mutants. Each alanine Scan mutant's $T_m$ assessed by DSF is presented here. It should be noted that data for E83A is not available. Most alanine scan mutants reduce the $T_m$ of the pentamer to approximately 75 °C and others seem to cause more drastic reductions in $T_m$. This suggests that making mutations at position 17, 41, 45 and 84 would produce structurally deleterious effects to the protein.

## 3.2.4 Comparison of Computational and Experimental Alanine Scanning

The FoldX 'alanine scan' function was used to predict a $\Delta\Delta G_{stability}$ upon mutation compared to the wild-type CTB. The predictions were compared directly to the $T_m$ determined from the DSF experiments. It should be noted that this comparison is rather superficial as the thermostability of protein and the free energy change from folded to unfolded protein have a convoluted physical relationship. However, there does seem to be a degree of correlation between the computational and experimental values. Figure 3-11 shows a scatter graph of the $\Delta\Delta G_{stability}$ compared to the $T_m$ derived from DSF and a bar chart presenting the same data (note that the sign on the $\Delta\Delta G_{stability}$ value has been inverted for ease of comparison). The scatter graph (Figure 3-11a) suggests a degree of trending in the data. Figure 3-11b shows a bar chart of the DSF-derived $T_m$ compared to the FoldX derived $\Delta\Delta G_{stability}$ in kcal/mol. The axis of the $\Delta\Delta G_{stability}$ values were set so that 0 kcal/mol aligns with a $T_m$ of 80 °C to represent the relative wild-type data. This depiction of the data emphasises which mutations have a $T_m$ that is reduced in accordance with the predicted

loss in $\Delta G_{stability}$ predicted by FoldX. Obviously these data do not correlate exactly but it should be expected that high speed computational assessment such as the FoldX 'alanine scan' function will have difficulty matching experimental data confidently. Nevertheless the majority of the data agrees well. It should be noted that protein was not obtained from the G45A mutant and G45A has the lowest $\Delta\Delta G_{stability}$ among those assessed. Additionally $T_m$ data for E83A was not available. The destabilisation of T19A and I47A substitutions were overestimated by FoldX and the estimations for T6A, Y18A, F25A, S26A, K81A, K84A and H94A were underestimated.

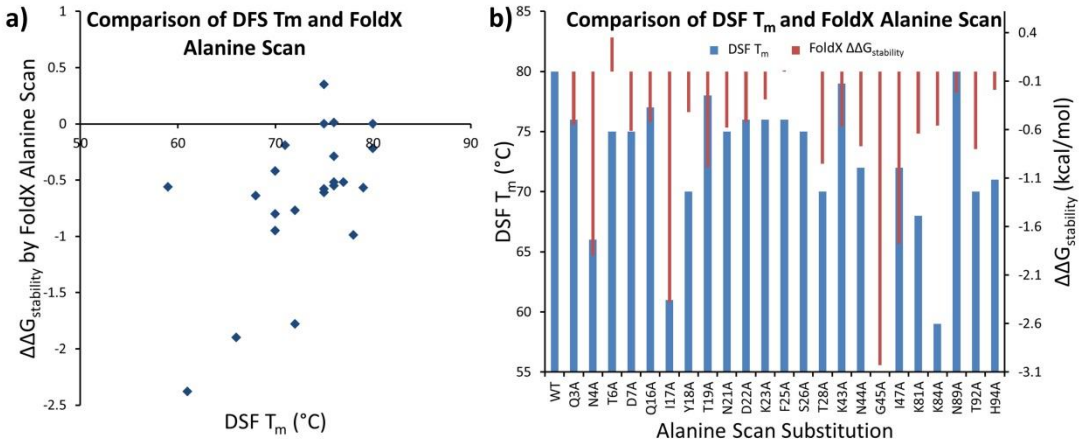



**Figure 3-11:** Comparison of computational and experimental stability assessments. **a)** A scatter graph of the DSF derived $T_m$ against the FoldX 'alanine scan' derived $\Delta\Delta G_{stability}$ in kcal/mol. **b)** The same data compared by way of a bar chart. The $\Delta\Delta G_{stability}$ of 0 kcal/mol is aligned to a $T_m$ of 80 °C, to represent the wild-type data; the ΔΔG axis was then manipulated to best agree with the DSF $T_m$. White space and overlaps represent areas of disagreement between the data sets, however, most of the data between the sets agrees well.

# 3.3 Concluding Remarks

A synthetic gene construct was obtained from Genscript and spliced into the plasmid pMAL p5x. This new plasmid pSAB2.0 allowed the production of two derivative plasmids pSAB2.1 and pSAB2.2, which allowed the expression of the MBP-$AB_5$ complex or CTB respectively. Expression and purification procedures for both of these products were developed to yield a pure protein product. The MBP-$AB_5$ complex and CTB were assayed by DSF in order to determine melting temperatures for the protein. Although this process required a large degree of refinement for CTB, reliable melting temperatures were achieved when concentrations of CTB above 100 µM were assessed.

Residues were chosen which lay on the side of the CTB pentamer, where we propose our interaction to take place. These residues were chosen to cover the breadth of the side face of the protein and used amino acids whose side-chains interacted with the solvent. These chosen residues were substituted for alanine residues by either site-directed mutagenesis or through assembly PCR of a new CTB gene. The alanine scan mutants were then expressed, purified and unboiled samples were visualised by SDS PAGE. The majority of the mutants maintained pentameric stability on the gel, however, some were found not to express, e.g. G45A, and others dissociated into monomers on the gel, e.g. I17A. Each of the mutants was then assessed by DSF and their $T_m$ was elucidated.

A computational alanine scan was performed on a group of CTB conformation retrieved from a CHARMM molecular dynamics simulation on the 3CHB.pdb structure. These conformations were assessed by the FoldX function 'alanine scan' which, when compared to the wild-type, gave a $\Delta\Delta G_{stability}$ in kcal/mol for the mutant protein compared to the stability of the wild-type. On comparison of the experimentally derived $T_m$ for the alanine scan mutants with the $\Delta\Delta G_{stability}$ produced by FoldX, a reasonable correlation in the data was observed. It should be considered that FoldX has no resource for determining the likelihood of the sequence to fold into the correct conformation and can consider only the stability of the structure submitted to it. Based on these constraints to the direct comparison of computational data to experimental data, we concluded that the observed agreement is satisfactory for future use.

# Chapter 4

# Intrinsic Scaffolds

# 4.1 CTB with C-terminal Coiled-Coil Peptides

## 4.1.1 Overview

As described in the project summary, the *C*-termini of the CTB pentamers are extended with coiled-coil fusions. The display of these oligomerisation sites at the correct geometry should allow construction of an enclosed particle. A number of design considerations are highlighted here. Due to the tapering of the CTB pentamers and the location of the *C*-terminal coiled-coils (Figure 4-1), it should be more energetically favourable for the coiled-coils to be housed on the interior of the capsid. Therefore the GM1 binding sites should be displayed on the exterior of the particle. A particle scaffolded on its interior requires that the total cross-sectional area of the scaffold is less than the total surface area of a sphere with a diameter corresponding to the distal end of the coiled-coil projection to ensure that steric clashes do not occur between the coiled-coils. Thus, using coiled-coils of three heptads or more precludes the formation of a dodecahedron and forces the geometry of the particle into a more expanded structure. In order to create a more geometrically complex, expanded particle, the association between pentamers must be somewhat promiscuous as the interfaces cannot bind to one another in only one conformation as seen with a dodecahedron. This point requires consideration for both the scaffold and for the CTB interface. Finally, the association of the particle by way of homomeric scaffolding requires 'inter-pentamer' interactions to be more energetically favourable than 'intra-pentamer' interactions between coiled-coils.

## 4.1.2 GM1 binding site presentation.

The location of the *C*-terminus of the CTB monomer is on the face opposite to the GM1 binding site (Figure 4-1a). The edge length of the GM1 binding face measures 37.5 Å, measuring between H13-$C_\beta$, whereas the edge length of the *C*-terminus face measures 31.9 Å, measuring between N44-$C_\alpha$. This makes the proposed interaction face tapered with respect to the top and bottom face of the pentamer (Figure 4-1b). This tapering, given the orientation shown in Figure 4-1c/e, brings the *C*-termini into close proximity such that protruding coiled-coils can be designed to interact in an inter-pentamer arrangement among the entire length of the coiled-coil. In contrast, the CTB orientation shown in Figure 4-1d/f would require either a flexible linker region or a portion of unbound coiled-coil to bridge the distance between the *C*-termini of the interacting pentamers. As any

homomeric scaffolded assembly would be required to compete with intra-pentamer associations of the extensions; the former orientation provides a thermodynamic advantage to assembly compared to the latter. Thus, the orientation of the assembly formed should be controlled thermodynamically using coiled-coils of a reduced length.

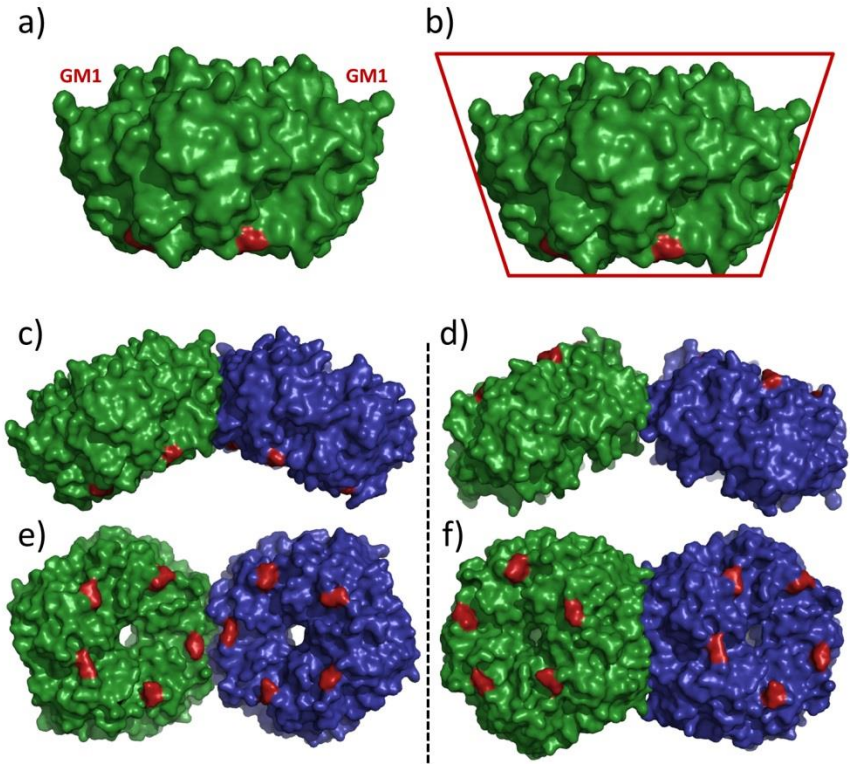



**Figure 4-1:** The *C*-terminus and CTB tapering. The location of the *C*-terminus of CTB should make coiled-coil fusions favourable for constructing enclosed particles. **a)** The side profile of CTB with the GM1 binding site located at the top and the *C*-terminal residue highlighted in red. **b)** The CTB pentamer with native tapering of the molecule highlighted by a wire frame. **c)** and **e)** The proximity of the C-termini given an orientation of CTB pentamers that would lead to an enclosed particle presenting the GM1 binding face on the exterior. **d)** and **f)** The proximity of the C-termini given an orientation of CTB pentamers that would lead to an enclosed particle presenting the GM1 binding face on the interior. Produced from pdb 3CHB.

### 4.1.3 An Expanded Dodecahedron

The addition of coiled-coils with lengths of three or more heptads should preclude the formation of the dodecahedron platonic solid (Figure 4-2a) as the internal space of this dodecahedron is insufficient to contain the volume occupied by the 60 extensions (Figure 4-2b). However, increasing the diameter of the capsid will in turn increase the internal volume relative to the volume occupied by the coiled-coils. A more precise interpretation of this theory is that the dihedral angle between pentagons in a dodecahedron is 116.56°, by expanding the capsid this dihedral angle increases. Steric clashes between neighbouring coiled-coils is dependent on both the extensions' length and the particle's dihedral angle:

increasing this angle towards 180° (where infinite extension length would result) eases the likelihood for a clash to occur. Larger capsids will have a greater number of extensions but also a greater dihedral angle in order to accommodate the extensions where smaller capsids would not.

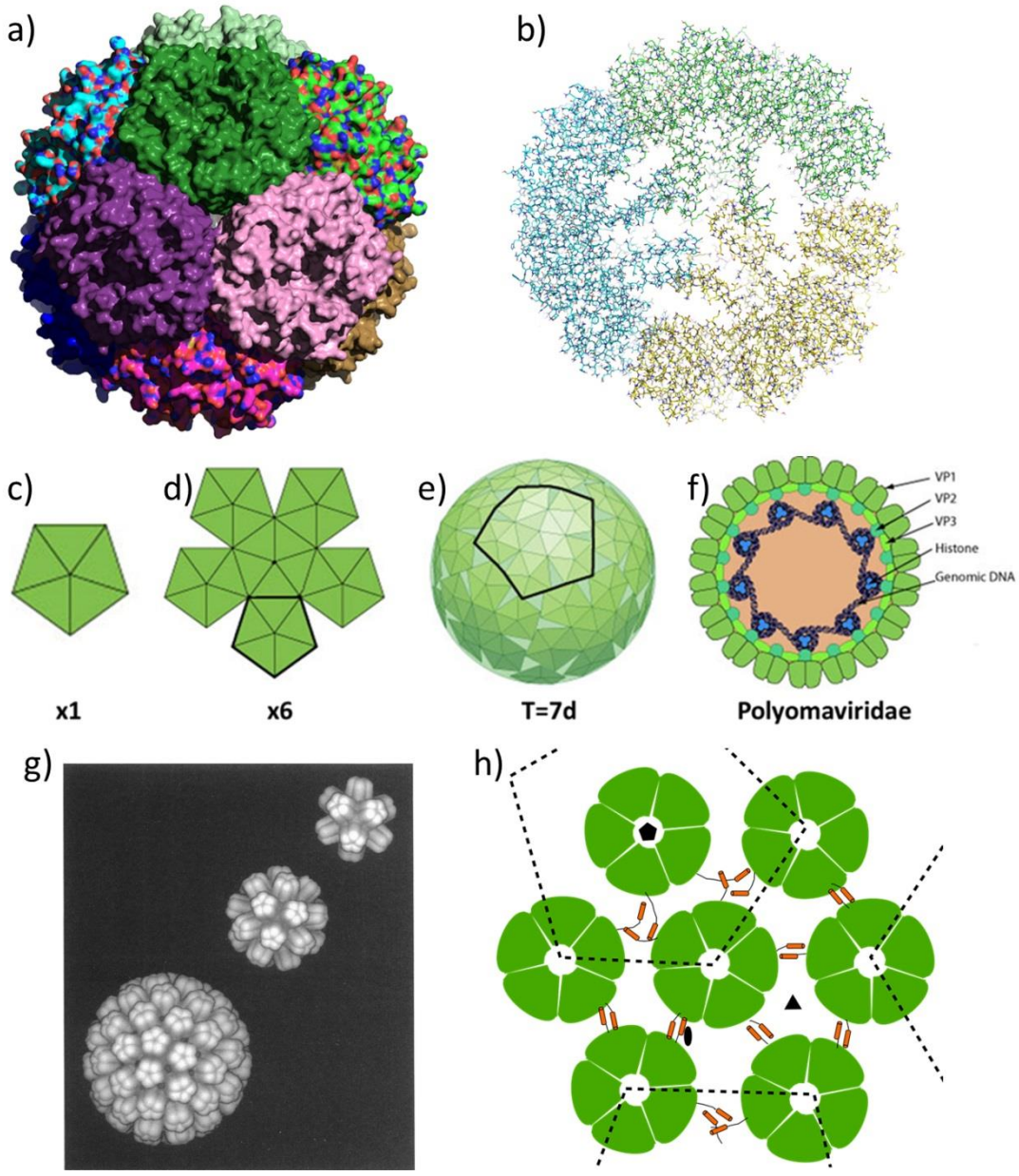



**Figure 4-2:** Pentagon tiling and virus subunit assembly. **a)** A proposed dodecahedron of CTB pentamers. **b)** The interior of a CTB dodecahedron scaffolded with coiled-coils of three heptads in length: only 2/5 of the coils are present in this image, which exemplifies the steric clashes of the particle interior. a) and b) produced from pdb 3CHB. **c)** A pentagon. **d)** Tiling arrangement of six pentagons to give one large pentagon, 'super pentagon' with triangular gaps. **e)** The super pentagon arranged in a distorted dodecahedron with T=7d symmetry, giving a super dodecahedron. **f)** A biological example, Polyomaviridae, which assembles 360 subunits (72 pentamers) into to a capsid with T=7d symmetry. Image modified from ViralZone[159], by permission of Oxford University Press. **g)** A dodecahedron, an intermediate, octahedral-based symmetrical particle and the T=7d expanded capsid. Reprinted from[109]. Copyright 1989, with permission from Elsevier **h)** A schematic representing the five different interactions required between subunits in the T-7d expanded capsid, based on a figure from [105].

What options are available for larger capsids to be made from pentagonal subunits? Figure 4-2c-e shows steps in the assembly of a pentagonal building block by arranging five pentagons around a central pentagon, a super pentagon. This super pentagon is similar to a flattened hemi-dodecahedron, Figure 4-2d, where the dihedral angle of this association is greater than 116.56° and gaps are formed between the peripheral pentagons. The dihedral angle is increased to the point where the external interaction faces of the peripheral pentamers become parallel, in a hyperbolic plane, with the external face of the neighbouring peripheral pentamers, creating a single large, yet curved, pentagon made from six smaller pentagons. This can then assemble into an expanded dodecahedron (Figure 4-2e) which possess T=7d symmetry. An intermediate sized particle of octahedral symmetry is also accessible for construction from pentagonal subunits (Figure 4-2g). Interesting biological examples of viruses with T=7d symmetry are the *polyomaviridae* and *papillomaviridae* families, which are both dsDNA viruses with capsids of 45 and ~54 nm diameter, respectively. As discussed in the introduction, these viral families have been observed to have a dynamic nature of protein-protein interactions in capsid assembly, more so when manipulated in the laboratory.

The interfaces of the T=7d capsid do not associate in just one arrangement. In order to create a more complex, expanded particle, the association between pentamers must be somewhat promiscuous, specifically pentagamous (a marriage of five!), where five different conformations of the interface are possible (Figure 4-2h). This is exemplified in the structure of Simian virus 40[105] (Figure 4-3). These interfaces are determined by a donor strand exchange scaffolding structure, positioning the globular units in differing interface orientations. However, the globular subunits must not interact repulsively across any of these interfaces. The 5-fold axis pentamer interacts with its periphery pentamers via a 3-way domain swap, and an identical globular interface. The intra-super-pentagon interaction of the peripheral pentamers use the same 3-way donor strand exchange, but the globular interface deviates from the 5-fold interface in such a way that a triangular gap is formed between them. Peripheral pentamers interact inter-super-pentamer in a further three ways, about the 2-fold axis, the 3-fold axis and an additional interface. Each of these interactions are defined by 2-way donor strand exchanges and each requires a different PPI for the globular subunit. This diversity in binding conformations is highlighted in Figure 4-3b which shows dimers of pentamers representative of the interactions found in Figure 4-3a. Each dimer of pentamers is aligned about the lower pentamer and the upper pentamer is allowed to sit according to the position of the lower pentamer. As is distinctly

apparent a range of binding conformations are available between the pentamers, with the red dimer of pentamers, representing the 5-fold axis interaction, sitting to the extreme left and the orange structure, representing the 3-fold axis interaction, sitting to the extreme right.

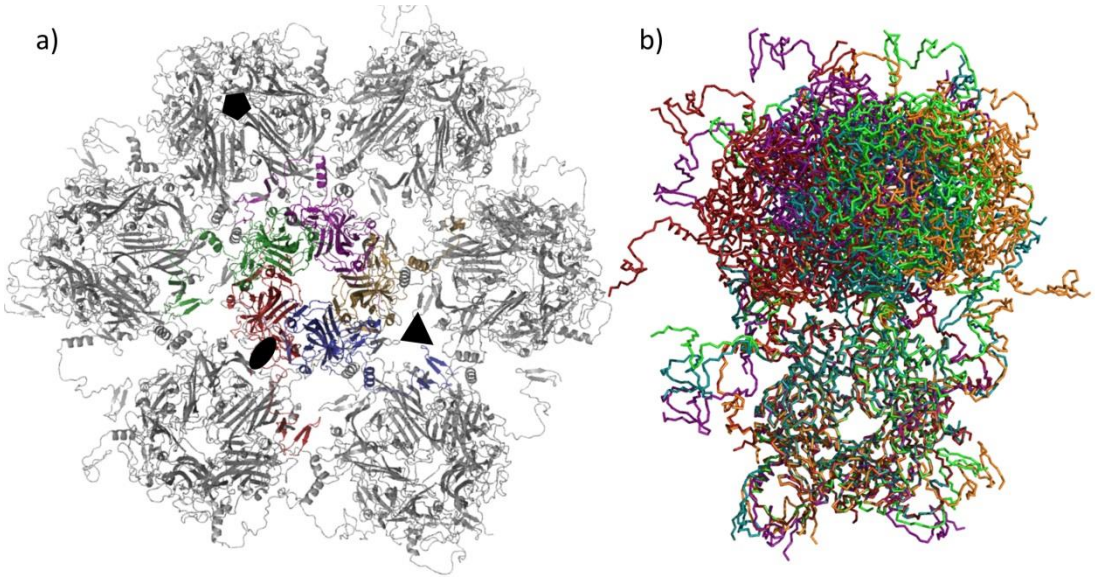



**Figure 4-3:** Simian virus 40 tiling and pentamer superposition[105]. **a)** The crystal structure of Simian virus 40 pentamers positioned as highlighted in the schematic in Figure 4-2h. **b)** An alignment of dimers of pentamers from the Simian virus 40 crystal structure. Showing the polydiverse nature of the binding orientations of the upper pentamers compared to the lower pentamer. Produced from pdb 1SVA

#### 4.1.3.1 Calculating the Expected Capsid Size

Estimating the size of the capsids formed from CTB pentamers by the tiling theory above requires a number of assumptions. The only structural information available are the dimensions of the CTB pentamer, these can be compared to the dimensions of the other T=7d forming pentamers such as the SV40 coat protein. In order to standardise the measurement, as edge lengths of the T=7d capsid vary, the centre of mass was used to make measurements. This measurement provided a rough figure for comparison with particle sizes observed with TEM and DLS. The Simian virus 40 capsid (from PDB 1SVA) measures 48.4 nm across its diameter, the average distance between the centre of mass of neighbouring pentamers measures 9.7 nm and the distance between the centre of mass of monomers within pentamers measures 3.4 nm. Using these parameters we can scale a hypothetical capsid size for the CTB particle with T=7d symmetry using the distance between monomers in the pentamer, measuring 2.2 nm, which gives a scaling ratio of 1.53. Application of this ratio to the sizes observed in SV40 estimates a particle size of 31.6 nm.

### 4.1.3.2 Coiled-Coil Quaternary Structure

The C-terminus of CTB lies between the suggested 2-fold and 3-fold axes, of a putative dodecahedron. It is thus in an ideal position to create dimeric and trimeric scaffolding interactions which could in turn produce particles from pentagon shaped subunits. Trimeric coiled-coil extensions could assemble about the 5-fold axis pentamer (Figure 4-4a); likewise, dimeric coiled-coil extensions could assemble about the 2-fold and 3-fold axis (Figure 4-4b). This heterovalency creates a degree of proximal binding freedom allowing experimentation with both types of coils in the assembly mutants. Moreover, to meet the criteria of expanded particle construction from pentagons, the use of coiled-coils which can form both dimeric and trimeric interactions is the most advantageous. Additionally, as the assembly will be mediated to some respect by the CTB interface, i.e. electrostatics, shape complementarity, hydrophobic regions etc., these coiled-coil interactions might be forced into non-ideal binding conformations. These individual coiled-coil motifs, once initial nucleation of the assembly has occurred, will be held by the assembly in the proximity and polarity for further subunit association. Given two homotrimeric coiled-coil motifs held in proximity, it is rational to assume that it is more energetically favourable for a dimeric complex to form from these putative trimeric coils, in the absence of a third binding partner. Indeed it is also rational, though possibly less energetically favourable than above, that if three homodimeric coiled-coil motifs are held in proximity that they could form trimeric complexes over the more favourable dimeric complex based on the satiation of the unbound coiled-coil motif's binding requirements. Many coiled-coils indeed have a preference for a particular oligomerisation state but are shown experimentally to exist in a dynamic equilibrium between monomer, dimer, trimer etc. In order to assess the ratio of multimerisation of the coils selected we performed a number of studies to deduce the quaternary structure of the coiled-coils used, which is discussed below.

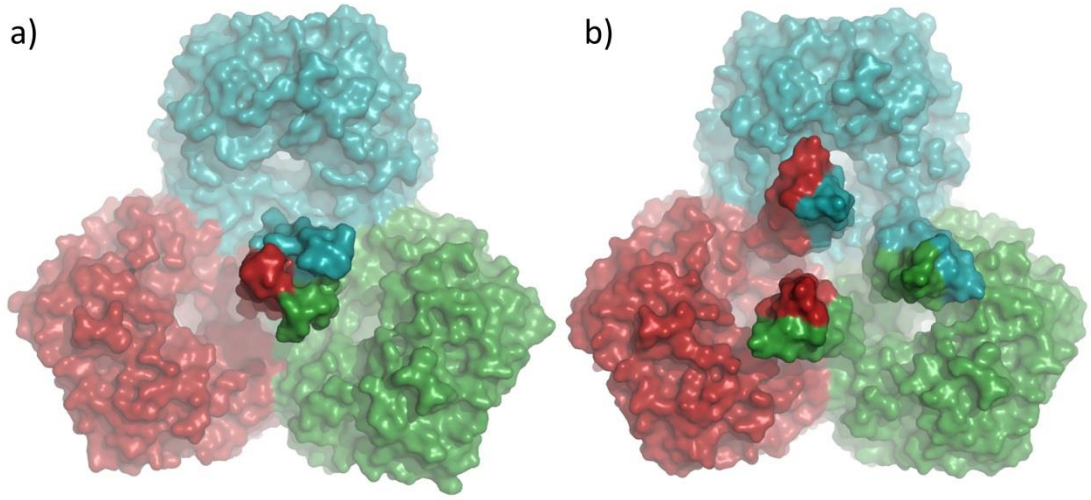



**Figure 4-4:** Trimeric and dimeric coiled-coil extensions of CTB. **a)** *C*-terminal trimeric coiled-coil extensions of CTB can complex about the putative 3-fold axis of a dodecahedron. **b)** *C*-terminal dimeric interactions could also form around the 3-fold symmetry axis. Produced from pdb 3CHB.

Coiled-coils are ideally suited to provide a platform for an interaction surface capable of forming both dimeric and trimeric interactions, dependent on the predicted oligomeric state and proximity of binding partners. This in turn makes them ideal as scaffolds which can bridge the gap between structurally supporting platonic solids (homovalent coiled-coils) as well as more complicated assemblies which contain polymorphic PPI sites between subunits (heterovalent coiled-coils).

## 4.1.4 Choosing a Starting Point for Design

De Crescenzo et al. (2003)[64] designed a parallel heterodimeric coiled-coil with the canonical heptad sequence VSALKEK for the 'coil K' and VASLEKE for the 'coil E'. This sequence included the hydrophobic residues valine and leucine in the 'a' and 'd' positions and electrostatic interactions formed between the dimers from glutamate and lysine residues at the 'e' and 'g' positions. The interaction between differing heptad lengths was measured by Surface Plasmon Resonance and the five heptad versions of coil E and coil K were found to have a 6.3 pM $K_d$[64]. The high stability made these sequences an attractive starting point for the design of coiled-coil scaffolds for CTB assembly. However, a homomeric coiled-coil would be advantageous to bring about the assembly of a closed spherical object as it requires just one protein to be expressed, rather than expressing a

pair required for heterodimerisation, which in turn may lead to assembly difficulties via electrostatic clashing between two coils of the same type (i.e. K to K, or E to E).

The use of homomeric coiled-coils presented from the same face of the CTB protein potentially provides the complication of intra-pentamer self-association. Given large coiled-coils, such as a five heptad repeat, the most entropically favourable arrangement is for the coiled-coils to associate with additional coiled-coil extensions emanating from the same CTB pentamer. However, the first heptad emanating from the *C*-terminus of CTB is unable to contribute to an intra-pentamer coiled-coil (Figure 4-5a) as the distance between adjacent *C*-termini is too large. However, the position of the *C*-terminus would allow length-long association between coiled-coil extensions for inter-pentamer interactions (Figure 4-5c). Thus, the inter-pentamer association is always the most enthalpically favourable from the point of view of coiled-coil formation. Therefore shorter coiled-coil motifs will form inter-pentamer interactions in preference over intra-pentamer coiled-coils. However, intra-pentamer interactions could prove problematic for larger coils. Three heptads repeats were chosen as a likely candidate to provide an effective inter-pentamer scaffold while deterring intra-pentamer associations with neighbouring extensions (Figure 4-5b). Although de Crescenzo et al. found that three heptad versions of the same coiled-coil bound too weakly to measure a $K_d$, here the multivalency of the system should allow these weakly binding coiled-coils to complex larger assemblies.

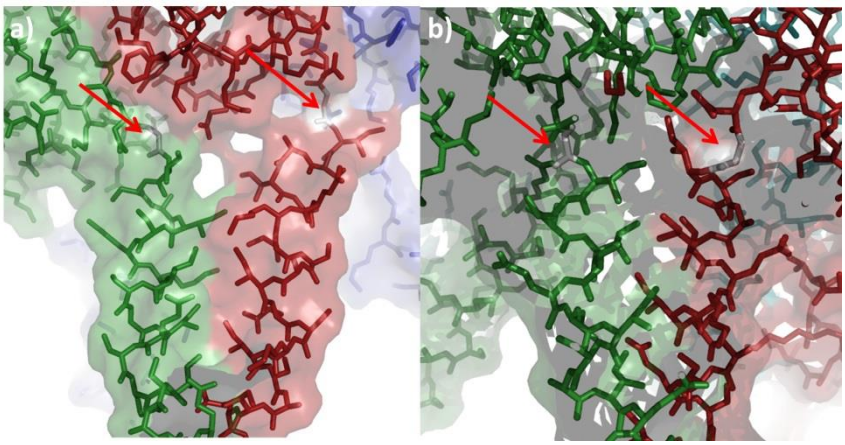



**Figure 4-5:** Intra-pentamer coiled-coil interactions compared with inter-pentamer coiled-coil interactions. **a)** Three-heptad coiled-coils associate intra-pentamer through two of the three heptads, while the first heptad binds with few residues. This model was constructed with foldit[160]. **b)** The inter-pentamer coiled-coil association however, allows coiled-coils to interact with one another directly as they protrude from the CTB *C*-terminus. The white residues highlighted by the red arrows are G104 which is the first residue of the extension. The inter-pentamer association should be favoured, enthalpically over the intra-pentamer association based on the greater interaction surface. Produced from pdb 3CHB.

### 4.1.5 Coiled-coil Designs

Two *de novo* intrinsic scaffolds were designed and evaluated for assembly of wild-type CTB, which are listed below with a brief description. Each sequence follows directly from residue 103 of CTB. Possible electrostatic interactions are indicated in the respective colours and changes to hydrophobic residues are green. The de Crescenzo (DeCr) coils are listed for reference only. Although only two new coiled-coils are described here, several other peptide sequences were investigated and are described in the appendix (10.2.4.1). The other sequences were less effective than the two described here for producing higher order assemblies.

```
DeCr-coilK GGGKVSALKEKVSALKEKVSALKEKVSALKEKVSALKEG
DeCr-coilE GGGEVSALEKEVSALEKEVSALEKEVSALEKEVSALEKG
JR-coil2   ----------------GVSALEKEVSALKEKVSALEF
JR-coil7   ---------------GGISAISEKISAIESKISAIES
```

**DeCr-coilK:** The coil K described in de Crescenzo et al. (2003)[64]. **DeCr-coilE:** The coil E described in de Crescenzo et al.

**JR-coil2:** In order to allow homodimerisation of the coil, the electrostatic interactions were mixed from coil E and coil K motifs. The *N*-terminal end of the coiled-coil was truncated to remove its electrostatic interaction in case this would interfere with the surface of CTB, which left two heptads-worth of electrostatic interactions and three heptads-worth of hydrophobic core. Originally it was thought that by introducing a *C*-terminal phenylalanine in to the coiled-coil the assembled coil would have a hydrophobic cap and potentially induce clustering of the other assembled coils in proximity. Upon reflection this was probably unlikely to help in assembly and more likely to create VLPs with a hydrophobic interior surface.

**JR-coil7:** This coil was designed to alleviate the potential for weak but potentially problematic antiparallel interactions, by reshuffling the electrostatic interactions. Additionally, the literature suggests that introducing isoleucine into the ‘a’ and ‘d’ positions of the heptad promotes trimeric interactions over dimeric interactions[161-163] as two coiled-coils are unable to adequately shield the larger hydrophobic centre from the bulk solvent. This coil implemented this theory to create a putative homotrimer.

# 4.2 Expression and Characterisation

## 4.2.1 MBP Coiled-Coil Fusions

The construction of the plasmid, the coiled-coil fusions and the expression and characterisation of the MBP-coiled-coil fusions was conducted by Michael Johnson, University of Leeds, for his Masters project. Plasmid maps and assembly part sequences can be found in the appendix (10.2.4.2).

In order to determine the oligomeric states of the designed coiled-coils, JR-coil2 and JR-coil7, these were made as *C*-terminal extensions of MBP. Chapter 3 describes the construction of the pSAB2.1 plasmid which is used to express MBP fusion proteins, possessing a TEV protease site, to the periplasm. This construct was used to clone the TEV protease domain into the pMal-c5x plasmid (which expresses protein in the cytosol), via digestion with SacI and PstI, the pSAB2.1 insert was spliced into the pMAL-c5x plasmid to give a new vector named pMaCo. This plasmid directed expression to the cytoplasm and allowed coiled-coil fusions to the *C*-proximal side of the TEV protease site. This allowed the option of removing the coiled-coil from the MBP fusion should it become necessary.

The genes for the coiled-coils were constructed by assembly PCR from short overlapping sections of ssDNA, which were subsequently spliced into the pMaCo plasmid using the BamHI and PstI restriction sites. Expression was induced and the fractionated cell lysate was purified by amylose affinity chromatography. Elutions from the amylose resin were analysed on SDS PAGE and revealed a high yield of MBP coiled-coil fusion which was adequately purified from the vast majority of the contaminating proteins (Figure 4-6a). Both JR-coil2 and JR-coil7 when expressed, concentrated and incubated at room temperature for 2 days, revealed no high weight oligomeric states after SEC. Therefore, 4-heptad versions of these coils were constructed and analysed. These 4-heptad versions gave a concentration-dependent decrease in SEC retention volume after incubation. Figure 4-6b shows the SEC for the 4-heptad version of MBP-JR-coil2. The red and black lines show MBP and MBP with a known trimeric coiled-coil attached[164]. The traces from orange to blue show increasing concentrations of the MBP coil fusion. With increasing concentrations, there was an increase in the size of the particle which suggests the presence of trimeric, dimeric and monomeric species. SEC for the 4-heptad version of MBP-JR-coil7 (Figure 4-6c) shows a major dimeric species and a smaller proportion of trimeric species. This result was somewhat unexpected as the major predispositions of the

coiled-coils were the opposite way round to expected. For example JR-coil7 which contained isoleucine at the a and d heptad positions was expected to favour trimeric interactions, however, the SEC suggests dimeric species are in the majority across the range of concentrations tested. In contrast, JR-coil2 with valine and leucine at the a and d positions respectively was presumed to favour dimeric interactions, however, given concentrations above 5 mg/ml these coils favoured trimeric interactions. Despite this inversion in the predicted behaviour of the coils, the purpose of this exercise was to find coiled-coils which could form both dimeric and trimeric interactions, which both coils do. Additionally as one coil prefers dimeric interactions and the other prefers trimeric interactions, these coils provide a good starting point for investigating which coil produces capsids more efficiently. These investigations were conducted on 4-heptad versions of the coiled-coils, however when investigating these coiled-coils on CTB it was decided that reverting to 3-heptad coiled-coils was more appropriate. This change in heptad length would reduce crowding on the capsid interior and would lead to a preference for fully formed capsids over partially formed capsid intermediates, which would likely have stability problems with just three heptads. Additionally as future mutations in the CTB-CTB interface will likely increase the assembly rate independent of the heptad length, a low level of initial assembly with the wild-type CTB is an advantageous position for determining the kinetic effect on capsid assembly of any introduced mutations.

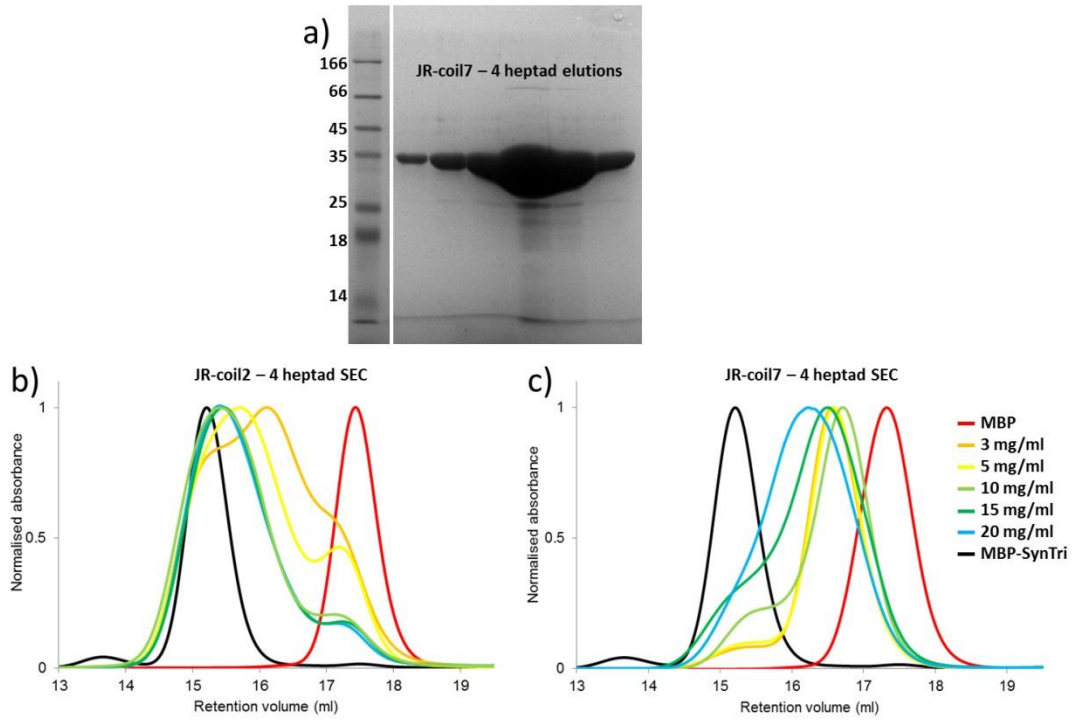



**Figure 4-6:** MBP fusions with designed coiled-coils. **a)** An example gel of the eluted fractions from the amylose affinity column: the MBP fusion is expressed to a high yield and is well purified from other contaminants. **b)** The SEC trace from the 4-heptad version of JR-coil2. The red line depicts the trace from MBP with no coiled-coil fusions. The black line shows the trace from an MBP fusion with a known trimeric coiled-coil, SynTri[164]. This trace shows that with increasing protein concentrations of the fusion that the propensity for trimeric oligomerisation increases, however there is always a portion of monomeric fusions as well as a proportion of dimers. **c)** The SEC trace from the 4-heptad version of JR-coil7, as described in b). With this coil there is a greater proportion of dimeric interactions than trimeric interactions, even when reaching concentrations such as 20 mg/ml, however there is a proportion of trimeric coils seen at all concentrations.

## 4.2.2 CTB-JR-coil2

Coiled-coil JR-coil2 was assembled with wild-type CTB and inserted into plasmid pSAB2.2, which was expressed and purified. In order to assess the assembly of the capsids, purified samples were concentrated to 10 mg/ml and incubated at room temperature for differing periods of time from two days to four weeks. Samples were then analysed by SEC (Figure 4-7a). CTB expressed without *C*-terminal extensions was also incubated for 4 weeks at room temperature and can be seen as the red trace. The 1 week, 2 week and 4 week traces produced by the CTB-JR-coil2 showed a high weight peak that gave higher yields with longer incubation time suggesting the accumulation of a high weight species with time. As seen in the inset panel, these high weight peaks occur just after the void volume of the column (left-hand shoulder of the peak).

In order to confirm the presence of the CTB-JR-coil2 protein in these high weight peaks, fractions were analysed, both boiled and unboiled, on SDS-PAGE (Figure 4-7b). This gel shows that CTB makes up the majority of the high weight peak. The gel also suggests that the protein in this high weight peak is not the product of aberrant aggregation of unfolded protein, as pentamers are clearly visible on the unboiled fractions. The SEC high weight peaks were combined and concentrated before being analysed by DLS.

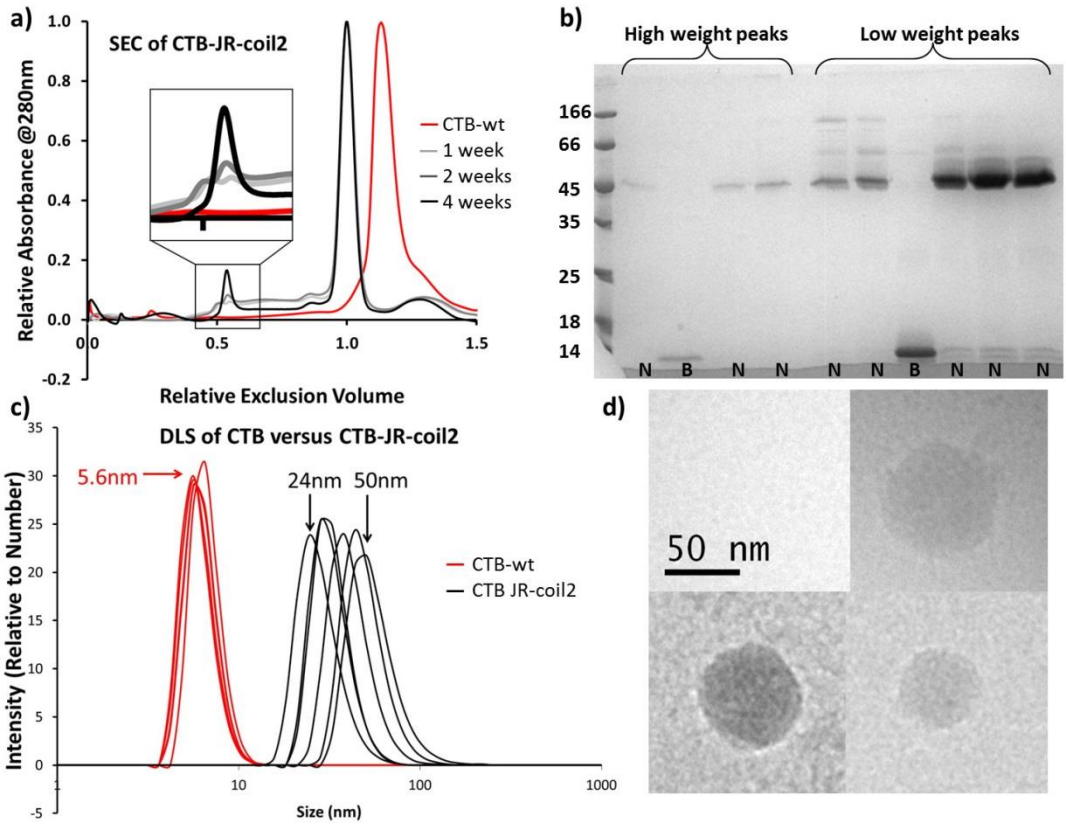



**Figure 4-7:** Characterisation of CTB with JR-coil2**. a)** Size exclusion chromatography of wild-type CTB in red and of CTB with JR-coil2 in grey to black. The darker the lines indicate extended periods of incubation at room temperature from 1-4 weeks, long incubations gave higher yields of high weight product. **b)** The high weight fractions from SEC are seen to contain pure CTB when subjected to electrophoresis, both boiled (B) and unboiled (N) and stained with coomassie blue. **c)** Dynamic light scattering of the wild-type CTB, red, suggests particle sizes of 5.6 nm, however the high weight peak from SEC of the CTB-JR-coil2, in black, gives particles of between 24 and 50 nm. **d)** Transmission electron microscopy of the high weight SEC fraction reveals particles ≥ 50 nm. Very few particles were observed.

DLS of CTB without coiled-coils gave 5.6 nm for the estimated spherical size of CTB (Figure 4-7c) in line with the pentamers width of 6.3 nm and height of 3.6 nm. However the CTB-JR-coil2 high weight component from SEC gave a range of particles sizes from 24 nm to 50 nm. The low concentration retrieved from the high weight peak (0.3 mg/ml) was close to the lower limit of concentrations acceptable for DLS. This could be a reason for the range of particle sizes seen by DLS; of course another reason could be the production of a range of particle of differing sizes.

The high weight peak from the size exclusion column was then applied to carbon grids, negatively stained with uranyl acetate and analysed by Transmission Electron Microscopy (TEM). Particles measuring from 54-78 nm in diameter with the appearance of spherical capsid-like structures were observed (Figure 4-7d). These particle sizes deviate from the data obtained through DLS. It is notable that these three capsid-like structures were found after a prolonged search. Given the yield purified from SEC, evident from its visibility on SDS-PAGE, a greater number of capsids were expected to have been found by TEM. It is thought that the low number of particles visible by TEM may be due to the intrinsic instability of the capsid structures. These capsids are held together solely through the interactions of the JR-coil2, which in turn must out-compete the unfavourable interactions which native CTB possess to prevent its own aggregation. The negative staining technique requires the hydrophilic, negative, charging of the carbon grid by exposure to UV light, application of the sample followed by air drying and then fixing with uranyl acetate, before exposure to the TEM under vacuum. This process exposes the capsids to a combination of potentially destabilising forces and it is possible that the preparation process itself resulted in the dissociation of a large number of the capsids.

### 4.2.3 CTB-JR-coil7

This coiled-coil was assembled with wild-type CTB, cloned into pSAB2.2, expressed and purified as described in the methods. Purified samples were concentrated to 10 mg/ml and incubated at room temperature for one week. Samples were then analysed by SEC (Figure 4-8a). CTB expressed with JR-coil2 was also incubated for one week at room temperature can be seen as the black trace and JR-coil7 as the blue trace. As is evident from the enlarged panel, after one week CTB-JR-coil7 produced a larger proportion of high weight product than CTB with JR-coil2. Notably, the high weight peaks of CTB-JR-coil2 makes up 1.8% of the total area of the SEC trace, whereas the high weight peak of CTB-JR-coil7 makes up 3.4% of the total peak area, suggesting greater oligomerisation with the JR-coil7.

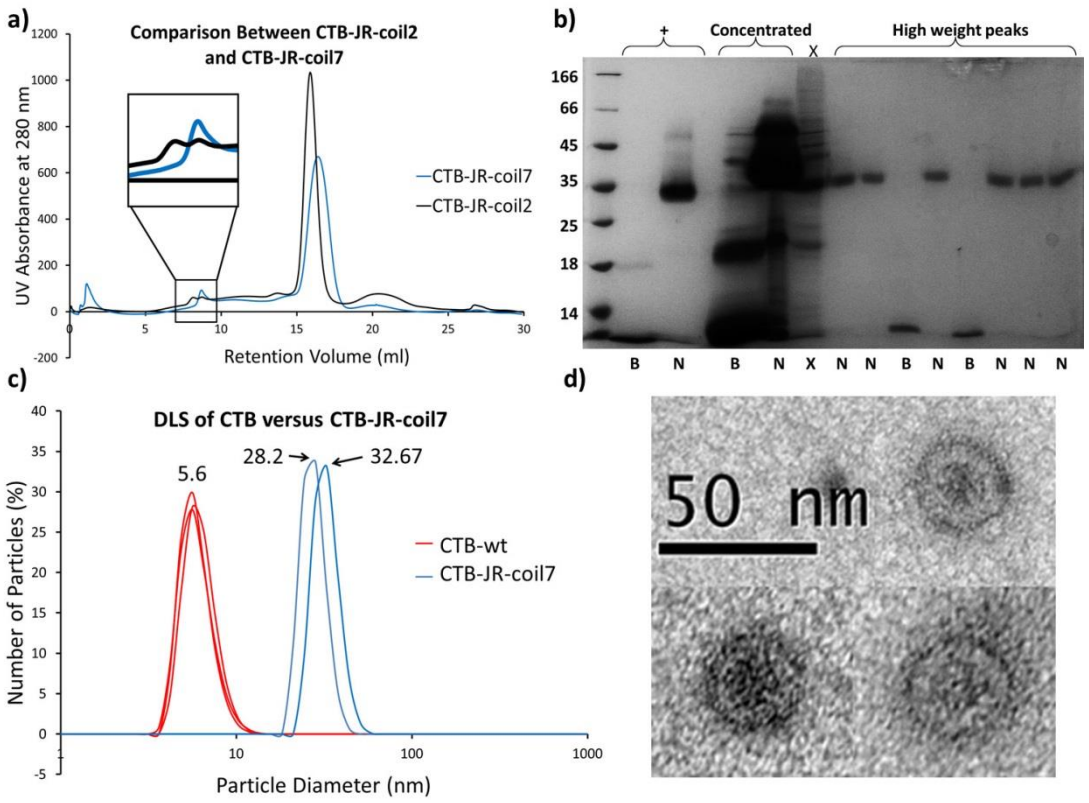



**Figure 4-8:** Characterisation of CTB with JR-coil7. **a)** Size exclusion chromatography of CTB JR-coil2 is in black and CTB JR-coil7 is in blue, both after one week incubation. Given the same period of incubation, the JR-coil7 extension seems to produce a larger yield of high weight product, additionally as discernible on the SEC of the pentameric peak, there is less overall pentamer, thus the ratio of assembled to unassembled is much greater. **b)** The high weight fractions from SEC are seen to contain pure CTB when run both boiled (B) and non-boiled (N) on SDS PAGE. The concentrated sample however shows a number of contaminants. **c)** Dynamic light scattering of the wild-type CTB, red, suggests particle sizes of 5.6 nm, however the high weight peak from SEC of the CTB-JR-coil7, in blue, gives particles of 28 to 32 nm. **d)** Transmission electron microscopy of the high weigh SEC fraction reveals particles of an agreeable size to the DLS results. These TEM structures were resolved more clearly than the particles for the CTB JR-cloi2 TEM and reveal the interior surface of the capsid. This is likely due to stain penetration of the capsid.

Figure 4-8b shows both boiled and unboiled samples analysed by SDS-PAGE for CTB-JR-coil7 with a positive control, (+) wild-type CTB. It is evident in the concentrated sample that despite CTB-JR-coil7 making up the largest component of the sample, other contaminating proteins are pulled through the purification procedure. However after SEC and isolation of the high weight peak only CTB-JR-coil7 is present. DLS of the high weight SEC fractions revealed particles of 28 to 33 nm, in blue (Figure 4-8c) in contrast to the particle size seen for wild-type CTB of 5.6nm in red.

Under observation with TEM, particles were found which possessed capsid like structures (Figure 4-8d) measuring from 29 to 36 nm in size. Dye penetration seems to have occurred more readily with the capsids produced from JR-coil7 than with JR-coil2 which revealed the appearance of the inside wall of the capsid. This might suggest that CTB-JR-coil7 may hold the CTB pentamers in a slightly different orientation, compared to JR-coil2. This potential

change in rotational orientation may lead to gaps in the capsid which allow leaching of uranyl acetate into the capsid interior. The capsids produced with JR-coil7 were more convincing examples of capsid like structures than those found with JR-coil2, which had poor contrast. Described below are a selection of TEM negative.

# 4.3 Electron Microscopy Controls

## 4.3.1 Buffer Control

In order to confirm that the particles seen under TEM relate to the CTB mutant protein found in the high weight peak from SEC, a number of control experiments were conducted. Firstly buffer taken from the SEC before the appearance of the high weight peak, 4 ml retention volume, was prepared for TEM as described in the methods (Figure 4-9).

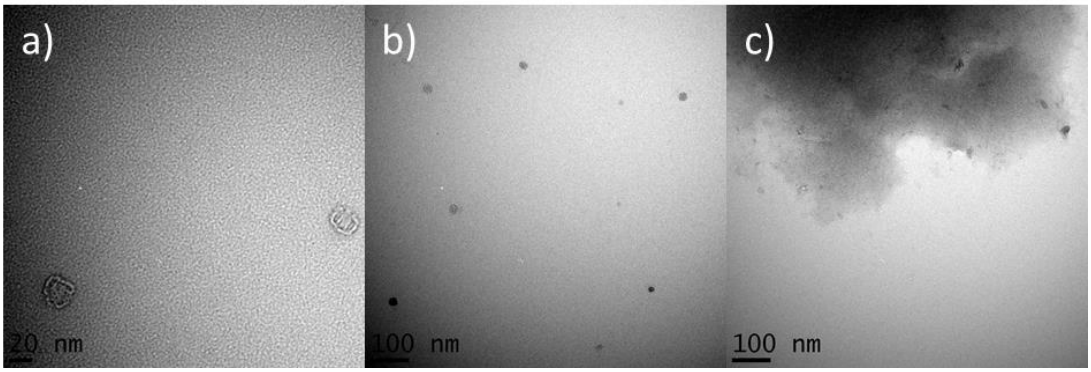



**Figure 4-9:** Buffer control from SEC. **a)** best attempts were made to find particles similar to those observed in the high weight peak. These particles are amorphic and do not appear similar to the CTB capsids. **b)** the particles seen in a) can be seen in the upper right of this micrograph with similar amorphic particles. **c)** The majority of the grid appeared as seen in this micrograph.

### 4.3.2 Pentamer Control

The pentamer peak from IGFBAA-coil7 was prepared for TEM to observe the appearance of the non-oligomerised CTB mutant particles (Figure 4-10). The CTB particles appear as donut shapes with a distinct difference in contrast compared to the buffer only grid.

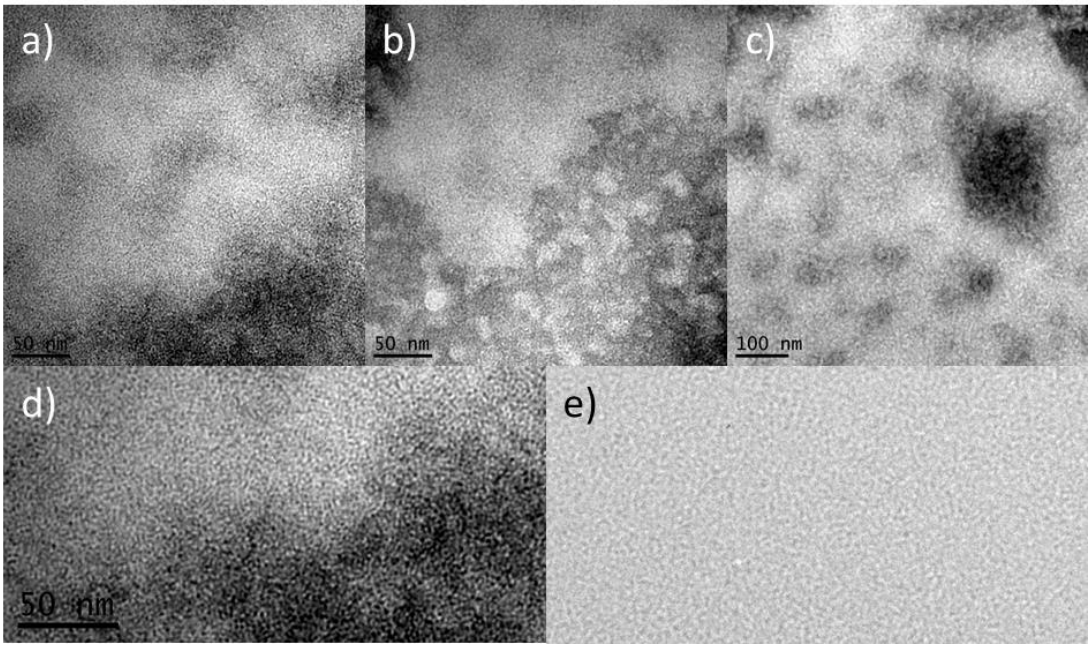



**Figure 4-10:** TEM of the CTB pentamer peak from IGFBAA. **a)-c)** generic micrographs displaying the appearance of the surface. Compared to the buffer control there is pronounced contrast to the surface which appears like small donut shapes reminiscent of CTB. **d)** an enlarged view of part of a). **e)** an enlarged view of the buffer only control at the same scale as d). There is a striking difference between these two images suggestive of CTB bound to the grid.

### 4.3.3 No-Plasmid Control

C41 cells which were not transformed with pSAB2.2 plasmid and thus will not express CTB, were treated with IPTG and processed in an identical fashion to cells containing a pSAB2.2 variant. The resultant nickel-purified precipitant fraction was concentrated to 2 mg/ml (of unknown protein) and incubated at room temperature for 2 weeks. No protein was found in or just after the void volume, where capsid structures are eluted. However there was a small peak at 12 ml retention volume which was prepared for TEM (Figure 4-11). These micrographs revealed the presence of polymorphic aspherical structures of which did not exceed 15 nm in size. Compared to the CTB capsids the overall shape of these structures was not spherical and appeared random.

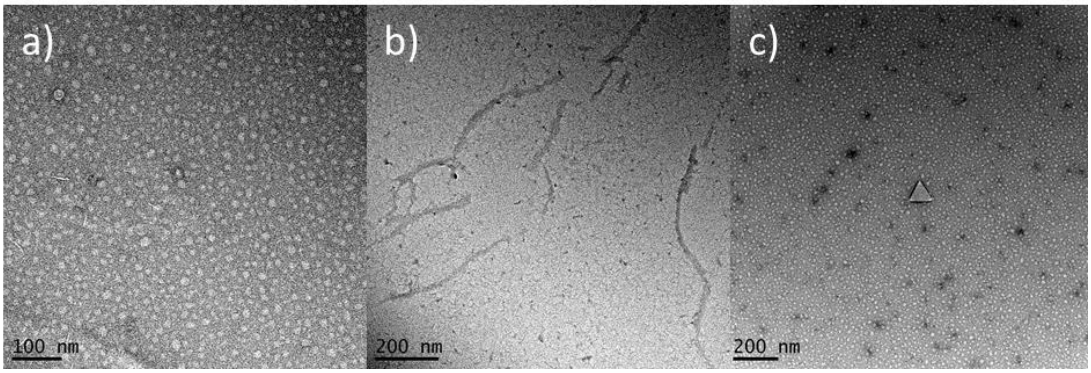



**Figure 4-11:** TEM of the high weight peak of the 'no plasmid' preparation of C41 cells. **a)** Particles were found in this preparation, however these sample particles are also observed under CTB mutant high weight peaks and are distinct from the particle thought to be constructed from CTB. The particles in this micrograph are amorphic and do not exceed 15 nm in size. **b)** The majority of the grid appeared as seen in this micrograph. **c)** We also found this triangle.

# 4.4 Concluding Statements.

Simple *C*-terminal extensions possessing coiled-coils were sufficient to create particles of a relatively discrete size. These particles had a more consistent morphology when the JR-coil7 was used as the scaffolding protein. This observation may be the result of JR-coil7 having a bias for dimeric interactions which are the dominant scaffolding interaction in the proposed T=7d geometry.

The MBP-coiled-coil studies suggested that the oligomeric state preference of the coils was inverted compared to that expected based on the peptide sequence. Although these data should be reliable enough, it is worth considering that these coils were extensions from MBP and, given the experiments described, it is not possible to rule out that an interaction of these coiled-coils with MBP may have caused an altered behaviour which is not present when these coils are attached to CTB. Fortuitously, and despite the inversion in predicted oligomeric state, both coils displayed a propensity to form both dimeric and trimeric interactions which could be desirable for the formation of the T=7d capsid. Despite the ubiquitously reported lack of a measurable binding affinity between 3-heptad coiled-coils in the literature, coiled-coils of this length were sufficient to promote assembly of capsids. Each additional pentamer that adds to the assembly after the first two reinforces the interactions of the initial two pentamers. Thus, providing a transient association is

maintained for long enough for the third pentamer to interact with the nucleating dimer of pentamers, then the total assembly becomes strengthened and addition of further pentamers becomes more favourable. Although longer coiled-coils, with their increased interaction affinity may have produced particles more quickly and in higher yields, these slowly assembling particles with low yields provide the ideal starting position for the investigation of CTB to CTB surface interaction mutations.

# Chapter 5

# Computational Interface Design

# 5.1 Computational Selection of Mutants

## 5.1.1 Overview

Modelling and simulation have been instrumental to the success of this project. The main utility of modelling has been to provide a rational strategy to select mutations that have the largest chance to increase the stability of the capsid assembly. In this chapter it is reported how target capsid states have been generated, how these have been optimised, how interfaces between pentamers have been characterised and how the free energy difference between capsid and pentameric states has been estimated. By determining a well-defined and robust protocol and applying it to a number of point mutants we have been able to perform high-throughput scoring of many more mutations than we could have experimentally handled.

### 5.1.1.1 Modelling the 3-Fold Axis in T=1 Particles

The atomistic simulation of the larger particles, such as the proposed protein cage described in the previous chapter are unfeasible given the current computational power of standard research computers. Constructing an in silico model which captures the non-identical interactions in the T=7d model would require a system of seven pentamers containing around 65000 atoms. However, given the symmetrical nature of such complexes, interactions at a given position can be assumed to be replicated in geometrically identical positions making simulations of entire capsids redundant.

In order to construct a simpler model, a dodecahedron of CTB pentamers is used *in silico*. There is much evidence in the literature, described in the introduction, that particles capable of forming more complicated spherical tiling geometries, such as T=7d, are also capable of forming more simplistic smaller capsids given different driving forces for assembly. Based on this theory we presume the modelling of an interface which would allow dodecahedral assembly would also promote, rather than deter, assembly of more complicated structures such as the T=7d particles, given that the driving force for assembly would be supplied by the coiled-coil extensions and that the mutations made would remove repulsive interactions. The dodecahedral model reduces the number of asymmetric modelled interfaces from five to one. This allows models containing fewer atoms which in turn reduce the computational cost of each assessment.

To further decrease processing times, quarter dodecahedrons, trimers of pentamers, were used in simulations. This trimer of pentamers captured interactions about the two-fold and three-fold axes of symmetry for dodecahedral assembly and brings the atom count to less than 30000. Furthermore, each model contains three interfaces which allows the assessment of three interactions for the price of one simulation. And thus reduces the number of ensemble conformations required.

#### 5.1.1.2 Determination of Interface Orientations

Given this reductionist model, the possible binding interfaces are described by the alignment of the pentamers to each other under T=1 symmetry constraints and rotated through 72° and a variety of z-axis translations, which define the distance from the centre of the dodecahedron. A computational approach was developed to estimate minimum energy conformations with perfect dodecahedral symmetry for all possible rotations and translations. This model provides a sort of "energy landscape" that allows identification of the optimal dodecahedral conformer.

#### 5.1.1.3 Ensembles in High-Throughput Interaction Determination

Once an energetically optimised initial dodecahedral conformer had been determined, a trimer of pentamer model from this conformer was simulated by molecular dynamics to probe their stability at finite temperature, and to determine an ensemble of representative structures compatible with the imposed symmetry; this also allowed for the initial structure to relax into more stable states and explore alternative conformations of the individual monomers in each pentamer that allow lower energy, higher entropy conformers to be explored with their appropriate Boltzmann weight. This step is very important for the estimation of the thermodynamic stability of designed mutants and an original aspect of the research presented here. Simulations were performed with CHARMM [142] and the EEF1 implicit solvent.

The stability of the capsid corresponds to the free-energy difference between the capsid state and a state where all monomers do not interact. Calculation of free-energy differences using molecular dynamic simulations is in principle possible, but requires

repeated simulations of the association and dissociation of the capsid, which is by no means practical for such large systems. However, a number of qualitative, empirical methods have been developed that provide an estimation of the free energy difference between two conformers of a protein, or between the bound and unbound state of a complex (assuming that in the monomeric state the components do not change conformation and are at infinite distance). One such method, that has been shown to provide good estimations for complexes with known stability is FoldX[145]. FoldX estimates the free energy difference as a sum of terms that represent the change in enthalpy between the bound and unbound forms and the entropy involved in the loss of conformational freedom. Such calculations could be performed using a single capsid conformation (e.g., that obtained by the energy minimisation procedure described above). However, the value obtained will depend largely on small changes in the conformation of the subunits and changes at their interfaces and would not be a good indicator of the stability of the complex. For this reason, the calculation of free energy was performed on the whole ensemble of 200 structures obtained from the simulations. While the values of the free energy difference vary considerably for different structures, ensemble averages converge rapidly to a well-defined value that we used to select the most stable mutants.

# 5.2 Computational protocol

## 5.2.1 Overview

A computational protocol was developed to allow the design of an interaction surface between three CTB pentamers based on dodecahedral symmetry (Figure 5-1). This routine performs a basic set of instructions three times, each referred to as a phase, and the whole routine can then be iterated again. The first of the three phases introduces a point mutation into a single subunit of a trimer of CTB pentamers and then identifies if the mutation improves $\Delta\Delta G_{stability}$ relative to the starting model. In the second phase, each chosen single mutation is applied to all subunits in a trimer of CTB pentamers and the $\Delta\Delta G_{interaction}$ between pentamers is assessed for each single mutation. In the final phase, the single point mutations are combined and $\Delta\Delta G_{interaction}$ between pentamers is assessed for the combinatorial mutations.

As seen in Figure 5-1, each phase contains five main steps. The methods for each step are described in the methods section and the appropriate scripts can be found on the accompanying DVD, however a brief description follows:

1. A starting model is constructed, either based on the wild-type sequence, or if the process is iterated, based on a successful mutant from the previous investigation.

2. The trimer of pentamers is minimised with CHARMM by finding a quick energy minima and removing unfavourable interactions brought about by the introduction of the new mutations.

3. Ensembles of the initial structure are generated by conducting a CHARMM molecular dynamics simulation and deconvoluting the coordinates from the simulation trajectory.

4. The conformations are passed to FoldX which, dependent on the phase, performs a stability-based analysis.

5. The results of the analysis are processed and ‘hits’ are identified.

These phases, steps and the appropriate considerations for each are described in more detail below.

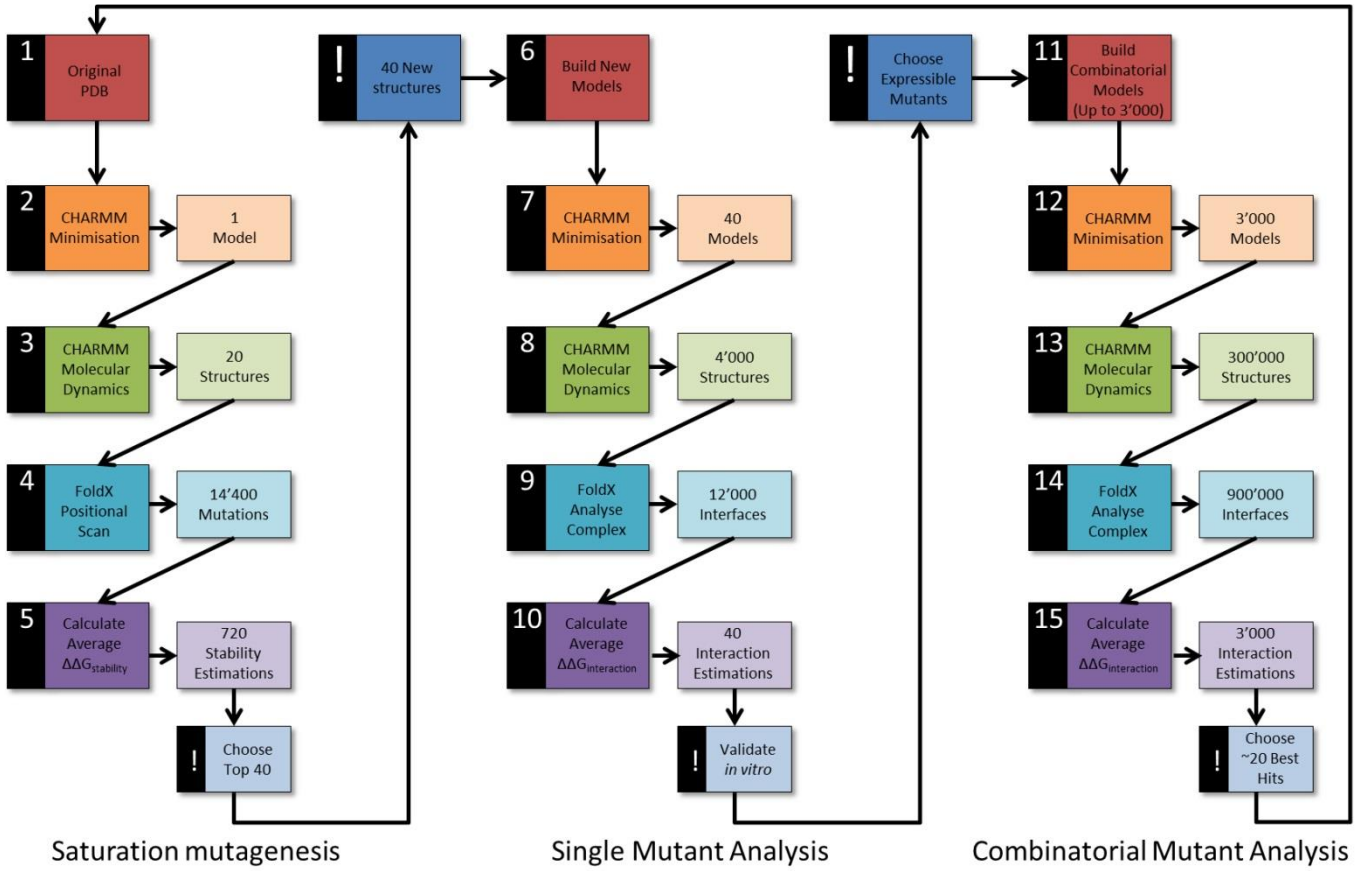



**Figure 5-1:** Overview of the computational routine, of which there are three mains parts. **Saturation mutagenesis**, where a FoldX 'position scan' is used to find mutations which increase the total stability of an ensemble of structures derived from a CHARMM trajectory of a starting complex. **Single Mutant Analysis**, in which ensembles of these 'single mutant' structures are analysed by FoldX 'Analyse Complex' to find changes in the free energy of interaction. Finally, **Combinatorial Mutant Analysis** takes place after a round of experimental validation, allowing the assembly of combinatorial mutants and their analysis in the same manner

## 5.2.2 Constructing the Starting Structure

This method was used to determine the template configuration of the wild-type protein, additional mutations suggested by the routine were applied to this model.

### 5.2.2.1 Symmetrical Docking

The interfaces at which the CTB-CTB interaction occurs in the CTB-JR-coil2 and CTB-JR-coil7 assemblies are unknown; furthermore, this interface is likely to be polymorphic. The coiled-coils hold the surfaces of the pentamers in a given orientation. With the wild-type protein this arrangement is naturally repulsive, given that CTB does not readily aggregate. The first step in the design process was to elucidate the most likely interaction surfaces given the constraints imposed by the coiled-coils and the shape and chemical complementarity of the surface. As discussed above, the most attainable estimation of capsid-like interactions was to construct an *in silico* dodecahedron.

A novel protocol was developed with CHARMM which allowed the symmetrical docking of target proteins using symmetry matrices obtained from other proteins in the PDB. The initial structure of a dodecahedral capsid was obtained by applying the crystal symmetry matrix for the biological unit from the PDB structure `1NQU` (Lumazine Synthase (LS) from *Aquifex aeolicus*), which forms a dodecahedron from pentamers of a similar size as CTB, to the coordinates from the PDB structure `3CHB` (Cholera Toxin B-subunit). The method aligns LS to the coordinate origin with its longest dimension along the x-axis, then the next longest perpendicular axis to the y-axis and records the total transformation. CTB is then aligned to the origin in the same manner and the recorded transformation from LS is inverted to align CTB to the original structure of LS on a geometric basis as opposed to more intricate structural alignment. As LS is a larger pentamer this initial placement leaves CTB 'out of contact' with its neighbouring CTB pentamers in the dodecahedron. CTB, and its dodecameric neighbours, were then systematically rotated through 360° in 1° increments and translated toward the centre of the dodecahedron, i.e. along the z-axis, in steps of 1 Å up to a total transformation of 10 Å. Each structure was then minimised by CHARMM. At each step CHARMM was asked to output a total $\Delta G$ (kcal/mol) for the system and FoldX was used to calculate an interaction energy in $\Delta\Delta G$ (kcal/mol). These results can be seen in Figure 5-2. A video of this process can be found on the accompanying DVD. The red lines in Figure 5-2 represent the pentamer at 0 Å z-axis translation.

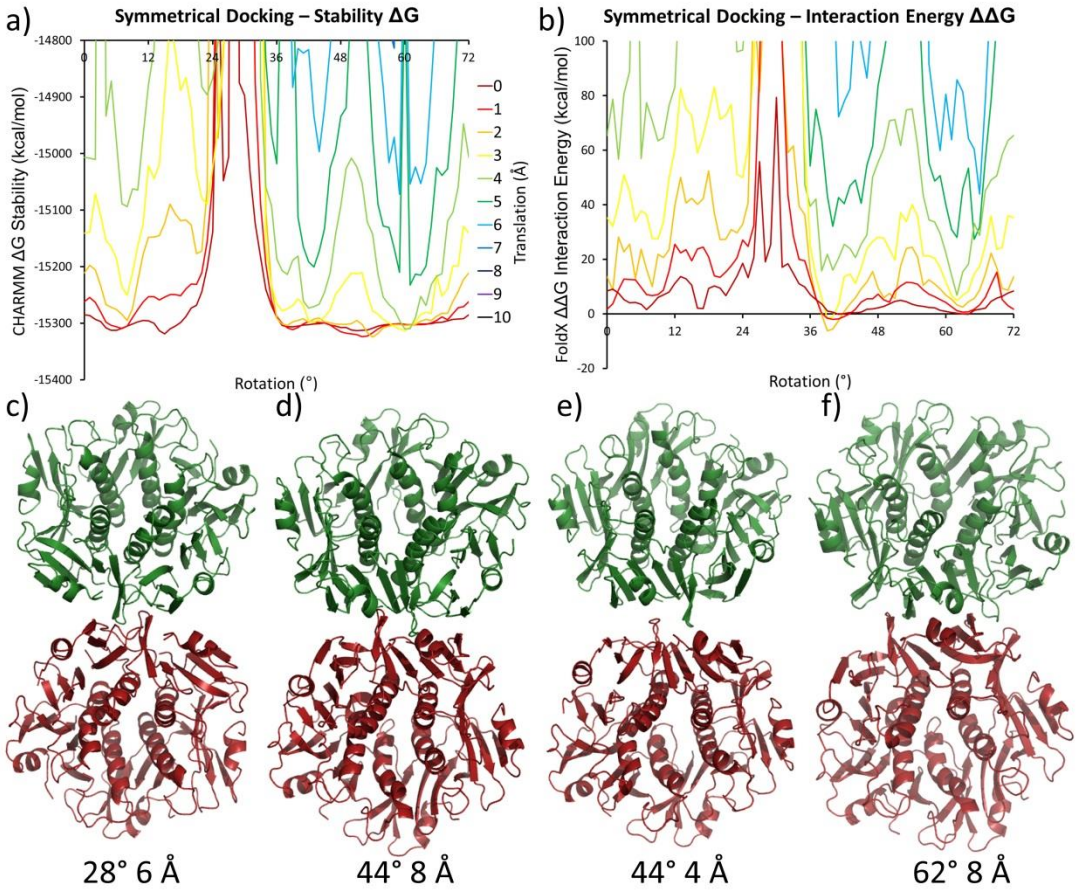



**Figure 5-2:** Symmetrical docking of CTB pentamers. **a)** CTB was placed in a range of orientations and transformations (relative to the original lumazine synthase cage structure) which would lead to a symmetrical particle, the total energy of the system was derived by CHARMM. **b)** These structure were then assessed by FoldX to elucidate an interaction energy. As the total size of the capsid is reduced, i.e. translation of the CTB, the CTB proteins start to clash sterically, however these transformations (~5-8 Å translation) show energy minima at around 44° and 62° rotation. This suggests that for interfaces with these rotations, there is reduced repulsion compared to other orientations of contact. **c)** An example of a repulsive interface at 28° rotation, the side chains, not shown, clash in this rotational conformation. **d)** An example of an energy minimum, which is less repulsive than most other orientations. **e)** An example of a less repulsive interface at a relatively expanded transformation of 4 Å, compared to d) which is the same rotation but translated by 8 Å. **f)** Another energy minimum rotational conformation. Structures produced from pdb 3CHB.

The CTB pentamers at the origin of rotation are not in contact with each other, creating an 'expanded capsid' with part of each pentagonal edge facing one another. As these pentamers are rotated, the vertices come into proximity with each other and make contact. A steric clash, and accompanying decrease in stability, is observed at 30° rotation in both $\Delta G_{stability}$ and $\Delta\Delta G_{interaction}$ as the vertices of the pentamers make contact with one another.

The total ΔG in Figure 5-2a between 40° and 60° at 0 Å translation can be regarded as the total energy of the unbound system, as the pentamers remain out of contact and the ΔG remains constant despite further contractile translations of up to 2 Å. Further contractions of the capsid in this rotational range start to cause deviations in the $\Delta G_{stability}$ which become

more evident as the contraction of the capsid increases. Notably, with translations of 3 Å to 6 Å and rotations of 44° and 62° from the origin, energy minima become distinct from the rest of each trace. This observation is replicated the $\Delta\Delta G_{interaction}$ energy calculated by FoldX for rotations of 41° and 62°. The emergence of these energy minima suggests areas of contact between the pentamers which, although they destabilise the total structure, the perturbation is less than that of neighbouring rotational orientations. As discussed earlier, native CTB is not expected to have an attractive binding affinity between pentamers and thus it is not surprising that the $\Delta\Delta G_{interaction}$ energies are positive. It is worth noting that as the capsid contracts the $\Delta\Delta G_{interaction}$ becomes less negative, indicating the increase in repulsive interactions. When the capsid is at 0 Å translation this $\Delta\Delta G_{interaction}$ is approximately zero. This is an indication of the separation of the pentamers, being too distant to interact, rather than a neutral proximal interaction.

#### 5.2.2.2 A Trimer of Pentamers

As seen in Chapter 4, the particles made by including the JR-coils are larger in size than would be predicted for dodecahedrons comprised of 12 pentamers. The supramolecular structure is unknown, so how can algorithms be applied to design this interface, given its unknown configuration? Based on the symmetrical docking data above, two interface configurations are clearly energy minimised with respect to the other rotational configurations; therefore they provide reasonable starting points for the design process.

Molecular dynamic simulations are renowned for being computationally demanding so rather than simulating an entire capsid, a trimer of pentamers was used as a base model for all simulations. This trimer comprised three pentamers and three interfaces, as opposed to a dimer of pentamers which would contain just one interface. The trimer of pentamers also captured interactions restricted to the three-fold symmetry axis. A trimer of pentamers model was constructed based on the docking data at 44° rotation and 8 Å contraction, (Figure 5-3), which was used as a starting position for future mutagenesis work. A second model considered to be a good starting position for investigation was at 62° rotation and 8 Å contraction. The two-fold axis of the 44° model contains a loop which extends into a cavity in the side face of CTB, seen in the centre of Figure 5-3c. This interaction was considered likely to increase the surface area for interaction and thus provide a number of design opportunities, as well as displacing water from the interface, which should allow the design algorithms to be more accurate. Additionally the interfaces are aligned in a flush manner. In contrast, the 62° model does not contain such a

protrusion and respective cavity or line up flush and the proposed interfaces are staggered, thus the total surface area for interaction is reduced. The three fold axis of the 44° model contains β-sheets running down the interface, it should be possible to mutate a number of these residues to add three-fold axis interactions without seriously perturbing the pentamer structure. The loop which extends into the cavity described on the two-fold axis of the 44° model extends into the 3-fold axis in the 62° model. This should also provide an adequate framework for mutation, however, the cavity described earlier is in an unfortunate position on the 62° models and would remain solvated upon assembly. Based on these discussions the 44° model was chosen as a suitable starting model for future computational work.

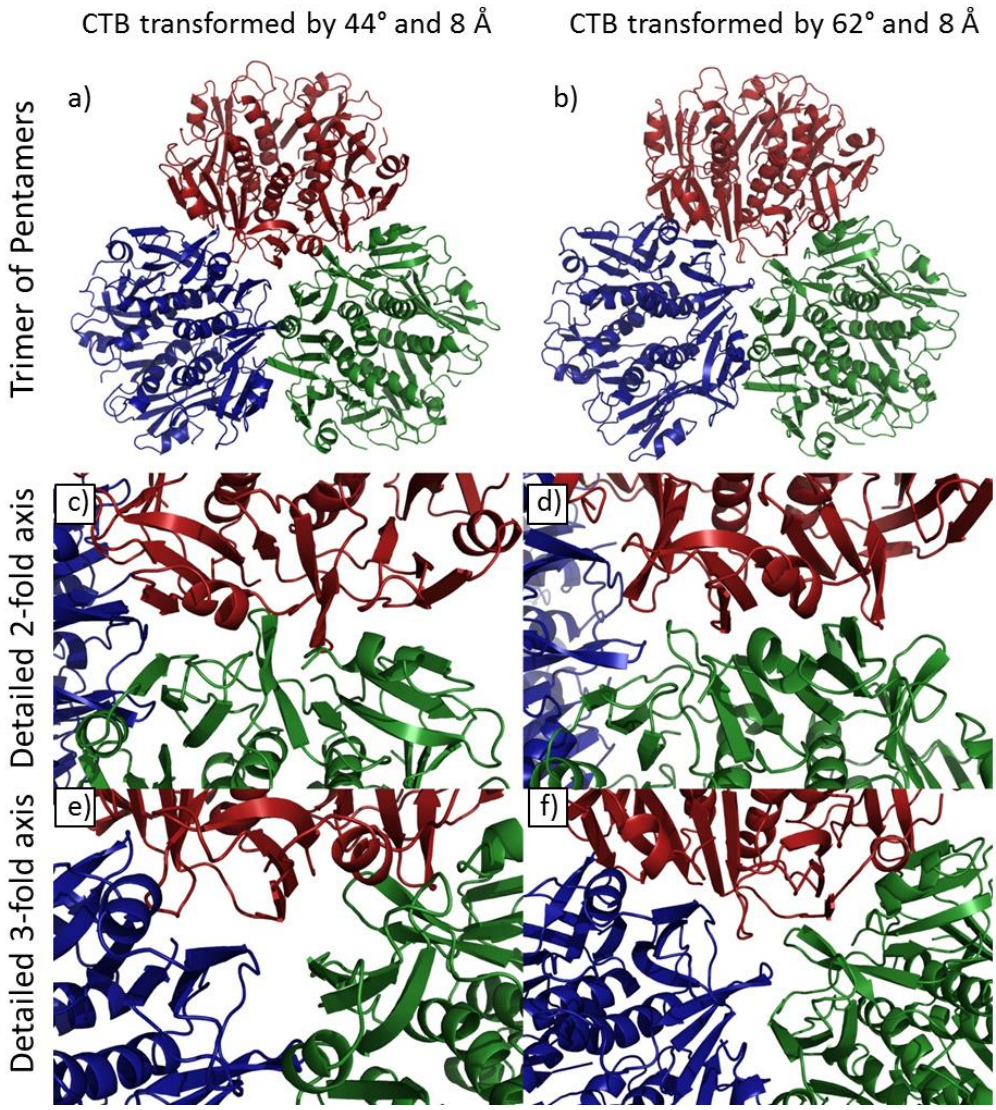



**Figure 5-3:** Structural views of the two best energy minima conformations (**a)** 44°and **b)** 62°) elucidated from the symmetrical docking studies. Overall views of the trimer of pentamers used in simulations as well as close up views of **c)** and **d)** the 2-fold and **e)** and **f)** 3-fold symmetry axes. Produced from pdb 3CHB.

### 5.2.3 Constructing Mutant Structures

In the first phase of the computational routine the wild-type protein is studied and saturation mutagenesis is performed by FoldX which introduces the relevant mutations. However, for the second and third phase of the computational routine mutations are introduced across the template structure. A script, named 'threading' was designed which implements the Pymol mutagenesis function. A library of target sequences or 'parts' of sequence are input into Pymol which in turn mutates those positions which differ from the wild-type template. The mutagenesis function of Pymol is efficient but not rigorous, thus in order to correct van der Waals clashes, the Pymol 'sculpting' function was implemented on the final structure. This technique was able to process large numbers of mutations in a timely manner; however, the final structures were often in energetically unstable conformations. These structures underwent minimisation by CHARMM before molecular dynamic simulations were conducted.

### 5.2.4 Minimisation and Molecular Dynamics

In order to produce a large ensemble set of conformations for each target protein, molecular dynamic simulations were conducted and the resultant trajectories deconvoluted to yield individual conformations per time step. Energy minimisation was conducted on the starting structure after the Pymol mutagenesis. Both minimisations and molecular dynamics were implemented from within the script 'run-fullcharmm' which performs the calculations on a list of supplied pdb files.

### 5.2.5 Analysis by FoldX

FoldX makes reductions in the total free energy of a protein by applying a rotamer library of amino acid side-chain conformations to specific sequence positions. FoldX then performs a fast energy calculation on the target protein structure and/or complex. In this project FoldX was used for the high throughput calculation of protein stability, $\Delta G$ (kcal/mol), and protein-protein interactions, $\Delta\Delta G$ (kcal/mol) as well as for making suggestions of side-chains which improve the total stability of the target structure.

#### 5.2.5.1 Ensemble construction

As FoldX makes calculations based only on the submitted structure and minor deviations of the structure can have a large impact on the result, large ensembles of structures were used to calculate an average $\Delta G_{stability}$. The ensembles were retrieved from the molecular dynamic simulations produced by CHARMM. These ensembles once calculated also provide an insight into the total entropy of the given target through the width of the distribution of energies observed.

#### 5.2.5.2 Analysis

A number of different functions of the FoldX program were used, however three functions were used routinely. The 'stability' function was used to derive a $\Delta G_{stability}$ in kcal/mol for a particular structure. The 'positional scan' function was used to perform saturation mutagenesis at selected positions of CTB. Finally the 'analyse complex' function was used to estimate the $\Delta\Delta G_{interaction}$ of a proposed interface.

##### 5.2.5.2.1 Stability

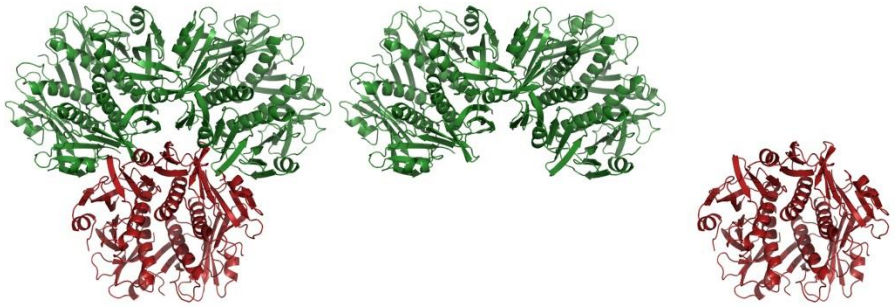
FoldX calculates the $\Delta G_{stability}$ of a target protein compared to a hypothetical unfolded version of the same sequence. This calculation forms the basis of all other FoldX calculations which measure the difference between a series of stability calculations, before and after mutations and given differing rotamers of side-chains. This calculation is described under the methods section.

##### 5.2.5.2.2 Position Scan

The position scan performs *in silico* saturation mutagenesis across the target protein at selected positions. The function first builds a wild-type reference structure, from which the chosen amino acid is removed. It then re-builds the initial wild-type residue and searches for the most energetically minimised rotamer. The program then tries to 'repair' surrounding side chains by refining their rotameric positions. FoldX returns a $\Delta G_{stability}$ in kcal/mol for the wild-type reference and then performs the same actions for each saturation mutant. The wild-type reference $\Delta G_{stability}$ is then compared against each of the remaining 19 amino acids in this position to give the $\Delta\Delta G_{stability}$, i.e., the change in $\Delta G_{stability}$ compared to the wild-type reference. Each round of position scan is conducted upon one CTB monomer which lies on the putative three fold axis in the centre of the trimeric model discussed above. In the second phase, the selected point mutations are replicated in all subunits of the trimer of pentamers prior to assessing the interaction energy of the complex.

#### 5.2.5.2.3 Analyse Complex

The analyse complex function calculates the $\Delta G_{stability}$ for the total complex and then calculates the $\Delta G_{stability}$ for the two components of the complex. These individual $\Delta G$ are subtracted from the total $\Delta G$ to give the $\Delta\Delta G_{stability}$ between the bound and unbound structures which is inferred to be the $\Delta\Delta G_{interaction}$ of the complex. Typically one pentamer in the trimer of pentamers is assessed for its $\Delta\Delta G_{interaction}$ against the two remaining pentamers in the trimer, Figure 5-4. Towards the end of this chapter FoldX was benchmarked against previous experimental data for both $\Delta G_{stability}$ and $\Delta\Delta G_{interaction}$ assessments.



$$\Delta\Delta G_{interaction} = \Delta G_{stability_{complex}} - \left(\Delta G_{stability_{component\ A}} + \Delta G_{stability_{component\ B}}\right)$$

**Figure 5-4:** The $\Delta\Delta G_{interaction}$ calculation. The sum of the $\Delta G_{stability}$ of the components is subtracted from the $\Delta G_{stability}$ of the whole complex. Structures produced from pdb 3CHB

# 5.3 Initial Computational Strategies

## 5.3.1 Overview

Three strategies were pursued to select mutants likely to increase the binding affinity of the proposed interface. In strategy #1.0 (S#1.0), which is described in this chapter, suggestions from FoldX which had the lowest predicted energy were used with minimal manual curation of the mutant list. This mutant set underwent a second round of *in silico* mutagenesis prior to experimental validation and thus incorporated a relatively large number of mutations. In strategy #2.0 (Chapter 7), however, suggestions made by FoldX were used as guidance for introducing a small number mutations through rational design based on the structural model. The subsets, S#1.1, S#2.1, S#2.2 all involved modest, yet computationally-driven second round mutations based on successful mutants from the first round libraries. Finally, all successfully expressible mutants from the previous two strategies were combined in strategy #3.0 (S#3.0) to derive combinations of these mutants.

## 5.3.2 Strategy 1.0 (S#1.0)

During the discussion of this set of mutants, a thorough description of the computational steps and evaluation of the data for a selection of mutants will be presented.

### 5.3.2.1 Objective

Strategy 1.0 (S#1.0) was designed to evaluate the reliability of pushing the computational routine through two rounds of mutagenesis prior to experimental validation. Effectively, the routine described in Figure 5-1 was followed twice, however different methods for generating the combinatorial mutants were used in each iteration. Furthermore, the first round of combinatorial selection was based solely on $\Delta\Delta G_{interaction}$ estimations from FoldX; no manual inspection of the structure of the proposed interaction surface was made.

The computational routine predicted a set of point mutations which were then assessed for their $\Delta\Delta G_{interaction}$ by FoldX. These mutants were manually combined into 39 combinatorial mutants, based solely on the FoldX analyse complex $\Delta\Delta G_{interaction}$ estimations (no *in vitro* expression testing was used in this iteration). Hits from these multi-mutant models were subjected to a second full iteration of the computational routine: the combinatorial mutations were re-evaluated for additional favourable point mutations which were added to the mutated sequence and assessed again for their $\Delta\Delta G_{interaction}$. During the course of this investigation the assembly PCR method was implemented in on-going molecular biology work, therefore the sequences of the computationally-derived mutants were broken down*, in silico*, to represent mutant 'parts' compatible with the assembly PCR method. These 'parts' were recombined *in silico* to generate a combinatorial library which was assessed for the best $\Delta\Delta G_{interaction}$. This strategy initially was investigating prolonged *in silico* rounds of the routine, however, the implementation of the assembly PCR 'part recombination' added an extra level of complexity.

### 5.3.2.2 Pass One

#### 5.3.2.2.1 Starting Structure

In S#1.0 the wild-type-44°-8Å model, described above, was used to initiate the routine.

#### 5.3.2.2.2 Saturation Mutagenesis

After minimisation of the starting structure, a 40 ps room temperature simulation was performed. An ensemble of 20 configurations was retrieved from the trajectory at 2 ps periods. These structures were passed to the FoldX function 'Position Scan' which performed saturation mutagenesis across the interaction surface residues. An example of the results from the FoldX position scan is presented in Figure 5-5. This position scan was performed across 36 positions in the protein, over 20 protein conformations, assessing 20 amino acid substitutions which gave 14'400 $\Delta G_{stability}$ estimations. These data were compared to the wild-type reference and averaged over the 20 conformations of the protein to give 720 averaged $\Delta\Delta G_{stability}$ values, Figure 5-6a. Of these data, 148 mutations predicted an improvement in $\Delta\Delta G_{stability}$ and, of these, 39 mutations gave a $\Delta\Delta G_{stability}$ change below a threshold of -1.0 kcal/mol (Figure 5-6b). FoldX suggested mainly hydrophobic residues (Figure 5-5, highlighted in green) from the position scan, however, some residues with hydrogen bonding potential were also identified (highlighted in yellow) along with three positively charged lysine residues (blue).

The position scan, apart from suggesting changes in $\Delta G_{stability}$, also discloses something more subtle. The number of mutants having a $\Delta\Delta G_{stability}$ less than 0.0 kcal/mol for a given residue position highlights which wild-type residues are particularly unfavourable for the hypothetical complex (Figure 5-7). Residues H18, D22, T41, E83, N89 and R94 each have between 8 and 14 suggested favourable mutations. Given that for these residues more than half of the suggested mutations result in a more stable conformation, it can be assumed that the wild-type residue is particularly unfavourable and that these positions should be targeted for mutagenesis. Additionally, these residues are biased towards charged residues which may present unfavourable electrostatic interactions across the proposed binding site. It should be noted that a small anomaly arose during the single mutant analysis section: Model #3, as seen in Figure 5-5 actually contains two mutations, a proline at position 3 and an isoleucine at position 6. Therefore, despite there being 39 mutations predicted, only 38 'single' mutants were analysed.

| Residue | Position | Average ΔΔG |
|---|---|---|
| PRO | 94 | -2.8 |
| LEU | 83 | -1.9 |
| GLY | 22 | -1.9 |
| ILE | 83 | -1.8 |
| SER | 22 | -1.8 |
| MET | 83 | -1.8 |
| PHE | 18 | -1.6 |
| VAL | 83 | -1.6 |
| ILE | 28 | -1.6 |
| PRO | 3 | -1.6 |
| MET | 41 | -1.5 |
| GLY | 103 | -1.5 |
| ALA | 22 | -1.5 |
| LEU | 94 | -1.5 |
| CYS | 22 | -1.5 |
| TYR | 18 | -1.4 |
| LEU | 18 | -1.4 |
| VAL | 28 | -1.4 |
| ILE | 47 | -1.3 |
| LEU | 41 | -1.3 |
| ASN | 22 | -1.3 |
| LEU | 28 | -1.3 |
| LEU | 3 | -1.2 |
| ILE | 18 | -1.2 |
| MET | 94 | -1.2 |
| MET | 18 | -1.2 |
| ILE | 6 | -1.2 |
| LEU | 6 | -1.2 |
| LYS | 83 | -1.2 |
| MET | 28 | -1.2 |
| LYS | 28 | -1.2 |
| PHE | 41 | -1.1 |
| VAL | 41 | -1.1 |
| TRP | 18 | -1.1 |
| ILE | 41 | -1.1 |
| LYS | 41 | -1.0 |
| PRO | 43 | -1.0 |
| PHE | 94 | -1.0 |
| MET | 1 | -1.0 |

| Model | Residue | Position | Average ΔΔG |
|---|---|---|---|
| 1 | MET | 1 | -1.0 |
| 2 | LEU | 3 | -1.2 |
| 3 | PRO | 3 | -1.6 |
| 3 | ILE | 6 | -1.2 |
| 4 | LEU | 6 | -1.2 |
| 5 | ILE | 18 | -1.2 |
| 6 | LEU | 18 | -1.4 |
| 7 | MET | 18 | -1.2 |
| 8 | PHE | 18 | -1.6 |
| 9 | TRP | 18 | -1.1 |
| 10 | TYR | 18 | -1.4 |
| 11 | ALA | 22 | -1.5 |
| 12 | ASN | 22 | -1.3 |
| 13 | CYS | 22 | -1.5 |
| 14 | GLY | 22 | -1.9 |
| 15 | SER | 22 | -1.8 |
| 16 | ILE | 28 | -1.6 |
| 17 | LEU | 28 | -1.3 |
| 18 | LYS | 28 | -1.2 |
| 19 | MET | 28 | -1.2 |
| 20 | VAL | 28 | -1.4 |
| 21 | ILE | 41 | -1.1 |
| 22 | LEU | 41 | -1.3 |
| 23 | LYS | 41 | -1.0 |
| 24 | MET | 41 | -1.5 |
| 25 | PHE | 41 | -1.1 |
| 26 | VAL | 41 | -1.1 |
| 27 | PRO | 43 | -1.0 |
| 28 | ILE | 47 | -1.3 |
| 29 | ILE | 83 | -1.8 |
| 30 | LEU | 83 | -1.9 |
| 31 | LYS | 83 | -1.2 |
| 32 | MET | 83 | -1.8 |
| 33 | VAL | 83 | -1.6 |
| 34 | LEU | 94 | -1.5 |
| 35 | MET | 94 | -1.2 |
| 36 | PHE | 94 | -1.0 |
| 37 | PRO | 94 | -2.8 |
| 38 | GLY | 103 | -1.5 |

**Figure 5-5:** A list of the 'hits' from the FoldX position scan run on the wild-type trimer of pentamers. The first list is ordered by $\Delta\Delta G_{stability}$ and the second list is ordered by residue position. The ΔΔGstability is measured in kcal/mol, where green denoted the best stabilisation and red, the worst. The residues chosen by FoldX have been highlighted in colours to represent their biochemical properties where green is hydrophobic, yellow is hydrogen bonding potential and red and blue are charged respectively.

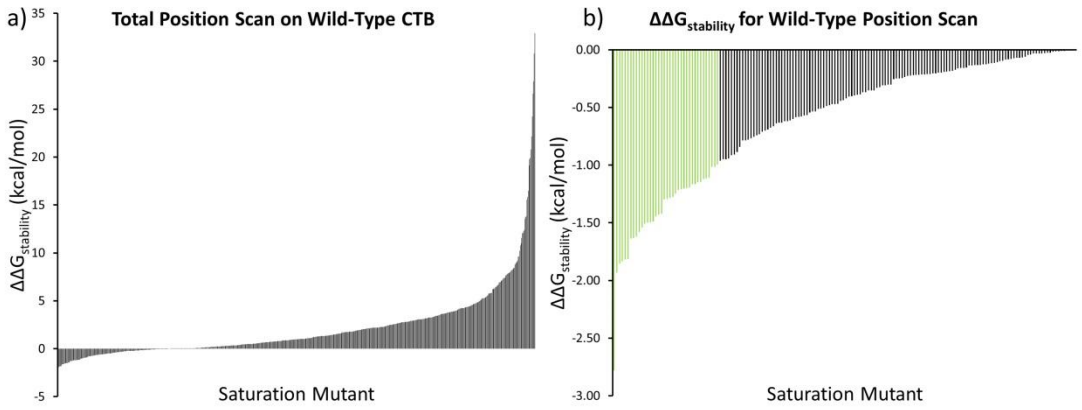



**Figure 5-6:** The average change in $\Delta G_{stability}$ for the trimer of pentamers upon saturation mutagenesis in a single CTB subunit. **a)** The energy change in kcal/mol for the 720 substitutions that were analysed which have $\Delta\Delta G_{stability}$ values ranging from 35 to -2.5 kcal/mol. **b)** The same data but presenting only those with a $\Delta\Delta G_{stability}$ below 0.0 kcal/mol. Those which fall below -1.0 kcal/mol are highlighted in green.

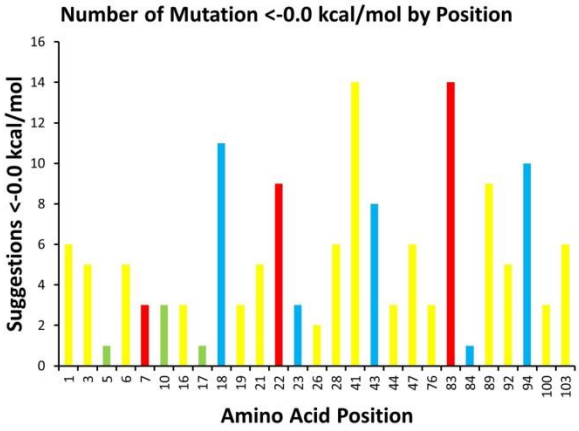



**Figure 5-7:** The frequency of mutations at each position which stabilise the complex. This graph shows the number of mutations at each position which are predicted to be more favourable than the wild-type residue. Those positions with many suggested mutations can be thought of as destabilising residues for complex formation. The biochemical property of the wild-type residue is highlighted in green for hydrophobic, yellow for hydrogen bonding potential and in red or blue for charged side-chains respectively.

#### 5.3.2.2.3 Single Mutant Analysis

The position scan revealed mutations which could improve the $\Delta G_{stability}$ of the complex. However, the process mutates only one subunit out of 15 in the trimer of pentamers model, therefore the hits need to be built into the remaining structures and simulated fully, which is done here. Each of the models was constructed using the threading script and minimised with CHARMM before conducting molecular dynamics simulations of 40 ps (Figure 5-8). Ensembles of 200 configurations were produced from these trajectories and analysed with the FoldX 'analyse complex' function. To account for the period of relaxation

seen at the beginning of the simulations (Figure 5-8a), an average of the final 20 ps (100 conformations) was used as the total $\Delta\Delta G_{interaction}$. Figure 5-8a shows the change in interaction energy over the course of the simulation for the wild-type CTB and one of the single point mutations, model #33 (E83V).

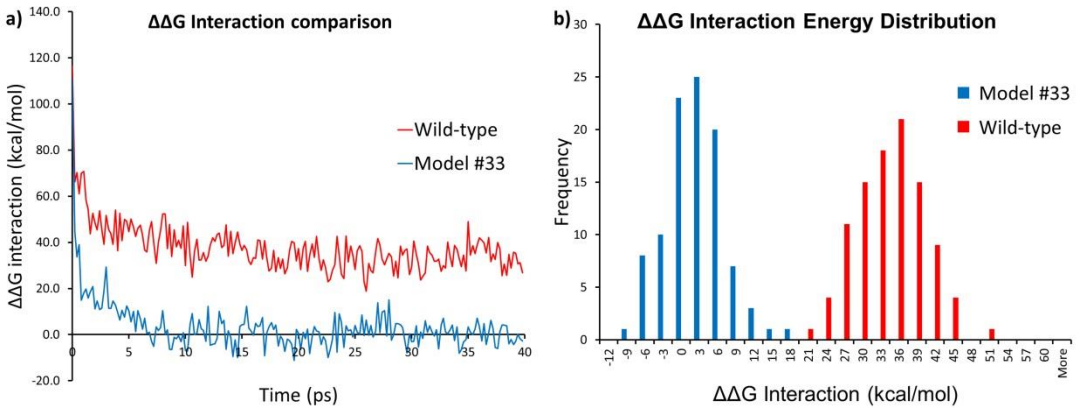



**Figure 5-8:** Analysis of molecular dynamic simulations of the mutant CTB. **a)** A plot of the $\Delta\Delta G_{interaction}$ of the complex across the structures retrieved from the CHARMM trajectory. The wild-type is in red, compared to the reduced energy of the mutant Model #33. **b)** A histogram of the interaction energies showing the distribution of ΔΔGinteraction over the 40 ps time frame.

Analysis of the wild-type model revealed that E83 and D22 lie in very close proximity across the proposed interface, therefore eliminating this charge clash removes a strong repulsive interaction. The total distribution of free energy differences for the final 20 ps is shown in Figure 5-8b. In each case the energy reaches a well-defined average value and fluctuates around it. Fluctuations are sizable and physically relevant (their amplitude reflects the variety of conformations making up the capsid state at room temperature). Such fluctuations also highlight the fact that free energy estimations performed on single structures are less relevant than those performed on ensembles of structures. In the present case, while the free energy distributions are large, the difference between the average $\Delta\Delta G_{interaction}$ for Model #33 and the wild-type are large relative to the width of the distributions. The length of the simulation, 40 ps was chosen such that the distributions of free energy differences converge. Table 5-1 shows the top ten mutants from the single mutant analysis as well as the wild-type. As discussed above, the best improvements in $\Delta\Delta G_{interaction}$ are obtained for mutants that remove the E83-D22 charge clash, therefore, these mutants populate the top end of the list.

| model | $\Delta\Delta G_{interaction}$ | 1 | 3 | 6 | 18 | 22 | 28 | 41 | 43 | 47 | 83 | 94 | 103 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | 0.69 | T | Q | T | H | D | T | T | K | T | V | R | N |
| 30 | 6.61 | T | Q | T | H | D | T | T | K | T | L | R | N |
| 29 | 11.45 | T | Q | T | H | D | T | T | K | T | I | R | N |
| 32 | 13.99 | T | Q | T | H | D | T | T | K | T | M | R | N |
| 11 | 16.13 | T | Q | T | H | A | T | T | K | T | E | R | N |
| 9 | 16.82 | T | Q | T | W | D | T | T | K | T | E | R | N |
| 3 | 16.85 | T | P | I | H | D | T | T | K | T | E | R | N |
| 8 | 16.88 | T | Q | T | F | D | T | T | K | T | E | R | N |
| 13 | 17.15 | T | Q | T | H | C | T | T | K | T | E | R | N |
| 15 | 17.47 | T | Q | T | H | S | T | T | K | T | E | R | N |
| WT | 33.07 | T | Q | T | H | D | T | T | K | T | E | R | N |

**Table 5-1:** Single mutants ranked by minimisation of $\Delta\Delta G_{interaction}$. The top ten hits from the single mutant analysis are shown along with their average $\Delta\Delta G_{interaction}$. The residue position is highlighted across the top and the wild-type residue is shown below. Each single mutant is highlighted in green.

#### 5.3.2.2.4 Combinatorial Mutants

A combinatorial library based on the hits from this round would consist of 216K mutant structures, which would require an infeasible amount of computation. Therefore, based on the hits from the first round point mutations, a small manually curated combinatorial library was developed as shown in Figure 5-9. This library was based around the two main combinatorial mutants, model 1 and model 17 which incorporate the best mutant hits (green) for unique positions from the single mutation analysis, namely `Q3P`, `T6I`, `H18W`, `D22A` or `D22S`, `K43P` and `E83V`, along with other variants including secondary (yellow), tertiary (orange) and quaternary (red) hits at the same positions. It should be noted that no reference to the structure of the protein was made when making choices for the combinatorial mutants, simply the $\Delta\Delta G_{interaction}$ energy, the initial residue and the substitution were considered.

As with the single mutants, 40 ps simulations were conducted and conformations from the last 20 ps were analysed and their $\Delta\Delta G_{interaction}$ was estimated by FoldX. Interestingly the mutants that incorporate the best mutations from the single mutation round, models 1 and 17, lie at 12$^{th}$ and 10$^{th}$ respectively in the $\Delta\Delta G_{interaction}$ ranking and mutants which substitute these primary mutations with secondary and tertiary mutations give the best interaction energies. This re-ranking following the combinatorial mixing step is interesting as it highlights that synergetic interactions between mutations are a factor for consideration in mutation design. Selected models from these combinatorial mutants were then passed through the computational routine a second time.

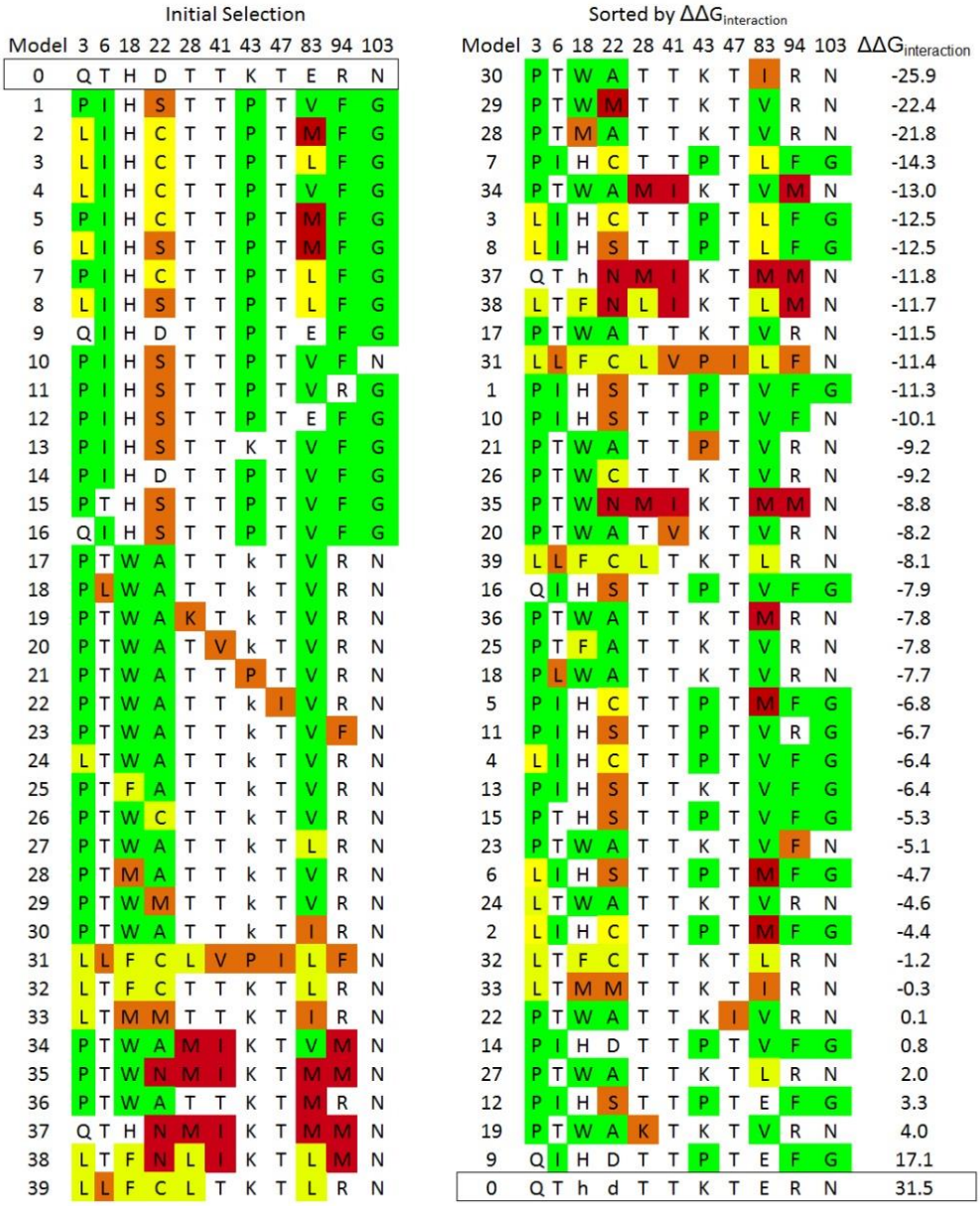



**Figure 5-9:** Combinatorial library selection and analysis. Model numbers run down the left hand side of each list and position numbers across the top. The wild-type residue is backed with white. Primary mutations, i.e. with the best $\Delta\Delta G_{interaction}$ for each position are highlighted in green, secondary in yellow, tertiary in orange and all others in red. The left hand list shows the mutants chosen. The right hand list shows the same combinatorial mutants ranked by $\Delta\Delta G_{interaction}$ in kcal/mol. Some mutation positions only contain secondary or tertiary mutations to suggest that these mutations only made minimal improvements during the single mutant analysis stage.

### 5.3.2.3 Pass Two

Three models from Figure 5-9, model 7, 18 and 31 were passed through the computational routine for a second time as these models housed a diverse range of different mutations

#### 5.3.2.3.1 Starting Structure

The progress of model 7 from Figure 5-9 is documented below.

#### 5.3.2.3.2 Saturation Mutagenesis

Saturation Mutagenesis was conducted as described above for the previous round. As mutations had already been made in the sequence which removed native repulsive interactions in the proposed binding surface, the output of the second round of saturation mutagenesis produced a more evenly distributed list of positions for mutagenesis. Additionally, the total number of substitutions which yielded $\Delta\Delta G_{stability}$ values of less than -1 kcal/mol was just 17, compared to 39 in the wild-type saturation mutagenesis investigation. The total number of substitutions follows a similar trend to the wild-type study; however, the positions which are highly populated are different to those seen for the wild-type protein (Figure 5-10).

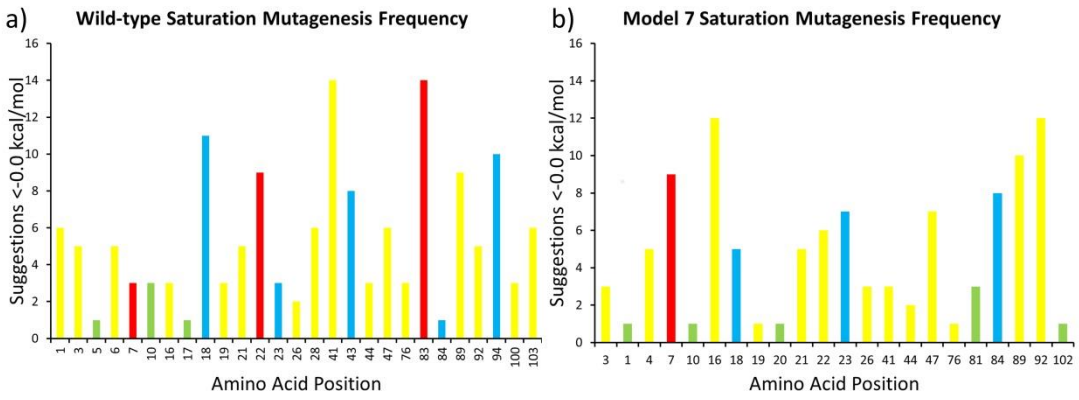



**Figure 5-10:** Comparison of positional mutation frequency between the wild-type and a first round mutant, model 7. As with Figure 5-7, these graphs show the number of mutations at each position which improve the total $\Delta G_{stability}$ of the complex. Here we compare the wild-type mutational frequency with that derived for model 7.

Analysis of the $\Delta\Delta G_{stability}$ results for this saturation mutagenesis experiment suggested the primary positions to mutate were K84, T47 and D7. These three positions populated the top ten positions of the saturation mutagenesis table (Figure 5-11). However it is noteworthy that these positions are not the top of the list for the frequency of mutation (Figure 5-10), which identifies positions 16, 89 and 92 as the most repulsive positions. Thus, mutation of positions 7, 47 and 84 may act toward promoting association rather than just reducing repulsion.

| Residue | Position | Average |
|---|---|---|
| MET | 84 | -2.8 |
| ILE | 47 | -2.0 |
| LEU | 47 | -1.7 |
| MET | 47 | -1.7 |
| ILE | 84 | -1.7 |
| LEU | 7 | -1.6 |
| LEU | 84 | -1.6 |
| VAL | 7 | -1.5 |
| ILE | 7 | -1.4 |
| VAL | 47 | -1.4 |
| GLY | 22 | -1.4 |
| GLY | 102 | -1.4 |
| MET | 7 | -1.2 |
| VAL | 84 | -1.2 |
| VAL | 41 | -1.2 |
| PHE | 18 | -1.1 |
| LEU | 18 | -1.0 |
| PRO | 23 | -1.0 |
| LEU | 89 | -0.9 |
| GLN | 84 | -0.9 |
| LYS | 47 | -0.9 |
| PHE | 23 | -0.8 |
| ALA | 26 | -0.8 |
| ILE | 89 | -0.8 |
| CYS | 26 | -0.8 |
| THR | 7 | -0.7 |
| ASP | 22 | -0.7 |
| ASP | 16 | -0.7 |
| LEU | 92 | -0.7 |
| MET | 92 | -0.7 |
| MET | 18 | -0.7 |
| PRO | 81 | -0.6 |
| ARG | 89 | -0.6 |
| TYR | 23 | -0.6 |
| CYS | 84 | -0.6 |
| VAL | 89 | -0.6 |
| MET | 1 | -0.6 |
| PHE | 21 | -0.6 |
| LYS | 92 | -0.5 |
| LEU | 1 | -0.5 |

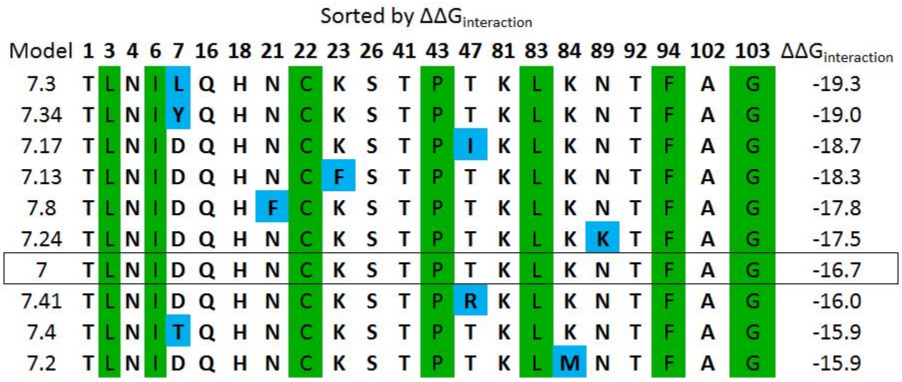
| Model | Residue | Position | Average |
|---|---|---|---|
| 1 | MET | 1 | -0.6 |
| 2 | LEU | 1 | -0.5 |
| 3 | LEU | 7 | -1.6 |
| 4 | VAL | 7 | -1.5 |
| 5 | ILE | 7 | -1.4 |
| 6 | MET | 7 | -1.2 |
| 7 | THR | 7 | -0.7 |
| 8 | ASP | 16 | -0.7 |
| 9 | PHE | 18 | -1.1 |
| 10 | LEU | 18 | -1.0 |
| 11 | MET | 18 | -0.7 |
| 12 | PHE | 21 | -0.6 |
| 13 | GLY | 22 | -1.4 |
| 14 | ASP | 22 | -0.7 |
| 15 | PRO | 23 | -1.0 |
| 16 | PHE | 23 | -0.8 |
| 17 | TYR | 23 | -0.6 |
| 18 | ALA | 26 | -0.8 |
| 19 | CYS | 26 | -0.8 |
| 20 | VAL | 41 | -1.2 |
| 21 | ILE | 47 | -2.0 |
| 22 | LEU | 47 | -1.7 |
| 23 | MET | 47 | -1.7 |
| 24 | VAL | 47 | -1.4 |
| 25 | LYS | 47 | -0.9 |
| 26 | PRO | 81 | -0.6 |
| 27 | MET | 84 | -2.8 |
| 28 | ILE | 84 | -1.7 |
| 29 | LEU | 84 | -1.6 |
| 30 | VAL | 84 | -1.2 |
| 31 | GLN | 84 | -0.9 |
| 32 | CYS | 84 | -0.6 |
| 33 | LEU | 89 | -0.9 |
| 34 | ILE | 89 | -0.8 |
| 35 | ARG | 89 | -0.6 |
| 36 | VAL | 89 | -0.6 |
| 37 | LEU | 92 | -0.7 |
| 38 | MET | 92 | -0.7 |
| 39 | LYS | 92 | -0.5 |
| 40 | GLY | 102 | -1.4 |

**Figure 5-11:** A list of the 'hits' from the FoldX position scan ran on model 7 from the first iteration of the computational routine. The first list is ordered by $\Delta\Delta G_{stability}$ and the second list is ordered by residue position. The $\Delta\Delta G_{stability}$ is measured in kcal/mol, where green denoted the best stabilisation and red, the worst. The residues chosen by FoldX have been highlighted in colours to represent their biochemical properties where green is hydrophobic, yellow is hydrogen bonding potential and red and blue are charged respectively. Notably the top ten mutations are comprised of just three positions.

#### 5.3.2.3.3 Single Mutant Analysis

Each of the proposed single mutations was applied to all subunits in the current starting model and the single mutant analysis process described above was conducted. A portion of the results from the single mutant analysis is presented in Figure 5-12. With the introduction of additional point mutations to the starting model only modest improvements in $\Delta\Delta G_{interaction}$ were made. In the wild-type single mutant analysis, of 38 mutants, the wild type was ranked 31$^{st}$ with a difference of -31 kcal/mol between its interaction energy and the mutant ranked 1$^{st}$. On this second round position scan, however, the starting model was ranked 7$^{th}$ of 40 analysed point mutations and the difference between the starting model and the best mutants was just -2.5 kcal/mol. This result is suggestive of an approach to a local energy minimum; however, further iterations of the routine were not performed on the model to test this hypothesis. Additional runs were performed, using different first round combinatorial mutants, however, the relative ranking of the starting models were 12$^{th}$ of 41 and 22$^{nd}$ of 45, which could also suggest the encroachment upon local energy minima.

Of Interest is the reproducibility of the $\Delta\Delta G_{interaction}$ values obtained. In the initial single mutation round, the wild-type model gave a $\Delta\Delta G_{interaction}$ of 33.1 kcal/mol with a standard deviation of 5.6 kcal/mol. In the combinatorial mutant analysis the average $\Delta\Delta G_{interaction}$ of the wild-type was 31.5 kcal/mol with a standard deviation of 5.6 kcal/mol. This compliance with previous data is replicated in model 7: in the combinatorial mutant analysis it was found to have a $\Delta\Delta G_{interaction}$ of -14.3 kcal/mol and standard deviation of 4.7 kcal/mol. In the second round single mutation analysis, using model 7 as the starting model, its $\Delta\Delta G_{interaction}$ was estimated at -16.7 kcal/mol with a standard deviation of 4.8 kcal/mol



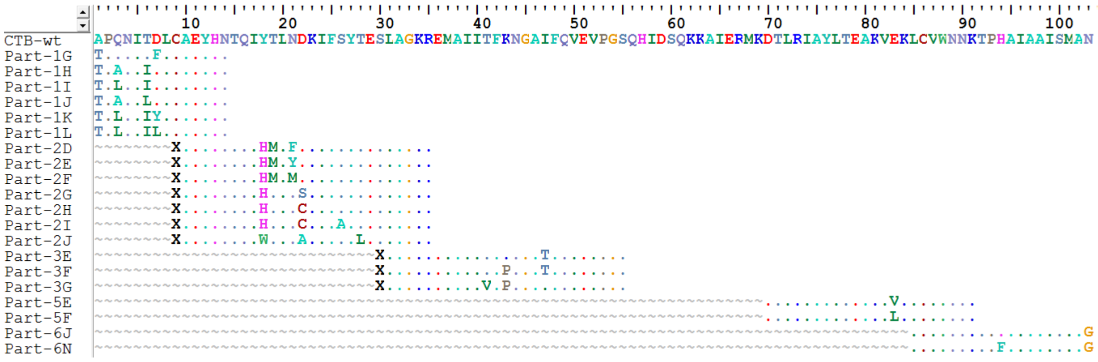
Sorted by $\Delta\Delta G_{interaction}$

| Model | 1 | 3 | 4 | 6 | 7 | 16 | 18 | 21 | 22 | 23 | 26 | 41 | 43 | 47 | 81 | 83 | 84 | 89 | 92 | 94 | 102 | 103 | $\Delta\Delta G_{interaction}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7.3 | T | L | N | I | L | Q | H | N | C | K | S | T | P | T | K | L | K | N | T | F | A | G | -19.3 |
| 7.34 | T | L | N | I | Y | Q | H | N | C | K | S | T | P | T | K | L | K | N | T | F | A | G | -19.0 |
| 7.17 | T | L | N | I | D | Q | H | N | C | K | S | T | P | I | K | L | K | N | T | F | A | G | -18.7 |
| 7.13 | T | L | N | I | D | Q | H | N | C | F | S | T | P | T | K | L | K | N | T | F | A | G | -18.3 |
| 7.8 | T | L | N | I | D | Q | H | F | C | K | S | T | P | T | K | L | K | N | T | F | A | G | -17.8 |
| 7.24 | T | L | N | I | D | Q | H | N | C | K | S | T | P | T | K | L | K | K | T | F | A | G | -17.5 |
| 7 | T | L | N | I | D | Q | H | N | C | K | S | T | P | T | K | L | K | N | T | F | A | G | -16.7 |
| 7.41 | T | L | N | I | D | Q | H | N | C | K | S | T | P | R | K | L | K | N | T | F | A | G | -16.0 |
| 7.4 | T | L | N | I | T | Q | H | N | C | K | S | T | P | T | K | L | K | N | T | F | A | G | -15.9 |
| 7.2 | T | L | N | I | D | Q | H | N | C | K | S | T | P | T | K | L | M | N | T | F | A | G | -15.9 |

**Figure 5-12:** Single mutant analysis of saturation mutagenesis on model 7. Model numbers are on the left hand side and positions across the top. Original mutations from model 7 are in green and new single mutations are highlighted in blue. Models are ranked by $\Delta\Delta G_{interaction}$ which are noted on the right hand side. The majority of the mutations do not provide increases in interaction energy, however six structures do provide a modest improvement.

### 5.3.2.3.4 Assembling Mutant Parts

Mutated sequences based on model 7 and the other mutants identified during the computational procedure thus far, were broken down into six 'parts' that were compatible with the assembly PCR strategy, described in the methods, and can be seen in Figure 5-13. Model 7, for example, can be constructed from parts `1I`, `2H`, `3F`, `5F` and `6N` with the wild-type part `4B` (most wild-type parts are labelled A and/or B). Therefore, we can rename model 7 by the parts used to create it, `IHFAFN`. This nomenclature will be used throughout the remainder of the thesis to refer to specific mutants in order to avoid confusion between mutants coming from different iterations of the computational protocol. Some mutations were modified at this stage if experimental difficulties were anticipated. For example, the mutation Q3P was replaced with Q3A to avoid potential proline isomerisation problems.



```
                10        20        30        40        50        60        70        80        90       100
CTB-wt   APQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVPGSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN
Part-1G  T.....F.......
Part-1H  T.A..I........
Part-1I  T.L..I........
Part-1J  T.A..L........
Part-1K  T.L..IY.......
Part-1L  T.L..IL.......
Part-2D  ~~~~~~~~X........HM.F.............
Part-2E  ~~~~~~~~X........HM.Y.............
Part-2F  ~~~~~~~~X........HM.M.............
Part-2G  ~~~~~~~~X........H...S............
Part-2H  ~~~~~~~~X........H...C............
Part-2I  ~~~~~~~~X........H...C...A........
Part-2J  ~~~~~~~~X........W...A.....L......
Part-3E  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~X................T........
Part-3F  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~X............P...T........
Part-3G  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~X..........V.P............
Part-5E  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~.............V........
Part-5F  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~.............L........
Part-6J  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~..................G
Part-6N  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~.........F........G
```

**Figure 5-13:** A sequence alignment of the mutant 'parts' ordered for assembly PCR. Different combination of these parts allows the construction of different full length mutants, i.e. model 7 can be constructed from parts 1I, 2H, 3F, 5F and 6N with the wild-type part 4A, not listed.

# 5.4 Experimental evaluation of mutant designs

## 5.4.1 Analysis of Individual Mutant Parts

DNA oligos for the individual parts were ordered and each of the mutant DNA parts were assembled by assembly PCR with the remaining five wild-type parts before cloning into the pSAB2.2 expression vector. These constructs were expressed and purified and the protein was run, unboiled, on SDS PAGE (Figure 5-14). Note, constructs containing parts 2I and 3G failed to progress in a timely manner through the cloning procedure and were not expressed. Parts 1I, 1J, 2G, 3E, 3F and 6J expressed to give comparable yields to the wild-type and were able to maintain their pentameric structure when run unboiled on SDS PAGE. Part 6N displayed a band at the appropriate size for CTB however its level of expression is severely reduced. There was also a suggestion of an appropriately size band for pentameric CTB when using part 5E, however, this is of negligible quantity. Parts 1H, 5E and 5F did not have pentameric bands on the gel, however a monomer band can be seen towards the dye front. As our purification process is known to bind only pentameric CTB, we can attribute the presence of only a monomer band to a reduced pentameric stability which is inadequate to maintain pentamers under the conditions of standard SDS PAGE. Of note are the smears emanating from the 1I, 1J and 2G part mutants which indicate a small proportion of pentamer dissociation during the SDS PAGE run. It is interesting to note that part 1H contains Q3A and T6I substitutions and does not express well, whereas 1I and 1J respectively substitute A3L and I6L into the 1H sequence and in each case expression is recovered. These examples provide a nice demonstration of recovery mutations in which a leucine substitution at either of the two deleterious positions restores near native expression.

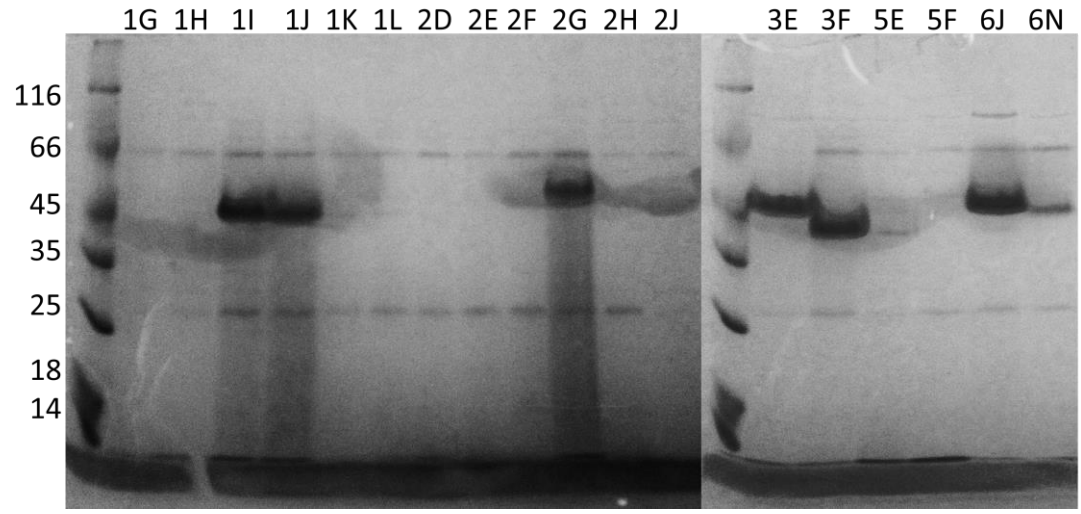



**Figure 5-14:** SDS PAGE of S#1.0 single part mutants. SDS PAGE was conducted on non-boiled samples of the mutant protein. As is evident a number of the mutants did not express any protein. Of additional interest some of the mutants were seen to run as monomers rather than pentamers, such as 5E and 5F. Also, some mutants display a 'smudge' from the pentamer band to the monomer band indicating the pentamer disassembling during the course of the SDS PAGE, such as 1I, 1J and 2G.

DSF was also performed in order to gain a better understanding of the impact of these mutations on the stability of the pentamers. Melting curves were not obtained for mutants containing parts 1G, 1H, 1K, 1L, 2D, 2E, 2F, 2H, 2J or 6N, which suggests the absence of stable folded protein. However 1I, 1J, 2G, 3E, 3F, 5E, 5F and 6J did produce melting curves, which are summarised in Figure 5-15. Figure 5-15 shows a correlation between morphology of the protein on SDS PAGE and its melting temperature by DSF. This data set is limited in size but there are other observations which also follow this trend. Part 5E shows a marginal amount of pentameric CTB whereas 5F shows none and their respective $T_m$ values of 62°C and 57°C agree with this observation. Additionally, the smear of 2G appears darker than that of 1I and 1J which each have a $T_m$ 2°C higher than 2G. This correlation is interesting as it allows us to make assumptions of relative stability based on the morphology of pentamers on SDS PAGE. For example the addition of multiple destabilising mutations is likely, though of course not inevitable, to increase the destabilisation beyond that of the single parts. Thus SDS PAGE can be used to inform us about the likelihood of expressing stable multiple mutants based on the stability of the single parts, but as described above, this assumption will not identify recovery mutations.

Ordered by Part

| Part | Gel Result | DSF Tm (°C) |
|---|---|---|
| 1I | Smear | 68 |
| 1J | Smear | 68 |
| 2G | Smear | 66 |
| 3E | Pentamer | 74 |
| 3F | Pentamer | 73 |
| 5E | Monomer | 62 |
| 5F | Monomer | 57 |
| 6J | Pentamer | 70 |

Ordered by DSF $T_m$

| Part | Gel Result | DSF Tm (°C) |
|---|---|---|
| 5F | Monomer | 57 |
| 5E | Monomer | 62 |
| 2G | Smear | 66 |
| 1I | Smear | 68 |
| 1J | Smear | 68 |
| 6J | Pentamer | 70 |
| 3F | Pentamer | 73 |
| 3E | Pentamer | 74 |

**Figure 5-15:** DSF comparison to stability during SDS PAGE. The DSF-derived $T_m$ values for single part mutants in this strategy are compared to the behaviour of the mutants during SDS PAGE. Those mutants with the lowest melting temperature are found to be unable to maintain pentamers during SDS PAGE, however those with higher melting temperatures maintain pentamers on the gel. Additionally those with intermediate $T_m$ are seen to create smears during SDS PAGE suggesting dissociation of the pentamer during SDS PAGE.

## 5.4.2 Analysis of Combinatorial Mutants

The mutant and wild-type parts were rearranged in all possible combinations having one member from each mutant part set 1-6, to give a combinatorial library of 2016 mutant CTB models. This saturation combinatorial library production removes human bias and is feasilble given a reduced number of initial parts (approximatly 20). Each model was assessed in the same manner as described earlier for the combinatorial mutant analysis. This intensive combinatorial approach ranked the interaction energy of the wild-type at 1932nd of the 2016 models, showing that in most case, any of these parts were predicted to give an improvement on the interaction relative to the wild-type.

By comparing every combination of the parts shown to express protein in the single mutant analysis, 72 combinatorial mutants are feasible (Figure 5-16). A number of mutants were selected which incorporated the parts `1I` or `1J` with `2G` and `3F`, as these expressed well, along with `5E`, `5F`, `6J` or wild-type parts. As described above, the mutant containing parts `1I`, `2G`, `3F`, `5F` with wild-type parts `4A` and `6A` is written as `IGFAFA`. From the combinatorial analysis, one mutant `IAFAEJ` gave a $\Delta\Delta G_{interaction}$ of -18.6 kcal/mol which predicts a much greater interaction energy than the next nearest mutant `JGFAAJ` at -7.4 kcal/mol, however, this mutant contained part `5E` which shows increased thermoinstability. Thus it was decided not to attempt to express this mutant. The mutants highlighted in Figure 5-16 were expressed with C-terminal coiled-coils, as described in chapter 4, these mutants are characterised in the next chapter.

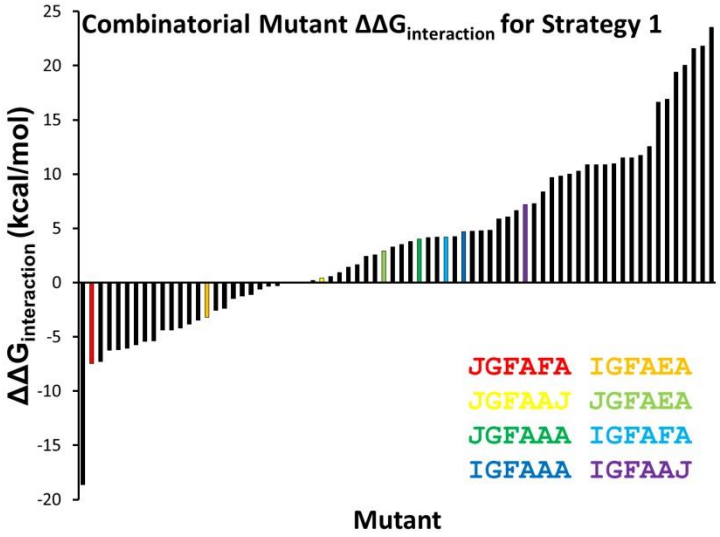



**Figure 5-16:** Combinatorial analysis of parts found to be experimentally expressible. The $\Delta\Delta G_{interaction}$ was assessed for the 72 combinatorial mutants built from parts which expressed protein. Highlighted mutants were expressed experimentally and described in Chapter 4.

The mutant IGFBAA (identical to IGFAAA above, as part-4B is a silent mutation of the wild-type) was chosen to produce a rotational and z-axis translational energy landscape based on dodecahedral symmetry as conducted previously for wild-type CTB and shown in Figure 5-2a & b. A comparison between this energy landscape for the wild-type and for the mutant IGFBAA can be seen in Figure 5-17. The initial structure, on which the IGFBAA mutant was designed, lay at 44° rotation and 8 Å z-axis translation. This energy landscape shows stabilisation of the 44° rotation area with respect to the rest of the landscape, including the competing minimum at around 62° rotation, which is evident from the 6 Å to 8 Å translations. This observation suggests that the method we have used effectively optimises the interface presented to the design routine. Although there is no evidence of a negative interaction energy from these mutations, the removal of repulsive residues has resulted in a lowering of the $\Delta\Delta G_{interaction}$ of the interface.

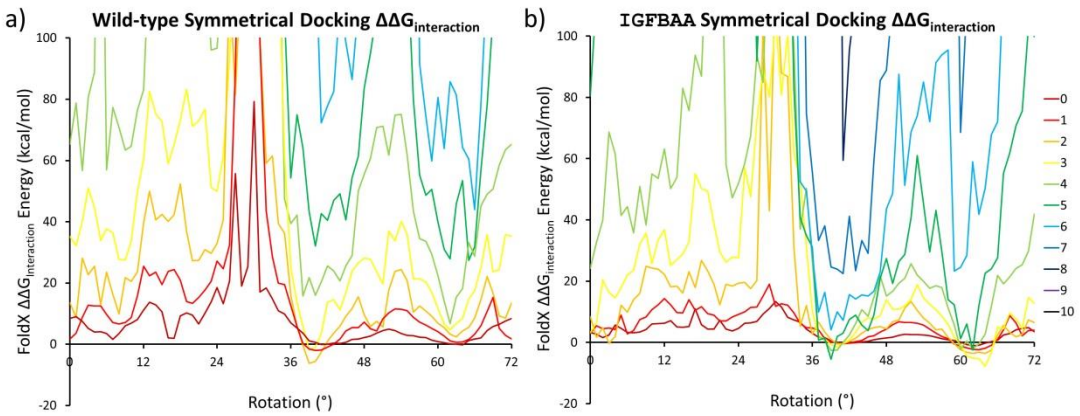



**Figure 5-17:** Wild-type and mutant symmetrical docking landscape. The $\Delta\Delta G_{interaction}$ energy as calculated by FoldX on a selection of structures produced by CHARMM encompassing 72° of rotation and 10 Å of capsid contraction (z-axis translation). **a)** Shown previously in Figure 5-2b, the dodecahedral docking results for the wild-type protein. As is evident, energy minima appear in the 4 Å to 7Å ranges at around 44° and 62°. **b)** After application of the suggested mutations from FoldX calculations, based on a model positioned at 44° rotation and 8 Å translation, the energy minimum at around 44° rotation is amplified with respect to the other rotations and also enriched with respect to the 62° energy minimum. Therefore, after mutation with respect to a specific orientation, that orientation is minimised with respect to the other possible orientations.

# 5.5 FoldX benchmarking

The Im9-E9 complex is a well-documented protein-protein interaction[165-168]. The E9 DNase is an endonuclease colicin belonging to a family of cytotoxically active bacterial toxins. The activity of E9 is inactivated by the host immunity protein Im9 upon complexation. An alanine scan was conducted on the Im9 protein in order to gain understanding into PPIs by Wallis et al.[169]. In that study, dissociation kinetics were determined by radioactive subunit exchange monitored by stopped-flow fluorescence. The stability of the complex was determined by fluorescence spectroscopy under guanidinium hydrochloride gradients. Their experiments revealed a number of thermodynamic parameters for the complex including the $\Delta\Delta G_{interaction}$ and the $\Delta\Delta G_{stability}$. These values were used to retrospectively benchmark the performance of FoldX. A 2 ps MD simulation of the pdb `1BXI` was conducted in order to generate ensemble structures. These structures were passed to FoldX to perform an alanine scan (Figure 5-18a/c) which simply truncates side-chains from $C_\beta$ as described in the methods. The $\Delta\Delta G_{stability}$ data provides a reasonable agreement with the experimental data. The mutagenesis function of FoldX was used to create models containing the appropriate alanine mutation. These structures were then analysed by the 'analyse complex' function to give $\Delta\Delta G_{interaction}$ energies for the complex (Figure 5-18b/d).

However the $\Delta\Delta G_{interaction}$ energies show little correlation between the experimental values and the computational predictions. This data suggests that alternative computational methods may improve the frequency at which mutations which improve the $\Delta\Delta G_{interaction}$ are discovered.

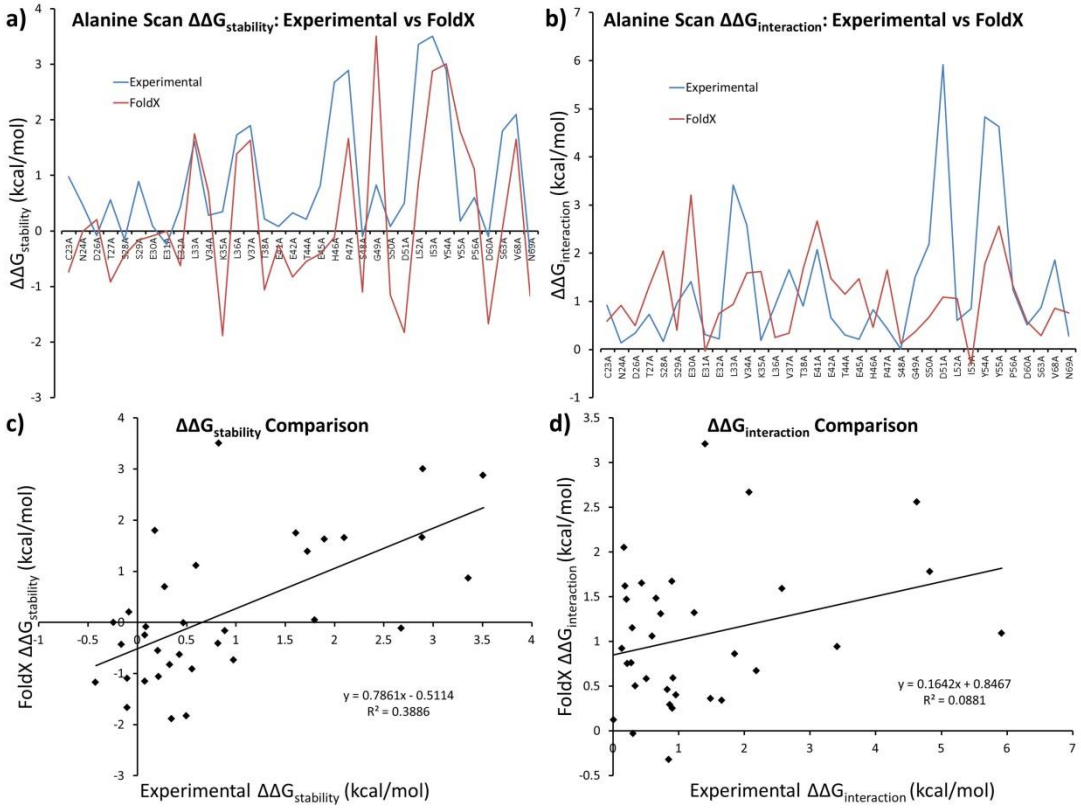



**Figure 5-18:** Experimental data vs. computational prediction for both $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{interaction}$ against the Im9-E9 complex. **a)** The $\Delta\Delta G_{stability}$ comparison, which shows a reasonable correlation between experimentally determined values and those predicted by FoldX. **b)** The $\Delta\Delta G_{interaction}$ comparison, which shows little correlation between the experimental and computational data. **c)** a scatter between computational and experiments assessments of $\Delta\Delta G_{stability}$, revealing R2 of 0.39. **d)** a scatter between computational and experiments assessments of $\Delta\Delta G_{interaction}$, revealing R2 of 0.09.

# 5.6 $\Delta\Delta G_{interaction}$ Benchmarking

Lumazine synthase in *Aquifex aeolicus* is a dodecahedral protein cage constructed from pentameric subunits of a slightly larger size than CTB. A recent paper by Chen and Woycechowsky (2012)[170] lists a number of mutations which prevent assembly of this protein cage to leave soluble pentamers. This provided an excellent test bed to compare our computational strategy to protein cage assembly experimental data. SEC revealed changes in oligomeric state upon mutation; some mutations resulted in the loss of dodecahedral assembly, while others caused no change in oligomeric state (Table 5-2).

| Nomenclature | Mutations | Pentamers Assembled |
|---|---|---|
| w1 | Wild-Type | 12 |
| w2 | R40E/H41E | 12 |
| w3 | R21E/R40E/H41E | 12 |
| w4 | H41R/L121R | 1 |
| w5 | R40E/H41E/L121E | 1 |
| w6 | R40E/H41E/I125E | 1 |
| w7 | R40S/H41S/I125S | 1 |
| w8 | R40S/H41L/L121E | 1 |

**Table 5-2:** Description of the nomenclature and oligomeric state of Lumazine synthase mutations found by Woycechowsky et al[170]. The mutants w2 and w3 maintain dodecahedral assembly. the remaining mutants cannot assemble into higher order structures.

All-atom CHARMM MD simulations of the mutants were conducted on a single pentamer with periodic boundary conditions applied to emulate simulation of the entire capsid. FoldX analysis of the interactions was conducted, the MD and FoldX analysis was conducted by a first year Wellcome Trust PhD rotation student, Sam Hickman. Analysis with Pisa (CCP4) was conducted upon these CHARMM simulations by Dr Emanuele Paci and Rosetta was applied to evaluate the CHARMM simulations by me.

Unlike CTB, the LS simulations were conducted on one pentamer and periodic boundary conditions were applied to estimate the interactions of the complete capsid. As with the computational routine developed for CTB, the trajectories from these simulations were used to generate 600 ensemble structures of each simulation. Figure 5-19 presents the data from this investigation. Foremost, the $\Delta\Delta G_{interaction}$ presented is as the Relative $\Delta\Delta G_{interaction}$, compared to the wild-type, which is a negative value.

$$Relative\ \Delta\Delta G_{interaction}\ \frac{\Delta\Delta G_{interaction_{wild-type}}}{\Delta\Delta G_{interaction_{Mutant}}}$$

Therefore an increased in Relative $\Delta\Delta G_{interaction}$ shows an improvement in $\Delta\Delta G_{interaction}$. There are a number of interesting features. The expected outcome was that w2 and w3 would show an improved $\Delta\Delta G_{interaction}$ over the destabilised mutants, yet the wild-type would maintain the best $\Delta\Delta G_{interaction}$. For both Pisa and Rosetta this is mostly true with the exception of mutant w7. However for FoldX this does not follow suit and w3 has one of the worst $\Delta\Delta G_{interactions}$ of the set. Additionally, Pisa and Rosetta follow almost an identical

ranking of the mutants, which was exciting to see. FoldX however did not rank the mutants in a similar order. The fact that all three programs ranked w7 better than the mutants which maintain capsid assembly was an interesting find as the three programs were simultaneously incorrect which may point to underlying problems with which computational programs assess $\Delta\Delta G_{interaction}$ energies, or problems with the method of experimental assessment of the oligomeric state. In dashed lines on the graph are potential threshold values for capsid assembly.

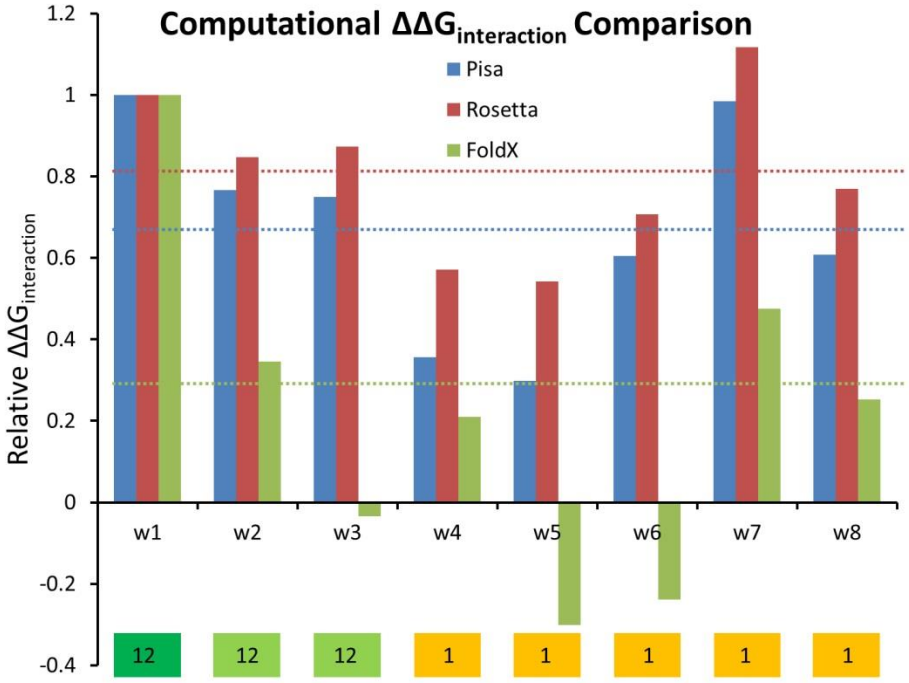



**Figure 5-19:** Software comparison for calculating the $\Delta\Delta G_{interaction}$ of Lumazine synthase. The $\Delta\Delta G_{interaction}$ (in kcal/mol) was scaled relative to the wild-type (w1). The boxes underneath the bar chart signify the number of pentamers which oligomerise according to experiment, where dark green (w1) is the wild-type LS, pale green (w2 & w3) are the mutants which still formed dodecahedrons and orange (w4 - w8) are the mutants which did not assemble. Dashed lines are potential assembly threshold $\Delta\Delta G_{interaction}$ energies for each of the investigated techniques.

This investigation suggests that programs such as Pisa and Rosetta are better equipped to calculate accurate $\Delta\Delta G_{interaction}$ energies as their results agree with the experimental data more so than FoldX does. Another exciting outcome of this investigation is how close the relative predictions of Pisa and Rosetta are to each other.

# 5.7 Closing Statements

The routine described in this chapter successfully predicts reductions in $\Delta\Delta G_{interaction}$ across the proposed interface *in silico*. This is achieved by selecting mutants which reduce the total $\Delta G_{stability}$ for one monomer in a trimer of pentamers model; building this mutation into each subunit and assessing the $\Delta\Delta G_{interaction}$; and finally constructing and assessing combinatorial mutants. It was found that the majority of the saturation mutants that were suggested translated into mutations which improved the assessed $\Delta\Delta G_{interaction}$. Of course, validating an *in silico* technique by using the same technique could be seen as a rather obvious and somewhat redundant method for determining a correlation. One argument against this view is the transfer of the assessed structures between Pymol, CHARMM and FoldX, i.e. before each assessment the models are constructed in Pymol and simulated by CHARMM before being analysed by FoldX. However the calculations are made by FoldX and thus any problematic residues or interactions would be miss-calculated in both the saturation mutagenesis and the PPI analysis.

Surprisingly approximately half of the suggested mutations did not produce a workable yield of CTB, however the respective alanine substitutions in these positions did allow expression. This observation is suggestive that the bulky residues introduced are having a greater negative effect on the pentameric stability than just the removal of the native amino acid side-chain. In the next chapter we discuss the improvement these mutations make up on the assembly of the capsids.

# Chapter 6
# Characterisation of Scaffolded Mutants

# 6.1 Mutated and Scaffolded CTB VLPs

## 6.1.1 Overview

As discussed in the project summary in the introduction, It was anticipated that the addition of coiled-coils to CTB pentamers would provide an assembly scaffold for creating CTB virus-like particles. These particles were not expected to be stable as CTB does not readily aggregate in solution. Therefore, the plan was to remodel the proposed interface to make the formation and final stability of these particles more favourable. The steps discussed previously are reiterated below, before presenting the VLPs that have been constructed.

## 6.1.2 Mutating Wild-Type CTB

Chapter 3 revealed the mutability of CTB, I.e., that many of the positions on the proposed interactions surface are mutable to alanine without destabilising the pentamer. However certain positions were identified to be quite destabilising. Differential Scanning Fluorimetry was employed to measure the $T_m$ of alanine mutants and, in chapter 5, other mutations. This technique revealed a correlation between morphology of the pentamers on SDS PAGE and the overall thermostability of the protein. Therefore, the morphology of the pentamers on SDS PAGE can be used as a qualitative measure of the pentameric stability in solution.

## 6.1.3 Scaffolding Wild-Type CTB

In chapter 4 the scaffolding of the wild-type CTB to produce capsids was discussed. The two coiled-coils, JR-coil2 and JR-coil7, were able to make both dimeric and trimeric interactions, but each with a preference for a different oligomeric state. These peptides were found to scaffold the assembly of CTB into capsids with varying success, but as the coiled-coils are short, their interaction affinities are weak. The pentavalency introduced via fusion to CTB, however, should allow enhanced binding interactions based on multivalency after the initial nucleation event. These weak binding interactions should make partial assemblies relatively unstable and hopefully lead from pentamers to assembled capsids with a relatively low number of kinetic assembly traps.

## 6.1.4 Mutating the Interface

It was argued that the slow assembly rate and yield, and relatively low stability of the particles were a product of unfavourable interface interactions between the wild-type CTB surfaces. Removal of these repulsive interactions should allow the assembly rate and yield

of product to be improved. Starting from a discrete low yielding product is an optimal position to investigate the effects of future mutations as it allows the effect of the removed repulsive interaction to be more discernible against the 'native' assembly rate. Chapter 5 described a computational routine developed to remove repulsive interactions from the proposed CTB interface. However this routine has little feedback until some experimental evidence suggests if the chosen mutations have any effect on the assemblies.

### 6.1.5 Scaffolding the Mutants.

Here, in chapter 6, the scaffolding of the mutants with coiled-coils is discussed. The computational design strategy developed in chapter 5 is evaluated experimentally. The introduction of JR-coil2 is discussed first, followed by JR-coil7. The majority of the analysis in this chapter is performed by SEC and TEM as these methods allow a quantitative yield of product to be calculated (through integration of the SEC 280 nm trace) and direct visualisation of the final product.

# 6.2 Characterising First Generation Mutants (S#1.0)

### 6.2.1 Strategy S#1.0

The computational routine to suggest mutations was performed on wild-type CTB, described in chapter 5 A number of combinatorial mutants were chosen for expression, of which the mutants in Figure 6-1 were assembled in the pSAB2.2 plasmid, expressed and purified as described previously and in the methods.

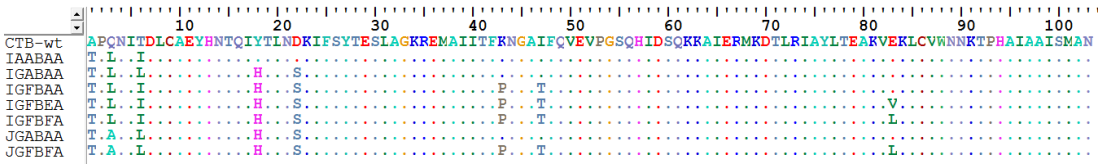



**Figure 6-1:** The amino acid sequence for combinatorial mutants from S#1.0. The expressible mutants from this round of single mutants were combined. Only seven were chosen for extension with coiled-coils.

These mutants were first expressed with the JR-coil2 coiled-coil as a *C*-terminal fusion, and then selected mutants which showed the production of high weight species under SEC were reassembled with the JR-coil7. For future reference the wild-type CTB is also referred to as AAAAAA and as BAABAA, based on the parts used for assembly PCR, part-1B contains

a A1T mutation but as this is a known natural CTB biotype mutation, it will still be referred to as the wild-type.

#### 6.2.1.1 Characterisation of Mutants with JR-coil2

The chosen mutants from the computational screen were assembled from parts as described in the methods, each with the addition of the JR-coil2 coiled-coil sequence. Following expression, the proteins were precipitated and purified by nickel affinity chromatography. Figure 6-2a shows a gel of a selection of these mutants expressed and purified. Lanes 3 to 10 show the nickel column elution followed by the ammonium sulphate precipitate fraction. As is evident, only IGFBAA maintains pentameric stability on the SDS PAGE. Mutant CTB is not discernible in the other mutant elutions. JGABAA does not appear to express. IGFBAA elution fractions were combined, when ran on SDS PAGE boiled and non-boiled in lanes 13 and 14. Lanes 15-20 contain the boiled and non-boiled elutions from IGFBAA, IGFBEA and IGFBFA which have been concentrated. As is evident in lanes 5 and 7, there is very little protein for IGFBEA and IGFBFA, however, after concentration the mutant CTB can be seen, although the purity is compromised.

These mutated CTB molecules containing three or more mutant parts were not expressed well, therefore mutants containing fewer mutant parts were constructed. These mutants expressed to a higher degree than those with more mutant parts (gel not shown). Figure 6-2b displays the SEC trace for a selection of these modest mutants, IAABAA, IGABAA and IGFBAA after differing lengths of incubation. IAABAA displayed a standard pentamer peak at 0.85 relative retention volume (rrv), but with a negligible high weight peak at 0.4 rrv despite the addition of the JR-coil2 sequence. IGABAA showed a high weight peak at 0.4 rrv which increased over longer incubations, however IGABAA also showed an increase in a peak at 1.15 rrv, which is thought to be monomeric CTB, suggesting that this mutant's pentameric stability is compromised. The part 2G contains a D22S mutation (as well as the bio-type variant Y18H) which removes an intra-subunit salt bridge between D22 and K43. This may cause monomer instability which in turn may lead to the disassembly of the pentamer. Surrounding D22 are another two lysine residues: K23 and K81. The removal of D22 may dramatically alter the local electrostatics of the area thus resulting in these three lysine residues repelling each other. IGFBAA however, expressed better and maintains its pentameric stability over extended periods of incubation of up to two months at room temperature. IGFBAA also showed a pronounced high weight peak after seven days of incubation (Figure 6-2c). This result was exciting as the ratio of high weight product

formed was increased compared to the wild-type CTB also containing the JR-coil2 extension. Therefore, the mutations contained in this mutant are presumably responsible for the increased level of assembly. Part 3F contains the mutation K43P and the bio-type variant I47T. As discussed for IGABAA, part 2G makes a destabilising D22S substitution, however, in mutant IGFBAA the free K43 is substituted with a proline which recovers the pentameric stability, which could be due to the relief of the local electrostatic field.

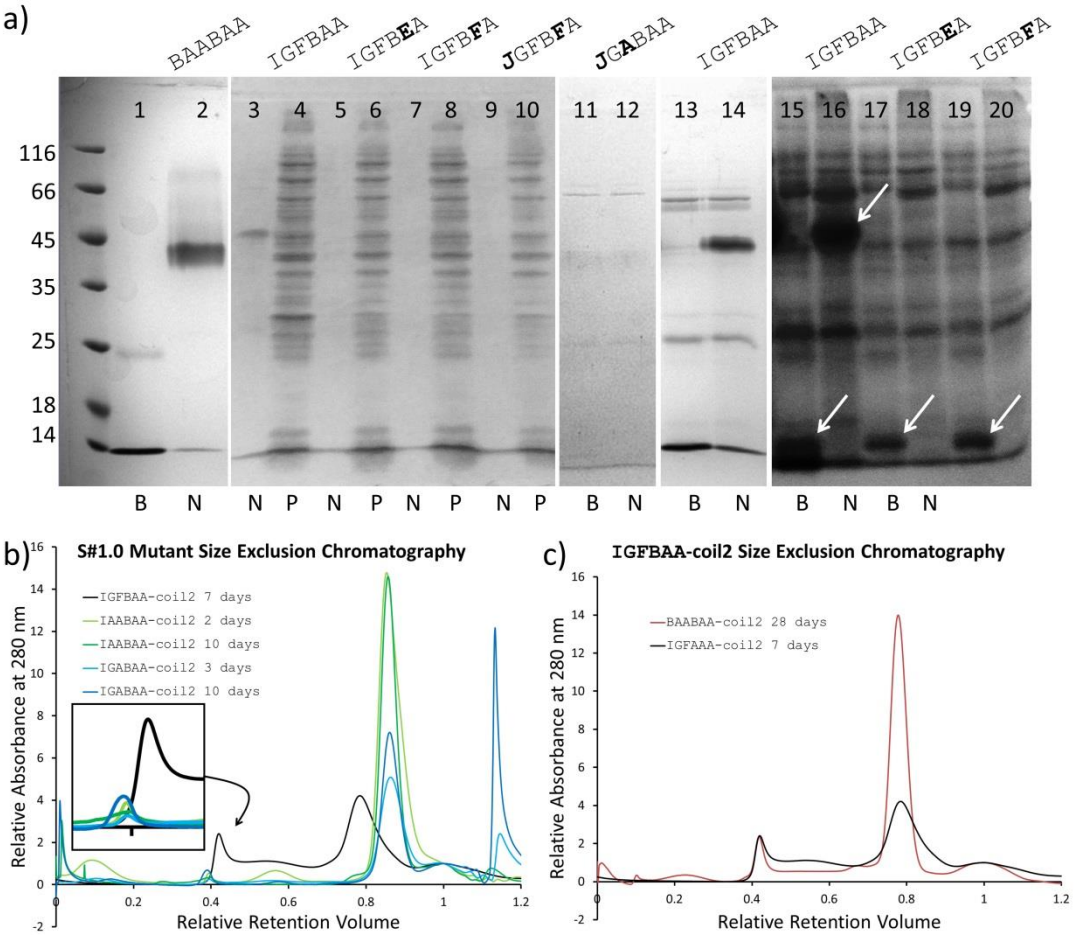



**Figure 6-2:** Purification, concentration and SEC of the combinatorial mutants from S#1.0. **a)** A selection of SDS PAGE of the selected mutants. Mutants are identified at the top of the gel and contain boiled (B), non-boiled (N) nickel elutions, boiled clarified precipitate fractions (P). Lanes 3-12 show expression of the mutants, only IGFBAA produced viable pentamers on SDS PAGE. After concentration of the nickel elutions CTB monomers are visible, lanes 15-20, however only IGFBAA has a higher concentration of CTB than contaminants. **b)** SEC of the expressed mutants. In green is IAABAA at 2 and 10 days incubation. There is very little assembly of these mutants although in the 2 day sample there does appear to be a peak at a relative retention volume (rrv) of 0.5; however this does not translate into high order structures with prolonged incubation. In blue is IGABAA at 3 and 10 days incubation. These show a modest increase in assembly given prolonged incubation, however there is also evidence of an increase in CTB monomers in solution over time, at 1.15 rrv. The mutant IGFBAA (in black) at 7 days incubation. Unfortunately due to changes in chromatography conditions and equipment this trace differs to other mutants, however there is evidence of a large high weight species and intermediate products. **c)** A comparison between the wild-type BAABAA and the mutant IGFBAA both with JR-coil-2 attached. Despite the reduced overall expression and the shorter incubation time, IGFBAA assembles into a similar yield of high weight product. Note, SEC traces have been normalised using the peak at 1 rrv in order to show differences in yields and ratios of assembly.

### 6.2.1.2 Characterisation of Mutants with JR-coil7

As the wild-type CTB-JR-coil7 produced higher quality capsids (as judged by TEM) than wild-type CTB-JR-coil2 proteins, JR-coil7 was also appended to the IGFBAA set of mutants for visualisation under TEM. The IGFBAA mutant gene with the JR-coil7 sequence was constructed by assembly PCR before expression, purification and biophysical analysis. IGFBAA-coil7 expressed less well than the wild-type, BAABAA-coil7, as evident from the gel in the Figure 6-3a. Thus when IGFBAA-coil7 was concentrated to 1.6 mg/ml, this sample contained a higher concentration of contaminating species. SEC performed on this sample showed an increase in the rate of assembly compared to the wild type (Figure 6-3b). The wild-type BAABAA-coil7 gave a high weight peak that was 9.7% the size of the pentameric peak; IGFBAA-coil2 produced a high weight peak 30.4% the size of the pentamer peak; and the high weight peak of IGFBAA-coil7 was 20.4% the size of the pentamer peak. Despite the JR-coil2 variant producing more high weight product, as the SEC trace baseline seemed somewhat elevated throughout the latter stages it was thought that coil7 may actually have produced more high weight species. The high weight fractions were pooled, concentrated and analysed by DLS. As seen in Figure 6-3c, the mutant IGFBAA-coil7 produced particles between 30 and 36 nm in diameter, which was similar to wild-type CTB with coil7. Of note is the widths of the curves in the IGFBAA-coil7 trace are much narrower than those for the wild-type which suggests a more discreet range of particle sizes. When analysed by TEM (Figure 6-3d), the IGFBAA-coil7 mutant could be seen as spherical particles which measured between 30 and 37 nm, with an average of 33 nm, which correlates well with the particle size discerned from DLS which varied between 30 and 36 nm. The capsid in Figure 6-3d(1) is rather ill-defined but displays a different morphology from the capsids in Figure 6-3d(2-6). It is larger in size and possesses a more elliptical geometry. Figure 6-3d(7) shows a capsid under lower magnification. This capsid also possesses a more elliptical geometry but in addition appears to house two smaller capsid structures. These smaller capsid structures each have an average diameter of 31 nm which correlates with the structures seen elsewhere. The larger capsid has an average diameter of 70 nm.

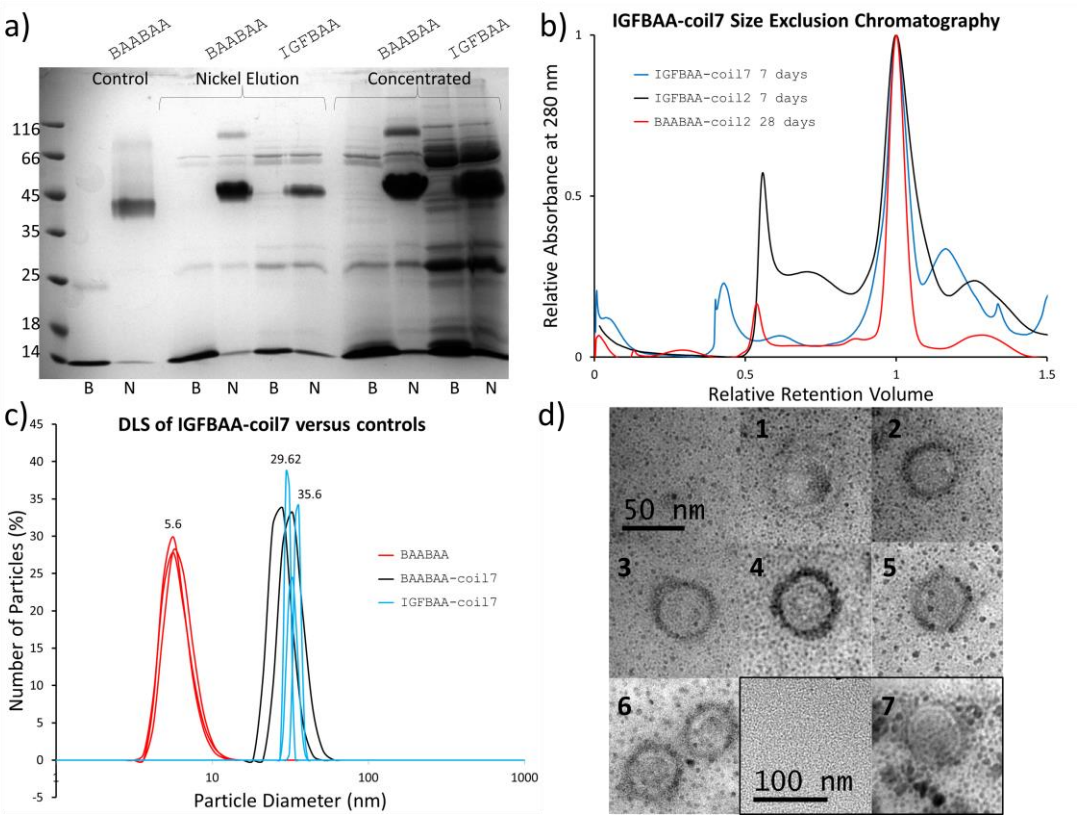



**Figure 6-3:** Biophysical analysis of the IGFBAA-coil7 mutant. **a)** SDS PAGE analysis of IGFBAA-coil7 compared to the wild-type BAABAA. The control lanes are BAABAA without JR-coil7, boiled samples (B), non-boiled samples (N). The total expression of IGFBAA is compromised with respect to the wild-type however IGFBAA still forms the majority of the product. In order to achieve comparable concentrations of IGFBAA to BAABAA the relative concentration of the contaminants becomes increased. **b)** SEC of IGFBAA-coil7 compared to the wild-type and the coil2 equivalent. IGFBAA shows an increased assembly ratio compared to the wild-type, however less assembly than IGFBAA-coil2. **c)** DLS comparison between IGFBAA-coil7 and the wild-type, with and without JR-coil7. The presence of the mutations in IGFBAA seems to make the particle size distribution more discrete, as seen by the narrower DLS traces. Particles of a comparable size to those achieved for the wild-type are seen. **d)** TEM of the IGFBAA particles from the SEC high weight peak. Panels 1-6 are at the same magnification as the 50 nm scale bar, whereas panel 7 matches the 100 nm scale bar. Particles measure approximately 33 nm which falls in line with DLS observations and theoretical predictions, note some stain deposits are present.

It was also interesting to note that the number of particles visible under TEM was increased compared to the wild-type assemblies. This could be due to an increased level of capsid stability inferred by the mutant interface. The increase in capsid yield is exemplified in Figure 6-4 which shows two representative micrographs, with four structures in a 100 nm scaled electron micrograph and five structures in a 200 nm scaled micrograph. In comparison, the wild-type protein produced far fewer capsids.

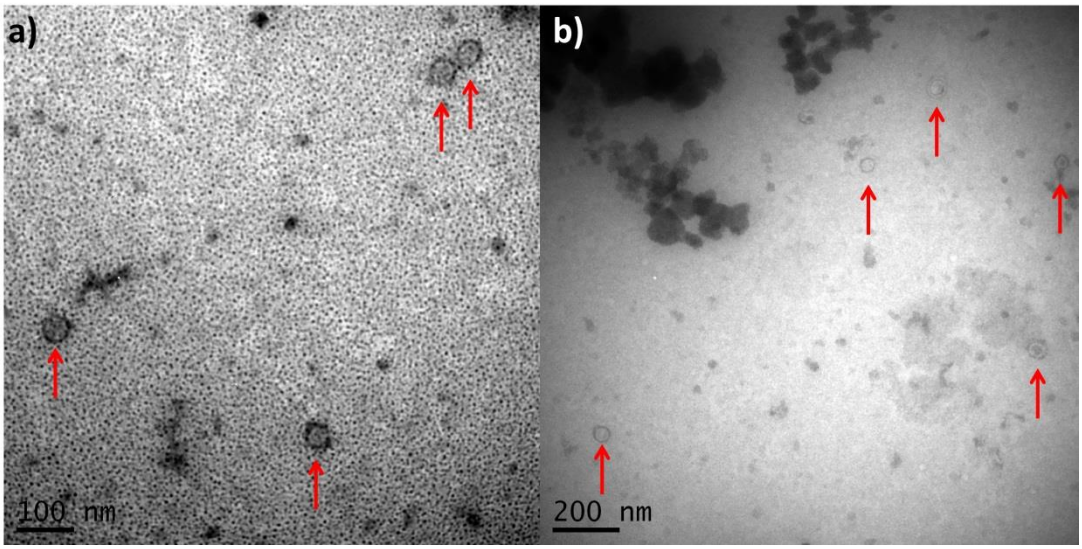



**Figure 6-4:** Representative electron micrographs of IGFBAA-coil7 displaying frequency of capsids visualised on populated areas.

### 6.2.1.3 Characterisation of Mutants with de Crescenzo coils.

The de Crescenzo coiled-coils described in Chapter 4 were also appended to the IGFBAA mutant. Two variants were constructed containing five heptad variants of either DeCr-coilE or DeCr-coilK. These constructs were expressed, purified and concentrated as per the standard technique described in the methods. Samples of the individual mutants and mixtures of both in ratios 1:5, 1:1 and 5:1 were left to incubate at room temperature for 2 days, before being analysed by SEC.

IGFBAA-coilE expressed as well as the IGFBAA-coil7 variant providing 0.6 g/L and was concentrated to 2.9 mg/ml, however, as seen in Figure 6-5a, IGFBAA-coilK expressed poorly allowing the retrieval of 0.2 g/L which was concentrated to 1 mg/ml . The reduced expression level is intriguing as the only difference between the two constructs is an EKE motif and a KEK motif, respectively. Translating this charge difference along the 5 heptad coiled-coil and around each pentamer creates a net change in charge of 50*e* between the two variants. This charge difference may affect the folding and assembly kinetics of the pentamer or it may in some way restrict the export of the pentamers into the media. The black arrows in Figure 6-5a show the pentameric and monomeric bands for IGFBAA-coilE. The red arrows highlight the position of the IGFBAA-coilK. This highlighted band in not

present in the non-boiled lane of IGFBAA-coilE, however it is present in the IGFBAA-coilK and as this is the only difference between the two samples, it is presumably attributable to IGFBAA-coilK. The SEC traces can be seen in Figure 6-5b and displays some intriguing properties. The concentration of the coil-K variant is very low as discernible by the SEC trace, however there is a small peak at the pentamer retention volume, additionally there is a significant high weight peak. There is a peak for a species below the size of the pentamer at retention volume of 22 ml; however, this peak is only present in the mixtures of the two IGFBAA variants and not in the samples of individual variants. The greatest yield of high weight product comes from the 1:1 mix of E and K variants. It is interesting to note that this sample displays one of the lowest peak heights for the pentamer peak which suggests that the majority of the product has formed larger complexes. There is also a peak at 15 ml which could indicate the presence of a dimer of pentamers. Figure 6-5c displays a representative electron micrograph from capsid populated areas of a grid with the 1:1 mixture of IGFBAA variants. As the histogram in Figure 6-5d reveals, these particles are relatively monodisperse averaging 27 nm in size. In addition to the presence of these 27 nm diameter particles, there was also a small population of larger capsids measuring from 30-40 nm in size. A selection of these larger particles displaying evidence of a 3-fold symmetry axis can be seen in Figure 6-6. The presence of a 3-fold symmetry axis in particles made from pentamers indicates a closed particle as 2D tiling of pentagons cannot form 3-fold symmetry axes[107].

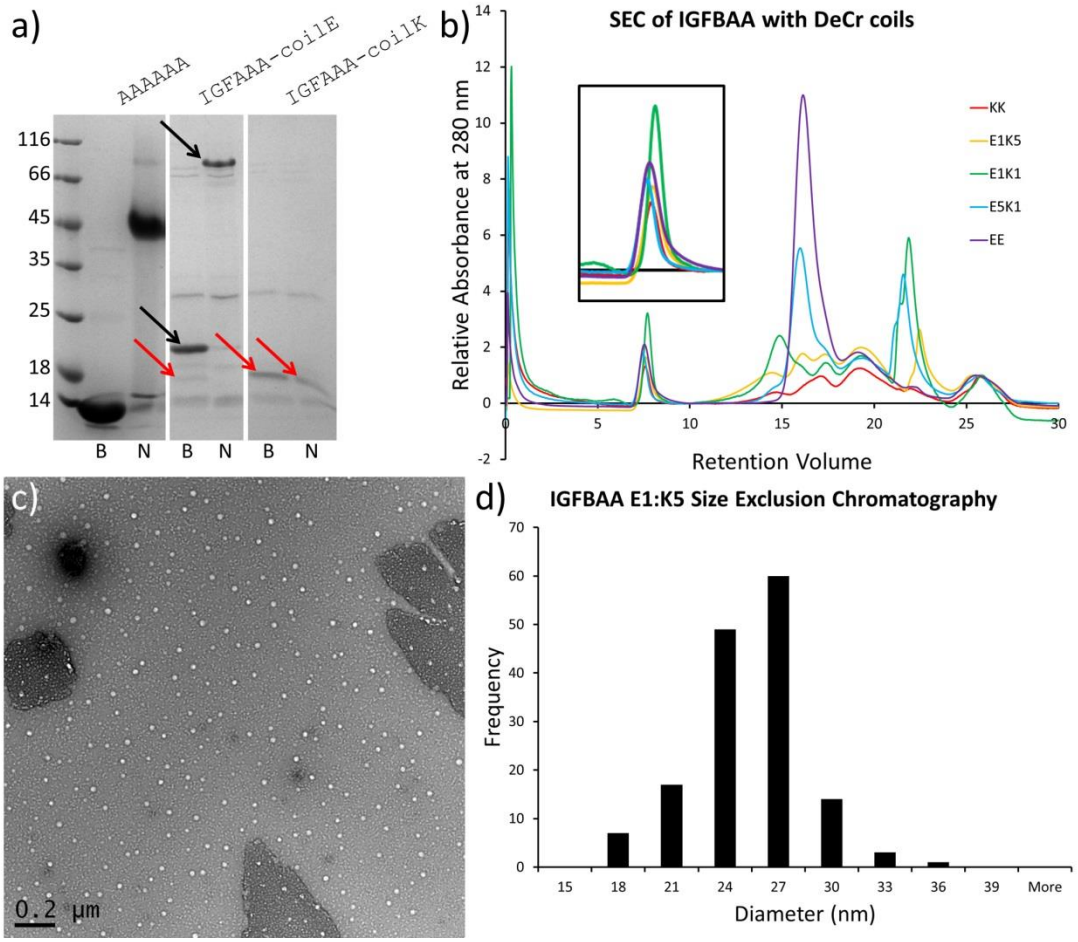



**Figure 6-5:** SDS PAGE, SEC and TEM of the mutant IGFBAA. **a)** Mutants are noted at the top of the gel and boiled samples (B) and non-boiled (N) samples are noted below. The coil-E variant expressed well however the coil-K variant expresses poorly. **b)** SEC of various ratios of IGFBAA-coilE and IGFBAA-coilK are indicated in the key: KK = pure K, E1K5 = 1:5 ratio of E:K, etc. The 1:1 ratio of coil-E to coil-K variants gives the largest yield of high weight product. The traces were normalised using the peak at 26 ml retention volume. **c)** A representative micrograph of capsid populated areas of the IGFBAA-E1K1 grid. **d)** The E1K1 mix when analysed by TEM revealed a number of capsid-like structures with an average diameter of 27 nm; this figure shows the particle size distribution of these particles.

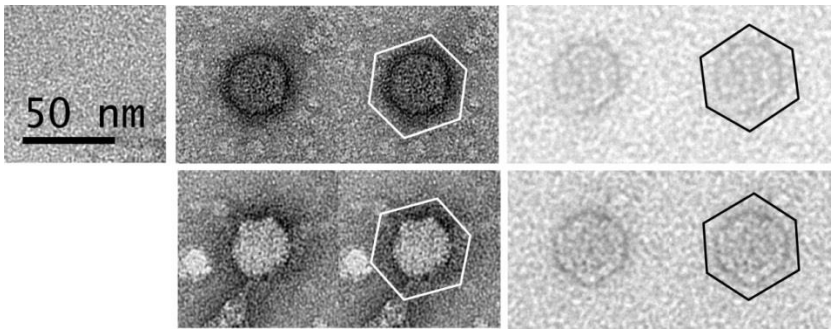



**Figure 6-6:** TEM of IGFBAA E1K1 variants. These particles were found on the same grid as Figure 6-5c. Some of the larger particles found displayed hints of a 3-fold axis of symmetry which is indicative of a closed particle when made from pentagonal subunits.

# 6.3 Gel Contaminants

A few contaminants were present on SDS PAGE after nickel affinity chromatography; it was a point of interest to identify these contaminants. Additionally as some of the SEC traces revealed species of CTB which appeared to elute at a similar retention volume to the non-coiled CTB it was suspected that a degree of proteolysis was occurring during the incubation step of capsid assembly. In order to answer some of these questions an SDS PAGE of IGFBAA-coilE and a mix of IGFBAA-coilE and IGFBAA-coilK were submitted for tryptic MS-MS analysis (Figure 6-7) which was conducted by Dr James Ault, University of Leeds. The methods are available in the appendix as a personal communication from James Ault. The mixture of IGFBAA-coilE/K formed part of the experiment where the positive and negatively charged coiled-coils were mixed together to elucidate if assembly occurred. The expression of IGFBAAA-coilK was very low compared to IGFBAA-coilE so it was also of interest to find the IGFBAA-coilK mutant in the mixed sample.

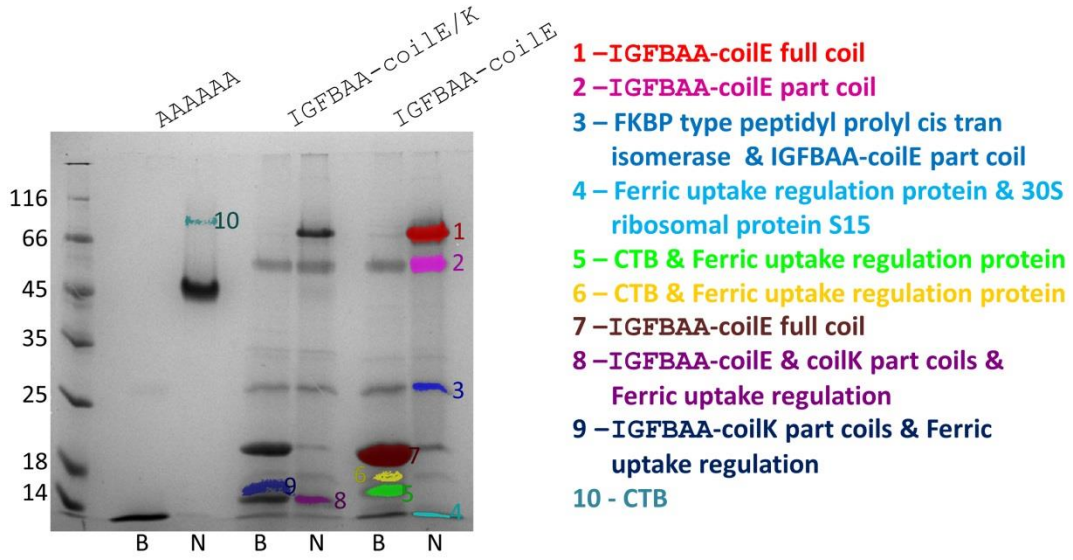



**Figure 6-7:** SDS PAGE trypsin digest followed by MS-MS. Proteins are labelled above the gel and boiled (B) and non-boiled (N) samples are noted below the gel. The IGFBAA-coilE/K is a sample with a 1:1 mix of both IGFBAA-coilE and IGFBAA-coilK. Pixel intensities were re-coloured using GIMP (GNU Image Manipulation Program at various thresholds).

Band 1 reveals IGFBAA and a number of coilE segments which due to the heptad nature of the coiled-coil repeat can account for the entire coil barring the *C*-terminal Gly residue. However in band 7 this final digest product, with a *C*-terminal Gly, is present. As this Gly is unique to the *C*-terminus we can confirm that band 7 contains the full length coiled-coil on CTB which infers that band 1 also contains the full length coil (Table 6-1).

| Band | Sequence |
|---|---|
| IGFBAA-E | TPLNIIDLCAEYHNTQIHTLNSKIFSYTESLAGKREMAIITFPNGATFQVEVPGSQHIDSQKKAIERMKDT |
| Band 1 | LRIAYLTEAKVEKLCVWNNKTPHAIAAISMAGGGEVSALEKEVSALEKEVSALEKEVSALEKEVSALEKG |
| IGFBAA-E | TPLNIIDLCAEYHNTQIHTLNSKIFSYTESLAGKREMAIITFPNGATFQVEVPGSQHIDSQKKAIERMKDT |
| Band 7 | LRIAYLTEAKVEKLCVWNNKTPHAIAAISMAGGGEVSALEKEVSALEKEVSALEKEVSALEKEVSALEKG |

**Table 6-1:** Tryptic MS-MS of IGFBAA-coilE from band 1 and band 7. The sequence in red and green were identified from the MS-MS analysis. The unique part of the sequence between band 1 and band 7 is in green, this C-terminal glycine in band 7 reveals the full length sequence is present.

Band 2 corresponds to IGFBAA but does not contain fragments of the canonical heptad repeat. However, one of the fragments contains the C-terminal portion of IGFBAA followed by one whole heptad of coilE, up to the first Lys of the repeat. In hindsight the band adjacent to band 2, in the boiled sample, should have been submitted as this band is a contaminant seen in many gels and should not contain CTB due to the boiling (this is currently being processed). Band 3 contained the *E. coli* FKBP-type peptidyl-prolyl cis-trans isomerase (SlyD-PPIase), 20 kDa. PPIase contains 196 amino acids, of which 19 are histidines, the sequence includes a region of HVHGAHDHHH DHDHDGCCGG HGHDHGHEHG, thus it is unsurprising that such a protein was captured by the nickel column. Bands 4, 5, 6, 8 and 9 contain the *E. coil* ferric uptake regulation protein, 17 kDa, of which 27% coverage was observed in bands 6, 8 and 9 and 16% coverage in bands 4 and 5, this is odd as this protein appears to run at multiple size on both boiled and non-boiled samples. Bands 5, 6, 8 and 9 also contain CTB. Band 8 also confirms the presence of both IGFBAA-coilE and IGFBAA-coilK (Table 6-2), whereas band 9 confirms only the presence of the coilK mutant. Band 4 also contains the *E. coli* 30S ribosomal protein S15, 10 kDa. Of interest, wild-type CTB is hypothesised to run as a dimer of pentamers and this is confirmed by analysis of band 10 which runs at twice the size of the pentamer, is not found in boiled samples of CTB and contains CTB peptide fragments.

| Band 8 | Sequence |
|---|---|
| IGFBAA-coilK | TPLNIIDLCAEYHNTQIHTLNSKIFSYTESLAGKREMAIITFPNGATFQVEVPGSQHIDSQKKAIERMKDT |
| | LRIAYLTEAKVEKLCVWNNKTPHAIAAISMAGGGKVSALKEKVSALKEKVSALKEKVSALKEKVSALKEG |
| IGFBAA-coilE | TPLNIIDLCAEYHNTQIHTLNSKIFSYTESLAGKREMAIITFPNGATFQVEVPGSQHIDSQKKAIERMKDT |
| | LRIAYLTEAKVEKLCVWNNKTPHAIAAISMAGGGEVSALEKEVSALEKEVSALEKEVSALEKEVSALEKG |

**Table 6-2:** Tryptic MS-MS of band 8. The sequence in red and green were identified from the MS-MS analysis. The unique parts of the sequences in green confirm both proteins are in the mixture. However, as the rest of the coiled-coil is not identified, proteolysis may have occurred.

The three contaminating proteins are interesting as it was initially assumed these would be periplasmic proteins which leach into the media due to an overcrowding of the periplasm from CTB expression. However all three protein are cytosolic in origin, including the SlyD PPIase, of which other PPIase are common in the periplasm. The presence of these cytosolic proteins raises interesting questions regarding CTB periplasmic expression in *E. coil*; was their presence in the media suggestive of cell death? This information suggests that alternative expression methods may help produce higher yields and reduce the level of contaminants. Selected trypsin MS-MS data is available in the appendix.

# 6.4 Further Optimisation via Computational Design (S#1.1)

## 6.4.1 Overview

In this section the IGFBAA mutant from S#1.0, which showed promising high order oligomerisation, was subjected to a further round of the computational routine. The basic routine follows the flow chart in chapter 5, however, combinatorial mutant analysis was only performed retrospectively due to expression limitations and will not be discussed.

### 6.4.1.1 Objective

The chosen hit from S#1.0, IGFBAA, was bought forward to act as the starting model for a second round of computational interface design. This mutant contains the following substitutions A1T, Q3L, T6I, Y18H, D22S, K43P and I47T. It is worth noting that the A1T, Y18H and I47T mutations are biotype variant mutations, thus this mutant actually only contains four *de novo* mutations.

### 6.4.1.2 Saturation Mutagenesis

Saturation mutagenesis was carried out with FoldX as described in the methods on the mutant IGFBAA. This investigation returned 35 substitutions with a $\Delta\Delta G_{stability}$ less than -1.0 kcal/mol.

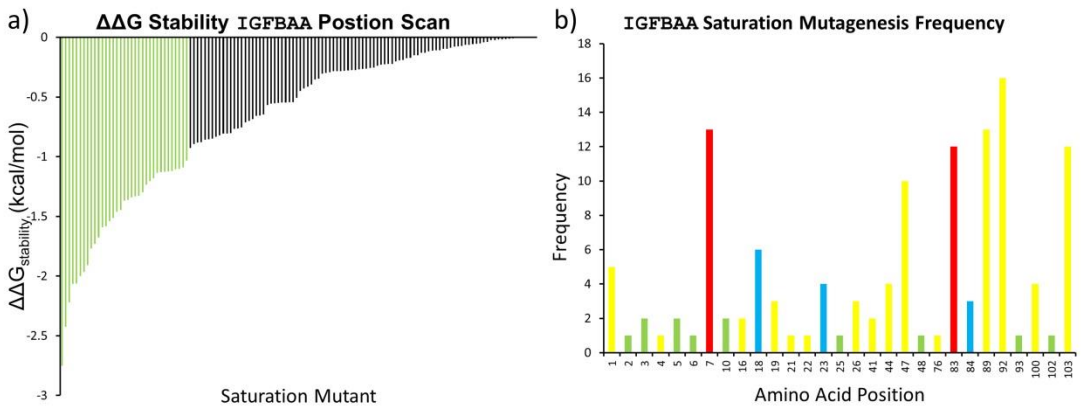



**Figure 6-8:** Results of saturation mutagenesis on the first round mutant IGFBAA. **a)** The position scan revealed a number of mutations which reduced the total $\Delta\Delta G_{stability}$ of the trimer of pentamers below -1.0 kcal/mol, presented in green. **b)** The frequency of suggested mutations by position. Notably only six positions are recommended to be changed more than ten times. The colours represent the wild-type protein residue being hydrophobic (green), hydrogen bonding (yellow) charged (red or blue respectively).

The saturation mutagenesis highlighted positions 7, 47, 83, 89, 92 and 103 as residues which are potentially repulsive to the assessed complex, each with 10 or more suggested substitutions. Of the top ten substitutions, only five positions were suggested for mutation. Position 7 and 47 were best represented in the top ten with four and three suggestions respectively. The mutations with $\Delta\Delta G_{stability}$ less than -1.0 kcal/mol in Figure 6-8 were applied to the CTB-44°-8Å model and the single mutation analysis was performed.

### 6.4.1.3 Single Part Mutants

25 mutations were selected from the 38 predictions from FoldX. Cysteines were removed from the list of mutations as well as the worst performing substitutions at over-populated positions. Residues with hydrogen bonding potential were included preferentially over hydrophobic residues. Ensembles of these mutations were analysed by FoldX as described previously.

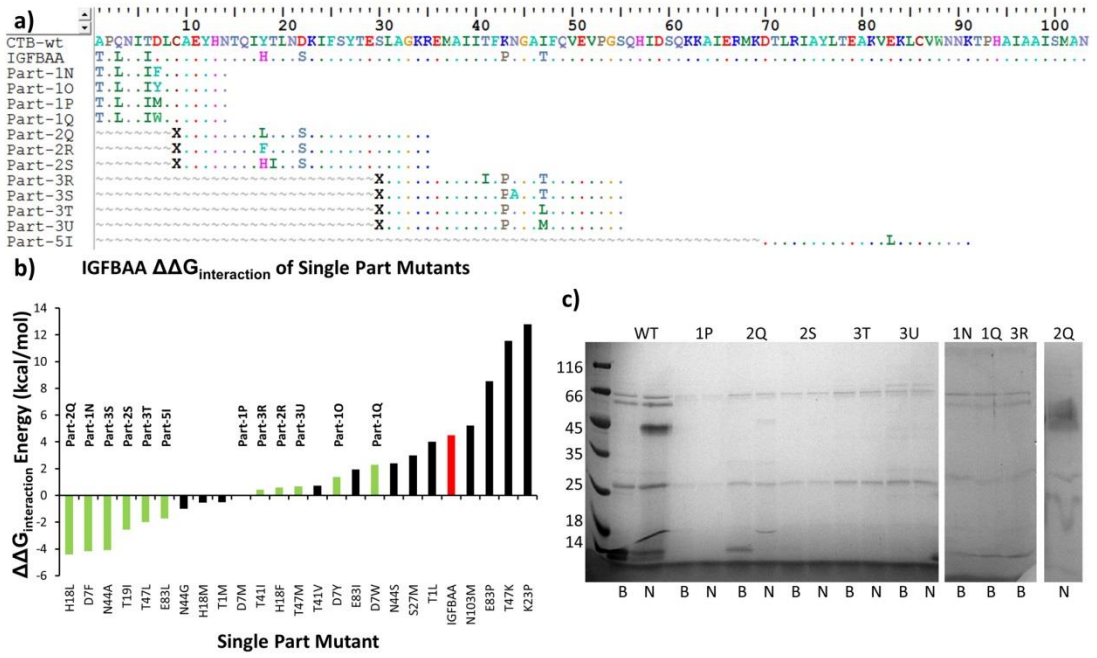



**Figure 6-9:** Second round mutants of the S#1.0 hit IGFBAA. **a)** The amino acid sequence of the wild-type compared to IGFBAA as well as the second round single part mutants. **b)** The $\Delta\Delta G_{interaction}$ energy in kcal/mol for the single part mutants compared to IGFBAA in red. Mutants selected for experimental validation are highlighted in green. **c)** Expression of the selected single part mutants, notably, almost all mutants cause loss of CTB expression, barring part 2Q.

The single mutant analysis suggested the majority of the chosen mutations should increase $\Delta\Delta G_{interaction}$ compared to the IGFBAA first round mutant as seen in Figure 6-91b, where IGFBAA is highlighted in red. Those highlighted in green were investigated experimentally, however, only 1N, 1P, 1Q, 2Q, 2S, 3R, 3T and 3R proceeded through the cloning stage to protein expression. It became evident that only part 2Q expressed to workable concentrations (Figure 6-9c). Perhaps fortuitously mutant IQFBAA (containing part 2Q) was predicted to have the best estimated improvement in interaction energy amongst those mutants assessed. The final lane of the gel in Figure 6-9 shows IQFBAA concentrated to 2mg/m., Although this gel was somewhat smeared, the pentameric band is the dominant band on the gel.

### 6.4.1.4 Combinatorial Mutants

A combinatorial mutant selection was performed *in silico*, however, due to the reduced expression of the single mutants further mutations were deemed too destabilising to experiment with further.

# 6.5 Characterising Second Generation Mutants (S#1.1)

## 6.5.1 Expression and Characterisation of IQFBAA with JR-coil2

Mutant IQFBAA, the sequence of which is shown in Figure 6-10, was expressed and purified as per the standard technique described above.

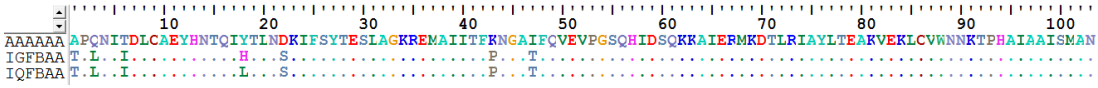



**Figure 6-10:** The amino acid sequence for the single part mutant hit from S#1.1, IQFBAA based on the S#1.0 hit, IGFBAA. The single substitution is H18L. The initial substitution in this position, Y18H, is CTB biotype variant, thus the 18L mutation introduces a novel amino acid at this position.

IQFBAA was then analysed on SDS PAGE and when concentrated, passed through SEC (Figure 6-11). IQFBAA gave very little yield of high weight product compared to the parent IGFBAA-coil2 mutant. As the expression of this mutant was compromised the total yield of contaminants was greater than with the parent mutant, as seen on the SEC trace in Figure 6-11b, where the contaminant peak at 1.2 rrv is much larger for IQFBAA. Due to the decline in production of the high weight peak found by SEC, no further analysis of this mutant was performed.

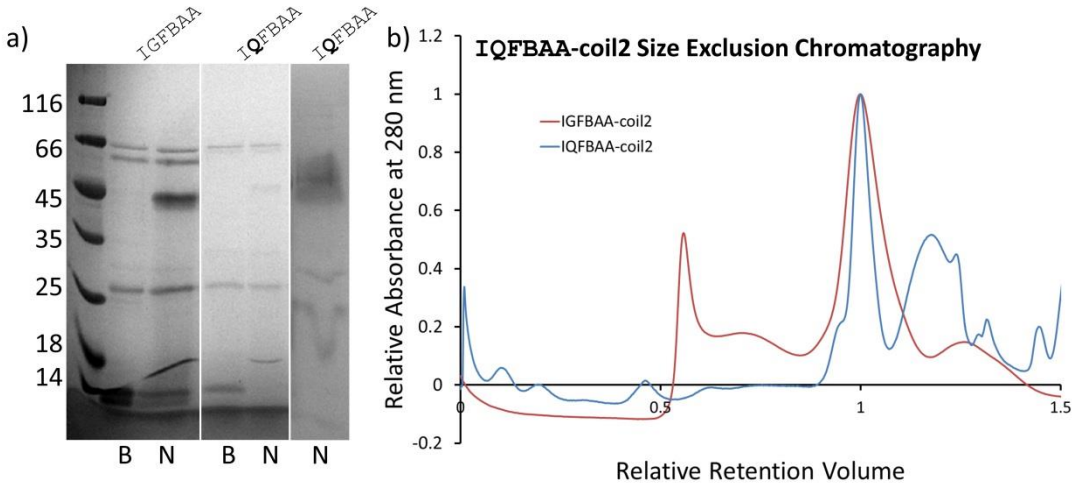



**Figure 6-11:** SDS PAGE and SEC of IQFBAA. **a)** IQFBAA showed a reduced expression compared to its parent mutant IGFBAA, suggesting the H18L is a deleterious mutation. **b)** Additionally 18H appears to be an important residue in the assembly of the capsid as the production of a high weight species of IQFBAA is seriously perturbed.

# 6.6 Concluding Statements

## 6.6.1 Computational Routine

The computational process used to select the mutations has been experimentally validated and shown to change the behaviour of the assembling subunits. The IGFBAA mutant was shown to assemble faster and to give higher yields of VLPs than were observed by SEC for the wild-type protein. Under TEM, the capsids were observed in greater numbers than were found with the wild-type. These results are a direct result of the computational studies conducted which show that our method has the ability to pick mutations which result in increased affinities. However, it should also be noted that the computational method sometimes also makes suggestions that do not express well and mutations which do not increase affinity.

## 6.6.2 Capsid Structures

The capsid structures observed appear similar to those found in the wild-type protein but are more numerous. In addition, a larger, more elliptical capsid was observed as well as a large capsid enclosing two smaller capsids. This observation leads to the conclusion that despite most capsids having a similar size, there is a degree of disordered assembly which gives rise to these larger structures. Furthermore, there must presumably be a degree of self-association between the interior and exterior faces of the capsid to allow encapsulation of small capsids within larger capsids.

## 6.6.3 The de Crescenzo Coiled-Coils

The five heptad coilE and coilK designs described by de Crescenzo were extended from the IGFBAA mutant CTB. These capsid structures assembled rapidly to form particles with an average diameter of 27 nm. They were also found to create structures between 30 and 40 nm, similar to those using the JR-coils. The speed of the assembly and the appearance of these 27 nm capsids could be evidence of a kinetic trap during the assembly process, however as highlighted in the introduction, these structures may be forming structural polymorphs mediated by the higher affinity of the de Crescenzo coils as seen with differing scaffolds and the assembly of SV40[110]

### 6.6.4 Expression Yields

The IGFBAA mutant expressed reasonably well, however, none of the other combinatorial mutants seemed to express well and even the single parts which were combined to make IGFBAA did not express to a reasonable level. As discussed, this may be due to the substitution of the K43 residue, as this may alleviate electrostatic field tension in that area of the protein which could reduce the yield of stable protein. The reduced expression highlights a design concern not addressed by the computational studies. The selection of combinatorial mutants contained four or more substitutions: it may be more feasible to approach the mutagenesis from a more modest starting point.

The next chapter describes mutations which were suggested and assessed by the same computational routine. However, in this case, they were biased by human design to pick interactions which would form on a molecular level according to the proposed orientation of binding discussed in chapter 5.

# Chapter 7

# Further Capsid Mutant Strategies

In this chapter two further design strategies are discussed (S#2 and S#3). A round of computational investigations for S#2.0, is followed by experimental validation of capsid structures and then by a second round of mutations: S#2.1 and S#2.2. Finally a third strategy is discussed which involved no computational assessment: Strategy S#3.0 involved combining expressible mutations from the previous investigations (S#1 and S#2).

# 7.1 Strategy S#2.0

Strategy 1.0 used a degree of human interjection in selecting the initial combinatorial mutants, of which the parts were ultimately recombined in the second iteration of the computational protocol to form the combinatorial library. The decision making was guided solely by $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{interaction}$ estimations, as well as occasionally omitting troublesome residues such as proline or cysteine. In strategy 2.0, however, the decision making was more structurally biased. the literature suggests most protein-protein interfaces maintain a degree of hydrogen bond donating and accepting residues within the interface to aid solubility when uncomplexed. In order to enrich the library of S#2.0 with more hydrogen bonding residues, the saturation mutagenesis data from S#1.0 was directly compared to the structure of the proposed protein-protein interface and mutants were chosen which satisfied both structural constraints of the interface and the $\Delta\Delta G_{stability}$ estimations from FoldX. This approach relied on the symmetrical docking program to have made adequate estimations of the interface orientation. The CTB-44°-8Å model used to initiate the routine in S#1.0 was used again here. Rational mutant substitutions were guided by the FoldX position scan.

#### 7.1.1.1 Starting Structure

In S#2.0 the wild-type-44°-8Å model was used to initiate the routine.

## 7.1.2 Position Scan

The saturation mutagenesis data from strategy S#1.0 was used to suggest mutants, however, in S#2.0 these suggestions were then rationally considered by inspection of the interface structure. For example, the FoldX position scan data recommended `T41` substitution with `M`, `L`, `P`, `V`, `I` or `K` and `T47` with `I`. These are predominantly hydrophobic residues; however upon observation of the CTB-44°-8Å model, these two positions are in close proximity across the interface, which should allow the introduction of interacting hydrophilic residues. The introduction of an asparagine residue in each of these positions should mediate hydrogen bond formation across the putative interface. However, MD

simulations of the mutant (Figure 7-1) show a hydrogen bond between an N47 of one pentamer and an N41 of the next pentamer, this interaction is potentially supported by further intra-pentamer interaction from the uncomplexed N41 and N47 residues. T41N gives a $\Delta\Delta G_{stability}$ improvement of -0.19 kcal/mol from the FoldX estimation, however, this does not account for the potential complimentary interactions including the same residue position interacting across the interface.

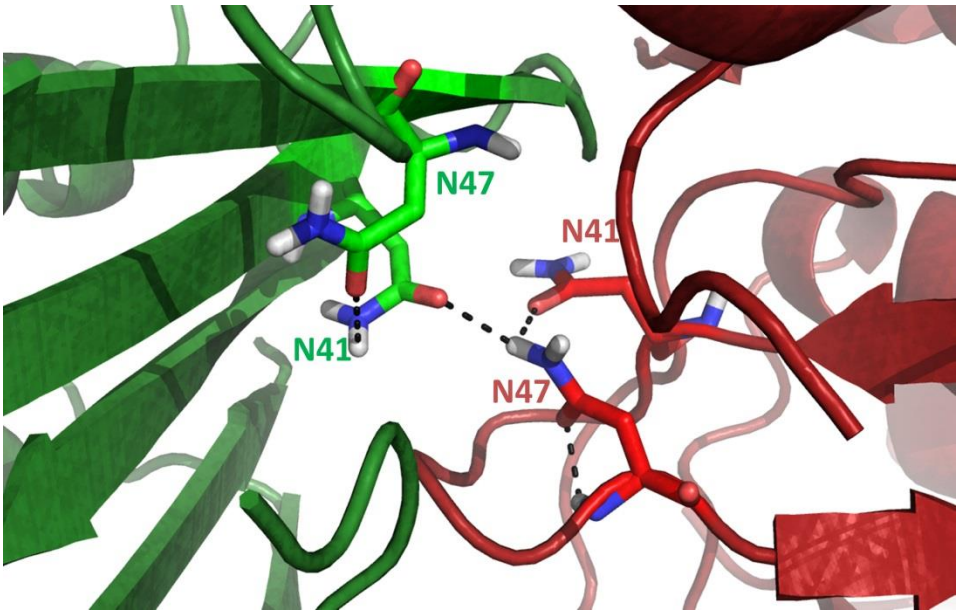



**Figure 7-1:** Example of a rational mutation based on suggestions by FoldX data. The introduction of asparagines at positions 41 and 47 may lead to a hydrogen bonding inter-pentamer interaction. Produced from a modified structure of pdb 3CHB.

The mutations chosen for the screen are shown in Figure 7-2. Of the choices made, eight were hydrophobic residues and ten were hydrogen bonding residues and a further three were charged. This is quite a different ratio of hydrogen bonding residues to hydrophobic residues from S#1.0 which contained 15 hydrophobic residues and only six with the potential to form hydrogen bonds.

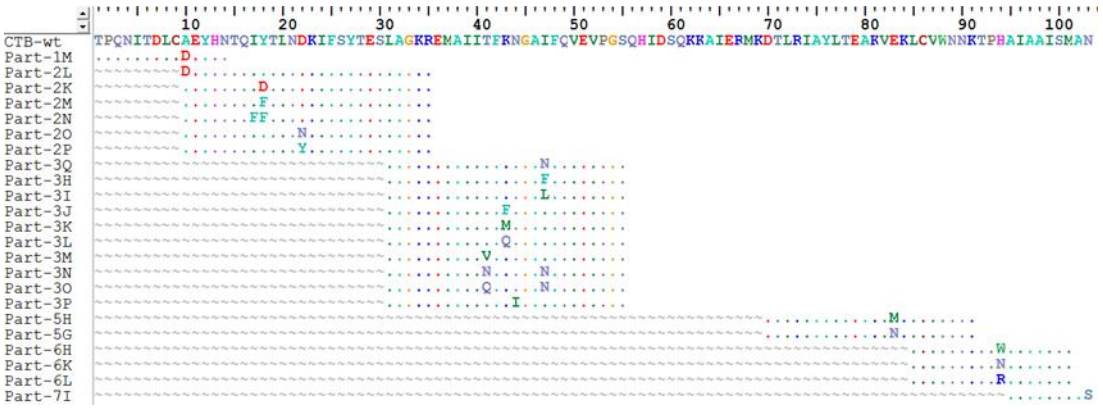



**Figure 7-2:** A sequence alignment of the mutant 'parts' ordered for assembly PCR in S#2.0. A different combination of these parts allows the construction of different full length mutants. These parts allow the construction of 3168 combinatorial mutants.

## 7.1.3 Single Part Mutants

Single mutants were not computationally assessed in this strategy as they were picked by hand; however verification of expression was determined experimentally. The individual parts were assembled with the respective wild-type parts by PCR and cloned into the expression vector pSAB2.2. Part 1M and 2L contain a complimentary mutation in their overlapping region; these parts were assembled together with the remaining wild-type parts. The constructs were expressed, purified and ran on SDS PAGE, examples of which can be seen in Figure 7-3. Mutants 2M, 2O, 2P, 3N, 3P, 4H, 6K, 6L and 6H did not express or expressed poorly. Mutant 3O and 5G produced large smears between the pentamer and monomer band suggesting pentameric dissociation during SDS PAGE.

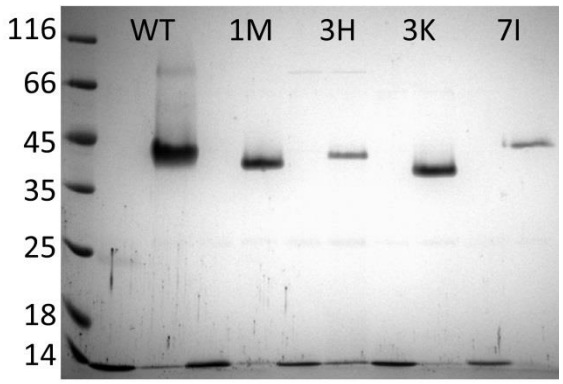



**Figure 7-3:** SDS PAGE of S#2.0 single part mutants. SDS PAGE was conducted on non-boiled samples of the mutant protein. The single mutant parts all display reduced expression compared to the wild-type CTB.

## 7.1.4 Combinatorial Mutants

The computational combinatorial mutant analysis was performed in the same manner as described in S#1.0, combining each mutant part in such a way to create 1848 mutants. This analysis was streamlined to include only those containing expressible parts. Of these potentially expressible mutants a further selection was made base on the morphology of the single part mutants on the gel. Those highlighted in Figure 7-5b were assembled by assembly PCR, cloned and expressed.

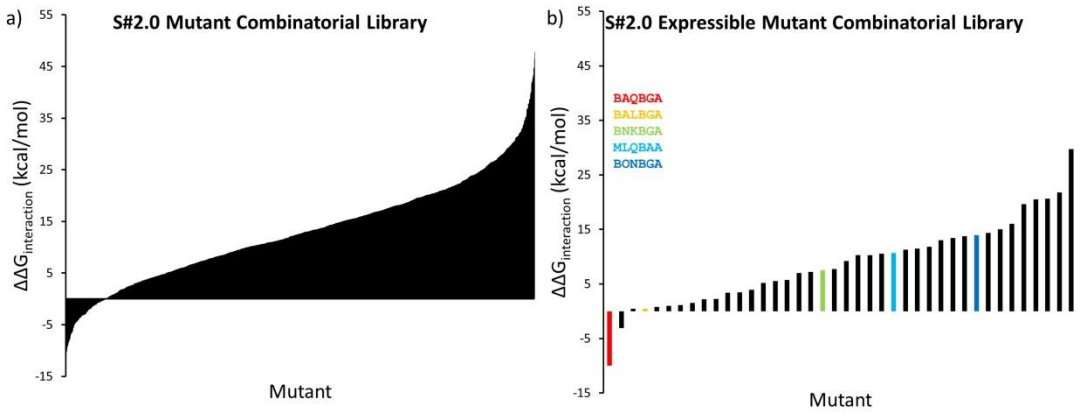



**Figure 7-4:** Combinatorial mutant selection. **a)** All possible combinatorial mutations assessed for their $\Delta\Delta G_{interaction}$ by FoldX. **b)** Expressible parts combinatorial library assessed for their $\Delta\Delta G_{interaction}$ by FoldX. Those highlighted in colour were experimentally validated.

### 7.1.4.1 Characterisation of Mutants with JR-coil2

Five mutants from the combinatorial computational selection were carried through for protein expression, shown in Figure 7-5. Due to time constraints these mutants were only fused to the JR-coil2 coiled and not to the JR-coil7 variant. These mutants were not analysed by DLS or EM, however mutants with promising SEC high weight peaks were passed on for a second round of computational design and analysis.

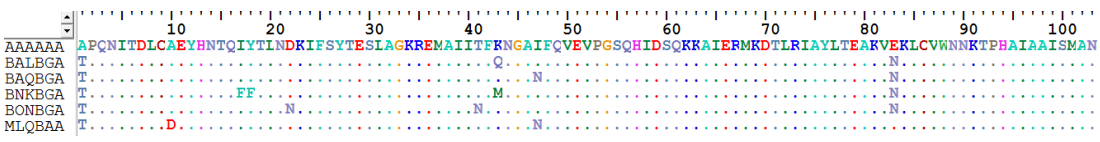



**Figure 7-5:** The amino acid sequence for combinatorial mutants from S#2.0. The expressible mutants from this round of single mutants were combined. Only 5 were chosen for extension with coiled-coils.

Each of the mutant proteins was produced as described for S#1.0 mutants. Comparatively, this set of mutants contained fewer substitutions than mutants in S#1.0 and as a result three of the five tested mutants maintained their pentameric stability on SDS PAGE, as seen in Figure 7-6a. However it is notable that the total protein yield was somewhat perturbed by these substitutions. Predictably, as these mutants had a reduced number of substitutions, they assembled into higher order species at a slower rate compared to the S#1.0 mutants (Figure 7-6b). Nevertheless, one mutant, BAQBGA, showed a reasonable amount of high order species when analysed by SEC after 10 days incubation at room temperature. Additionally, the mutant MLQBAA showed an increase in the total high weight product after 10 days when compared to 4 days incubation. This mutant also displayed what is thought to be monomeric CTB at 22 ml retention volume. This is of interest as it is suggestive of pentameric instability. However the ratio of this monomeric peak compared to pentameric peak decreases over prolonged incubation which is suggestive of either pentameric reassembly or of the production of some high weight amorphous aggregation.

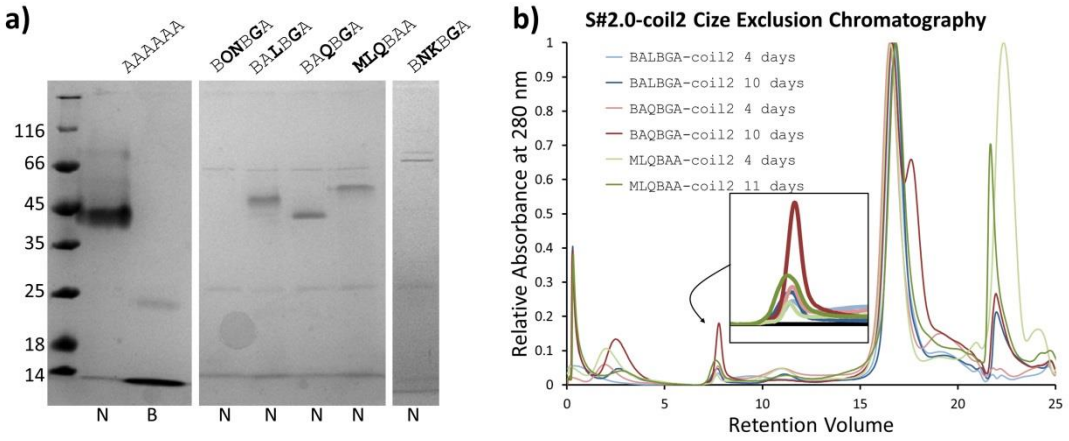



**Figure 7-6:** SDS PAGE and SEC for the S#2.0 combinatorial mutants. **a)** Mutants are listed at the top of the gel with the wild-type AAAAAA, no coils, as the reference control, boiled samples (B), non-boiled samples (N). Of the five combinatorial mutants selected for expression with the JR-coil2, BALBGA, BAQBGA and MLQBAA revealed viable pentamers on SDS PAGE and were at higher concentration than the contaminants. **b)** Each of the expressible mutants was incubated for either 4 or 10-11 days before SEC was conducted. BALBGA, in blue, displayed the lowest rate of assembly despite there being an increase in high weight species over the prolonged incubation. MLQBAA, in green, showed a greater increase in assembly and greater yield than BALBGA. However, this mutant also displayed a peak at the CTB monomer retention volume of 22 ml. Oddly this peak reduces given longer incubations. BAQBGA (red) displays the greatest increase in high weight product over time. However, the right-hand shoulder of the pentamer peak suggests that proteolysis has occurred resulting in the cleavage of some portion of the coiled-coil. Additionally, as seen with all of the prolonged incubations there is a peak for monomeric CTB.

BAQBGA and MLQBAA both showed an increase in the high weight species over time (although the increase of the MLQBAA peak was marginal). Thus these two mutants were selected for a second pass through the computational routine to select additional mutations for analysis. Optimisations S#2.1 and S#2.2 used the MLQBAA and BAQBGA sequences, respectively as the starting structures. Interestingly the mutant BAQBGA was observed to give the greatest increase in capsid assembly, this mutant was also found to give the best $\Delta\Delta G_{interaction}$ energy of -10.0 kcal/mol by the computational analysis. Before assessing these mutants by DLS and TEM it was decided to put the mutants through a second round of computational design to attempt to increase the yield of high weight product.

# 7.2 Strategy 2.0, Iteration 0.1 (S#2.1)

The mutant MLQBAA, which contains the substitutions A10D and T47N compared to the wild-type, was found to increase the assembly of high weight products by SEC and form capsids under TEM. In this iteration of the computational strategy, the mutant was assessed for additional mutations likely to increase the $\Delta\Delta G_{interaction}$ for the proposed interface. As the total stability and expression of the single mutants was impaired, the combinatorial library analysis was not used.

## 7.2.1 Position Scan

The position scan revealed a number of mutations which improved the $\Delta G_{stability}$ compared to the wild-type residue. The top seven $\Delta\Delta G_{stability}$ hits were at position 83, with E83G providing the best energy reduction of -4.6 kcal/mol. MLQBAA does not contain a mutation at either D22 or at E83, discussed earlier as possible electrostatically repulsive residues against capsid formation, so it is unsurprising that these predictions follow the same trend as the wild-type position scan. The remaining hits below -1.0 kcal/mol populated positions 3, 6, 22, 47, 83 and 94 which is similar to the wild-type position scan, albeit that the wild-type scan also included suggestions at position 18, 28 and 41, which were under-represented in the MLQBAA position scan. The 1M/2L parts of MLQBAA corresponds to an A10D mutation which appears in the mutation frequency distribution graph (Figure 7-7b). Despite there being ten energy-minimising mutations at this position, the best improvement for a D10 substitution is ranked 51$^{st}$ against other possible mutations (a reduction of -0.89 kcal/mol). Re-substitutions of D10 were thus not considered in this study. The 3Q part of MLQBAA corresponds to an I47N substitution. The position scan

predicted that N47P would give a -2.56 kcal/mol reduction in $\Delta\Delta G_{stability}$. Another four, mainly hydrophobic substitutions, were also suggested for this position. However, the cavity at the interface into which this side chain protrudes has a number of backbone hydrogen bond donors/accepters, thus the original asparagine substitution was retained at this position.

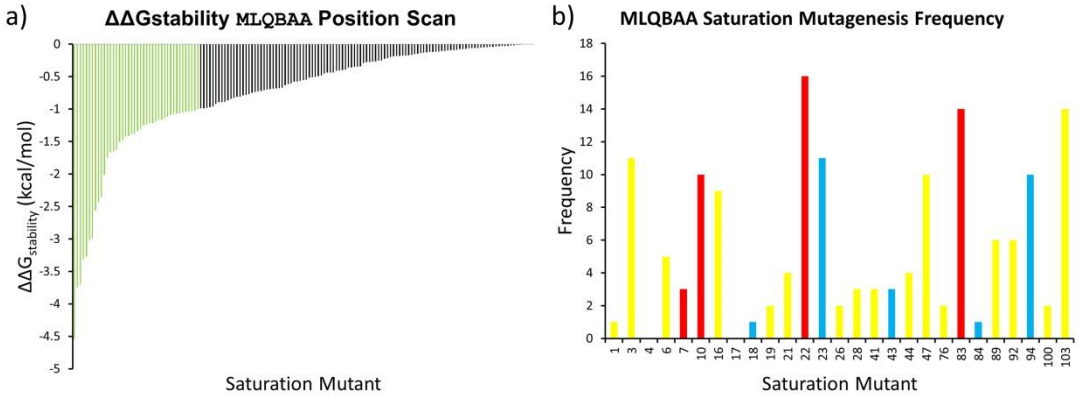



**Figure 7-7:** Results of saturation mutagenesis on the first round mutant MLQBAA. **a)** The position scan suggested a number of mutations which should give a $\Delta\Delta G_{stability}$ below -1.0 kcal/mol for the trimer of pentamers (presented in green). **b)** The frequency of suggested mutations by position. Notably eight positions are recommended to be changed more than ten times, one of which was the original first round mutation A10D.

## 7.2.2 Single Mutants of MLQBAA

Selected mutations predicted from the position scan were modelled, minimised and CHARMM MD trajectories were performed. Of these, most gave a $\Delta\Delta G_{interaction}$ energy improvement over the initial mutant MLQBAA. The position scan $\Delta\Delta G_{stability}$ values suggested that the best energy improvements would come from the E83 substitutions, however, as seen in Figure 7-8b, the best improvements in $\Delta\Delta G_{interaction}$ were predicted to arise from D22 substitutions. The parts highlighted in green on Figure 7-8 were ordered, constructed and expressed. As is evident from the example gel in Figure 7-8c there was a very low yield of protein from the few constructs which expressed. It should be noted that the only mutants in this round that did express were parts 1U, 1V and 5G (5G is part previously used in S#2.0, but was re-predicted in this round of mutagenesis).

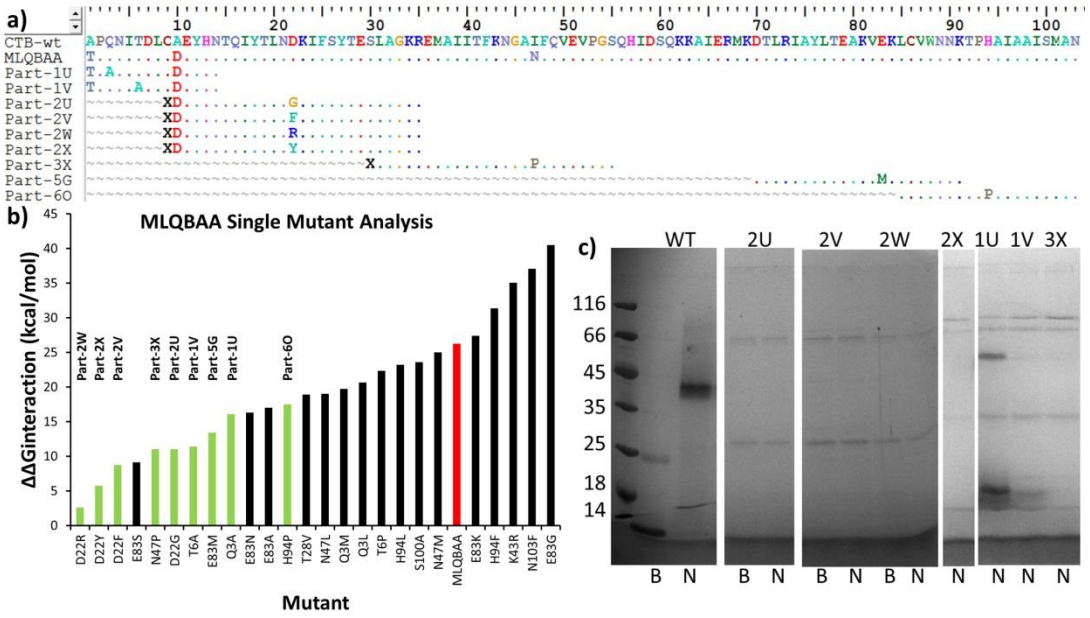



**Figure 7-8:** Second round mutants of the hit MLQBAA. **a)** The amino acid sequence of the wild-type compared to MLQBAA as well as the second round single part mutants. **b)** The $\Delta\Delta G_{interaction}$ energy in kcal/mol for the single part mutants compared to MLQBAA in red. Mutants selected for experimental validation are highlighted in green. **c)** Expression of the single part mutants. It is notable that almost all mutants cause loss of CTB expression, barring parts 1U and 1V.

## 7.2.3 Capsid Formation for Mutants from Strategy S#2.1

Sequences for the second round derivatives of mutant MLQBAA, ULQBAA and VLQBAA, are displayed in Figure 7-9. These mutant genes were assembled with the JR-coil2 sequence at their C-termini (see chapter 4 or the appendix).

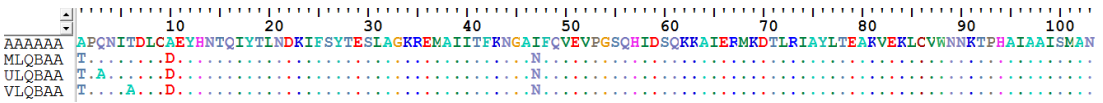



**Figure 7-9:** The amino acid sequence of single part mutants ULQBAA and VLQBAA, derived from the parent mutant MLQBAA from S#2.0. Each mutation is an alanine substation of 3Q or 6T respectively.

### 7.2.3.1 Characterisation of mutants with JR-coil2

Figure 7-10a shows the SDS PAGE from the MLQBAA mutants, ULQBAA and VLQBAA. ULQBAA, containing a Q3A substitution, expressed to form pentamers stable under SDS PAGE however, VLQBAA, containing a T6A mutation, dissociated into monomers during SDS PAGE. Interestingly, from a qualitative perspective, ULQBAA seemed to give a more intense band by SDS PAGE than MLQBAA (Figure 7-6). MLQBAA was previously noted to contain monomers in solution when studied by SEC. In contrast, when ULQBAA was analysed by SEC (Figure 7-10b), there was a distinct lack of the monomer peak. This observation suggests that the Q3A daughter mutation rescues the instability of the parent mutant. The ULQBAA mutant also provided an increased yield of protein compared to the parent MLQBAA mutant. ULQBAA, however, showed little increase in the high weight peak compared to the parent mutant. Unfortunately, a SEC trace for VLQBAA is unavailable due to the corruption of the output file during a fault with the AKTA HPLC. This SEC trace had revealed a modest increase in the high weight peak for VLQBAA and hence VLQBAA was selected for TEM over ULQBAA. The TEM of VLQBAA produced a limited yield of capsid structures similar to the quantity produced for the wild-type protein. The morphology of these structures differed substantially, some of which were barely viable rings, which could be a staining issue or artefacts of collapsed capsids. Other particles were solid structures which showed no internal detail and a third set resembled the wild-type structure in which the contrast agent has penetrated the capsid. The capsid in Figure 7-10c(7) and (8) are shown again in Figure 7-10d. Figure 7-10d(8) seems to display a hexagonal angles in its lower half. This hexagonal shape is also observed in Figure 7-10d(7) and (7.1), suggesting a 3-fold symmetry axis. As the mutant VLQBAA showed no real improvement on the yield of capsids compared to the wild-type CTB-coil2, this mutant was not investigated with the JR-coil7 coiled-coil.

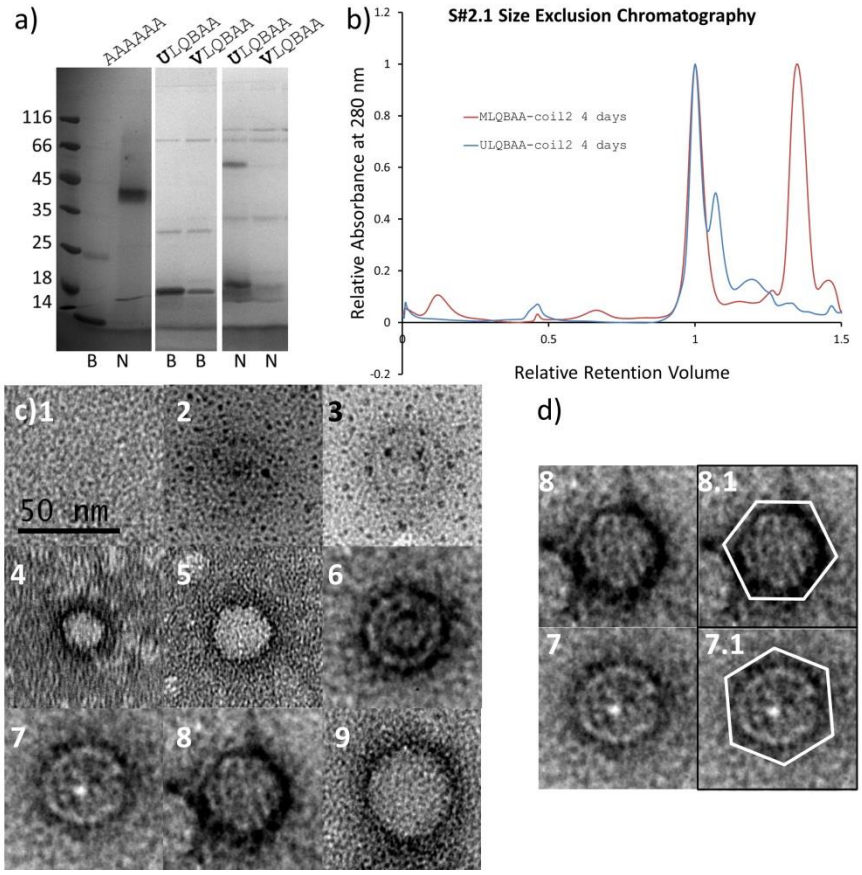



**Figure 7-10:** SDS PAGE, SEC and TEM of the S#2.1 mutants, ULQBAA and VLQBAA. **a)** Mutants are identified above the gel, boiled sample (B) and non-boiled samples (N) are labelled below the gel. ULQBAA, a Q3A mutation of MLQBAA expresses better than VLQBAA and maintains its pentameric structure on the gel. **b)** Comparison of the SEC trace of ULQBAA with the parent mutant MLQBAA reveals a very modest enhancement of high weight species However, the right-hand shoulder of the pentamer peak also suggests a degree of proteolysis of the coiled-coils. ULQBAA show no monomeric peak, suggesting the mutation has recovered a degree of stability compared to the parent. **c)** TEM of the high weight peak from VLQBAA SEC. This mutant revealed only a few capsids under TEM and these were polydisperse in size and polymorphic. However, some of the structures revealed interior walls and some seemed to possess 3-fold symmetry.

# 7.3 Strategy 2.0, Iteration 0.2 (S#2.2)

The mutant which gave the best increase in high weight product from S#2.0 was BAQBGA which contained the mutations A1T, I47N and E83N. This mutant was passed through the computational routine for a second time in order to elucidate substitutions which will increase the rate and total stability of the assembly. As the total stability and expression of the single mutants was impaired, the combinatorial library analysis was not used.

## 7.3.1 Position Scan

The position scan performed on the BAQBGA mutant suggested 32 point mutations which reduced the $\Delta\Delta G_{stability}$ by more than -1.0 kcal/mol (Figure 7-11a). Six of these suggestions were D22 substitutions. No other position was suggested more than three times and these suggestions were spread across positions 3, 6, 7, 16, 44, 47, 84, 89 and 94 (Figure 7-11b). The top ranking point mutants were N44G (-3.1 kcal/mol), K84L (-2.5 kcal/mol) and Q3L (-2.2 kcal/mol). The wider distribution of suggestions for positions to mutate could be the result of the substitution of E83 in round S#2.0 which was thought to provide repulsive interactions across the putative interface. As before, models of the hits from the position scan were constructed and passed to CHARMM for molecular dynamic simulations.

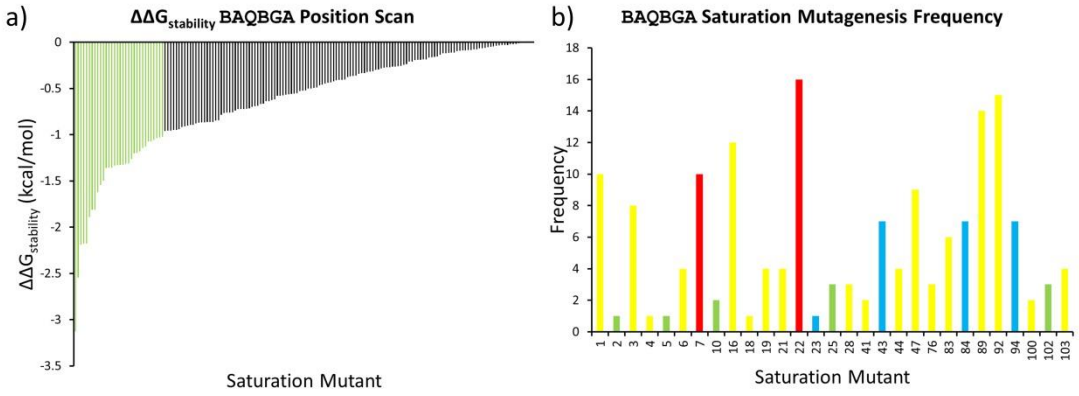



**Figure 7-11:** Results of saturation mutagenesis on the first round mutant BAQBGA. **a)** The position scan revealed a number of mutations which reduced the total $\Delta\Delta G_{stability}$ of the trimer of pentamers below -1.0 kcal/mol, presented in green. **b)** The frequency of suggested mutations by position. Notably just six positions are recommended to be changed more than ten times, highest ranking is position D22, interestingly position 83 is no longer recommended to be mutated since its substitution to Asn.

## 7.3.2 Single Mutants of BAQBGA

Conformations pulled from CHARMM trajectories of the point mutants were analysed by the FoldX 'analyse complex' function to give a $\Delta\Delta G_{interaction}$ for the interface. The best of the single part mutations was Q3L (part-1R, Figure 7-12b). The Q3L substitution had previously been proposed and implemented in S#1.0, however, in S#1.0 this mutation was accompanied by a T6I substitution (part-1I). T6I mutation was also suggested among the hits from this investigation (part-1S), therefore part-1I is a combination of the two parts suggested here, 1R and 1S. Interestingly expression of part-1R with the BAQBGA (creating mutant RAQBGA) resulted in a loss of pentameric CTB seen under SDS PAGE,

Figure 7-12c, as previously discussed this is indicative of a loss of stability. This dissociation during SDS PAGE is not seen for the analogous mutant containing the T6I substitution (part-1I). Again, few of these second round mutants were found to express efficiently; only parts 1R, 1S, 3V and 3W produced protein and only the latter two remained as pentamers on SDS PAGE. Additionally, part-5E is a part previously selected in S#1.0.

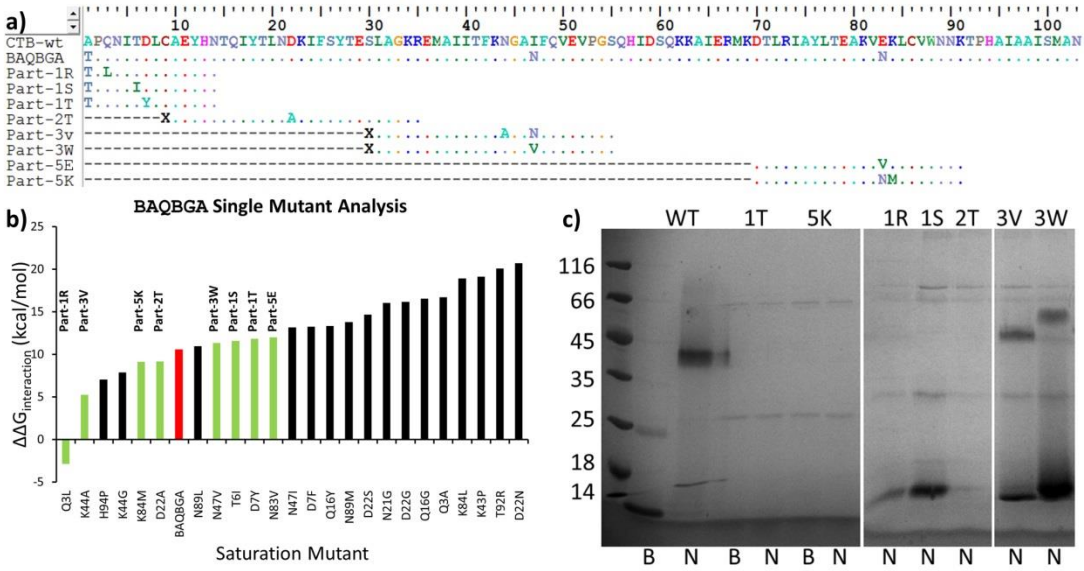



**Figure 7-12:** Second round mutants of the hit BAQBGA. **a)** The amino acid sequence of the wild-type compared to BAQBGA as well as the second round single part mutants. **b)** The $\Delta\Delta G_{interaction}$ energy in kcal/mol for the single part mutants compared to BAQBGA in red. Mutants selected for experimental validation are highlighted in green. **c)** Expression of the experimentally trialled single part mutants, notably, almost all mutants cause loss of CTB expression, barring part 1R, 1S, 3V and 3W.

## 7.3.3 Capsid Formation for Mutants from Strategy S#2.2

Figure 7-13 shows the mutant choices from the computational investigation that could be expressed. Those which did not express are not shown. Part-3W and part-5E are also interesting as these substitute out the 'rational' mutations that were introduced in S#2.0, replacing these hydrogen bonding residues with hydrophobic side-chains. The mutants in the next section were single part mutants expressed with JR-coil2 extensions. Additionally, three hand chosen mutants were picked to be expressed with JR-coil7, which follow the JR-coil2 mutants.

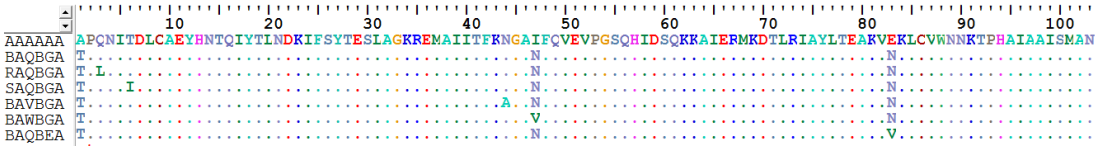



**Figure 7-13:** The amino acid sequence of single part mutants from S#2.2 single part computational assessment on the first round hit BAQBGA from S#2.0. Interestingly part-1R and part-1S contain single mutations, which if combined make part-1I, seen in S#1.0 and S#1.1. Also part-5E, seen in S1.0, was reselected.

### 7.3.3.1 Characterisation of mutants with JR-coil2

Each mutant gene was constructed by assembly PCR with the 3' JR-coil2 sequence. The Figure 7-14a shows a gel for the expressible point mutations, based on the parent sequence BAQBGA. Notably only the N44A and N47V mutants (BAVBGA and BAWBGA) form pentamers under SDS PAGE. Both of these mutants also increase the total yield of CTB produced compared to the parent mutant. BAWBGA also presents a slight smear across the gel suggesting this mutant is more destabilising than the parent mutant.

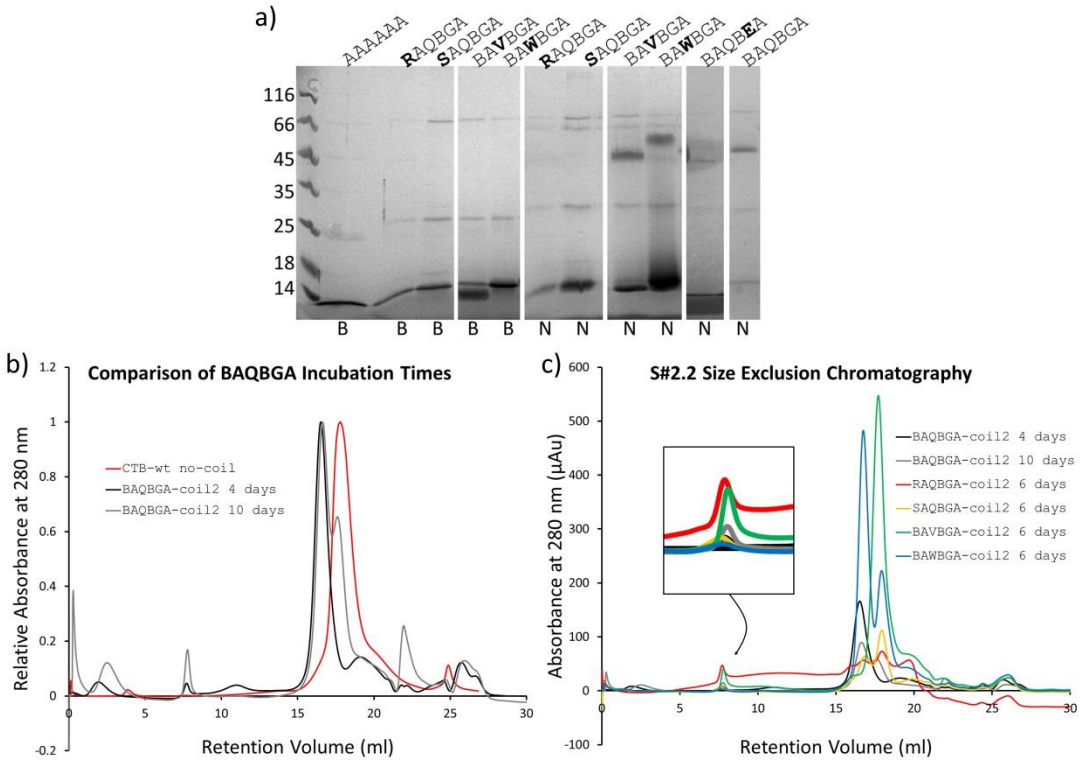



**Figure 7-14:** SDS PAGE and SEC of the S#2.2 mutants bearing the JR-coil2 sequence at their C-termini. **a)** Mutants are noted above the gel. Whether the lanes contain boiled samples (B) or non-boiled samples (N) is noted below. For reference, the parent mutant BAQBGA is in the last lane, and the wild-type CTB is in the first. Both of the part-1 mutants seem to perturb expression; however, both of the part-3 mutations seem to recover a degree of expression. **b)** A comparison of the parent mutant to the wild-type CTB without coils. notice the right-hand shoulder seems to have the same retention volume as CTB without coils, suggesting proteolysis of the coils. **c)** This putative proteolysis is evident in some of the mutants in S#2.2. Two mutants seem to give an improved rate of high weight assembly, RAQBGA and BAVBGA, however RAQBGA expresses poorly and BAVBGA seems to lose its coiled-coil extensions rapidly in the non-assembled pentamer peak.

An interesting observation is that the parent construct for this set of mutants, BAQBGA-coil2, develops a right hand shoulder in the SEC trace after incubation for extended periods, such as 10 days (Figure 7-14b). This additional SEC peak for BAQBGA seems to have a retention volume similar to the wild-type CTB without coiled-coil extensions which could be suggestive some proteolysis of the coiled-coils. Given this hypothesis, it is logical that assembled capsids would be sterically protected from proteolysis due to the coiled-coils being present on the interior of the capsid. Figure 7-14c shows the rest of the SEC traces from mutants in the S#2.2 set. These traces show a disappointing level of assembly and a number of mutants, BAVBGA, BAWBGA and SAQBGA display SEC traces in line with pentamers that do not have C-terminal extensions. Despite this potential proteolytic cleavage, assembly of BAVBGA and RAQBGA appeared to increase compared to other mutants, particularly RAQBGA, for which the high weigh peak constitutes 10% of the total CTB yield (considering the high weight peak and both pentamers with and without coiled-coils). BAVBGA displays a 4% total yield of high weight species.

Two of these mutants were selected for TEM: RAQBGA and BAVBGA. Samples were recovered from the SEC high weight and prepared for TEM as described in the methods.

#### 7.3.3.1.1 RAQBGA Transmission Electron Microscopy

RAQBGA expressed reasonably poorly and was unable to maintain pentamers during SDS PAGE, however when concentrated to 2 mg/ml total protein, incubated for six days and subjected to SEC, the ratio of high weigh peak to pentamer was greater than with other mutants. The total yield of the high weight peak, however, is not so different from other successful mutants, such as the parent BAQBGA. The high weight peak was isolated and applied to an electron microscopy grid and stained as described in the methods. Figure 7-15 shows the particles found under TEM. Larger particles are shown in Figure 7-15(2-15). These particles ranged in size from 45 nm to 105 nm. These larger particles all seem to possess electron shielding particles in the middle of the structures, this maybe an artefact of the TEM process, but these dense black particles do not appear apart from on the inside of the particles. These artefacts are not found in any other TEM of CTB capsids, which may be suggestive of a discrepancy in the preparation of the grid. The small electron shielding particles are pure black as far as pixel intensity is concerned, haloed with rings of pure white. It is possible that these black areas could be quantum dots of metal aggregates. The serendipitous presence of these particles apparently encapsulated is encouraging for the prospect of future targeting of molecules to the capsid interior. Figure 7-15(16-17) show

typical populated areas of the TEM grid, where these larger round structures are dispersed among many smaller capsids which range in size from 18-33 nm. Figure 7-16, displays a histogram of the capsid sizes found on the EM grid. The larger round structures have sizes that are too polydiverse and too few in number to make any statistical claims; however, the smaller particle size range seems to fit a Gaussian distribution.

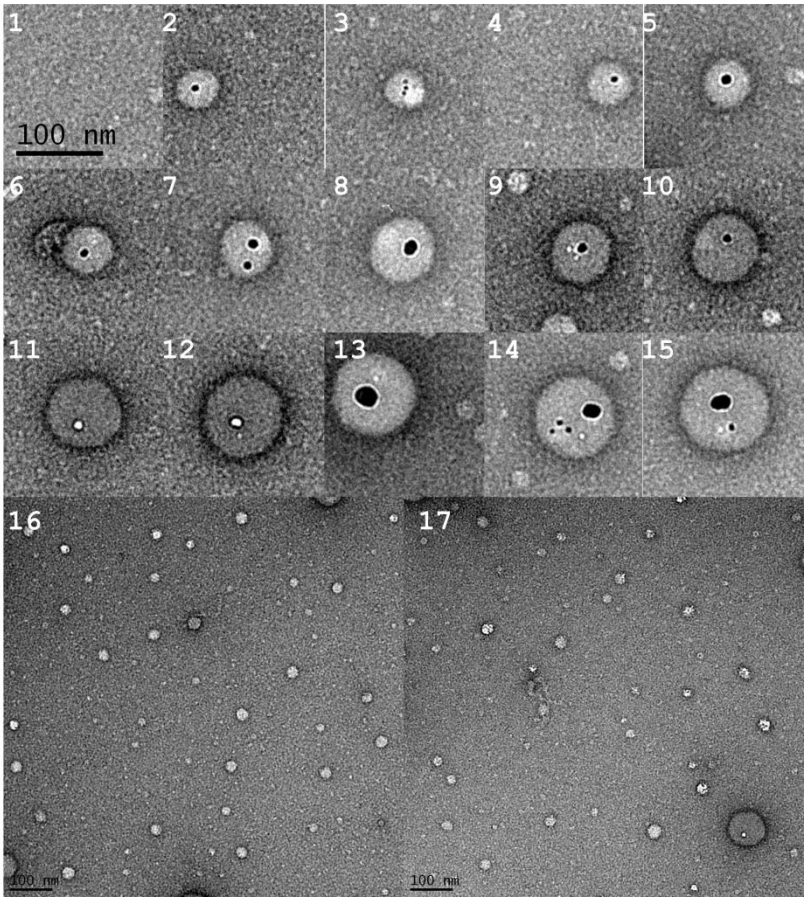



**Figure 7-15:** TEM of mutant RAQBGA. Micrographs **16** and **17** are representative of populated areas of the grid and contain objects of approximately 25 nm in size across a range of 20-30 nm. However, as seen in the bottom right corner of **17**, larger spherical structures are also produced. Micrographs **2-15** show the range of sizes exhibited by these larger round structures. Strangely, even though larger capsids are seen in other mutants, the large structures in this sample possess electronically-shielded structures, presumably in the interior, as these are only present within the confines of larger round structures. This may be due to an inconsistency in sample preparation.

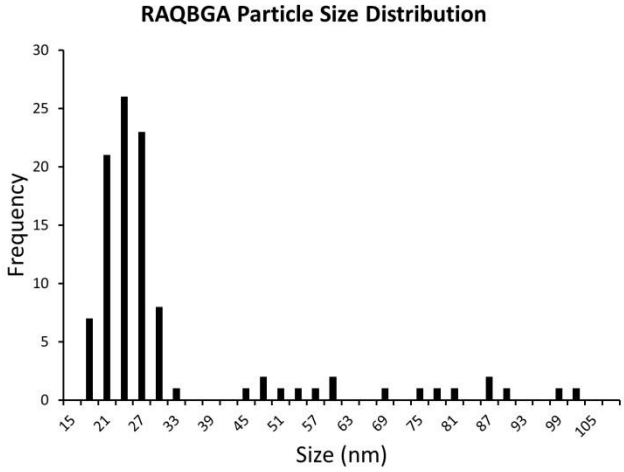



**Figure 7-16:** Particle size distribution for the mutant RAQBGA. The smaller particles seem to fit a Gaussian distribution, however there are too few of the larger particles to make an estimate to size groupings.

#### 7.3.3.1.2 BAVBGA Transmission Electron Microscopy

BAVBGA produced a small but significant high weight peak by SEC. TEM of this high weight peak revealed a number of capsid structures of differing sizes. Compared to RAQBGA, BAVBGA produced a large amount of medium sized particles between 40 and 60 nm in size as well as the smaller capsid size of between 20 and 30 nm. Figure 7-17a shows a collection of structures sized between 40 and 85 nm in diameter and Figure 7-17b & c show typical images from the capsid-populated areas of the TEM grid. The histogram in Figure 7-18 shows the distribution of particle size found among the grid. It is notable that the smaller particles appear to have a similar range and frequency as RAQBGA. However, the BAVBGA mutant also seems enriched in the medium-sized particles between 40 and 60 nm, as is evident from the histogram (Figure 7-18). It is of interest that a modest number of mutations at the proposed interface of the capsid subunits can make such dramatic differences in particle size. Compared to RAQBGA, BAVBGA contains a L3Q and a N44A substitution, the result of which is the creation of a mixture of two distinct particles sizes.

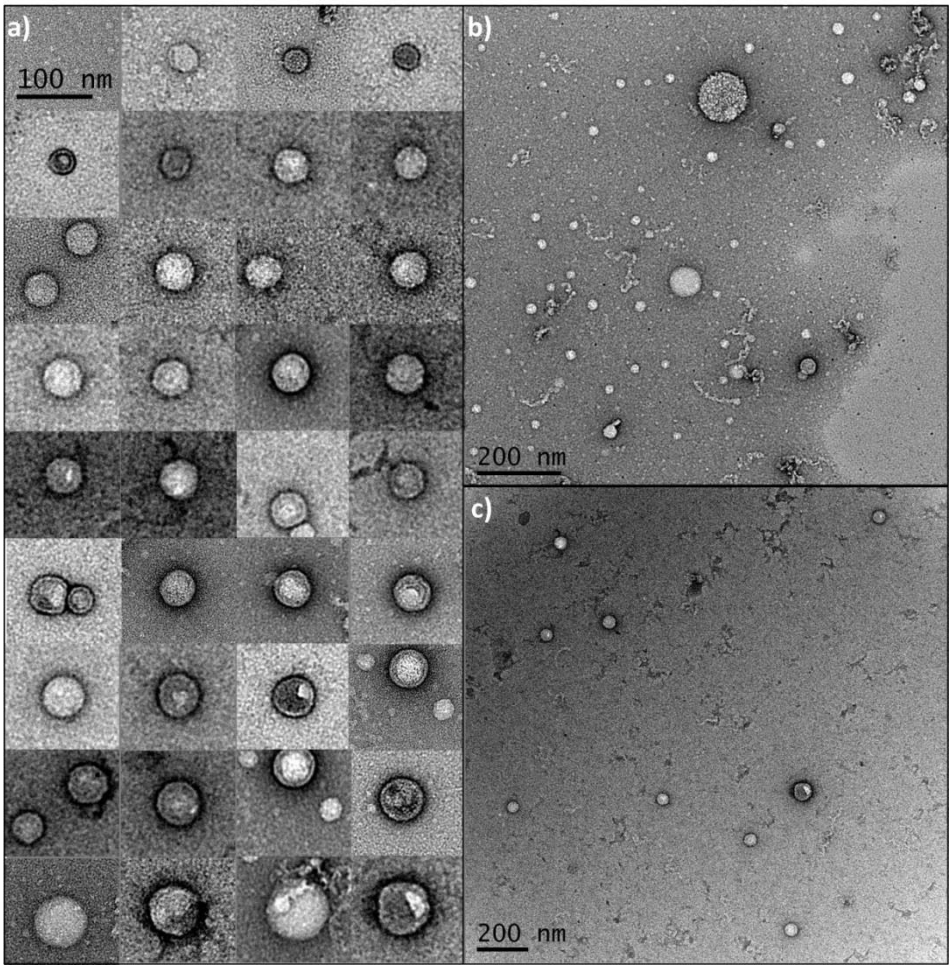



**Figure 7-17:** TEM of mutant BAVBGA. **a)** A selection of capsid structures ranging in size from 20 to 55 nm. Perhaps due to inconsistencies in the stain penetration, some capsids appear solid, while in others the interior wall of the capsid is visible. **b) & c)** Representative micrographs of capsid populated areas of the grid.

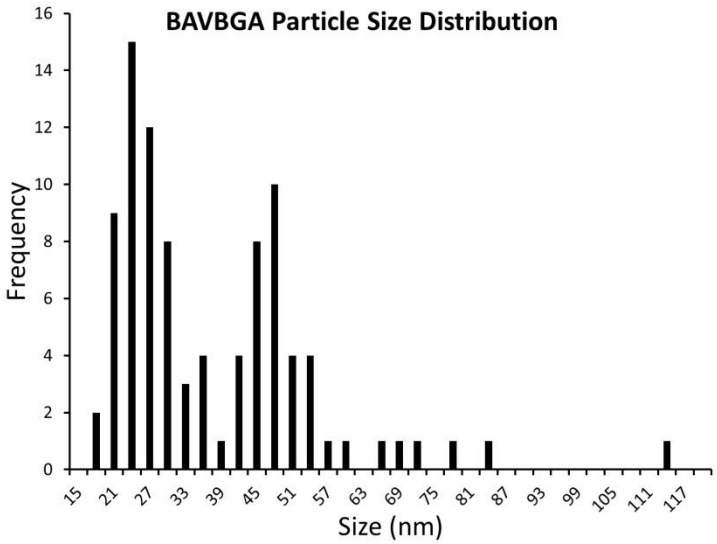



**Figure 7-18:** Particle size distribution for the mutant BAVBGA. This mutant produced two distinct particle sizes, about 25 nm and 47 nm in size. Each distribution appears to fit a Gaussian distribution.

### 7.3.3.2 Characterisation of mutants with JR-coil7

Based on the success of RAQBGA and BAVBGA in producing capsids, these mutations were combined to create the mutant RAVBGA. Furthermore, as part-5E (previously discussed in S#1.0) was also suggested to be beneficial for BAQBGA, thus the mutant RAQBEA was created to test the combinatorial success of this mutant. The sequences are shown below in Figure 7-19.

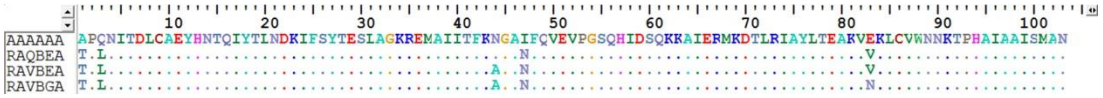



**Figure 7-19:** The amino acid sequence of the S#2.2 combinatorial mutants. Each mutant is based on the success of single part mutants of the parent mutant BAQBGA from S#2.0

Each of these mutants was constructed from scratch via assembly PCR with JR-coil7 extension. Expression of these mutants revealed that RAVBGA and RAQBEA maintained similar levels of expression to BAQBGA however RAVBEA expressed less efficiently. High weight peaks were identified during SEC analysis and so these samples were used to prepare TEM grids (unfortunately, the AKTA data became corrupted and is no longer available).

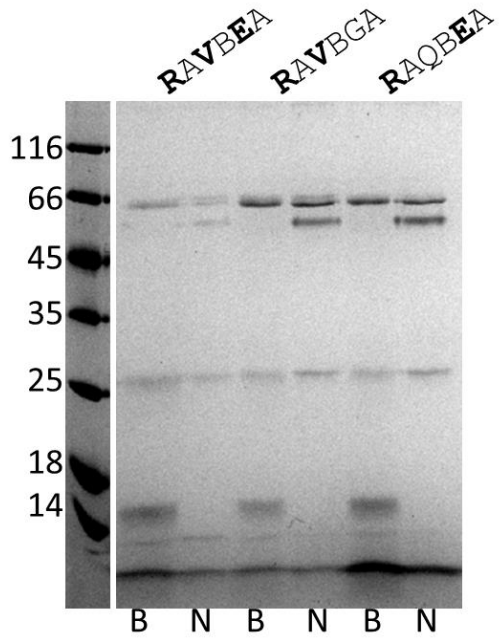



**Figure 7-20:** SDS PAGE of S#2.2 combinatorial mutants. Mutants are noted above the gel, boiled samples (B) and non-boiled samples (N) are noted below. RAVBEA seems to express less efficiently than the other two combinatorial mutants RAVBGA and RAQBEA.

#### 7.3.3.2.1 RAQBEA Transmission Electron Microscopy

The RAQBEA mutant produced few capsids of discrete sizes. Figure 7-21a & b show two capsid structures of approximately 70 and 50 nm, respectively; the 70 nm capsid appears to house another 50 nm capsid. Figure 7-21c shows a number of highly polymorphous structures which were found more commonly than the spherical capsid structures. As this mutant failed to produce structures appropriate to our intentions, further investigations into this mutant were not conducted.

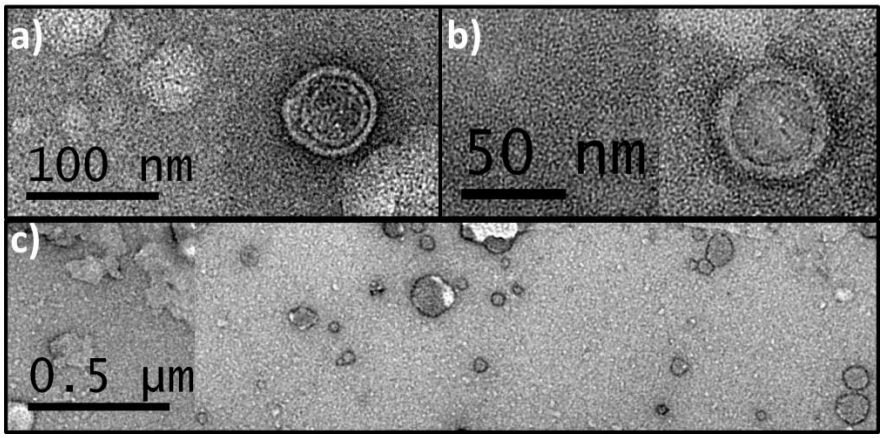



**Figure 7-21:** TEM of the mutant RAQBEA. **a)** & **b)** Few capsid-like structures were seen under TEM. **c)** Other areas of the grid revealed amorphous structures.

#### 7.3.3.2.2 RAVBGA Transmission Electron Microscopy

The RAVBGA mutant produced a large number of capsid structures which could be found throughout the grid. Many allowed the negative stain to penetrate the capsid and TEM thus revealed the inside capsid wall, whereas others resisted stain penetration and appear without discernible internal features. Figure 7-22 shows a range of capsid structures measuring from 25 to 65 nm in diameter. Figure 7-22b & c display typical TEM micrographs of capsid-populated areas of the grid. Figure 7-23 shows a histogram of the particle size distribution for this mutant which show most capsids have a size about 36 and 48 nm in size, with few larger capsids present. The morphology of these capsids seemed more uniform despite the broad size distribution.

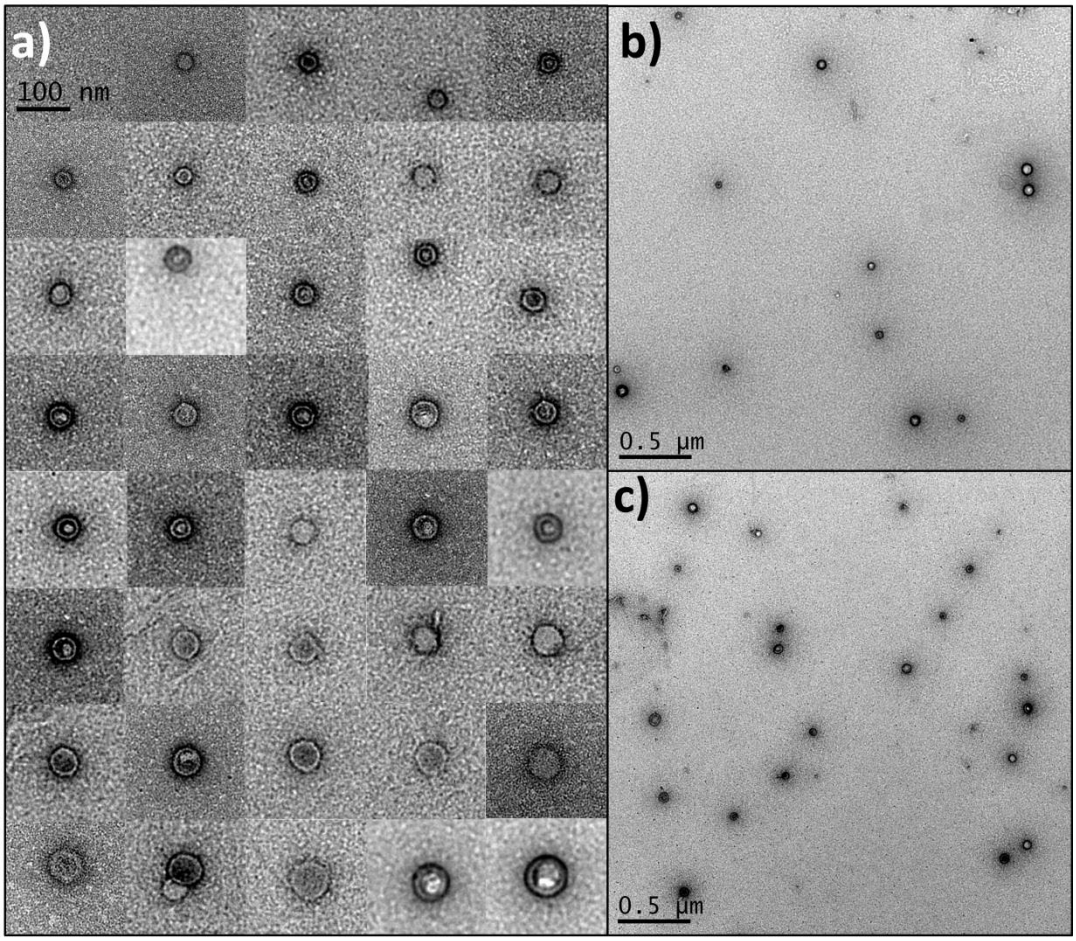



**Figure 7-22:** TEM of RAVBGA. **a)** A selection of capsid structures ranging in size from 27 to 63 nm. Many of the particles are stained to reveal the presence of the inside of the capsid wall. **b) and c)** Representative micrographs of the capsid populated areas of the grid.

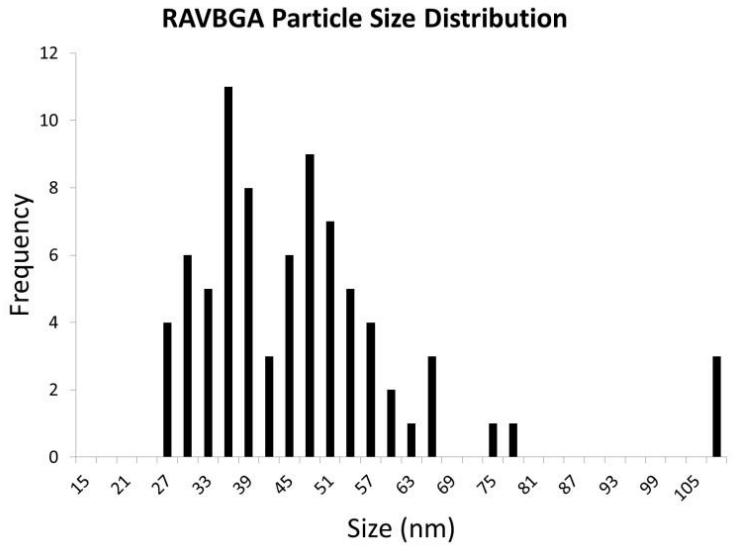



**Figure 7-23:** Particle size distribution for the mutant RAVBGA. This mutant seems to produce less discrete sizes of particles, although there are enrichments of particles with 36 and 48 nm diameters.

# 7.4 Strategy S#3.0

A final combinatorial selection of mutants was made by hand. This set of mutants combined substitutions which had led to capsid assembly from differing strategies. Parts from the S#1.0 mutant IGFBAA were combined with the S#2.0 mutant BAQBGA and the S#2.2 mutant RAVBGA. Figure 7-24 shows the amino acid sequences of the two mutants successfully expressed.

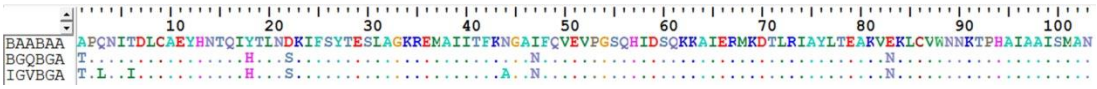



**Figure 7-24**: The amino acid sequence of hand chosen mutants of S#3.0. Part-1I and part-2G come from S#1.0, part-3Q and part-5G come from S#2.0 and part-3V comes from S#2.2.

## 7.4.1 Characterisation of mutants with JR-coil7

Both of the mutants expressed poorly: in Figure 7-25a IGVBGA barely shows a pentameric band and BGQBGA is only marginally better. As the overall expression of the target proteins is reduced, the contaminants now form the majority of these samples. This is of course not favourable, however, due to time constraints at this late stage of the project another purification method was not investigated. Each mutant was concentrated to ca. 1.8 mg/ml and incubated at room temperature for four days. When these two mutants were investigated by SEC (Figure 7-25b), both mutants gave very similar traces, however BGQBGA gave a substantially greater yield of high weight particles. Unfortunately, due to another AKTA problem, the BGQBGA SEC run was not collected and the sample went to waste. Nevertheless IGVBGA was analysed by TEM (Figure 7-25c). The average particle size was 33 nm and the morphology of the capsids seemed to be relatively constant. One the images displays the interior walls of the capsid, and others seem to possess a 3-fold symmetry axis. A histogram of the particle measurements (Figure 7-25d), displayed a narrow Gaussian-like particle size distribution.

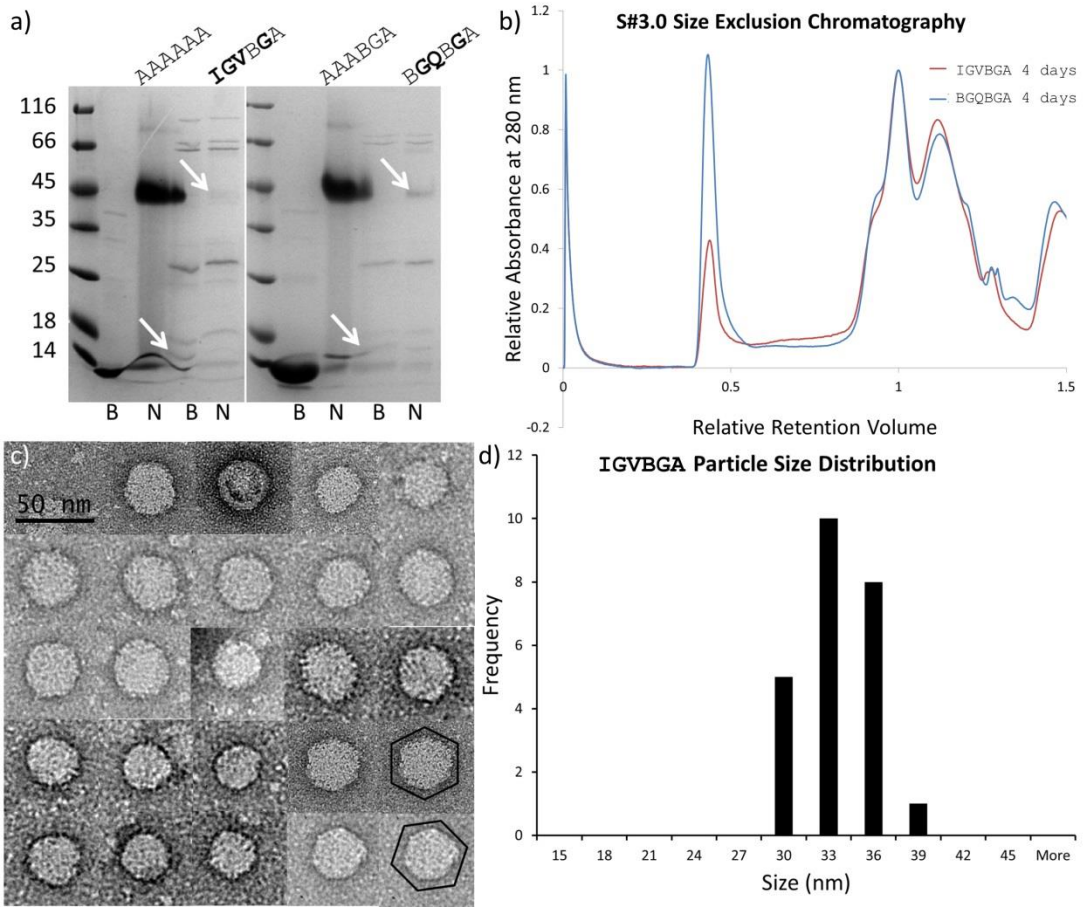



**Figure 7-25:** SDS PAGE, SEC and TEM of S#3.0 mutants. **a)** SDS PAGE of IGVBGA and BGQBGA: boiled samples (B) and non-boiled samples (N) are noted below the gel. Both mutants expressed very poorly. BGQBGA is viable as a pentamer on the gel, however IGVBGA is barely visible. Note that the monomer of IGVBGA is distinct in the boiled lane as opposed to the non-boiled lane. **b)** During SEC, the BGQBGA mutant produced more of the high weight product than IVGBGA after 4 days incubation. **c)** IGVBGA was analysed by TEM and a number of capsid structures were visible. The images did not often display the interior walls of the capsid. However some of them do appear to display clear evidence of a 3-fold axis of symmetry indicative of enclosed particles. **d)** A histogram of the particle size distribution for IGVBGA which averages about 33 nm in size.

# 7.5 Structural Polymorphisms

The structures visualised above and in the previous chapters by TEM are representative of the majority of the capsid populated areas of the grid. However additional structures were found on the CTB grids which resembled part assembled VLP and other aberrant structures. In the introduction, structural polymorphisms are discussed in which natural virus coat proteins can assemble into aberrant structures. Here we present a number of CTB pentamer polymorphisms (Figure 7-26). Oversized capsids, malformed capsids, incomplete capsids, capsid-enclosed capsids and tubes are observed which appear similar to the aberrant structures seen with *in vitro* viral capsid assembly.

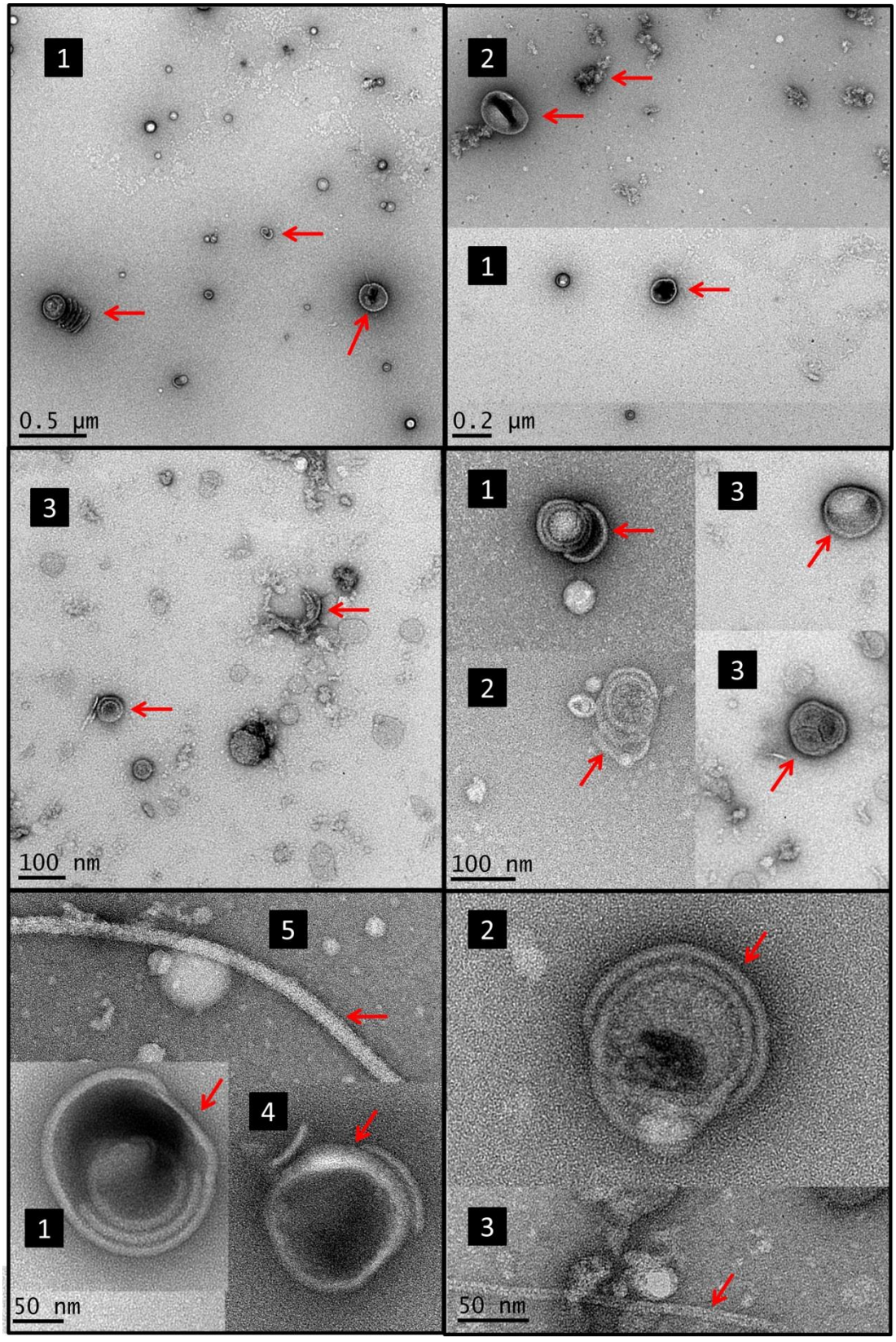



**Figure 7-26:** Structural polymorphs observed under a variety of TEM magnifications. Showing a selection of oversized capsids, malformed capsids, incomplete capsids, capsid-enclosed capsids and tubes. **1)** RAVBGA-coil7. **2)** IGFBAA-coilE. **3)** IGFBAA-coilE/coilK 1:5 ratio. **4)** IGFBAA-coilE/coilK 1:1. **5)** BAVBGA-coil2.

# 7.6 Concluding Statements

In this chapter the computational routine was used to make informed hand-picked mutations based on the saturation mutagenesis data, this provided alternative starting structures for a second iteration of the computational routine.

One of the parent mutants initially suggested, MLQBAA, expressed poorly and the majority of further mutations to the sequences resulted in a complete lack of expression. The daughter mutant VLQBAA produced some capsid structures but these were structurally polymorphic and low in number. The frequency of encountering a capsid was similar to that of the wild-type. Thus no improvement was made with this set of mutants.

The parent mutant BAQBGA, which was shown to have the most improved $\Delta\Delta G_{interaction}$, also produced the greatest yield of high weight species. Four substitutions suggested by the second iteration of the computational routine further improved the $\Delta\Delta G_{interaction}$. Of these suggestions, the daughter mutants RAQBGA and BAVBGA were extended with the JR-coil2. These two mutants produced a variety of well-populated capsid structures. In both cases the highest population of capsids had diameters of 27 nm. However, the BAVBGA mutant also produced a high population of capsids with a diameter of approximately 50 nm (Figure 7-27). The part-1R made a Q3L substitution whereas the part-3V made a N44A mutation. This subtle difference in the biochemistry of the interacting faces seems to be responsible for the production of the expanded capsid structures. Two further daughter mutants RAQBEA and RAVBGA which were extended with JR-coil7 also produced capsids. While RAQBEA produced only a few capsids structures, RAVBGA which combines both the Q3L and N44A mutation produced a large number of capsids. Though some of these capsids had a diameter of 27 nm, there was a shift in population size of the smaller capsids to approximately 36 nm diameter. Additionally, the proportion of 50 nm diameter capsids was larger compared to BAVBGA. Each of the above mutants produced other structures with diameters up to 114 nm, however these were too few to make any estimations of distinct size distributions.

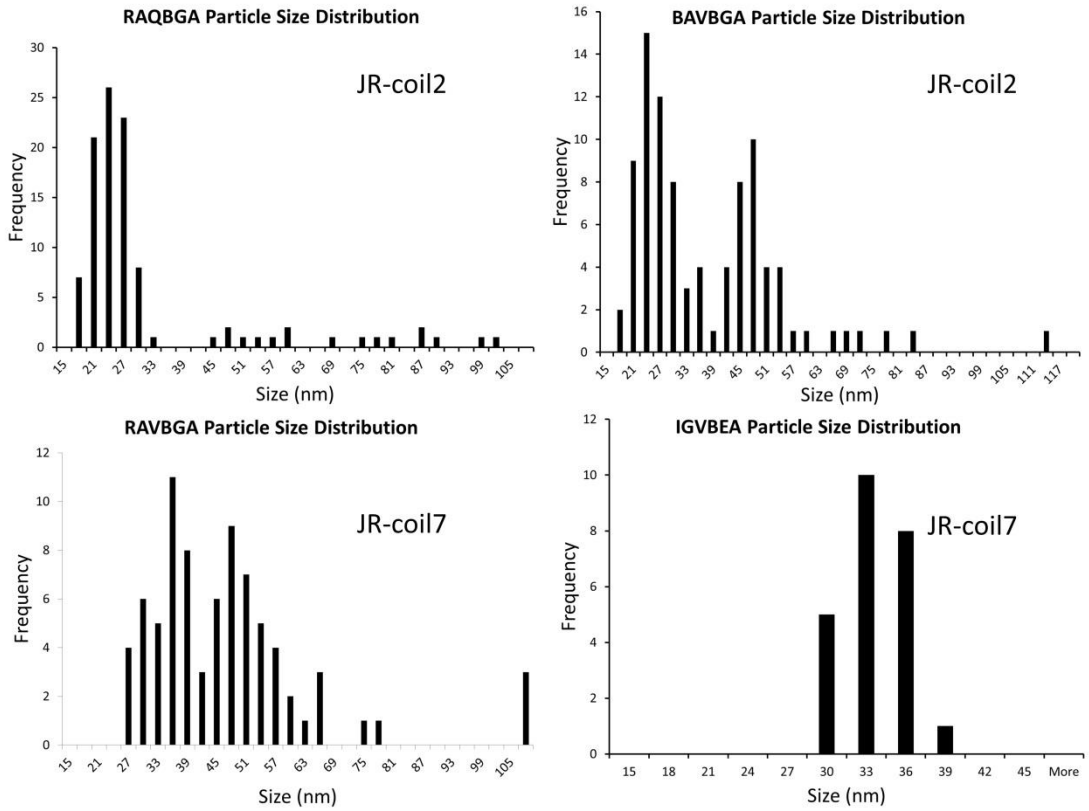



**Figure 7-27:** The particle size distribution observed with different mutants and differing coiled-coils.

In the final investigation, the mutant IGVBEA was constructed with the JR-coil7. Like RAVBGA, this mutant contains the Q3L and N44A mutations as well as T6I, Y18H and D22S. This mutant also showed a lack of 27 nm diameter particles, with most capsid diameters between 33 and 36 nm. This mutant displayed no evidence of greater sized particles.

These data show that the size of the structures assembled is determined by changes to the residues on the interacting surfaces and to the coiled-coils used (Figure 1-26). Although the coiled-coils are probably responsible for the assembly of the capsid structures, the morphology of the capsid is still influenced by the residues at the interface. JR-coil7 seems to preferentially form capsids of 33 nm and over, whereas JR-coil2 seems to mostly give rise to capsids of 27 nm diameter. At present no data is available to suggest the mechanism by which the residues and coils confer this polymorphism. Though, for the coils at least, the propensity to form either dimers (coil7) or trimers (coil2) may be partly responsible for the differing populations observed.

# Chapter 8
# General Conclusions and Future Perspectives

# 8.1 Expression of CTB in E. coli

## 8.1.1 Plasmid Construction

Although there are other species, such as *Vibrio sp.60*, which can be used to express CTB to a good yield, *E. coli* provided a more convenient standardised platform, for the molecular biology techniques used here, to generate sufficient quantities of the chosen mutations. The use of the SAB2.0 gene construct was instrumental in this process through the construction of a plasmid which could provide different expression solutions to unexpected expression problems for the MBP-CTA2 fusion (pSAB2.1) and as a construct to provide CTB expression alone (pSAB2.2). Assembly PCR was very successful in producing combinatorial mutants with ease as well as adding *C*-terminal extensions to the CTB protein. The use of this procedure, excluding *C*-terminal extensions, has yielded a potential combinatorial library of just under 5 million mutants from only 100 designed DNA parts.

Although the assembly PCR method was fast and convenient, it presented a number of reliability issues. Firstly, the initial assembly seemed to vary in success, during some experiments the final target gene was produced as the major species with few intermediates. However in other experiments the intermediates dominated with little or no final product visible. In both cases an amplification using just terminal primers produced a pure final product indicating the presence of the final product in poor assemblies. These deviations in reliable amplification seemed independent of the number of parts used and independent of batches of oligos, suggesting that minor deviations in the reaction conditions may be responsible. Secondly, due to the large size of the oligos, up to 95nt, there was a reliable number of coupling synthesis errors. Due to the assembly PCR method there was a number of miss-assemblies. A total of 282 assembly PCR reactions were carried out where 14 of the assembled genes (5%) were found to be missing parts and 59 of the assembled genes (21%) were found to contain synthesis errors. This resulted in the re-picking of colonies from many of the transformation plates to sample alternate assembled genes from the reactions and in a number cases a repeat of the assembly PCR.

The reliance on the traditional restriction-cut, ligase-paste approach was a stumbling block for progressing mutants from assembled genes into full plasmids. However, new techniques such as Gibson assembly[171] and the Golden Gate strategy[172,173] have been receiving a lot of attention lately. Designing a set of overlapping DNA sequences for use with such techniques would allow future mutations in this system to be generated with greater efficiency.

### 8.1.2 Protein Yield

Another stumbling block in the project was the total yield of CTB produced. As mutations were successively introduced into the structure, the level of expression was successively reduced. The fractionation step of the CTB production was inefficient, requiring large amounts of centrifugation time after a lengthy ammonium sulfate incubation step which challenged the throughput of the procedure. CTB is expressed to the periplasm where folding is mediated by disulfide bond isomerase, after which CTB is found in the media suggesting export or leaching of CTB. Over expression of secreted proteins puts strain on the SEC translocon which perturbs effective translocation of proteins required for nutrient uptake and cell division[174], as well as straining the outer membrane. An intracellular expression of CTB may alleviate the need to extract CTB from the media. *E. coli* strains such Origami2 cells express a cytosolic version of disulfide bond isomerase which may allow cytosolic expression of CTB, and hence allow more traditional methods of fractionation such as cell disruption, this in turn may lead to enhanced yields.

Despite this yield problem, many of the proteins trialled did express to workable yields and as a general conclusion the strategy employed for the production and purification of these CTB mutants was successful for the investigations required in this project.

## 8.2 CTB Mutability

### 8.2.1 Mutational Space Map

CTB is 103 residues long and natural homologs of the CTB protein provide 85 possible substitutions to this sequence. Four previous literature examples highlight the effect of a further 83 mutations, some expressible and some deleterious to the pentameric structure. In this work we have made a further 73 novel mutations to the CTB structure (Figure 8-1). As these mutations are sometimes incorporated in combinatorial parts or single parts, with either one or multiple mutations per part, the success of expression has not been indicated

as this is complex and often dependent on which other mutations are present in the structure containing the mutation of interest. Additionally, the mutations we present are concentrated to an area for which little work has been conducted previously, as more attention has been paid to the GM1 binding site. However an additional oligosaccharide binding site lies within the target PPI[122,123]. Therefore many of the mutations may have changed the oligosaccharide specificity for this additional site. The mutants produced in this work could be used in future screens for sugar binding capacity which may highlight a number of under-investigated cellular targets of CTB.

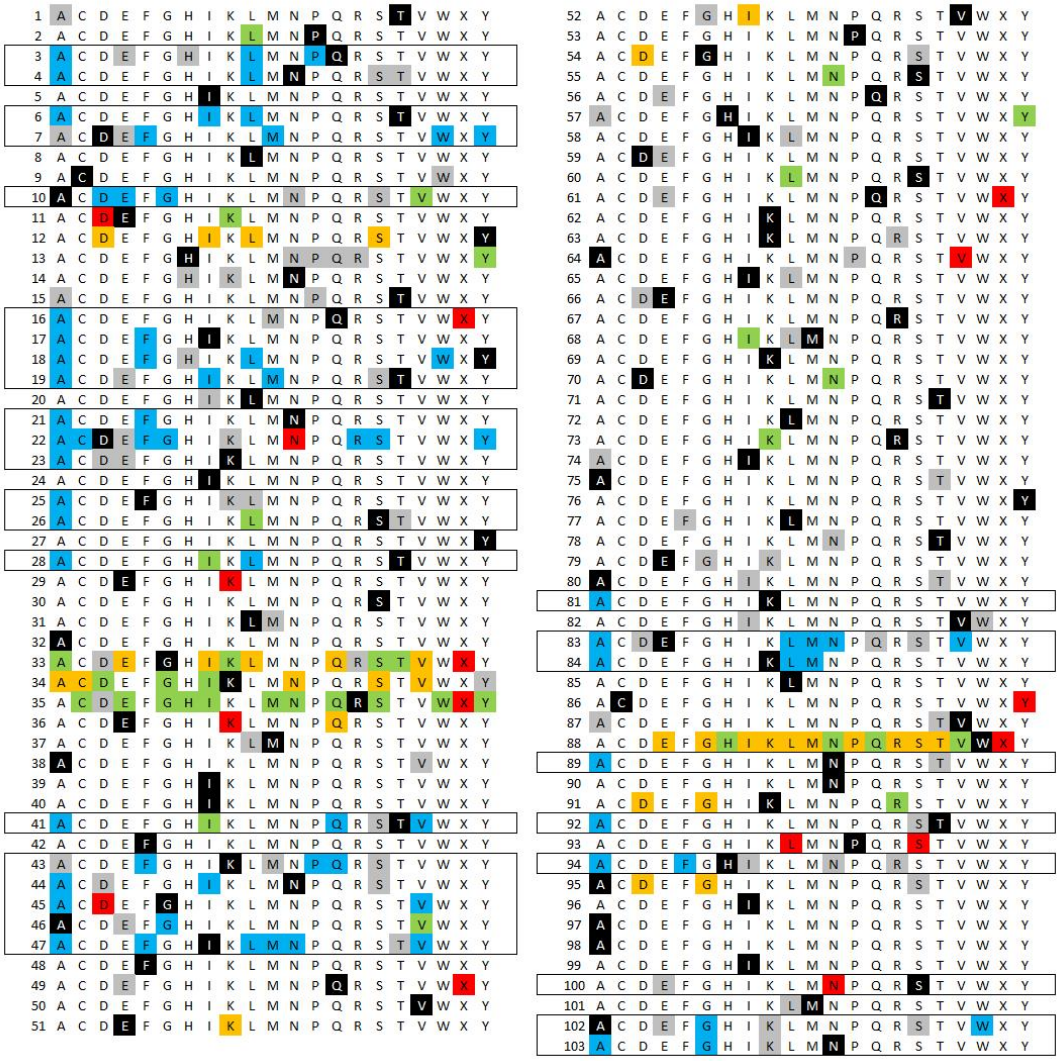


**Figure 8-1:** Updated mutational space map. Black shows the wild-type el tor CTB sequence, grey displays substitutions in natural homologs and strains. Red show inexpressible mutants, orange show mutations which produce pentamers, green shows mutations proven to still form the holotoxin. Light blue shows mutations made to the sequence during this investigation. The boxes represent positions of interest for PPI design. X represents truncation mutants at the appropriate position.

### 8.2.2 Alanine Scan

The alanine scan conducted on the wild-type CTB revealed that the majority of positions of interest were mutable without serious perturbation of the pentameric stability. The SDS PAGE of unboiled mutant samples together with the $T_m$ derived from DSF experiments showed that only a selection of alanine substitutions caused a reduction in $T_m$ of more than 10 degrees. This shows that CTB is a relatively good choice of protein for mutagenesis and that the interface chosen should be highly mutable. As bacterial toxins are more regularly becoming a choice protein for biotechnological applications this alanine scan should provide guidance for future biotech endeavours with this protein.

### 8.2.3 Designer Mutations

The designed mutations seemed to confer a higher degree of instability to the pentamer compared to the alanine scan mutations. This observation could be due to the fact many of the mutations introduced bulky hydrophobic residues at the surface and in many cases replaced intra-pentamer salt bridges which were identified as deleterious to capsid formation. Nonetheless, many of the mutants yielded intact pentamers which went on to increase the rate of capsid assembly.

## 8.3 Coiled-coils

The use of coiled-coils as a scaffold of choice in these investigations was instrumental in the success of the project. Many other scaffolding choices exist but none exhibit the same high degree of literature characterisation. Coiled-coils uniquely possess a large protein-protein interface to volume ratio. Thus the use of coiled-coils in bionanotechnology is useful when space conservation is important, such as here, on the interior of a capsid. The fact that coiled-coils can form both dimers and trimers (as well as many other oligomeric states), again make them uniquely suitable for assembling complex structures which may rely on these types of interaction in the overall assembly. Additionally, the pH dependency of coiled-coil association[175-177] could make these appropriate scaffolds for drug delivery vehicles, as following endocytosis, the intracellular vesicle is likely to follow the lysosomal pathway. A lysosome/endosome-based reduction in the pH of the capsid environment could be a design feature for capsid disassembly and resultant drug delivery.

# 8.4 Computational Strategy

## 8.4.1 Ensembles

The computational approach used generated ensemble structures for analysis with FoldX. This ensemble approach was used to alleviate the short comings of the FoldX algorithm, which makes no attempt to move the backbone of the protein, dealing only with side-chain rotamers. However this ensemble approach is now frequently built into programs such as Rosetta in order to determine more accurate energetic properties. Of course, in nature no proteins exist in a single conformation so it would be counter intuitive not to use ensembles for the assessment of protein and complex structures. The application of MD simulation to produce ensembles from a protein provides a realistic estimation of the solution state of the proteins conformational heterogeneity based on first principles. This method, as opposed to the ensembles produced from programs such as Rosetta, is more accurate from a physical perspective. The Rosetta ensembles, like FoldX side chain sampling, allow main-chain torsional reconfigurations, however the movements, rather than step wise, are bulk random movements. Rosetta then discards high energy conformations. MD however determines each new structure from the previous allowing time dependent observation of the proteins energetics which is not captured by Rosetta.

## 8.4.2 FoldX

Although the structural data for the interface orientations is lacking, the increases in the rates of capsid assembly demonstrate that the computational routine succeeded in providing an improvement in the binding affinity of the PPI. Additionally these mutations directly impacted the size distribution of the particles assembled, which can only be mediated by changes in the affinity/specificity of the interface. Although the coiled-coils most likely provide the majority of the binding affinity, the mutations at the interface have likely reduced native repulsive interactions present in the wild-type. Possibly the most powerful function of FoldX is the position scan (saturation mutagenesis), as this identifies both individually attractive mutations as well as wild-type repulsive residues. The additional studies conducted on the Im9-E9 complex demonstrate that FoldX struggles identifying small changes in the $\Delta\Delta G_{interaction}$. However, the increase in capsid assembly observed in this project shows that FoldX is capable of suggesting mutations which improve the $\Delta\Delta G_{interaction}$ energy.

Future studies could use the methods described here in order to design protein-protein interactions. As exemplified with the lumazine synthase study, the application of periodic boundary conditions to a protein of icosahedral symmetry can allow the MD simulation of large protein complexes. The interaction between these subunits can then be assessed via MD ensembles to elucidate an estimate of the binding energy of the interface. By combining this technique with in silico saturation mutagenesis or ab initio protein design, mutants can be selected which alter the binding affinities of the PPI.

# 8.5 Capsid Assembly

The incorporation of C-terminal coiled-coils to the CTB pentamer allowed the formation of spherical particles to occur. The design of an interface, based on a putative binding orientation and using MD simulation followed by *in silico* mutagenesis to reduce the $\Delta\Delta G_{interaction}$ energy, promoted the rate at which these capsids assemble. A number of coiled-coils were shown to promote capsid assembly and a number of mutations were identified which increase the rate of particle formation.

While the CTB mutants with C-terminal coiled-coil fusions mostly assembled into a specific size of VLP, other sizes of capsids were also found in many cases. This feature is also observed when natural viral proteins assemble *in vitro*[108,109,178]. Additionally, the modified CTB also appeared to form aberrant structures which emulates the behaviour of natural virus coat proteins outside of their optimum physiological environment; for example, malformed capsids, incomplete capsids, capsid-enclosed capsids and tubes. These observations suggest that we have successfully mimicked the assembly properties of viral coat proteins. Use of SEC multi-angle static light scattering (SEC-MALS) could help define the retention volume at which the differing sized capsids elute in order to captures capsids of more discrete sizes. Refinement of the mutants, and further experimentation with coiled-coils should yield modified CTB that assemble readily and to completion under defined conditions. The final structure of the VLP was not elucidated during the course of this work, however, as the capsid structures are strong enough to survive the harsh conditions of negative stain TEM they should also survive the cryo EM environment. Reconstruction of particles imaged by cryo EM could provide an envelope for the fitting of CTB pentamers and reveal structural features which suggest the symmetry at which the pentamers associate. The use of immunogold labelling would also be advantageous to discriminate between the CTB VLPs and artefacts of the TEM procedure or contaminating proteins.

This work demonstrated that pentameric bacterial toxins can be modified to form putative drug delivery vehicles which, in the future, may provide a more robust system than current viral systems.

# 8.6 Drug delivery

At present, studies have not been conducted into the ability of these particles to enter cells. However the molecular biology has been completed to emulate the process described by Seebeck et al.[179], in which a GFP molecule was encapsulated by the protein cage of Lumazine synthase via electrostatic attraction. Once complete this experiment will transport a modified mCherry protein into cells only if captured by capsid structures of CTB, which will provide proof that once capsids have assembled, their endocytic triggering mechanism is preserved. Given that the particles are around 30 nm wide and contain 3 heptad coiled-coils, they should have an internal diameter of ca. 16 nm and a volume of ca. 280 $nm^3$, which would be sufficient for the delivery of small molecule, RNA and protein therapeutics.

The GM1 binding site of CTB provides an exciting target for the mutagenesis. Close homologs of CTB such as the heat labile enterotoxins LT-IIa and LT-IIb have been shown bind a number of other gangliosides (Figure8.2), the shiga toxin and verotoxin . Each of these proteins contain minor mutations at the oligosaccharide binding site which tune the binding specificity, thus introducing binding specificity for alternate oligosaccharide should be attainable goal with CTB and the capsid forming mutants.

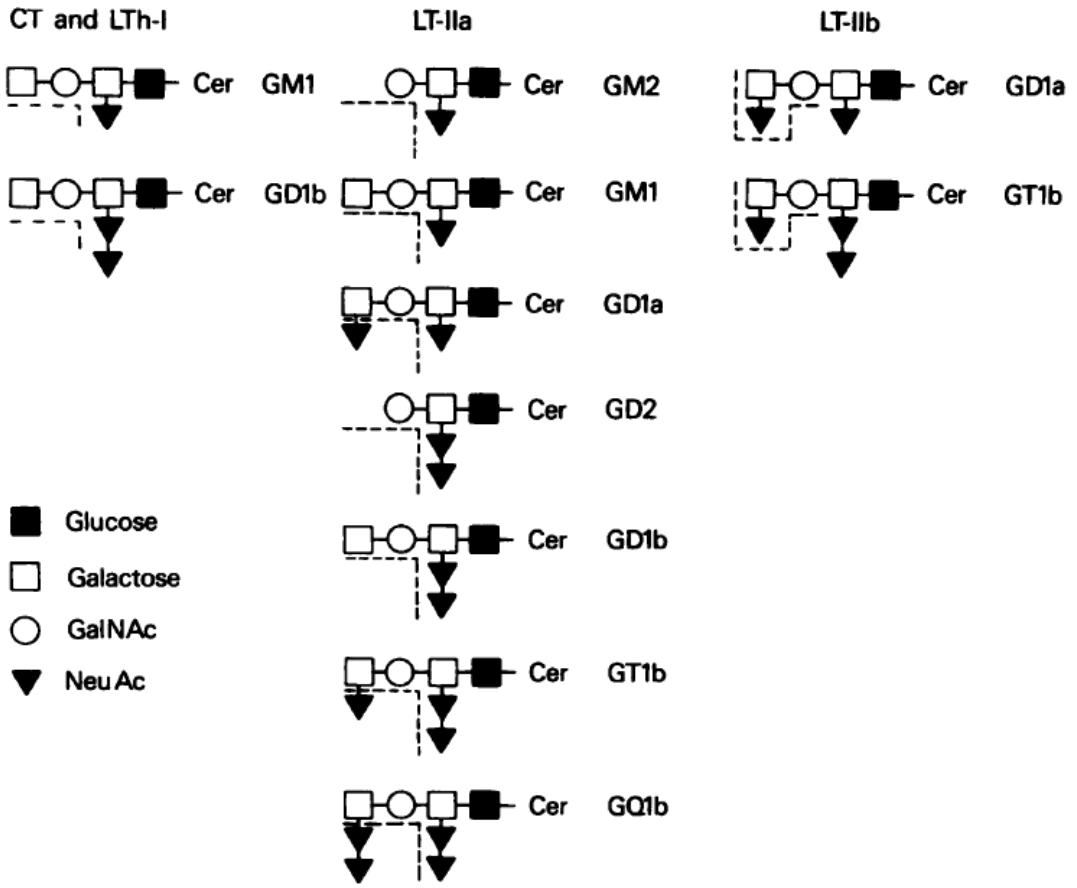



**Figure 8-2:** Ganglioside structures which bind to CTB and CTB like bacterial toxins. The probable binding site is highlighted by the dotted line. This exemplifies the range of specificity for oligosaccharide binding which may be achieved by mutating the GM1 binding site[180].

With the wide range of potentially designable ganglioside binding sites, a wide range of cell types become available for the therapeutic VLP's to target. For example the GD1a ganglioside is abundant in erythrocytes, bone marrow, testes, spleen and liver, however, high concentrations of the GM4 ganglioside are found in in the kidney, GM1 in erythrocytes, GM2 in bone marrow and GM3 in the intestine.

The CTA2 peptide, which threads through the CTB central hole, contains the C-terminal KDEL sequence which targets endosomes containing the CTA2/B complex to the golgi and on to the endoplasmic reticulum. CTB endocytosis without the CTA2 peptide traffics endosomes through the lysosome pathway. This should allow transport of the VLP to these different sub-cellular locations. The targeting of further subcellular compartments is also possible through nuclear localisation tags and mitochondrial targeting sequences. Development of differing disassembly conditions could then allow release of the therapeutic agents at the final desired destination.

Targeting specific cell types in the body and localising therapeutics to a sub-cellular level will reduce off target drug interactions, mitigating side effects. Furthermore, with precise delivery of the therapeutic agents to the site of interest the total quantity of the therapeutic agent required would be severely reduced, reducing the total cost of manufacturing.

There are also many uses for self-assembling proteinaceous cages outside of the context of drug delivery. One such alternate use of protein cages is as bioreactors for the production, in vivo, of commercially interesting fine chemicals and solvents. Many of these kinds of chemicals are toxic to cells and require compartmentalisation for efficient in vivo synthesis. These types of compartments, such as metabolosomes, also address the enzymes required in the production of the chemical to the compartment, which give large increases in the kinetics of multistep enzymatic synthesis. Although the particles produced in this report are of an insufficient size for a bioreactor, the principles underlying its construction are transferable to bioreactor studies.

# Chapter 9

# Appendix

# 9.1 Plasmids and Part Sequences

## 9.1.1 Plasmids

A number of plasmids were used in this work, all of which derive from pMAL-p5x. This plasmid was cloned with a designed insert for the expression of CTB.

### 9.1.1.1 The pMal-p5x Plasmid

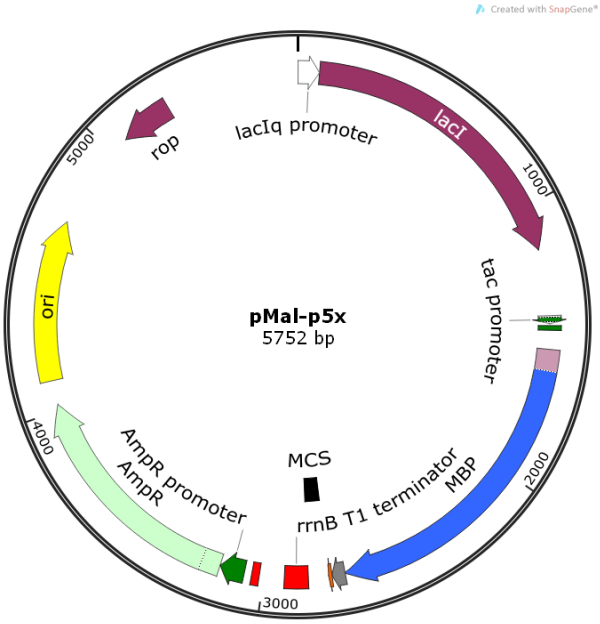



**Figure 9-1:** The pMAL-p5x plasmid from New England Biolabs. This plasmid is used for the gene fusion of MBP to a protein of choice. Expression is controlled by the lac operon.

### 9.1.1.2 The pMAL-p5x Multiple Cloning Site

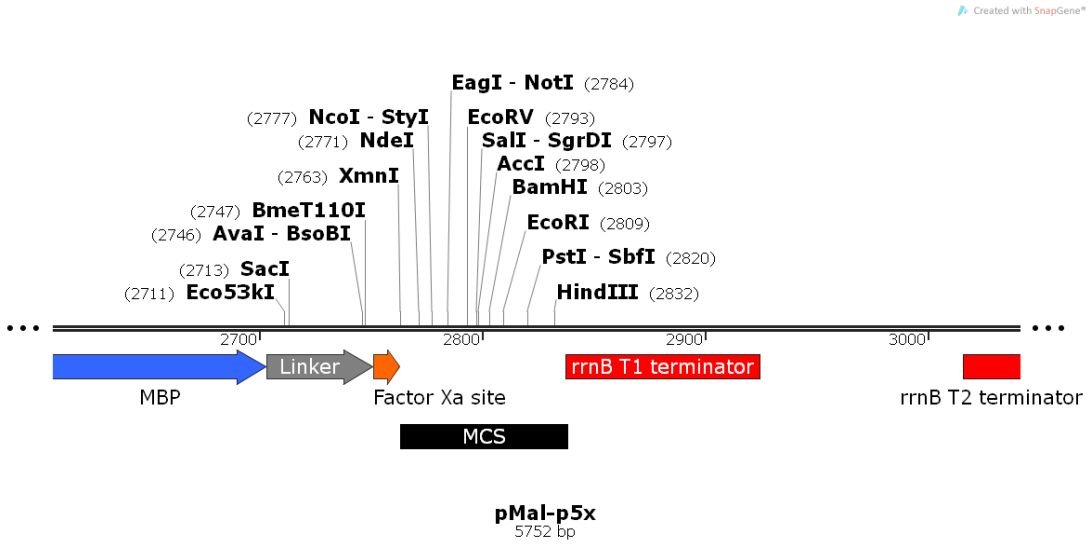



**Figure 9-2:** A detailed view of the pMAL-p5x multiple cloning site (MCS), downstream from the MalE gene which encode MBP. The linker region is translated to polyasparagine, a FactorX protease site is encoded after the linker, before the MCS.

### 9.1.1.3 The SAB2.0 Synthetic Gene Construct

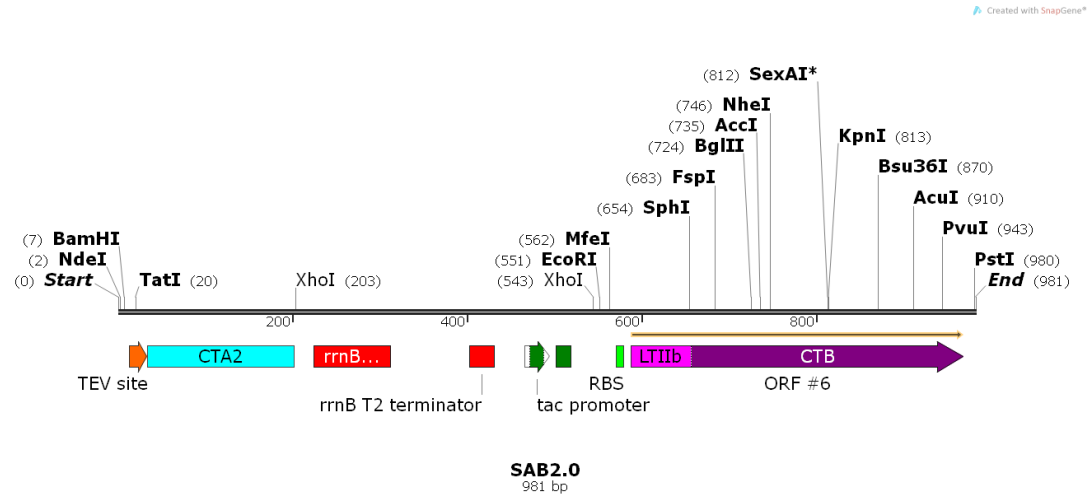



**Figure 9-3:** The synthetic gene construct SAB2.0 purchased from Genscript. This construct contains the CTA1 gene which when cloned into pMAL-p5x will express as a *C*-terminal fusion of MBP. This is followed by a transcriptional termination region (rrnb T2 terminator) and a further promoter and ribosome binding site, which initiate transcription and translation, respectively, of the CTB gene with the LTIIb periplasmic leader sequence.

### 9.1.1.4 Detailed Overview of the SAB2.0 Gene Construct

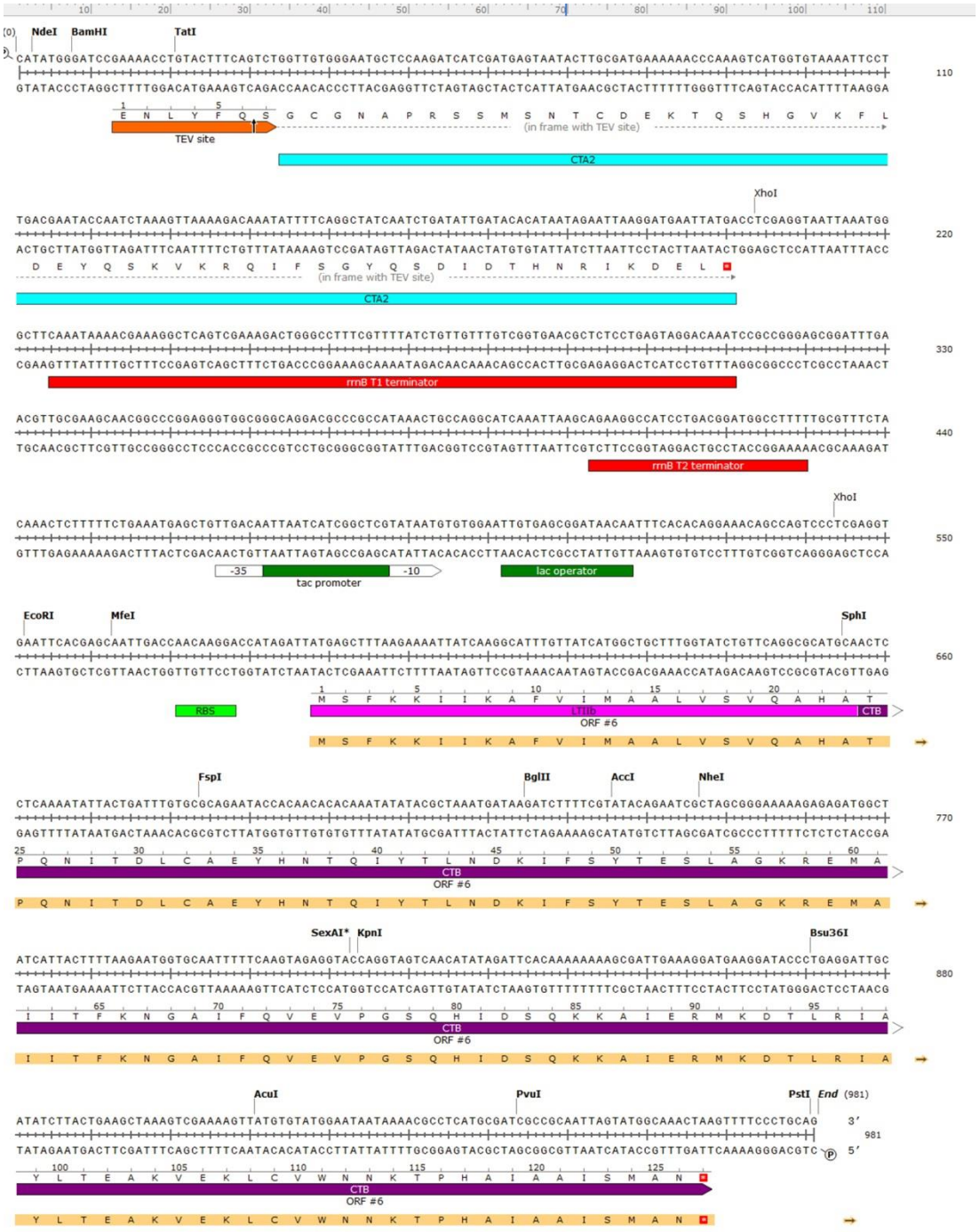



**Figure 9-4:** The sequence of SAB2.0 with highlighted features, restriction sites and the translated amino acid sequence. Restriction sites were intentionally spread throughout the CTB gene in areas where mutations were not designed, this allowed the option of using restriction enzymes to reassemble genes if desired.

### 9.1.1.5 The pSAB2.0 Plasmid

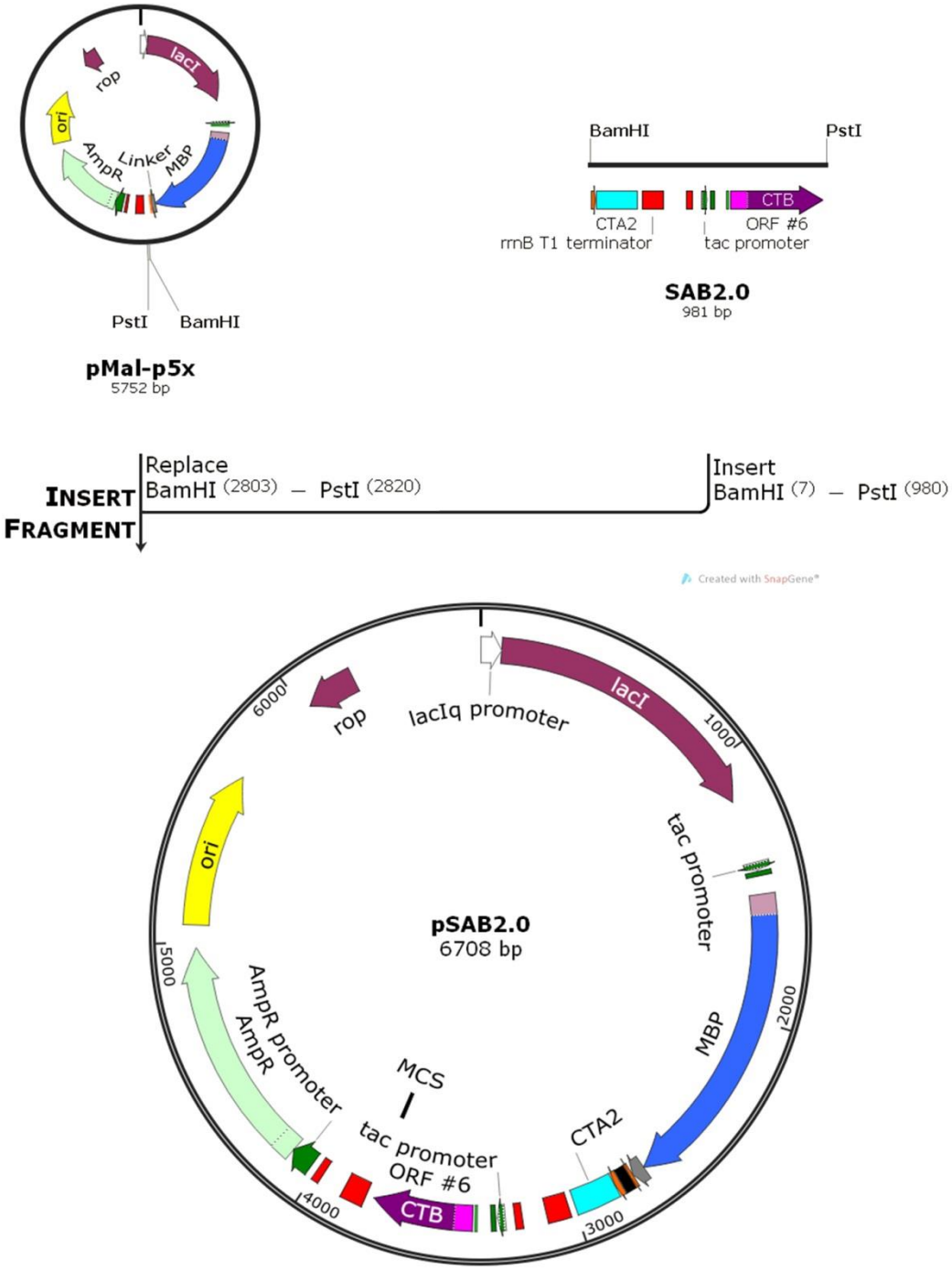



**Figure 9-5:** The pSAB2.0 plasmid. This is created by cloning the SAB2.0 gene construct into pMAL-p5x via digestion with BamHI and PstI and subsequent ligation. This plasmid expresses the MBP-CTA1 gene fusion and CTB from two separate mRNA

### 9.1.1.6 The pSAB2.1 Plasmid

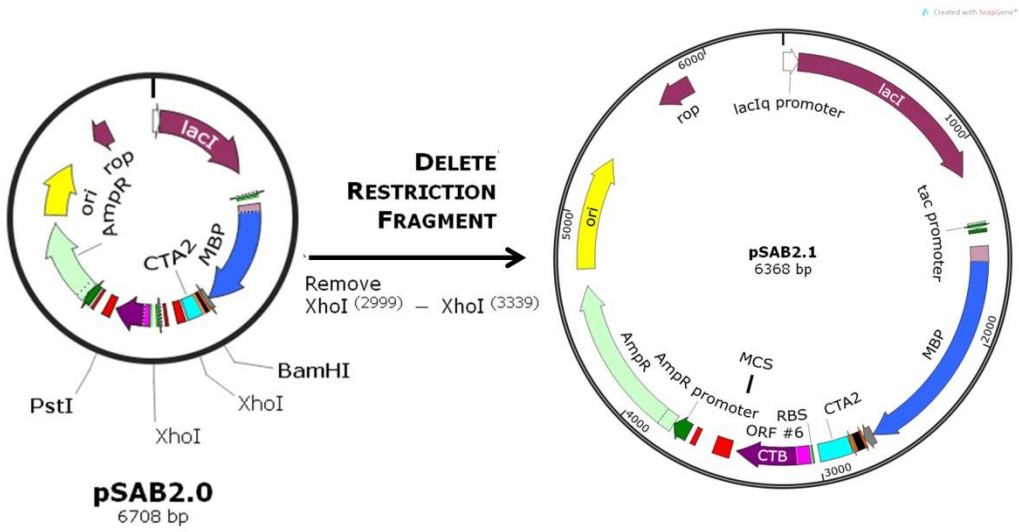



**Figure 9-6:** The pSAB2.1 plasmid. This construct is created by the digestion of pSAB2.0 with XhoI followed by re-ligation. This plasmid expresses the MBP-CTA1 gene fusion and CTB from a single polycistronic mRNA

### 9.1.1.7 pSAB2.2

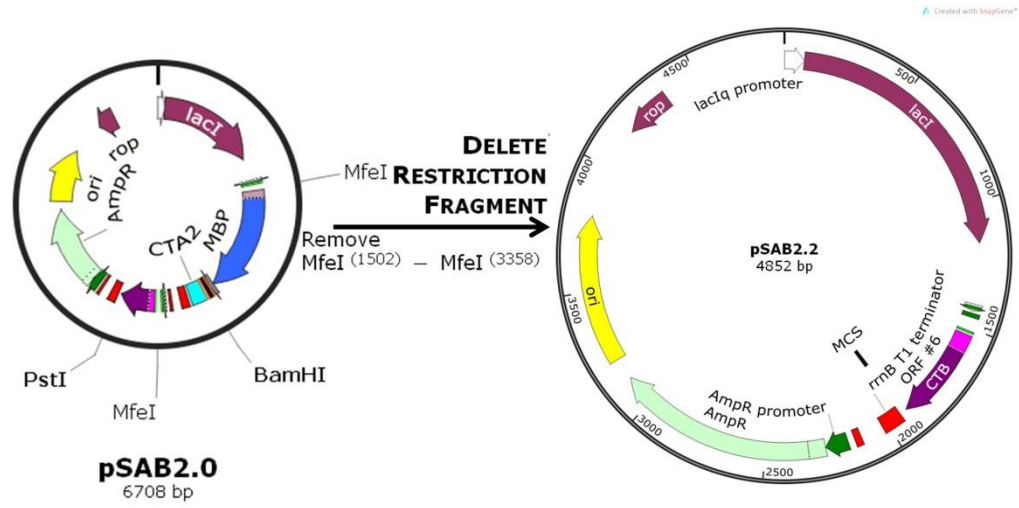



**Figure 9-7:** The pSAB2.2 plasmid. This construct is created by the digestion of pSAB2.0 with MfeI followed by re-ligation. This plasmid expresses CTB from a single mRNA

## 9.1.2 Parts

Oligonucleotides which were involved in the assembly of CTB mutants are listed below, first by nucleotide sequence and then by amino acid sequence.

### 9.1.2.1 Nucleic acids

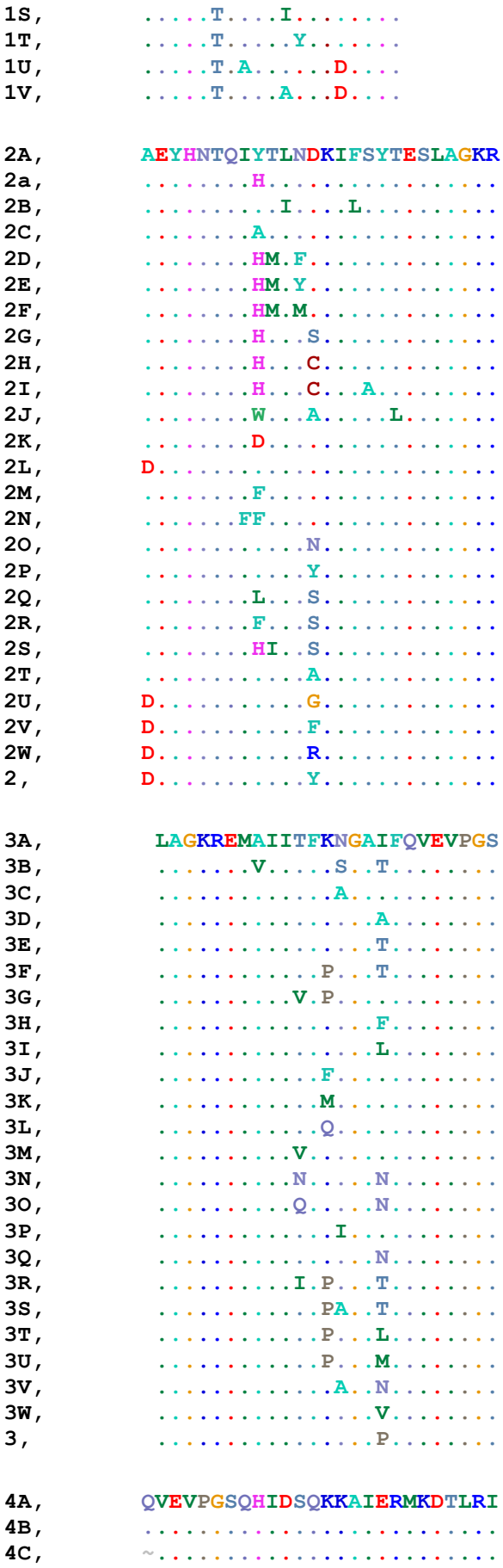
```
1A        CTGTTCAGGCGCATGCAGCTCCTCAAAATATTACTGATTTGTGCGCAGAATACCACAACAC
1B        CTGTTCAGGCGCATGCAACTCCTCAAAATATTACTGATTTGTGCGCAGAATACCACAACAC
1C        CTGTTCAGGCGCATGCAACTCCTCAGTCTATTACAGAACTATGCGCAGAATACCACAACAC
1D        CTGTTCAGGCGCATGCAACTCCTGCAAATATTACTGATTTGTGCGCAGAATACCACAACAC
1E        CTGTTCAGGCGCATGCAACTCCTCAAGCTATTACTGATTTGTGCGCAGAATACCACAACAC
1F        CTGTTCAGGCGCATGCAACTCCTCAAAATATTGCTGATTTGTGCGCAGAATACCACAACAC
1G        CTGTTCAGGCGCATGCAACTCCTCAAAATATTACTTTCTTGTGCGCAGAATACCACAACAC
1H        CTGTTCAGGCGCATGCAACTCCTGCAAATATTATTGATTTGTGCGCAGAATACCACAACAC
1I        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTGATTTGTGCGCAGAATACCACAACAC
1J        CTGTTCAGGCGCATGCAACTCCTGCAAATATTCTTGATTTGTGCGCAGAATACCACAACAC
1K        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTTATTTGTGCGCAGAATACCACAACAC
1L        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTCTCTTGTGCGCAGAATACCACAACAC
1M        CTGTTCAGGCGCATGCAACTCCTCAAAATATTACTGATTTGTGCGATGAATACCACAACAC
1N        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTTTCTTGTGCGCAGAATACCACAACAC
1O        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTTACTTGTGCGCAGAATACCACAACAC
1P        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTATGTTGTGCGCAGAATACCACAACAC
1Q        CTGTTCAGGCGCATGCAACTCCTCTAAATATTATTTGGTTGTGCGCAGAATACCACAACAC
1R        CTGTTCAGGCGCATGCAACTCCTCTTAATATTACTGATTTGTGCGCAGAATACCACAACAC
1S        CTGTTCAGGCGCATGCAACTCCTCAAAATATTATTGATTTGTGCGCAGAATACCACAACAC
1T        CTGTTCAGGCGCATGCAACTCCTCAAAATATTACTTATTTGTGCGCAGAATACCACAACAC
1U        CTGTTCAGGCGCATGCAACTCCTGCAAATATTACTGATTTGTGCGATGAATACCACAACAC
1V        CTGTTCAGGCGCATGCAACTCCTCAAAATATTGCTGATTTGTGCGATGAATACCACAACAC

2A        CGCAGAATACCACAACACACAAATATATACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2a        CGCAGAATACCACAACACACAAATACATACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2B        CGCAGAATACCACAACACACAAATATATACGATAAATGACAAGATACTATCATATACGGAATCGCTAGCGGGAAAAAGAG
2C        CGCAGAATACCACAACACACAAATAGCTACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2D        CGCAGAATACCACAACACACAAATACATATGCTATTTGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2E        CGCAGAATACCACAACACACAAATACATATGCTATATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2F        CGCAGAATACCACAACACACAAATACATATGCTAATGGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2G        CGCAGAATACCACAACACACAAATACATACGCTAAATTCTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2H        CGCAGAATACCACAACACACAAATACATACGCTAAATTGTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2I        CGCAGAATACCACAACACACAAATACATACGCTAAATTGTAAGATCTTTGCGTATACAGAATCGCTAGCGGGAAAAAGAG
2J        CGCAGAATACCACAACACACAAATATGGACGCTAAATGCTAAGATCTTTTCGTATCTAGAATCGCTAGCGGGAAAAAGAG
2K        CGCAGAATACCACAACACACAAATAGATACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2L        CGATGAATACCACAACACACAAATATATACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2M        CGCAGAATACCACAACACACAAATATTTACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2N        CGCAGAATACCACAACACACAATTCTTTACGCTAAATGATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2O        CGCAGAATACCACAACACACAAATATATACGCTAAATAATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2P        CGCAGAATACCACAACACACAAATATATACGCTAAATTATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2Q        CGCAGAATACCACAACACACAAATACTTACGCTAAATTCTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2R        CGCAGAATACCACAACACACAAATATTTACGCTAAATTCTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2S        CGCAGAATACCACAACACACAAATACATATCCTAAATTCTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2T        CGCAGAATACCACAACACACAAATATATACGCTAAATGCTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2U        CGATGAATACCACAACACACAAATATATACGCTAAATGGTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2V        CGATGAATACCACAACACACAAATATATACGCTAAATTTTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2W        CGATGAATACCACAACACACAAATATATACGCTAAATCGTAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG
2         CGATGAATACCACAACACACAAATATATACGCTAAATTATAAGATCTTTTCGTATACAGAATCGCTAGCGGGAAAAAGAG

3A        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3B        GCTAGCGGGAAAAAGAGAAATGGTTATCATTACATTTAAGAGCGGCGCAACATTTCAAGTAGAGGTACCAGGTAGTC
3C        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGGCTGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3D        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCAGCTTTTCAAGTAGAGGTACCAGGTAGTC
3E        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCAACTTTTCAAGTAGAGGTACCAGGTAGTC
3F        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTCCGAATGGTGCAACTTTTCAAGTAGAGGTACCAGGTAGTC
3G        GCTAGCGGGAAAAAGAGAGATGGCTATCATTGTTTTTCCGAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3H        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCATTTTTTCAAGTAGAGGTACCAGGTAGTC
3I        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCATTGTTTCAAGTAGAGGTACCAGGTAGTC
3J        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTCTTCAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3K        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTATGAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3L        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTCAGAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3M        GCTAGCGGGAAAAAGAGAGATGGCTATCATTGTTTTTAAGAATGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3N        GCTAGCGGGAAAAAGAGAGATGGCTATCATTAATTTTAAGAATGGTGCAAATTTTCAAGTAGAGGTACCAGGTAGTC
3O        GCTAGCGGGAAAAAGAGAGATGGCTATCATTCAGTTTAAGAATGGTGCAAATTTTCAAGTAGAGGTACCAGGTAGTC
3P        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGATTGGTGCAATTTTTCAAGTAGAGGTACCAGGTAGTC
3Q        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCAAATTTTCAAGTAGAGGTACCAGGTAGTC
3R        GCTAGCGGGAAAAAGAGAGATGGCTATCATTATCTTTCCGAATGGTGCAACTTTTCAAGTAGAGGTACCAGGTAGTC
3S        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTCCGGCAGGTGCAACTTTTCAAGTAGAGGTACCAGGTAGTC
3T        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTCCGAATGGTGCACTGTTTCAAGTAGAGGTACCAGGTAGTC
3U        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTCCGAATGGTGCAATGTTTCAAGTAGAGGTACCAGGTAGTC
3V        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGGCTGGTGCAAATTTTCAAGTAGAGGTACCAGGTAGTC
3W        GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCAGTTTTTCAAGTAGAGGTACCAGGTAGTC
3         GCTAGCGGGAAAAAGAGAGATGGCTATCATTACTTTTAAGAATGGTGCACCGTTTCAAGTAGAGGTACCAGGTAGTC

4A        CAAGTAGAGGTACCAGGTAGTCAACATATAGATTCACAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGC
4B        CAAGTAGAGGTACCAGGTAGTCAACATATAGATTCACAAAAAAAGGCAATCGAACGTATGAAGGATACCCTGAGGATTGC
4C        ~~~GTAGAGGTACCAGGTAGTCAACATATAGACTCCCAAAAAAAGGCAATCGAACGTATGAAGGATACCCTGAGGATTGC

5A        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAA
5B        GATACCCTGAGGATTGCATATCTGACCGAGACCAAAATTGATAAGCTATGTGTATGGAATAACAAG
5C        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGCAAAGTTATGTGTATGGAATAATAAA
5D        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGAAGCGTTATGTGTATGGAATAATAAA
5E        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGTAAAGTTATGTGTATGGAATAATAAA
5F        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAATCCTAAAGTTATGTGTATGGAATAATAAA
5G        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCAATAAGTTATGTGTATGGAATAATAAA
5H        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCATGAAGTTATGTGTATGGAATAATAAA
5I        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCCTGAAGTTATGTGTATGGAATAATAAA
```

```
5J        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCGTTAAGTTATGTGTATGGAATAATAAA
5K        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCAATATGTTATGTGTATGGAATAATAAA
5L        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCAACAAGTTATGTGTAAAGAATAATAAA
5M        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCAACATGTTATGTGTAAAGAATAATAAA
5N        GATACCCTGAGGATTGCATATCTTACTGAAGCTAAAGTCAACAAGTTATGTGTATGGAATAATAAA

6A        TTATGTGTATGGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATGGCAAACTAAGTTTTCCCTGCAGGTAATTCG
6B        CTATGTGTATGGAATAACAAGACGCCTAATGCGATCGCCGCAATTAGTATG
6C        CTATGTGTATGGAATAACAAGACGCCTAATGCGATCGCCGCAATTAGTATGGAAAACTAAGTTTTCCCTGCAGG
6D        TTATGTGTATGGAATAATAAAGCGCCTCATGCGATCGCCGCAATTAGTATGGCAAACTAAGTTTTCCCTGCAGGTAATTCG
6E        TTATGTGTATGGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATGGGAAACTAAGTTTTCCCTGCAGGTAATTCG
6F        TTATGTGTATGGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATGGCAGCCTAAGTTTTCCCTGCAGGTAATTCG
6G        TTATGTGTATGGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATGGGAAACTAAGTTTTCCCTGCAGGTAAT
6I        TTATGTGTATGGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATGGCAGCCTAAGTTTTCCCTGCAGGTAAT
6J        TTATGTGTATGGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATGGCAGGC
6K        TTATGTGTATGGAATAATAAAACGCCTAATGCGATCGCCGCAATTAGTATG
6L        TTATGTGTATGGAATAATAAAACGCCTCGCGCGATCGCCGCAATTAGTATG
6H        TTATGTGTATGGAATAATAAAACGCCTTGGGCGATCGCCGCAATTAGTATG
6M        TTATGTGTATGGAATAATAAAACGCCTTTTGCGATCGCCGCAATTAGTATG
6N        TTATGTGTATGGAATAATAAAACGCCTTTTGCGATCGCCGCAATTAGTATGGCAGGC
6O        TTATGTGTATGGAATAATAAAACGCCTCCGGCGATCGCCGCAATTAGTATG
6Q        TTATGTGTAAAGAATAATAAAACGCCTCATGCGATCGCCGCAATTAGTATG

7A        GCGATCGCCGCAATTAGTATGGCAAACGGTGTGTCGGCCCTGGAAAAAGAGGTATCAGCTTTGAAGGAGAAAGTCTCC
7C        GCGATCGCCGCAATTAGTATGGCAAACGGTGTGTCGGCCCTGTCTGAGAAAGTATCAGCTTTGGAGTCGAAAGTCTCC
7D        GCGATCGCCGCAATTAGTATGGCAAACGGTGGTGTGTCGGCCCTGTCTGAGAAAGTATCAGCTTTGGAGTCGAAAGTCTCC
7E        GCGATCGCCGCAATTAGTATGGCAAACGGTGGTATTTCGGCCATTTCTGAGAAAATTTCAGCTATCGAGTCGAAAATCTCC
7F        GCGATCGCCGCAATTAGTATGGCAGGCGGTGGTATTTCGGCCATTTCTGAGAAAATTTCAGCTATCGAGTCGAAAATCTCC
7G        GCGATCGCCGCAATTAGTATGGCAAACGGTGGTATTTCGGCCATTTCTGAGAAAATTTCAGCTATCGAGTCCTAAGTTTTCCCTGCAGG
7H        GCGATCGCCGCAATTAGTATGGCAAACGGTGGTATTTCTGAGAAAATTTCAGCTATCGAGTCCTAAGTTTTCCCTGCAGG
7I        GCGATCGCCGCAATTAGTATGGCATCTGGTGTGTCGGCCCTGTCTGAGAAAGTATCAGCTTTGGAGTCGAAAGTCTCC
7J        CTATGTGTATGGAATAACAAGACGCCTCATGCGATCGCCGCAATTAGTATGGCAAACGGTAACGGCGTGCCTGGTGTCGG
7K        CTATGTGTATGGAATAACAAGACGCCTCATGCGATCGCCGCAATTAGTATGGCAAACGGTAACGGCGTGCCTGGTCTGGG
7L        GCGATCGCCGCAATTAGTATGGAAAACGGTGTGTCGGCCCTGGAAAAAGAGGTATCAGCTTTGAAGGAGAAAGTCTCC
7M        GCGATCGCCGCAATTAGTATGGAAAACGGTGGTATTTCGGCCATTTCTGAGAAAATTTCAGCTATCGAGTCGAAAATCTCC
7N        GCGATCGCCGCAATTAGTATGGCAGGCGGTGGTAGGGTTGAGGCTCTTGAGAAGAAGGTCGCCGCTTTAGAGTC
7O        GCGATCGCCGCAATTAGTATGGCAGGCGGTGGTGAGGTTTCCGCGCTGGAAAAAGAGGTTTCTGCATTGGAGAAGGAAGTCTCCGCACTCG
7P        GCGATCGCCGCAATTAGTATGGCAGGCGGTGGTAAAGTGTCTGCCCTCAAGGAAAAGGTTTCGGCTCTTAAGGAGAAAGTCTCCGCACTCAAAGAG

8A        GCTTTGAAGGAGAAAGTCTCCGCACTCGAGTTCTAAGTTTTCCCTGCAGG
8B        GCTTTGAAGGAGAAAGTCTCCGCACTCGAGAAAGAGGTATCAGCTTTGAAGTTCTAAGTTTTCCCTGCAGG
8C        GCTTTGGAGTCGAAAGTCTCCGCACTCGAAAGCAAGGTATCAGCTTTGGAGTTCTAAGTTTTCCCTGCAGG
8D        GCTTTGGAGTCGAAAGTCTCCGCAAATGAAAGCAAGGTATCAGCTTTGGAGTTCTAAGTTTTCCCTGCAGG
8E        GCTTTGGAGTCGAAAGTCTCCGCACTCGAATCCTAAGTTTTCCCTGCAGG
8F        GCTATCGAGTCGAAAATCTCCGCAATTGAATCCTAAGTTTTCCCTGCAGG
8G        GCTATCGAGTCGAAAATCTCCGCAATTGAAGAATAAGTTTTCCCTGCAGG
8H        GCTATCGAGTCGAAAATCTCCGCAATTGAATCCGGCAATAACGGTCTGCCGGAAACCGGTGGCTAAGTTTTCCCTGCAGG
8I        GTCGCCGCTTTAGAGTCTAAAGTTCAGGCCCTGGAGAAGAAGGTTGAGGCACTCGAACATGGCTAAGTTCTGCAGGTAAGC
8J        GGAAGTCTCCGCACTCGAGAAAGAGGTGTCGGCCCTGGAAAAAGAGGTATCAGCTTTGGAGAAGGGCTAAGTTCTGCAGGTAAGC
8K        GTCTCCGCACTCAAAGAGAAGGTGTCGGCCCTGAAAGAAAAGGTATCAGCTTTGAAGGAGGGCTAAGTTCTGCAGGTAAGC
```

### 9.1.2.2 Amino Acid sequence.

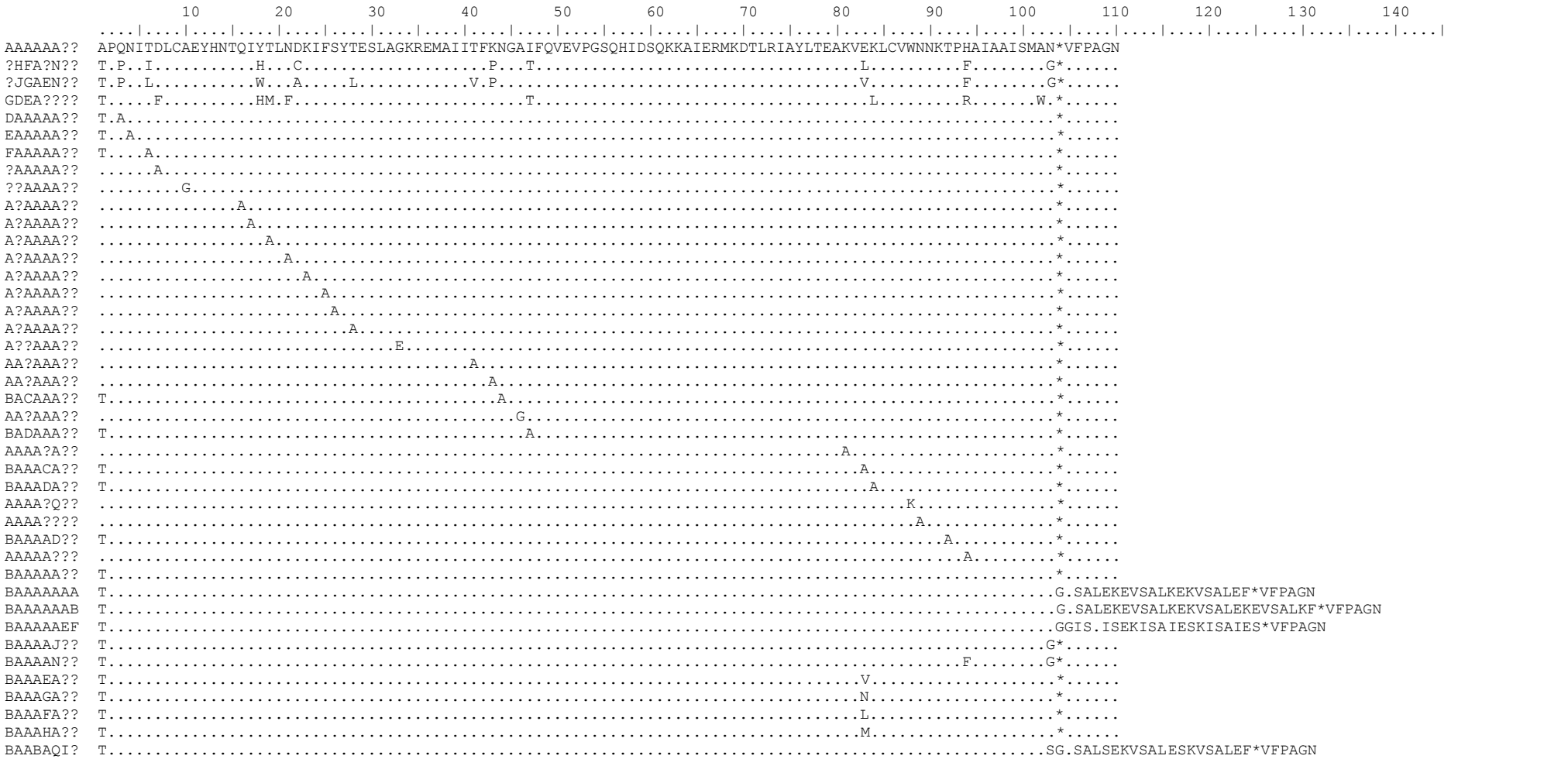
These sequences are displayed with '.' to represent the wild-type residue, thus highlighting the change made to the sequence. Some parts contain silent mutations, where the amino acid sequence is identical but the nucleotide sequence is different. Apart from part 4B, this is accidental. Parts 7 & 8 show the whole sequences as the parts differ substantially.

```
1A,       VQAHAAPQNITDLCAEYHN
1B,       .....T.............
1C,       .....T..S..E.......
1D,       .....T.A...........
1E,       .....T..A..........
1F,       .....T....A........
1G,       .....T.....F.......
1H,       .....T.A..I........
1I,       .....T.L..I........
1J,       .....T.A..L........
1K,       .....T.L..IY.......
1L,       .....T.L..IL.......
1M,       .....T.........D....
1N,       .....T.L..IF.......
1O,       .....T.L..IY.......
1P,       .....T.L..IM.......
1Q,       .....T.L..IW.......
1R,       .....T.L...........
```


1S,
1T,
1U,
1V,
2A, AEYHNTQIYTLNDKIFSYTESLAGKR
2a,
2B,
2C,
2D,
2E,
2F,
2G,
2H,
2I,
2J,
2K,
2L,
2M,
2N,
2O,
2P,
2Q,
2R,
2S,
2T,
2U,
2V,
2W,
2,
3A, LAGKREMAIITFKNGAIFQVEVPGS
3B,
3C,
3D,
3E,
3F,
3G,
3H,
3I,
3J,
3K,
3L,
3M,
3N,
3O,
3P,
3Q,
3R,
3S,
3T,
3U,
3V,
3W,
3,
4A, QVEVPGSQHIDSQKKAIERMKDTLRI
4B,
4C,

```
5A,       DTLRIAYLTEAKVEKLCVWNNK
5B,       ..........T.ID........
5C,       ............A........
5D,       ............A.......
5E,       ............V........
5F,       ............L........
5G,       ............N........
5H,       ............M........
5I,       ............L........
5J,       ............V........
5K,       ............NM.......
5L,       ............N....K...
5M,       ............NM...K...
5N        ............N........

6A,       LCVWNNKTPHAIAAISMAN*VFPAGNS
6B,       .........N.......
6C,       .........N.......E.*....
6D,       .......A..........*.......
6E,       ................G.*.......
6F,       .................A*.......
6G,       ................G.*......
6I,       .................A*......
6J,       ................G
6K,       .........N.......
6L,       .........R.......
6H,       .........W.......
6M,       .........F.......
6N,       .........F........G
6O,       .........P.......
6Q,       ...K............

7A,       AIAAISMANGVSALEKEVSALKEKVS
7C,       AIAAISMANGVSALSEKVSALESKVS
7D,       AIAAISMANGGVSALSEKVSALESKVS
7E,       AIAAISMANGGISAISEKISAIESKIS
7F,       AIAAISMAGGGISAISEKISAIESKIS
7G,       AIAAISMANGGISAISEKISAIES*VFPA
7H,       AIAAISMANGGISEKISAIES*VFPA
7I,       AIAAISMASGVSALSEKVSALESKVS
7J,       LCVWNNKTPHAIAAISMANGNGVPGV
7K,       LCVWNNKTPHAIAAISMANGNGVPGL
7L,       AIAAISMENGVSALEKEVSALKEKVS
7M,       AIAAISMENGGISAISEKISAIESKIS
7N,       AIAAISMAGGGRVEALEKKVAALE
7O,       AIAAISMAGGGEVSALEKEVSALEKEVSAL
7P,       AIAAISMAGGGKVSALKEKVSALKEKVSALKE


8A,       ALKEKVSALEF*VFPA
8B,       ALKEKVSALEKEVSALKF*VFPA
8C,       ALESKVSALESKVSALEF*VFPA
8D,       ALESKVSANESKVSALEF*VFPA
8E,       ALESKVSALES*VFPA
8F,       AIESKISAIES*VFPA
8G,       AIESKISAIEE*VFPA
8H,       AIESKISAIESGNNGLPETGG*VFPA
8I,       VAALESKVQALEKKVEALEHG*VLQVS
8J,       EVSALEKEVSALEKEVSALEKG*VLQVS
8K,       VSALKEKVSALKEKVSALKEG*VLQVS
```

## 9.1.3 Expressed sequences



```
BAABAACC  T.....G.SALSEKVSALESKVSALESKVSALEF*VFPAGN
                 10        20        30        40        50        60        70        80        90       100       110       120       130       140
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AAAAAA??  APQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVPGSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN*VFPAGN
BAABAACD  T.....G.SALSEKVSALESKVSANESKVSALEF*VFPAGN
BAABAAG?  T.....GGIS.ISEKISAIES*VFPAGN
BAABAAH?  T.....GGISEKISAIES*VFPAGN
BAEAAA??  T.....T.....*......
BAFAAA??  T.....P...T.....*......
BAHBAA??  T.....F.....*......
BAIAAA??  T.....L.....*......
BAJAAA??  T.....F.....*......
BAKAAA??  T.....M.....*......
BALBAA??  T.....Q.....*......
BALBGAAA  T.....Q.....N.....G.SALEKEVSALKEKVSALEF*VFPAGN
BAMAAA??  T.....V.....*......
BAOAAA??  T.....Q.....N.....*......
BAPAAA??  T.....I.....*......
BAQAAA??  T.....N.....*......
BAQBGAEF  T.....N.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
BAQBKAAA  T.....N.....NM.....G.SALEKEVSALKEKVSALEF*VFPAGN
BAVBGAAA  T.....A..N.....N.....G.SALEKEVSALKEKVSALEF*VFPAGN
BAWBGAAA  T.....V.....N.....G.SALEKEVSALKEKVSALEF*VFPAGN
BCAAAA??  T.....A.....*......
BGQBGAEF  T.....H...S.....N.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
BHAAAA??  T.....H...C.....*......
BIAAAA??  T.....H...C...A.....*......
BJAAAA??  T.....W...A....L.....*......
BMABAA??  T.....F.....*......
BNKBGAAA  T.....FF.....M.....N.....G.SALEKEVSALKEKVSALEF*VFPAGN
BOAAAA??  T.....N.....*......
BTQBGAAA  T.....A.....N.....N.....G.SALEKEVSALKEKVSALEF*VFPAGN
GAAAAA??  T.....F.....*......
IAABAAAA  T.L..I.....G.SALEKEVSALKEKVSALEF*VFPAGN
IAQBGAEF  T.L..I.....N.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
IAVBGAEF  T.L..I.....A..N.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
IGFAFAAA  T.L..I.....H...S.....P...T.....L.....G.SALEKEVSALKEKVSALEF*VFPAGN
IGFAAAAA  T.L..I.....H...S.....P...T.....G.SALEKEVSALKEKVSALEF*VFPAGN
IGFAEAAA  T.L..I.....H...S.....P...T.....V.....G.SALEKEVSALKEKVSALEF*VFPAGN
IGFBAAEF  T.L..I.....H...S.....P...T.....GGIS.ISEKISAIESKISAIES*VFPAGN
IGFBAAEG  T.L..I.....H...S.....P...T.....GGIS.ISEKISAIESKISAIEE*VFPAGN
IGFBGAEF  T.L..I.....H...S.....P...T.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
IGFBLQEH  T.L..I.....H...S.....P...T.....N....K.....GGIS.ISEKISAIESKISAIESGNNGLPETGG*VFPAGN
IGFBNAEH  T.L..I.....H...S.....P...T.....N.....GGIS.ISEKISAIESKISAIESGNNGLPETGG*VFPAGN
IGQBAAEF  T.L..I.....H...S.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
IGQBGAEF  T.L..I.....H...S.....N.....N.....GGIS.ISEKISAIESKISAIES*VFPAGN
IGRBAAAA  T.L..I.....H...S.....I.P...T.....G.SALEKEVSALKEKVSALEF*VFPAGN
IGSBAAAA  T.L..I.....H...S.....PA..T.....G.SALEKEVSALKEKVSALEF*VFPAGN
```

```
IGTBAAAA  T.L..I...........H...S.....................P...L.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
                 10        20        30        40        50        60        70        80        90       100       110       120       130       140
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AAAAAA??  APQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVPGSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN*VFPAGN
IGUBAAAA  T.L..I...........H...S.....................P...M.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
IGVBGAEF  T.L..I...........H...S.....................A..N.....................................N.....................GGIS.ISEKISAIESKISAIES*VFPAGN
IQFBAAAA  T.L..I...........L...S.....................P...T.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
IRFBAAAA  T.L..I...........F...S.....................P...T.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
ISFBAAAA  T.L..I...........HI..S.....................P...T.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
JAAAAA??  T.A..L.....................................................................................................*......
JGABAAAA  T.A..L...........H...S.....................................................................................G.SALEKEVSALKEKVSALEF*VFPAGN
JGFAFAAA  T.A..L...........H...S.....................P...T.....................................L.....................G.SALEKEVSALKEKVSALEF*VFPAGN
MLQBAAAA  T........D.....................................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
MLQBAAAA  T........D.....................................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
MLQBAAEF  T........D.....................................N.....................................................GGIS.ISEKISAIESKISAIES*VFPAGN
MLXBAAAA  T........D.....................................P.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
MUQBAAAA  T........D...........G.........................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
MVQBAAAA  T........D...........F.........................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
MWQBAAAA  T........D...........R.........................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
MXQBAAAA  T........D...........Y.........................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
NGFBAAAA  T.L..IF..........H...S.....................P...T.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
PGFBAAAA  T.L..IM..........H...S.....................P...T.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
QGFBAAAA  T.L..IW..........H...S.....................P...T.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
RAQBEAEF  T.L.....................................................N.....................................V.....................GGIS.ISEKISAIESKISAIES*VFPAGN
RAQBEAEF  T.L.....................................................N.....................................V.....................GGIS.ISEKISAIESKISAIES*VFPAGN
RAQBGAAA  T.L.....................................................N.....................................N.....................G.SALEKEVSALKEKVSALEF*VFPAGN
RAQBGAEF  T.L.....................................................N.....................................N.....................GGIS.ISEKISAIESKISAIES*VFPAGN
RAVBEAEF  T.L.....................................A..N.....................................V.....................GGIS.ISEKISAIESKISAIES*VFPAGN
RAVBGAEF  T.L.....................................A..N.....................................N.....................GGIS.ISEKISAIESKISAIES*VFPAGN
RAVBGAEG  T.L.....................................A..N.....................................N.....................GGIS.ISEKISAIESKISAIEE*VFPAGN
RAVBLQEH  T.L.....................................A..N.....................................N....K..............GGIS.ISEKISAIESKISAIESGNNGLPETGG*VFPAGN
SAQBGAAA  T....I...................................N.....................................N.....................G.SALEKEVSALKEKVSALEF*VFPAGN
TAQBGAAA  T.....Y..................................N.....................................N.....................G.SALEKEVSALKEKVSALEF*VFPAGN
ULQBAAAA  T.A......D...............................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
ULQBAAAA  T.A......D...............................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
VLQBAAAA  T....A...D...............................N.....................................................G.SALEKEVSALKEKVSALEF*VFPAGN
??GAAA??  T........E.............................V.P.....................................................*......
IGQBG???  T.L..I...........H...S.....................N.....................................N.....................ES*VFPAGN
MUQBA??A  T........D...........G.....................N.....................................................RANGVSALEKEVSALKEKVSALEF*VFPAGN
IGF?FAAA  T.L..I...........H...S.....................P...T...........N.....................L.....................G.SALEKEVSALKEKVSALEF*VFPAGN
```

# 9.2 Supplementary Information

## 9.2.1 Chapter 1-Introduction

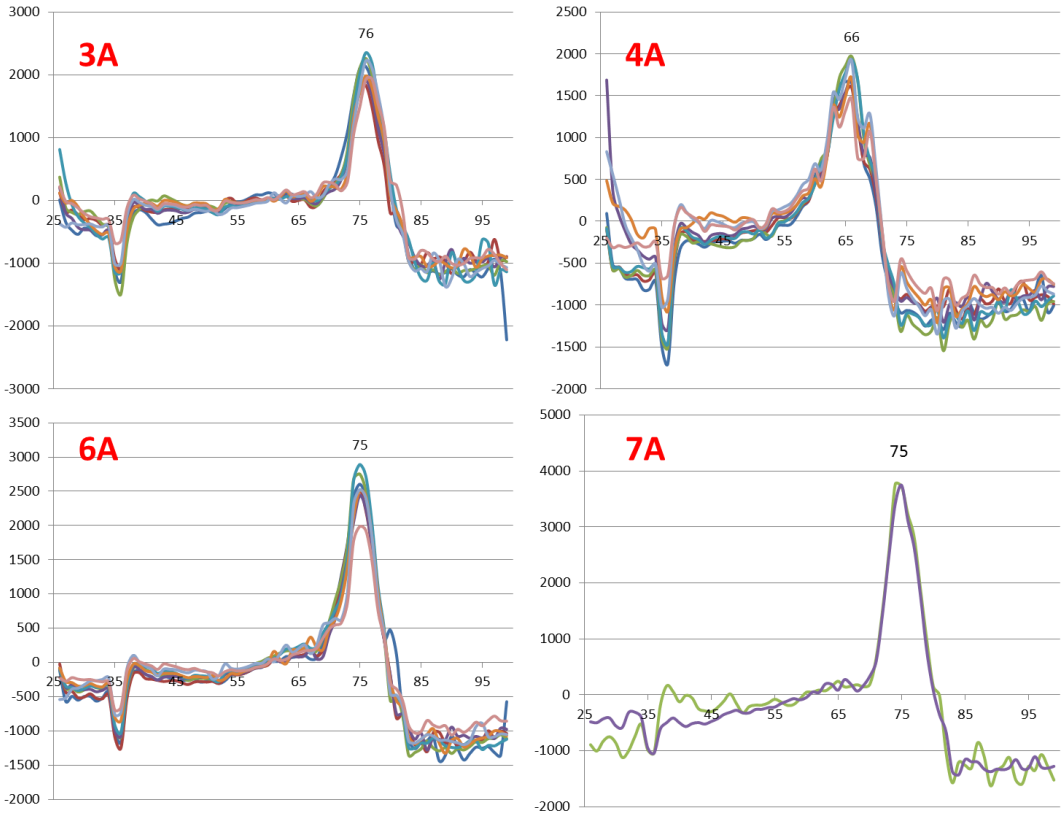
No additional information.

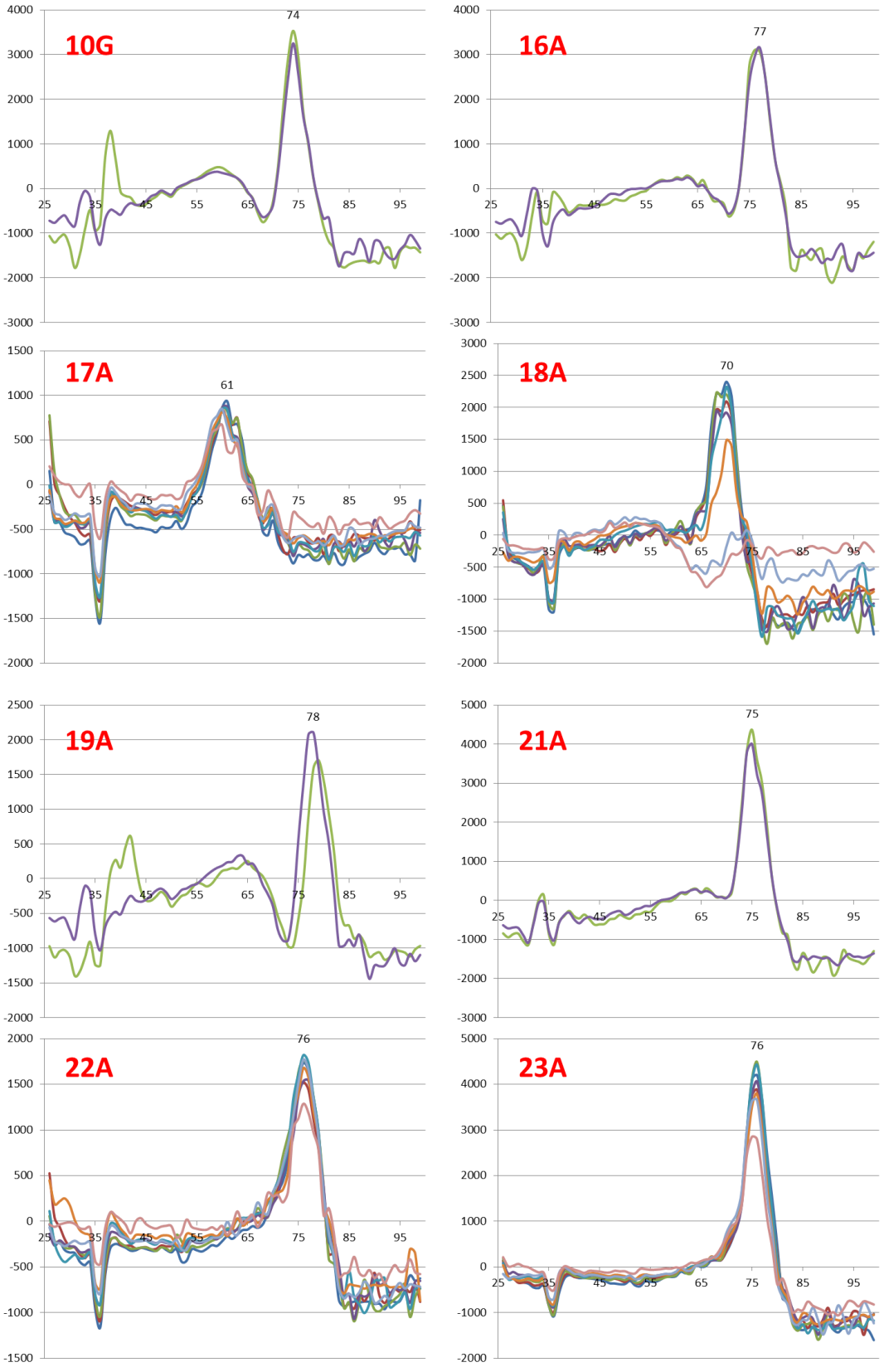
## 9.2.2 Chapter 2-Materials and Methods

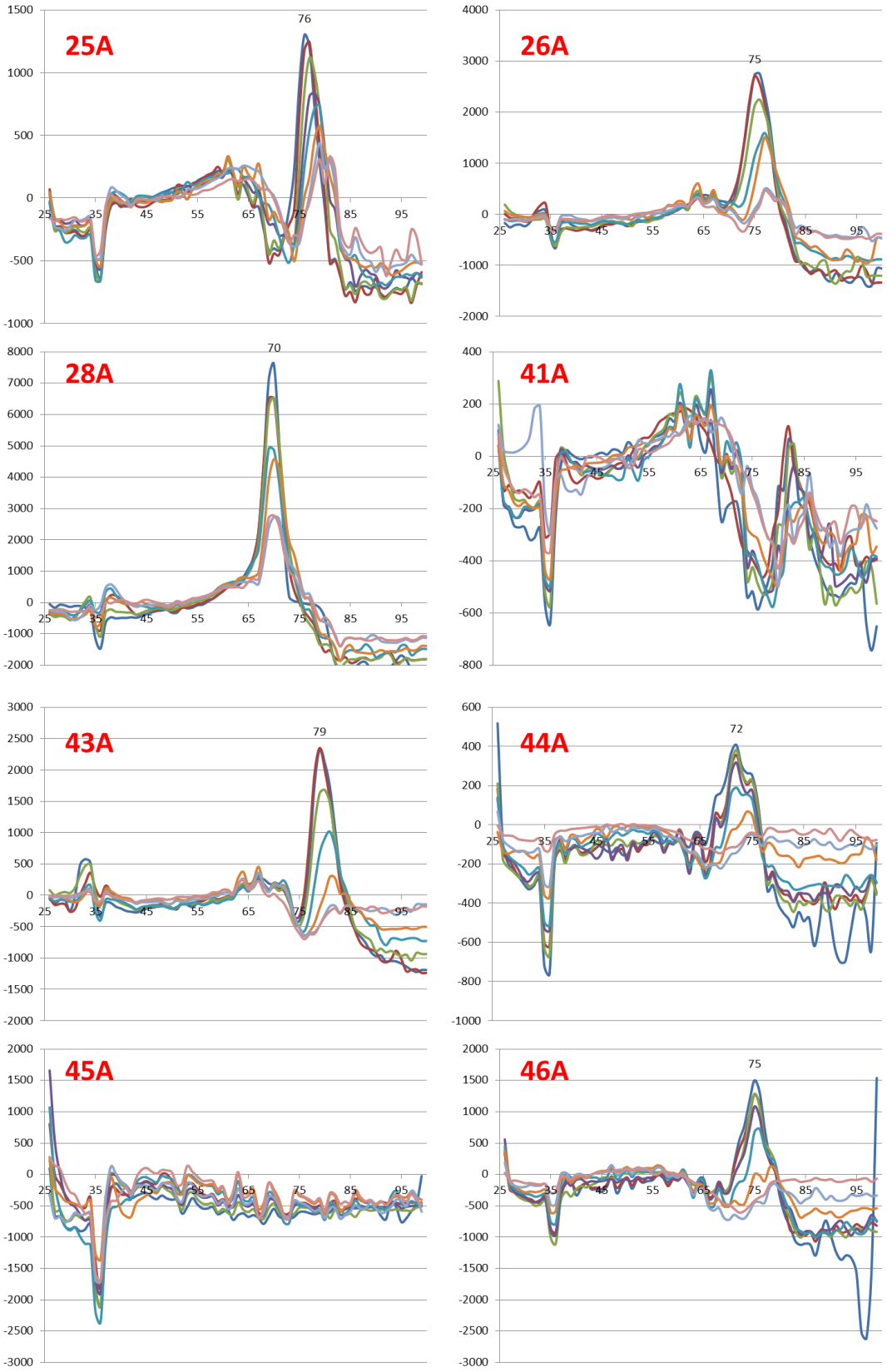
No additional information.

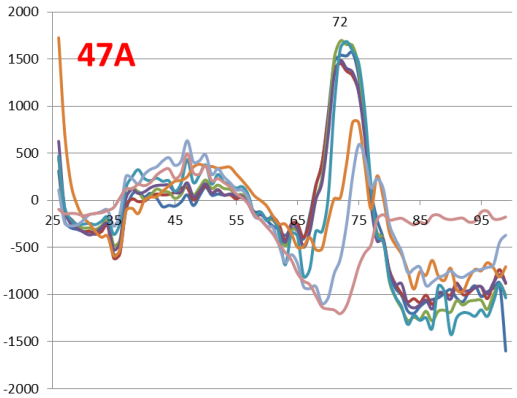
## 9.2.3 Chapter 3-Assessing CTB Stability

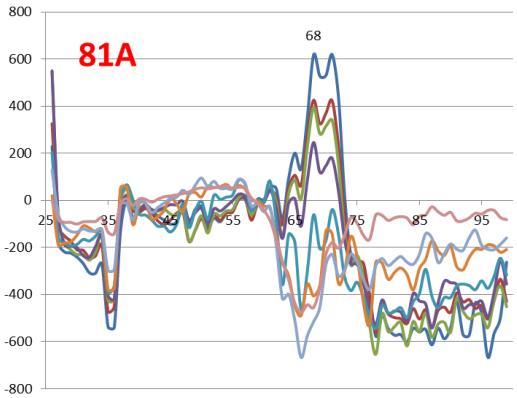
The DSF traces for the alanine scans are presented here.

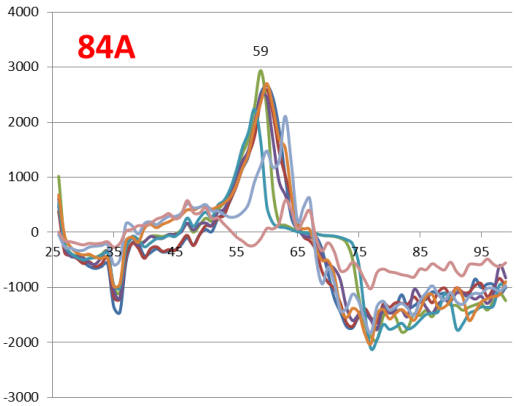


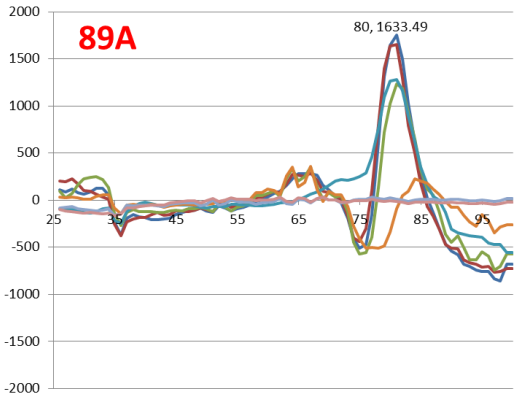

10G
74
16A
77
17A
61
18A
70
19A
78
21A
75
22A
76
23A
76

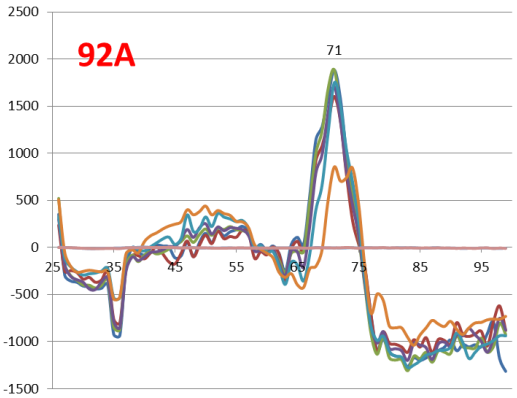

25A
76
26A
75
28A
70
41A
43A
79
44A
72
45A
46A
75

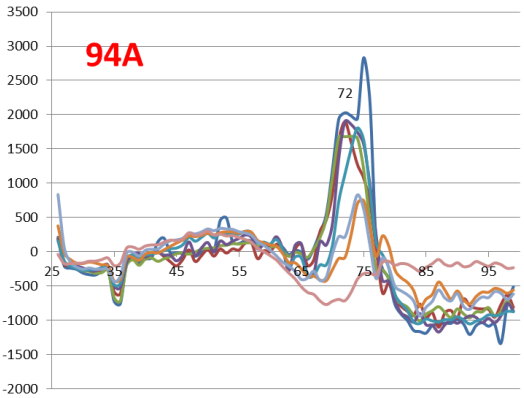

47A
72
2000
1500
1000
500
0
-500
-1000
-1500
-2000
25
35
45
55
65
75
85
95

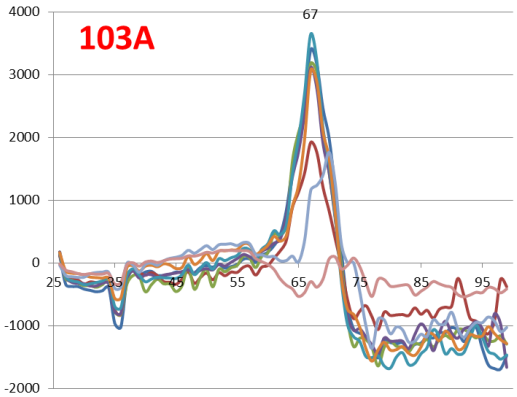

81A
68
800
600
400
200
0
-200
-400
-600
-800
25
35
45
55
65
75
85
95


84A
59
4000
3000
2000
1000
0
-1000
-2000
-3000
25
35
45
55
65
75
85
95


89A
80, 1633.49
2000
1500
1000
500
0
-500
-1000
-1500
-2000
25
35
45
55
65
75
85
95


92A
71
2500
2000
1500
1000
500
0
-500
-1000
-1500
25
35
45
55
65
75
85
95


94A
72
3500
3000
2500
2000
1500
1000
500
0
-500
-1000
-1500
-2000
25
35
45
55
65
75
85
95


103A
67
4000
3000
2000
1000
0
-1000
-2000
25
35
45
55
65
75
85
95

## 9.2.4 Chapter 4-Intrinsic Scaffolds

### 9.2.4.1 Additional Coiled-Coils

In Chapter 4 we refer to a range of trialled coiled-coils on the wild-type CTB. The additional coils listed below were also tried but were less effective at producing high order assemblies. The DeCr-coils are listed for reference only.

```
DeCr-coilK GGGKVSALKEKVSALKEKVSALKEKVSALKEKVSALKEG
DeCr-coilE GGGEVSALEKEVSALEKEVSALEKEVSALEKEVSALEKG
JR-coil2   ------------------GVSALEKEVSALKEKVSALEF
JR-coil3   ----------GVSALEKEVSALKEKVSALEKEVSALKF
JR-coil4   ----------GVSALSEKVSALESKVSALESKVSALEF
JR-coil5   ----------GVSALSEKVSALESKVSANESKVSALEF
JR-coil7   ----------------GGISAISEKISAIESKISAIES
JR-coil8   ----------------------GGISAISEKISAIES
JR-coil9   --------------------------GGISEKISAIES
```

**JR-coil3:** In case JR-coil2 proves too short for assembly to occur a 'four heptad' coil, minus N-terminal electrostatics, was created which should promote increased rates of assembly. However an increase in coil length adds to the potential that internal overcrowding of the VLP will become problematic. This coil also contains electrostatic interactions which may allow competition between parallel and antiparallel interactions. **JR-coil4:** This coil was designed to alleviate the potential for antiparallel interactions by reshuffling the electrostatic interactions while also maintaining homodimerisation. An additional note for concern is as coiled-coil length increases, the five coils protruding from the C-terminal face of a CTB pentamer have more likely-hood to interact with each other. This is of little concern for DeCr-coilK and E variants due to their electronic repulsion, however with the homomeric coils this is a perceivable problem. **JR-coil5:** To reduce the affinity of the c-

terminal half of these coiled-coils a ‘d’ position asparagine was introduced into heptad three of four. This was hoped to reduce the overall affinity of the C-terminal half, promoting coiled-coil interaction inter-CTB rather than intra-CTB, as intra-CTB interactions cannot occur over the first heptad as they are not in proximity, based on the position of the C-terminus, however, inter-CTB interactions are able to form interactions across the length of the coiled-coil. **JR-coil8 and JR-coil9:** These coils are truncations of JR-coil7. As mutations, discussed in later chapters, are introduced to CTB and refined to mediated increases in assembly independent of the coiled-coils, these truncations will be used to bridge the gap between large coils and no coils.

### 9.2.4.2 Cloning for pMaCo

Cytosolic expression was required to the MBP-coiled-coil fusions, thus pMAL-c5x was used as a backbone plasmid (Figure 9-8). pSAB2.1 contains a modified linker region which was introduced into pMAL-c5x through a digestion and ligation of the SacI – PstI insert from pSAB2.1 into pMAL-c5x. This created the pMaCo plasmid. Assembly PCR inserts of the coiled-coil sequences where ligated into the pMaCo plasmid after BamHI – PstI digestion. This produced a plasmid which expressed MBP-coiled-coil to the cytosol.

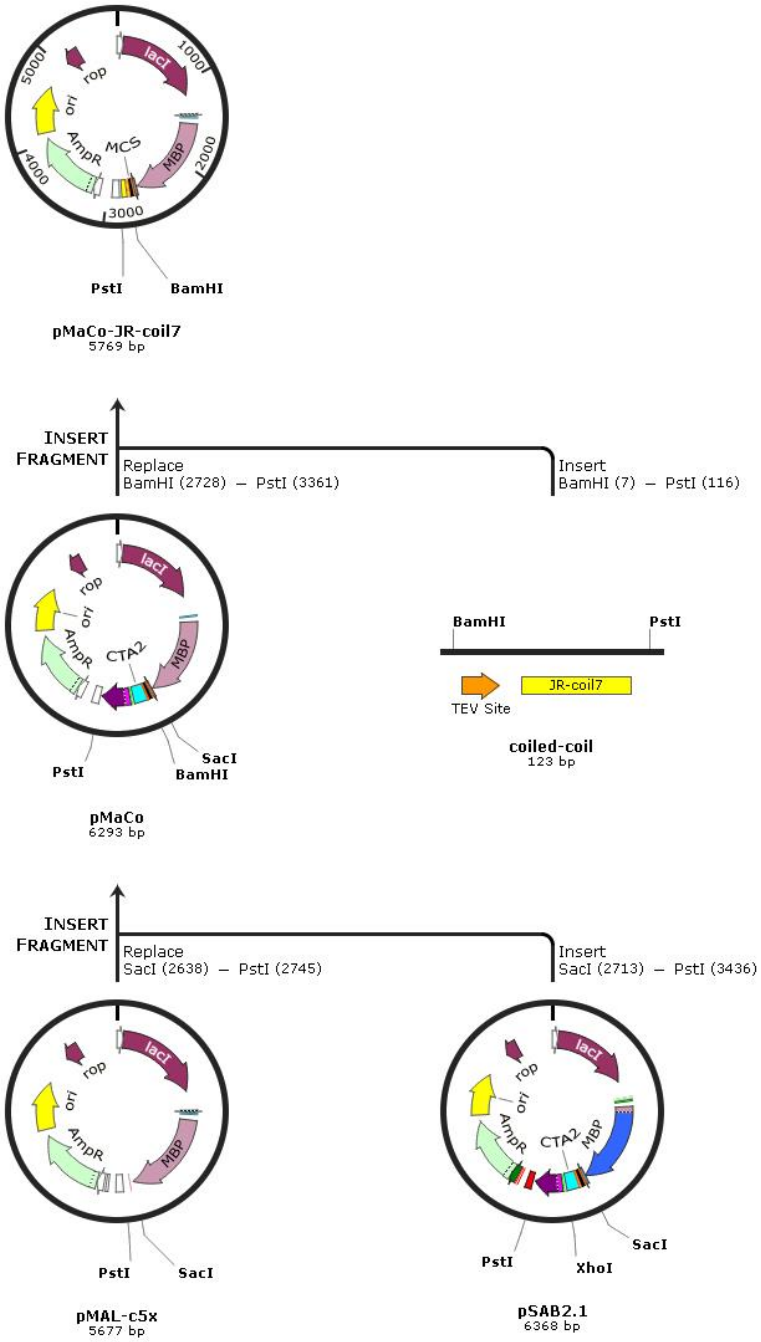



**Figure 9-8:** Cloning methodology for the construction of the pMaCo plasmid and splicing of coiled-coil fusions into pMaCo.

#### 9.2.4.2.1 pMaCo Assembly PCR Parts

Each assembly PCR used the terminal primers MaCoWtFT and MaCoWtRT with one part A, B and C. Part A (`MaCoWtA0`) was the same for all reactions. Parts B and C code for the coiled-coil. The Tb parts are for the JR-coiled-coil extensions requiring B and C relative to the number of the coil required. Dk parts construct the de Crescenzo coilK[64], De parts construct the de Crescenzo coilE, St parts construct the SynTri coil[164] To construct the JR-coil2 insert, assembly PCR is conducted with part-A `MaCoWtA0`, part-B `MaCoTbB2/3`,

part-C MaCoTbC3, and the terminal primers. The protocol for assembly PCR from the methods is followed with a 55 °C annealing temperature.

```
MaCoWtFT    CACATGGGATCCGAGAAC
MaCoWtRT    GCTTACCTGCAGAACTTAG
MaCoWtA0    CACATGGGATCCGAGAACCTGTACTTTCAGGGTAATAACGG
MaCoTbB2/3 GAGACTTTCTCCTTCAAAGCTGATACCTCTTTTTCCAGGGCCGACACACCGTTATTACCCTGAAAGTAC
MaCoTbC2    GCTTACCTGCAGAACTTAGAACTCGAGTGCGGAGACTTTCTCCTTCAAAGC
MaCoTbC3    GCTTACCTGCAGAACTTAGAACTTCAAAGCTGATACCTCTTTCTCGAGTGCGGAGACTTTCTCCTTCAAAGC
MaCoTbB4/5 GAGACTTTCGACTCCAAAGCTGATACTTTCTCAGACAGGGCCGACACACCGTTATTACCCTGAAAGTAC
MaCoTbC4    GCTTACCTGCAGAACTTAGAACTCCAAAGCTGATACCTTGCTTTCGAGTGCGGAGACTTTCGACTCCAAAGC
MaCoTbC5    GCTTACCTGCAGAACTTAGAACTCCAAAGCTGATACCTTGCTTTCATTTGCGGAGACTTTCGACTCCAAAGC
MaCoTbB7    CGATAGCTGAAATTTTCTCAGAAATGGCCGAAATACCACCGTTATTACCCTGAAAGTAC
MaCoTbC7    GCTTACCTGCAGAACTTAGGATTCAATTGCGGAGATTTTCGACTCGATAGCTGAAATTTTCTCAG
MaCoTbB8    CGATAGCTGAAATTTTCTCAGAAATGGCCGAAATACCCCGTTATTACCCTGAAAGTAC
MaCoTbB9    CGATAGCTGAAATTTTCTCAGAAATACCACCGTTATTACCCTGAAAGTAC
MaCoTbC8/9 GCTTACCTGCAGAACTTAGGACTCGATAGCTGAAATTTTCTCAG
MaCoDkB0    CTCTTTGAGTGCGGAGACTTTCTCCTTAAGAGCCGAAACCTTTTCCTTGAGGGCAGACACTTTACCGTTATTACCCTGAAAGTAC
MaCoDkC0    GCTTACCTGCAGAACTTAgccCTCCTTCAAAGCTGATACCTTTTCTTTCAGGGCCGACACCTTCTCTTTGAGTGCGGAGAC
MaCoDeB0    CGAGTGCGGAGACTTCCTTCTCCAATGCAGAAACCTCTTTTTCCAGCGCGGAAACCTCACCGTTATTACCCTGAAAGTAC
MaCoDeC0    GCTTACCTGCAGAACTTAgccCTTCTCCAAAGCTGATACCTCTTTTTCCAGGGCCGACACCTCTTTCTCGAGTGCGGAGACTTCC
MaCoStB0    GACTCTAAAGCGGCGACCTTCTTCTCAAGAGCCTCAACCCTCCCGTTATTACCCTGAAAGTAC
MaCoStC0    GCTTACCTGCAGAACTTAgCCATGTTCGAGTGCCTCAACCTTCTTCTCCAGGGCCTGAACTTTAGACTCTAAAGCGGCGAC
```

### 9.2.5 Chapter 5-Computational Interface Design

Scripts are included on the accompanying DVD for the computational design protocol.

### 9.2.6 Chapter 6-Characterisation of Scaffolded Mutants

#### 9.2.6.1 SDS PAGE Tryptic Digest.

Dr James Ault, University of Leeds, supplied the following description of the methodology for the SDS PAGE tryptic digest.

##### 9.2.6.1.1 Gel processing and tryptic digestion.

Gel bands were excised and chopped into small pieces ( ~ 1 $mm^3$ ), covered with 30 % ethanol in a 1.5 mL microcentrifuge and heated to 70 °C for 30 min with shaking. The supernatant was removed and replaced with fresh ethanol solution was again heated to 70 °C for 30 min. This was repeated until all coomassie stain was removed from the gel. The destain solution was replaced with 50 µL 25 mM ammonium bicarbonate and was vortexed for one hour. The wash solution was discarded and the gel slices covered with 25 mM ammonium bicarbonate/50% acetonitrile and vortexed for ten minutes. The gel slices were then covered with 100% acetonitrile and left for five minutes with vortexing before the supernatant was discarded and replaced with a fresh aliquot of acetonitrile. Acetonitrile was removed and the gel pieces were completely dried under vacuum centrifugation for 30 mins. Once dry, the gel slices were cooled on ice. The gel slices were then covered with ice cold trypsin solution (20 ng $\mu L^{-1}$ in 25 mM ammonium bicarbonate) and left on ice for 30 mins to rehydrate. Excess trypsin solution was removed and the gel slices were covered with a minimal amount of 25 mM ammonium bicarbonate. After briefly vortexing and centrifuging, the gel slices were incubated at 37 °C with shaking for 18 hours. The resulting

digest was vortexed, centrifuged and 50 µL water was added. Following vortexing for 10 mins, the supernatant was recovered and added to an eppendorf containing 5 µL acetonitrile/ water/ formic acid (60/35/5; v/v). 50 µL acetonitrile/ water/ formic acid (60/35/5; v/v) was added to the gel slices and vortexed for an additional 10 mins. The supernatant was pooled with the previous wash and one additional wash of the gel slices was performed. The pool of three washes was dried by vacuum centrifugation. The peptides were reconstituted in 20 µL acetonitrile/ water/ formic acid (2/97.9/0.1; v/v).

#### 9.2.6.1.2 Liquid chromatography-mass spectrometry.

LC separation of the peptide mixtures was performed on an Ultimate 3000 nano LC system. 2 µL of each sample in water was loaded onto a C18 guard column and washed with 2% acetonitrile/0.1% formic acid for 5 min at 25 µL $min^{-1}$. After valve switching, the peptides were then separated on a PepMap C18, 100 µm i.d. x 15 cm analytical column (Dionex, Amsterdam, NL) by gradient elution of 2-60% solvent B in A over 60 min. at 0.3 µL $min^{-1}$. Solvent A was 0.05% formic acid in 98% water/2% acetonitrile, solvent B was 0.05% formic acid in 20% water/80% acetonitrile.

The column eluant was directly interfaced to a quadrupole-ion mobility - orthogonal time of flight mass spectrometer (Synapt HDMS, Waters UK, Manchester) via a Z-spray nanoflow electrospray source. The MS was operated in positive TOF mode using a capillary voltage of 3.2 kV, cone voltage of 25 V, backing pressure of 2.47 mbar and a trap bias of 4 V. The source temperature was 80°C. Argon was used as the buffer gas at a pressure of $5.0 \times 10^{-4}$ mbar in the trap and transfer regions. Mass calibration was performed by a separate injection of sodium iodide at a concentration of 2 µg/ µl. GluFib was infused as a lock mass calibrant with a one second lock spray scan taken every 30 s during acquisition. Ten scans were averaged to determine the lock mass correction factor. Data acquisition was using data dependent analysis with a one second MS over *m/z* 350-3000 being followed by three 1 s MS/MS over *m/z* 50-2000 taken of the three most intense ions in the MS spectrum. CE applied was dependent upon charge state and mass of the ion selected. Dynamic exclusion of 60 s was used. Data processing was performed using the MassLynx v4.1 suite of software supplied with the mass spectrometer. Peptide MS/MS data were processed with ProteinLynx Global Server (Waters) and searched against UniProtKB/SwissProt database (release 2012_11).

The MS/MS data is verbose. Two examples will be presented here, CTB and a contaminant.

#### 9.2.6.1.3 Band 1

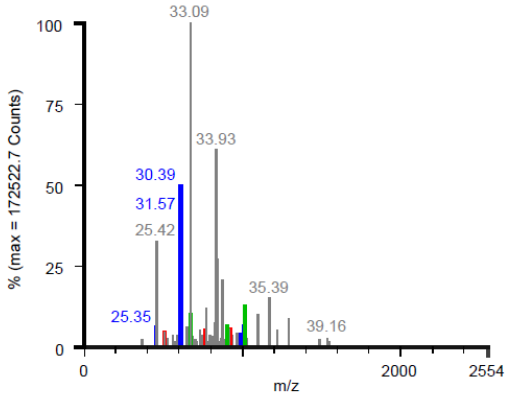
***Protein Summary***

| Accession Number | RMS Mass Error | Coverage | Number of Matches |
|---|---|---|---|
| 1310-165_2 | 9.414 | 84.3972 | 13 |
| TRYP_PIG | 2.5324 | 12.987 | 2 |
| CI117_DANRE | 58.4915 | 2.3355 | 1 |
| RD3_MOUSE | 28.3407 | 4.6154 | 1 |
| Y1321_PYRHO | 52.6535 | 4.1543 | 1 |
| YABN_BACSU | 20.6291 | 2.045 | 1 |

Where 1320-165_2 is the sequence for IGFBAA-DeCrE. Also identified were TRYP_PIG, the trypsin protease used and four false positives from extraneous organisms.

***1310-165_2***

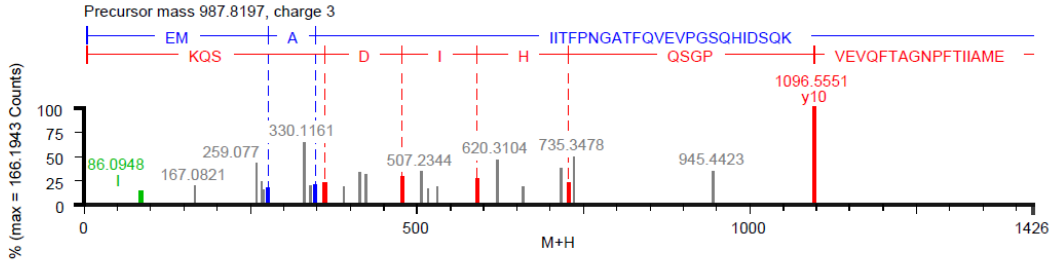
| | |
|---|---|
| *Description* | *James Ross mutant 2* |
| *Entry* | *1310-165_2* |
| *Coverage* | *84.3972* |
| *Molecular Weight* | *15433* |
| *No. Matches* | *15* |
| *RMS Error* | *4.6014* |
| *Spectrum* | *D:\Data\fbsjrau_131016_LC_02.raw* |



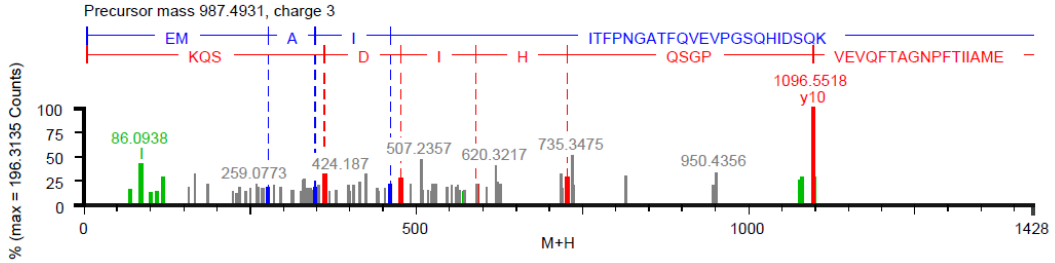

***Peptide Identifications***

| Mass | M/Z | Delta(ppm) | Modifications | Sequence |
|---|---|---|---|---|
| 2960.422 | 987.8197 | -4.5357 | Oxidation M (2), Deamidation N (9) | EMAIITFPNGATFQVEVPGSQHIDSQK |
| 2959.4382 | 987.4931 | -5.9396 | Oxidation M (2) | EMAIITFPNGATFQVEVPGSQHIDSQK |
| 2959.4382 | 987.4931 | -5.9396 | None | EMSIITFPNGATFQVEVPGSQHIDSQK |
| 2944.4272 | 982.4851 | -1.5754 | Deamidation N (9) | EMAIITFPNGATFQVEVPGSQHIDSQK |
| 2943.443 | 982.1567 | -1.1612 | None | EMAIITFPNGATFQVEVPGSQHIDSQK |
| 2765.4265 | 922.816 | 0.7063 | None | TPHAIAAISMAGGGEVSALEKEVSALEK |
| 2717.2886 | 906.7683 | 2.6056 | Phosphoryl STY (12) | TPLNIIDLCAEYHNTQIHTLNSK |
| 2025.02 | 676.0132 | 1.929 | Oxidation M (10) | TPHAIAAISMAGGGEVSALEK |
| 2009.0251 | 1005.5256 | -5.1647 | None | TPHAIAAISMAGGGEVSALEK |
| 1530.8141 | 511.278 | 2.3125 | None | EVSALEKEVSALEK |
| 1530.8141 | 511.278 | 2.3125 | None | EVSALEKEVSALEK |
| 1530.8141 | 511.278 | 2.3125 | None | EVSALEKEVSALEK |
| 1370.7194 | 457.912 | 4.9871 | None | IFSYTESLAGKR |
| 1214.6183 | 608.3181 | -1.9095 | None | IFSYTESLAGK |
| 907.5015 | 454.7553 | 7.1965 | None | IAYLTEAK |

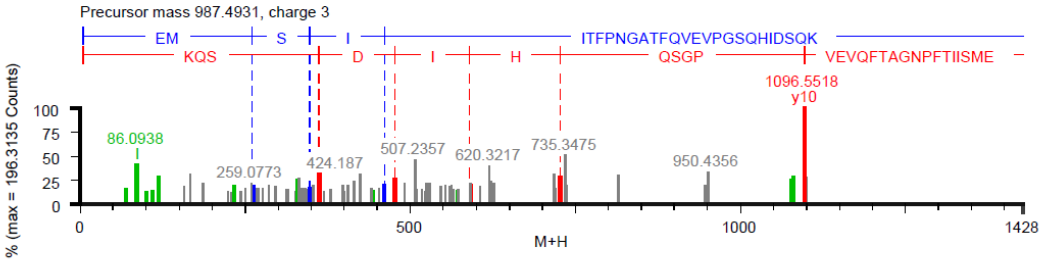
***Fragmentation for Peptide - 2960.422***



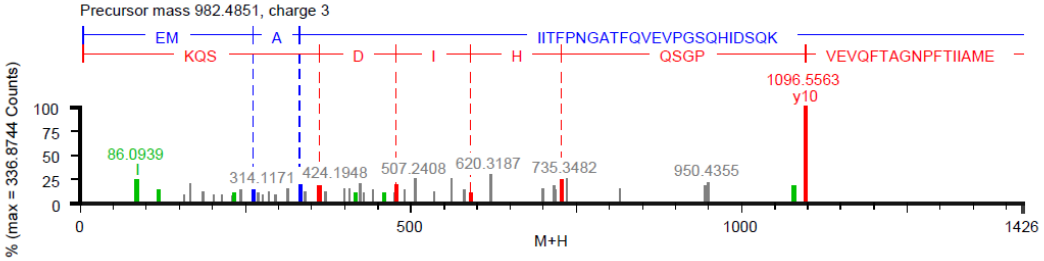

***Fragmentation for Peptide - 2959.4382***



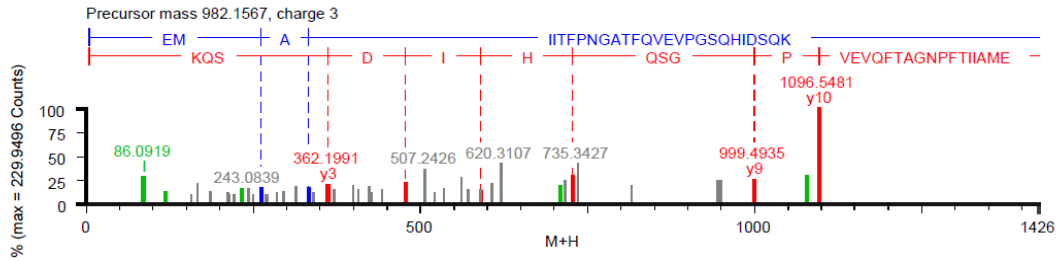

***Fragmentation for Peptide - 2959.4382***



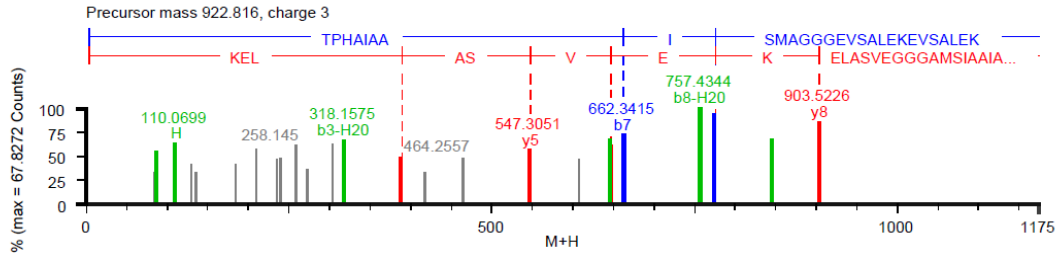

***Fragmentation for Peptide - 2944.4272***



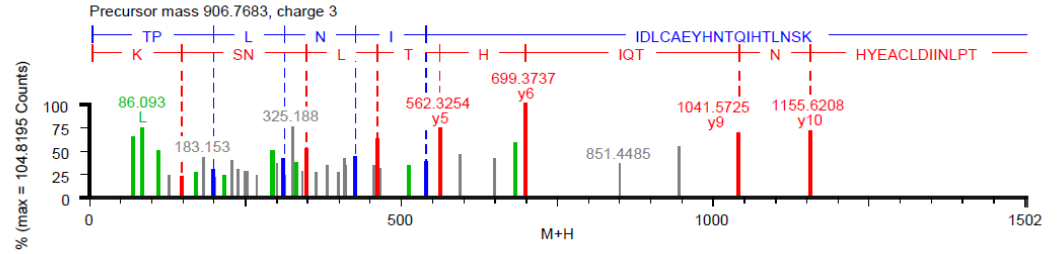
***Fragmentation for Peptide - 2943.443***



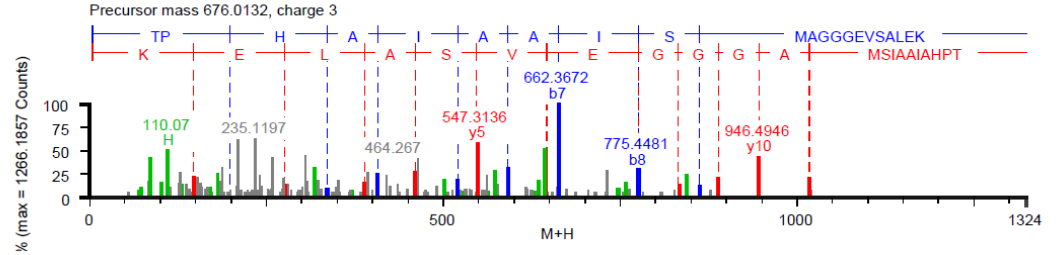

***Fragmentation for Peptide - 2765.4265***



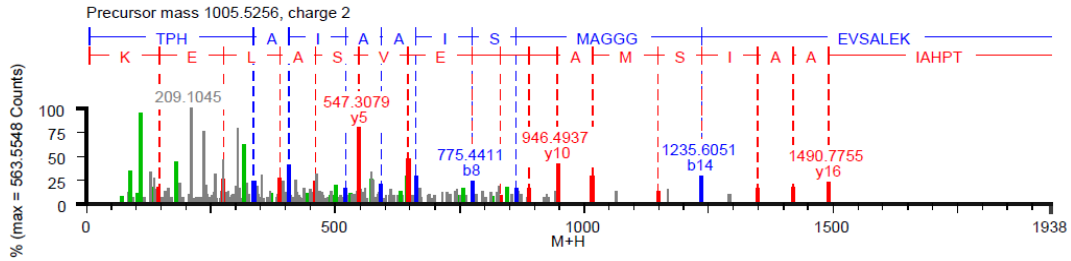

***Fragmentation for Peptide - 2717.2886***



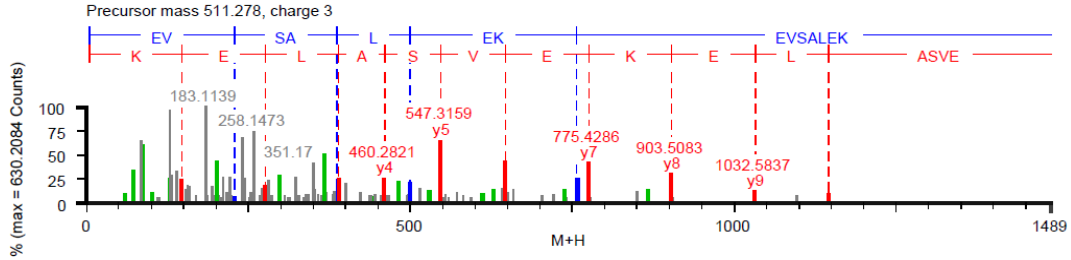

***Fragmentation for Peptide - 2025.02***



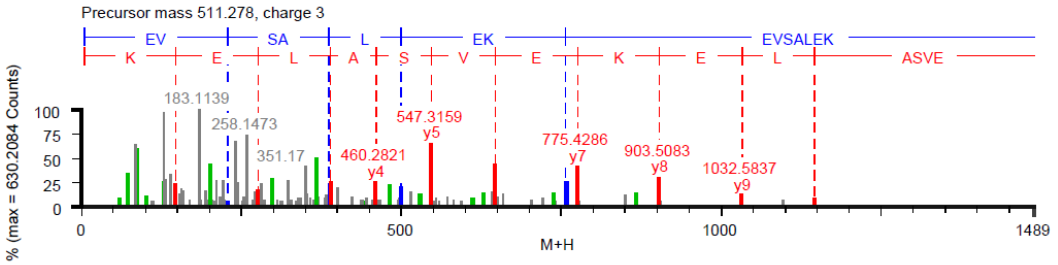
***Fragmentation for Peptide - 2009.0251***



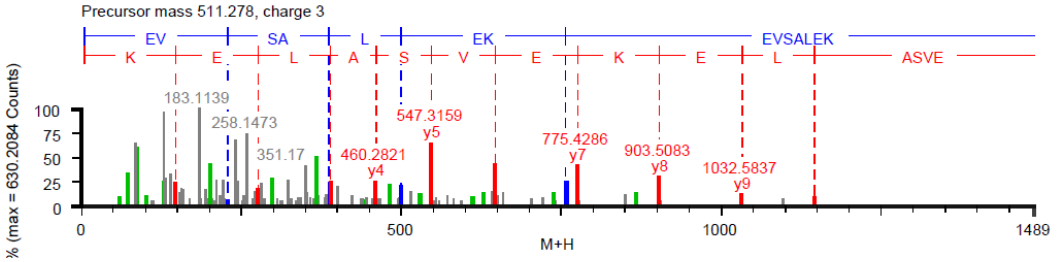

***Fragmentation for Peptide - 1530.8141***



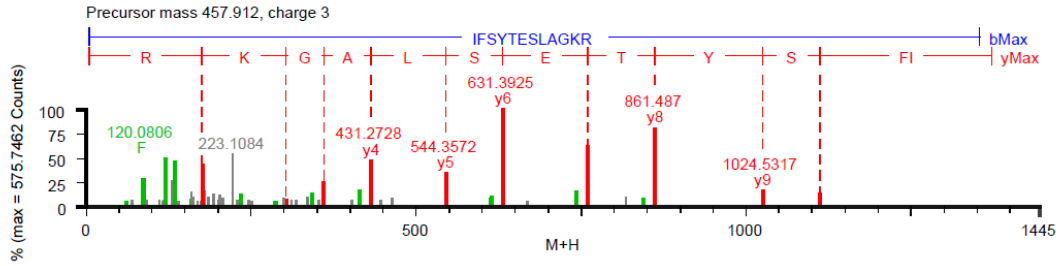

***Fragmentation for Peptide - 1530.8141***



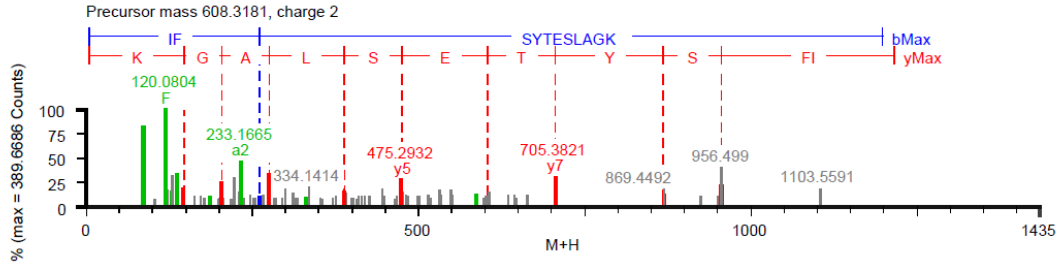

***Fragmentation for Peptide - 1530.8141***



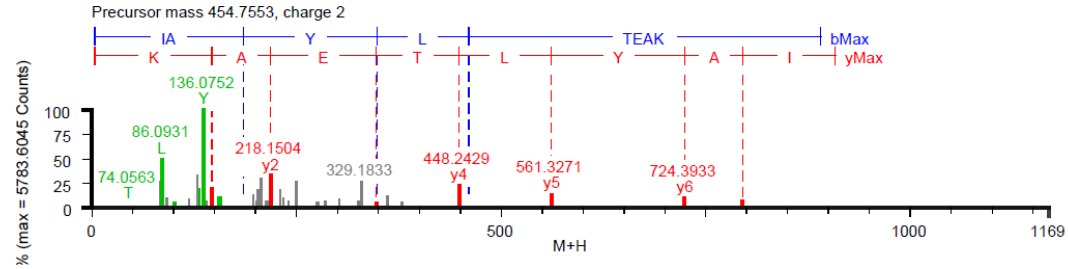
***Fragmentation for Peptide - 1370.7194***




***Fragmentation for Peptide - 1214.6183***




***Fragmentation for Peptide - 907.5015***



### 9.2.6.1.4 Band 4

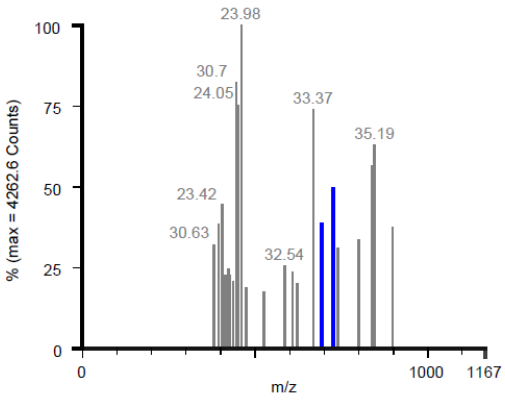
***Workflow:*** ***131018_1310-165_uniport***

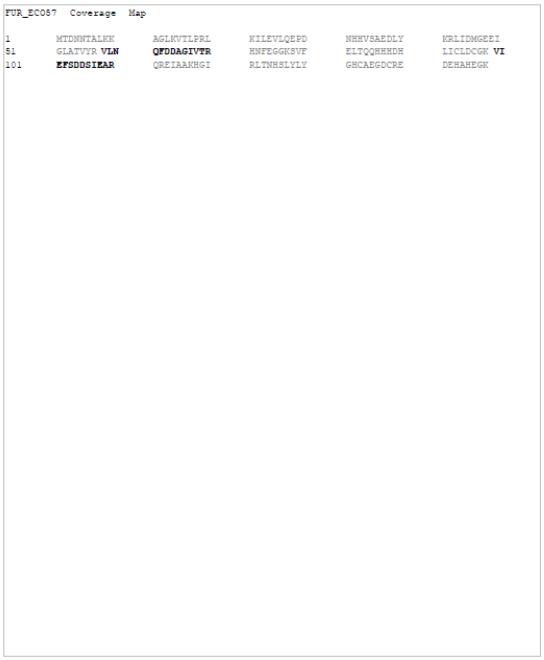
***Protein Summary***

| Accession Number | RMS Mass Error | Coverage | Number of Matches |
|---|---|---|---|
| FUR_ECO57 | 2.0296 | 16.8919 | 2 |
| RS15_ECO24 | 3.3855 | 15.7303 | 2 |
| SCP_CHIOP | 2.7562 | 100.0 | 1 |
| TRY1_CANFA | 0.1169 | 4.065 | 1 |

***FUR_ECO57***

| | |
|---|---|
| *Description* | *Ferric uptake regulation protein OS Escherichia coli O157 H7 GN fur PE 3 SV 1* |
| *Entry* | *FUR_ECO57* |
| *Coverage* | *16.8919* |
| *Molecular Weight* | *16784* |
| *No. Matches* | *2* |
| *RMS Error* | *2.0296* |
| *Spectrum* | *D:\Data\fbsjrau_131016_LC_06.raw* |

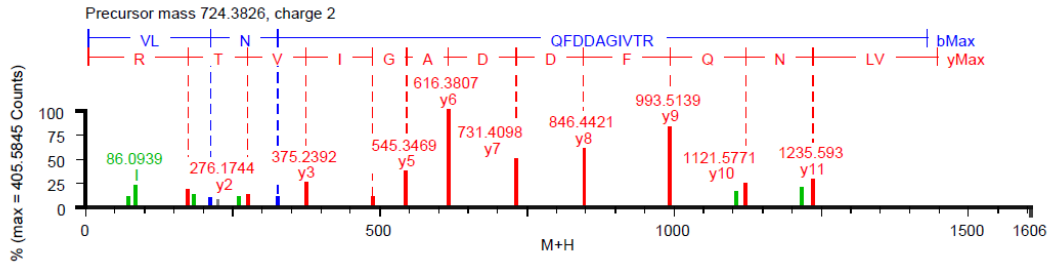


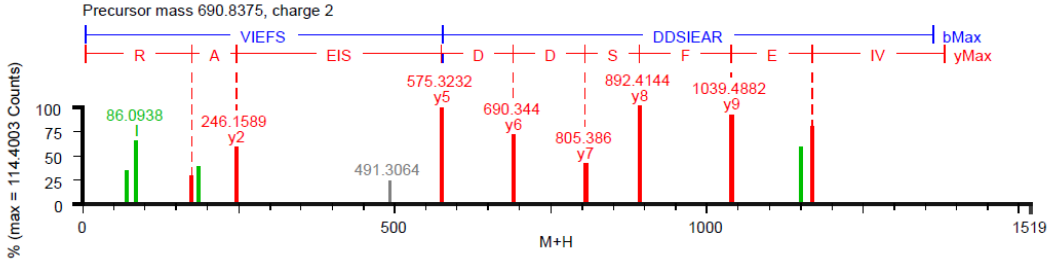

***Peptide Identifications***

| Mass | M/Z | Delta(ppm) | Modifications | Sequence |
|---|---|---|---|---|
| 1446.7466 | 724.3826 | -2.1094 | None | VLNQFDDAGIVTR |
| 1379.6567 | 690.8375 | -1.9465 | None | VIEFSDDSIEAR |

***Fragmentation for Peptide -*** ***1446.7466***

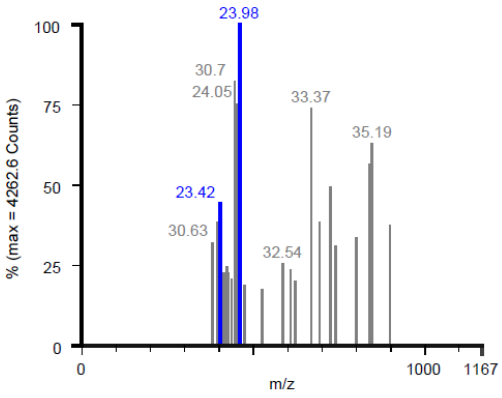


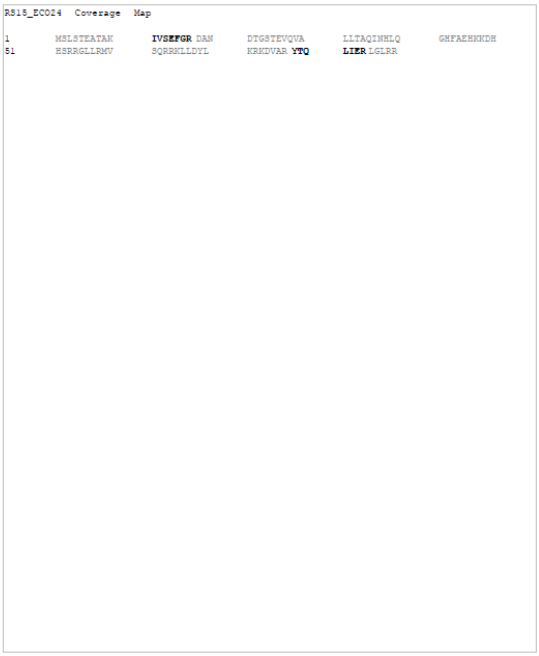

***Fragmentation for Peptide -*** ***1379.6567***



**RS15_ECO24**

| | |
|---|---|
| Description | 30S ribosomal protein S15 OS Escherichia coli O139 H28 strain E24377A ETEC GN rpsO PE |
| Entry | RS15_ECO24 |
| Coverage | 15.7303 |
| Molecular Weight | 10262 |
| No. Matches | 2 |
| RMS Error | 3.3855 |
| Spectrum | D:\Data\fbsjrau_131016_LC_06.raw |

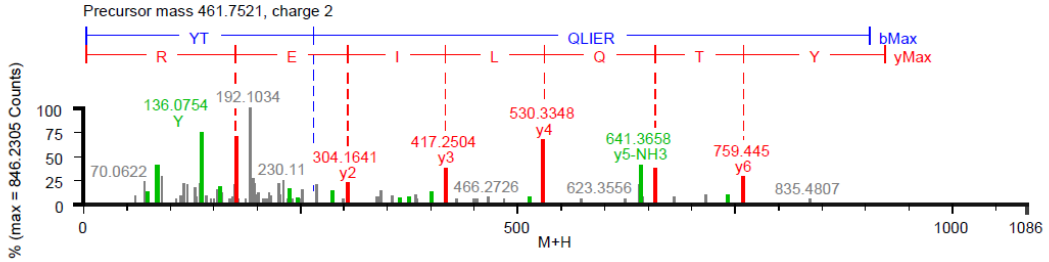



***Peptide Identifications***

| Mass | M/Z | Delta(ppm) | Modifications | Sequence |
|---|---|---|---|---|
| 921.492 | 461.7521 | 3.7092 | None | YTQLIER |
| 806.4286 | 404.2209 | 3.0274 | None | IVSEFGR |

***Fragmentation for Peptide - 921.492***



### 9.2.7 Chapter 7- Further Capsid Mutant Strategies

Mutants, mutations, molecular weights and yields.

| Name | Mutation | Coil | MW | Protein on Gel | Yeild (g/L) | conc (mg/ml) |
|---|---|---|---|---|---|---|
| AAAAAA-- | WT | none | 11615.3 | Yes | 15.6 | 9.5 |
| BAAAAA-- | 1T-WT | none | 11645.3 | Yes | 15.0 | 10.5 |
| BAAAAAAA | 1T-WT | JR-coil2 | 13861.9 | Yes | 5.1 | 4.7 |
| BAAAAAAB | 1T-WT | JR-coil3 | 14617.8 | n/d | n/d | n/d |
| BAABAACC | 1T-WT | JR-coil4 | 14493.6 | Yes | 0.9 | 4.2 |
| BAABAACD | 1T-WT | JR-coil5 | 14494.5 | Yes | 7.1 | 7.7 |
| BAAAAAEF | 1T-WT | JR-coil7 | 13817.8 | Yes | 8.1 | 15.0 |
| BAABAAG- | 1T-WT | JR-coil8 | 13089.0 | Yes | 5.4 | 7.8 |
| BAABAAH- | 1T-WT | JR-coil9 | 12817.6 | Yes | 5.7 | 9.0 |
| BAABAA-- | 1T-WT | JR-ELP2 | 15573.0 | yes | 1.5 | 0.3 |
| DAAAAA-- | 1T;3A | none | 11588.3 | Yes | 1.5 | 4.9 |
| EAAAAA-- | 1T;4A | none | 11602.3 | Yes | 0.9 | 2.9 |
| FAAAAA-- | 1T;6A | none | 11615.3 | Yes | 1.4 | 4.6 |
| ?AAAAA-- | 7A | none | 11571.3 | Yes | 1.4 | 2.7 |
| ??AAAA-- | 10G | none | 11601.3 | Yes | 1.9 | 3.8 |
| A?AAAA-- | 16A | none | 11558.3 | Yes | 1.6 | 3.2 |
| A?AAAA-- | 17A | none | 11573.2 | trace | 0.6 | 1.9 |
| BCAAAA-- | 1T;18A | none | 11553.3 | yes | 2.3 | 7.7 |
| A?AAAA-- | 19A | none | 11585.3 | Yes | 1.6 | 3.1 |
| A?AAAA-- | 21A | none | 11572.3 | Yes | 0.6 | 1.2 |
| A?AAAA-- | 23A | none | 11558.2 | Yes | 1.6 | 5.3 |
| A?AAAA-- | 25A | none | 11539.2 | Yes | 1.0 | 5.1 |
| A?AAAA-- | 26A | none | 11599.3 | Yes | 1.2 | 4.9 |
| A?AAAA-- | 28A | none | 11585.3 | Yes | 1.4 | 4.6 |
| A??AAA-- | 33E | none | 11687.4 | Yes | 0.8 | 2.5 |
| AA?AAA-- | 41A | none | 11585.3 | Yes | 1.3 | 6.7 |
| AA?AAA-- | 43A | none | 11558.2 | Yes | 1.1 | 6.2 |
| BACAAA-- | 1T;44A | none | 11602.3 | trace | 0.4 | 1.3 |
| AA?AAA-- | 45A | none | 11601.3 | no | 0.4 | 1.3 |
| AA?AAA-- | 46G | none | 11601.3 | Yes | 2.0 | 6.5 |
| BADAAA-- | 1T;47A | none | 11603.3 | Yes | 3.0 | 10.0 |
| BAEAAA-- | 1T;47T | none | 11633.3 | yes | 3.0 | 10.1 |
| AAAA?A-- | 81A | none | 11558.2 | Yes | 1.3 | 4.3 |
| BAAACA-- | 1T;83A | none | 11587.3 | Yes | n/d | n/d |
| BAAADA-- | 1T;84A | none | 11588.3 | Yes | 2.6 | 8.6 |
| AAAA?Q-- | 88K | none | 11557.3 | Yes | n/d | n/d |
| AAAA??-- | 89A | none | 11572.3 | Yes | 0.2 | 1.7 |
| BAAAAD-- | 1T;92A | none | 11615.3 | Yes | 2.0 | 6.8 |
| BAAAAJ-- | 103G | none | 11588.3 | Yes | 2.5 | 8.3 |
| BAFAAA-- | 1T;43P;47T | none | 11602.2 | yes | 2.5 | 8.4 |
| BAHBAA-- | 1T;47F | none | 11679.4 | yes | 6.5 | 1.3 |
| BAIAAA-- | 1T;47L | none | 11645.3 | yes | 9.5 | 7.3 |
| BAJAAA-- | 1T;43F | none | 11664.4 | yes | 4.5 | 1.7 |
| BAKAAA-- | 1T;43M | none | 11648.4 | yes | 8.2 | 4.7 |
| BALBAA-- | 1T;43Q | none | 11645.3 | yes | 13.5 | 2.7 |
| BAMAAA-- | 1T;41V | none | 11643.4 | yes | 1.0 | 1.6 |
| BAOAAA-- | 1T;41Q;47N | none | 11673.3 | yes | 7.4 | 9.0 |
| BAPAAA-- | 1T;44I | none | 11644.4 | no | 0.5 | 0.6 |
| BAQAAA-- | 1T;47N | none | 11646.3 | yes | 11.9 | 5.6 |
| BAAAEA-- | 83V | none | 11615.4 | yes | 1.4 | 4.8 |
| BAAAFA-- | 83L | none | 11629.4 | yes | 2.7 | 8.9 |
| BAAAGA-- | 83N | none | 11630.3 | yes | 13.9 | 9.8 |
| BAAAHA-- | 83M | none | 11647.4 | no | 2.7 | 1.3 |
| AAAAA?-- | 94A | none | 11549.3 | Yes | 1.7 | 5.5 |
| BAAAAN-- | 94F;103G | none | 11598.3 | yes | 1.1 | 3.8 |
| BAABAQI? | 1T;103S | JR-coil2 | 13751.7 | n/d | n/d | n/d |
| BALBGAAA | 1T;43Q;83N | JR-coil2 | 13846.9 | yes | 0.9 | 2.0 |
| BAQBGAAA | 47N;83N | JR-coil2 | 13847.8 | yes | 0.9 | 1.9 |
| BAQBGAEF | 47N;83N | JR-coil7 | 13803.7 | yes | n/d | n/d |
| BAQBKAAA | 1T;47N;83N;84M | JR-coil2 | 13850.9 | no | n/d | n/d |
| BAVBGAAA | 1T;43A;47N;83N | JR-coil2 | 13804.8 | yes | 5.4 | 5.3 |
| BAWBGAAA | 1T;47V;83N | JR-coil2 | 13832.9 | yes | 6.3 | 8.0 |
| BDAAAA-- | 1T;18A;19M;21F | none | 11682.4 | no | 0.7 | 2.2 |
| BEAAAA-- | 1T;18A19M;21Y | none | 11698.4 | no | 0.7 | 2.4 |
| BFAAAA-- | 1T;18A;19M;21M | none | 11666.5 | no | 0.7 | 2.3 |
| BGAAAA-- | 1T;18H;22S | none | 11591.3 | yes | 3.3 | 11.0 |
| BGQBGAEF | 1T;18H;22S;47N;83N | JR-coil7 | 13749.7 | yes | n/d | 1.8 |
| BHAAAA-- | 1T;18H;22C | none | 11607.4 | no | 0.8 | 2.7 |
| BIAAAA-- | 1T;18H;22C;26A | none | 11591.4 | no | n/d | n/d |
| BJAAAA-- | 1T;18W;22A;28L | none | 11636.4 | no | 0.5 | 1.5 |
| BKAAAA-- | 1T;18D | none | 11567.2 | yes | GOOD | n/d |
| BMABAA-- | 1T;18F | none | 11629.3 | - | - | - |
| BNAAAA-- | 1T;17F;18F | none | 11633.3 | yes | GOOD | n/d |

| Name | Mutation | Coil | MW | Protein on Gel | Yeild (g/L) | conc (mg/ml) |
|---|---|---|---|---|---|---|
| BNKBGAAA | 1T;17F;18F;43M;83N | JR-coil2 | 13867.9 | no | n/d | n/d |
| BOAAAA-- | 1T;22N | none | 11644.4 | yes | 0.6 | 2.1 |
| BPAAAA-- | 1T;22Y | none | 11663.4 | yes | low | n/d |
| BTQBGAAA | 1T;22A;47N;83N | JR-coil2 | 13803.8 | no | n/d | n/d |
| GAAAAA-- | 1T;7F | none | 11677.4 | no | 0.6 | 1.9 |
| HAABAA-- | 1T;3A;6I | none | 11600.3 | no | 0.3 | 1.1 |
| IAABAA-- | 1T;3L;6I | none | 11642.4 | yes | 3.7 | 12.4 |
| IAABAAAA | 1T;3L;6I | JR-coil2 | 13859.0 | yes | 1.8 | 3.2 |
| IAQBGAEF | 1T;3L;6I;47N;83N | JR-coil7 | 13800.8 | trace | n/d | n/d |
| IAVBGAEF | 1T;3L;6I;44A;47N;83A | JR-coil7 | 13757.8 | trace | 0.4 | 1.4 |
| IGABAAAA | 1T;3L;6I;18H;22S | JR-coil2 | 13804.9 | trace | 0.6 | 0.6 |
| IGFBAAAA | 1T;3L;6I;18H;22S;43P;47T | JR-coil2 | 13761.8 | yes | 1.3 | 2.7 |
| IGFAEAAA | 1T;3L;6I;18H;22S;43P;47T;83V | JR-coil2 | 13731.9 | yes/no | 0.7 | 0.4 |
| IGFAFAAA | 1T;3L;6I;18H;22S;43P;47T;83L | JR-coil2 | 13745.9 | no | 0.2 | 0.5 |
| IGFBAAEF | 1T;3L;6I;18H;22S;43P;47T | JR-coil7 | 13717.7 | yes | 2.0 | 5.0 |
| IGFBAAEG | 1T;3L;6I;18H;22S;43P;47T | JR-coil7-neg | 13759.8 | yes | n/d | 1.6 |
| IGFBAJN? | 1T;3L;6I;18H;22S;43P;47T | Syn-Tri | 14788.0 | trace | n/d | n/d |
| IGFBAJO? | 1T;3L;6I;18H;22S;43P;47T | DeCrK | 15443.6 | trace | n/d | 0.9 |
| IGFBAJP? | 1T;3L;6I;18H;22S;43P;47T | DeCrE | 15438.9 | yes | GOOD | 2.9 |
| IGFBGAEF | 1T;3L;6I;18H;22S;43P;47T;83N | JR-coil7 | 13702.7 | - | - | - |
| IGFBLQEH | 1T;3L;6I;18H;22S;43P;47T;83N;88K | JR-coil7-LPETG | 14541.6 | - | - | - |
| IGFBNAEH | 1T;3L;6I;18H;22S;43P;47T;83N | JR-coil7-LPETG | 14599.6 | - | - | - |
| IGQBAAEF | 1T;3L;6I;18H;22S;47N | JR-coil7 | 13761.8 | trace | n/d | n/d |
| IGQBGAEF | 1T;3L;6I;18H;22S;47N;83N | JR-coil7 | 13746.8 | trace | n/d | n/d |
| IGRBAAAA | 1T;3L;6I;18H;22S;41I;43P;47T | JR-coil2 | 13773.9 | no | n/d | n/d |
| IGSBAAAA | 1T;3L;6I;18H;22S;43P;44A;47T | JR-coil2 | 13718.8 | no | n/d | n/d |
| IGTBAAAA | 1T;3L;6I;18H;22S;43P;47L | JR-coil2 | 13773.9 | no | n/d | n/d |
| IGUBAAAA | 1T;3L;6I;18H;22S;43P;47M | JR-coil2 | 13791.9 | no | n/d | n/d |
| IGVBEAAA | 1T;3L;6I;18H;22S;44A;47N;83V | JR-coil2 | 13732.9 | yes | n/d | 2.0 |
| IGVBGAAA | 1T;3L;6I;18H;22S;44A;47N;83N | JR-coil2 | 13747.9 | trace | n/d | n/d |
| IGVBGAEF | 1T;3L;6I;18H;22S;44A;47N;83N | JR-coil7 | 13703.8 | trace | n/d | 1.1 |
| IQFBAAAA | 1T;3L;6I;18L;22S;43P;47T | JR-coil2 | 13737.9 | yes | 0.6 | 1.3 |
| IRFBAAAA | 1T;3L;6I;18F;22S;43P;47T | JR-coil2 | 13771.9 | no | n/d | n/d |
| ISFBAAAA | 1T;3L;6I;18H;19I;22S;43P;47T | JR-coil2 | 13773.9 | no | n/d | n/d |
| JAAAAA-- | 1T;3A;6L | none | 11600.4 | yes | 3.5 | 11.8 |
| JGABAAAA | 1T;3L;6L;18H;22S | JR-coil2 | 13762.9 | no | n/d | n/d |
| JGFAFAAA | 1T;3L;6L;18H;22S;43P;47T;83L | JR-coil2 | 13703.8 | no | 0.3 | 0.5 |
| KAAAAA-- | 1T;3L;6I;7Y | none | 11690.5 | no | 0.8 | 2.6 |
| LAAAAA-- | 1T;3L;6I;7L | none | 11640.5 | no | 0.7 | 2.2 |
| MLABAAAA | 1T;10D | JR-coil2 | 11689.3 | yes | 1.2 | 2.4 |
| MLQBAAAA | 1T;10D;47N | JR-coil2 | 13906.9 | yes | 1.2 | 2.5 |
| MLQBAAEF | 1T;10D;47N | JR-coil7 | 13862.8 | trace | n/d | n/d |
| MLXBAAAA | 1T;10D;47P | JR-coil2 | 13889.9 | yes | n/d | n/d |
| MUQBAAAA | 1T;10D;22G;47N | JR-coil2 | 13848.8 | no | n/d | n/d |
| MVQBAAAA | 1T;10D;22F;47N | JR-coil2 | 13938.9 | no | n/d | n/d |
| MWQBAAAA | 1T;10D;22R;47N | JR-coil2 | 13948.0 | no | n/d | n/d |
| MXQBAAAA | 1T;10D;22Y;47N | JR-coil2 | 13954.9 | no | n/d | n/d |
| NGFBAAAA | 1T;3L;6I;7F;18H;22S;43P;47T | JR-coil2 | 13793.9 | no | n/d | n/d |
| PGFBAAAA | 1T;3L;6I;7M;18H;22S;43P;47T | JR-coil2 | 13777.9 | no | n/d | n/d |
| QGFBAAAA | 1T;3L;6I;7W;18H;22S;43P;47T | JR-coil2 | 13833.0 | no | n/d | n/d |
| RAQBEAEF | 1T;3L;47N;83V | JR-coil7 | 13773.8 | yes | low | 2.0 |
| RAQBGAAA | 1T;3L;47N;83N | JR-coil2 | 13832.9 | yes | low | 2.0 |
| RAQBGAEF | 1T;3L;47N;83N | JR-coil7 | 13788.8 | yes | low | n/d |
| RAVBEAEF | 1T;3L;44A;47N;83V | JR-coil7 | 13730.8 | yes | low | n/d |
| RAVBGAEF | 1T;3L;44A;47N;83N | JR-coil7 | 13745.7 | yes | 2.6 | 1.5 |
| RAVBGAEG | 1T;3L;44A;47N;83N | JR-coil7-neg | 13787.8 | no | n/d | n/d |
| RAVBLQEH | 1T;3L;44A;47N;83N;88K | JR-coil7-LPETG | 14584.6 | trace | n/d | 1.6 |
| SAQBGAAA | 1T;6I;47N;83N | JR-coil2 | 13859.9 | yes | 3.0 | 2.9 |
| TAQBGAAA | 1T;7Y;47N;83N | JR-coil2 | 13895.9 | no | n/d | n/d |
| ULQBAAAA | 1T;3A;10D;47N | JR-coil2 | 13849.8 | yes | 3.8 | 3.6 |
| VLQBAAAA | 1T;6A;10D;47N | JR-coil2 | 13876.8 | yes | 3.2 | 1.6 |
| ??GAAA-- | 1T;10E;41V;43P (10E error) | none | 11670.4 | no | n/d | n/d |
| IGF?FAAA | see 10.1.3 (59N error) | JR-coil2 | 13744.9 | no | n/d | n/d |
| GDEA??-- | see 10.1.3 | none | 11821.7 | no | n/d | n/d |
| ?HFA?N-- | see 10.1.3 | none | 11482.3 | no | n/d | n/d |
| ?JGAEN-- | see 10.1.3 | none | 11507.4 | no | n/d | n/d |

### 9.2.8 Chapter 8- General Conclusions

No additional Information.

# Chapter 10

# Bibliography

1 He, Y., Ye, T., Su, M., Zhang, C., Ribbe, A., Jiang, W. & Mao, C. Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* **452**, 198-201, (2008).

2 Gibson, D. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52-56, (2010).

3 Heinemann, M. & Panke, S. Synthetic biology--putting engineering into biology. *Bioinformatics* **22**, 2790, (2006).

4 Chopra, P. & Kamma, A. Engineering life through synthetic biology. *In silico biology* **6**, 401-410, (2006).

5 Ball, P. Synthetic biology: starting from scratch. *Nature* **431**, 624-626, (2004).

6 Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536, (1996).

7 http://parts.mit.edu (29/10/2013).

8 http://biobricks.org/ (29/10/2013).

9 http://igem.org/Main_Page (29/10/2013).

10 http://insdc.org/ (29/10/2013).

11 Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics* **29**, 365-371, (2001).

12 http://www.rcsb.org/pdb/index.html (29/10/2013).

13 Webb, E. C. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. (Academic Press, 1992).

14 Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-531, (2003).

15 Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Res.* **31**, 418, (2003).

16 Ecker, J., Bickmore, W., Barroso, I., Pritchard, J., Gilad, Y. & Segal, E. Genomics: ENCODE explained. *Nature* **489**, 52-55, (2012).

17 Saeidi, N., Wong, C., Lo, T.-M., Nguyen, H., Ling, H., Leong, S., Poh, C. & Chang, M. Engineering microbes to sense and eradicate Pseudomonas aeruginosa, a human pathogen. *Mol. Syst. Biol.* **7**, 521, (2011).

18 Pinheiro, V. & Holliger, P. The XNA world: progress towards replication and evolution of synthetic genetic polymers. *Curr. Opin. Chem. Biol.* **16**, 245-252, (2012).

19 Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441-444, (2010).

20 Foo, J., Ching, C., Chang, M. & Leong, S. The imminent role of protein engineering in synthetic biology. *Biotechnol. Adv.* **30**, 541-549, (2012).

21 Kuhlman, B., Dantas, G., Ireton, G., Varani, G., Stoddard, B. & Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368, (2003).

22 Fleishman, S., Whitehead, T., Ekiert, D., Dreyfus, C., Corn, J., Strauch, E.-M., Wilson, I. & Baker, D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816-821, (2011).

23 Bryan, P. Protein engineering. *Biotechnol. Adv.* **5**, 221-234, (1987).

24 Drexler, K. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci. U.S.A.* **78**, 5275-5278, (1981).

25 Arnold, F. Combinatorial and computational challenges for biocatalyst design. *Nature* **409**, 253-257, (2001).

26 Venkatraman, J., Shankaramma, S. & Balaram, P. Design of Folded Peptides. *Chem. Rev* **101**, 3131-3152, (2001).

27 Tabor, S. & Richardson, C. A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl Acad. Sci. U.S.A.* **92**, 6339-6343, (1995).

28 Sismour, A., Lutz, S., Park, J.-H., Lutz, M., Boyer, P., Hughes, S. & Benner, S. PCR amplification of DNA containing non-standard base pairs by variants of reverse transcriptase from Human Immunodeficiency Virus-1. *Nucleic Acids Res.* **32**, 728-735, (2004).

29 Liu, D., Magliery, T., Pastrnak, M. & Schultz, P. Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. *Proc. Natl Acad. Sci. U.S.A.* **94**, 10092-10097, (1997).

30 Bromley, E., Channon, K., Moutevelis, E. & Woolfson, D. Peptide and protein building blocks for synthetic biology: from programming biomolecules to self-organized biomolecular systems. *ACS chem. biol.* **3**, 38-50, (2008).

31 Serrano, L. Synthetic biology: promises and challenges. *Mol. Syst. Biol.* **3**, (2007).

32 Fu, Y., Zeng, D., Chao, J., Jin, Y., Zhang, Z., Liu, H., Li, D., Ma, H., Huang, Q., Gothelf, K. & Fan, C. Single-step rapid assembly of DNA origami nanostructures for addressable nanoscale bioreactors. *J. Am. Chem. Soc.* **135**, 696-702, (2013).

33 http://www.parliament.uk/documents/post/postpn298.pdf (29/10/2013).

34 http://www.rcuk.ac.uk/documents/publications/SyntheticBiologyRoadmap.pdf (29/10/2013).

35 http://www.bbsrc.ac.uk/funding/opportunities/2007/synthetic-biology.aspx (29/10/2013).

36 http://www.biochemistry.org/Conferences/AllConferences/tabid/379/View/Conference/Filter/66%2050HD/MeetingNo/70HDN/Default.aspx (29/10/2013).

37 http://www.biochemistry.org/Conferences/AllConferences/tabid/379/Filter/64/Page/1/MeetingNo/SA143/view/Conference/Default.aspx (29/10/2013).

38 http://sb6.biobricks.org (29/10/2013).

39 Bordner, A. & Abagyan, R. Statistical analysis and prediction of protein-protein interfaces. *Proteins* **60**, 353-366, (2005).

40 Tsai, C. & Nussinov, R. Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. *Protein Sci.* **6**, 1426-1437, (1997).

41 Sheinerman, F. & Honig, B. On the role of electrostatic interactions in the design of protein-protein interfaces. *J. Mol. Biol.* **318**, 161-177, (2002).

42 Fernández, A. & Scheraga, H. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl Acad. Sci. U.S.A.* **100**, 113-118, (2003).

43 Kleanthous, C. *Protein-protein recognition*. Vol. 31 (Oxford University Press, USA, 2000).

44 Janin, J., Chothia, C., Shabb, J., Ng, L., Corbin, J., Bütikofer, P., Lin, Z., Chiu, D., Lubin, B. & Kuypers, F. The structure of protein-protein recognition sites. *Structure* **265**, (1990).

45 Young, L., Jernigan, R. & Covell, D. A role for surface hydrophobicity in protein protein recognition. *Protein Sci.* **3**, 717-729, (1994).

46 Conte, L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177-2198, (1999).

47 Jones, S. & Thornton, J. Principles of protein-protein interactions. *Proc. Natl Acad. Sci. U.S.A.* **93**, 13-20, (1996).

48 Bogan, A. & Thorn, K. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1-9, (1998).
49 Thorn, K. S. & Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**, 284, (2001).
50 Lichtarge, O., Bourne, H. & Cohen, F. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358, (1996).
51 Ofran, Y. & Rost, B. Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**, 377-387, (2003).
52 Lijnzaad, P. & Argos, P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins* **28**, 333-343, (1997).
53 Reichmann, D., Phillip, Y., Carmi, A. & Schreiber, G. On the contribution of water-mediated interactions to protein-complex stability. *Biochemistry* **47**, 1051-1060, (2008).
54 Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. U.S.A.* **98**, 4569, (2001).
55 Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K. & Boutilier, K. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-183, (2002).
56 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207, (2003).
57 Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422**, 216-225, (2003).
58 Alber, F., Eswar, N. & Sali, A. Structure determination of macromolecular complexes by experiment and computation. *Pract. Bioinf.*, 73-96, (2008).
59 Russell, R., Alber, F., Aloy, P., Davis, F., Korkin, D., Pichaud, M., Topf, M. & Sali, A. A structural perspective on protein-protein interactions. *Curr. Opin. Struc. Biol.* **14**, 313-324, (2004).
60 Fletcher, J., Boyle, A., Bruning, M., Bartlett, G., Vincent, T., Zaccai, N., Armstrong, C., Bromley, E., Booth, P., Brady, R., Thomson, A. & Woolfson, D. A basis set of de novo coiled-coil Peptide oligomers for rational protein design and synthetic biology. *ACS Synt. Biol.* **1**, 240-250, (2012).
61 Crick, F. The packing of-helices: simple coiled-coils. *Acta Crystallogr.* **6**, 689-697, (1953).
62 Phillips, G. Tropomyosin crystal structure and muscle regulation. *J. Mol. Biol.* **192**, 111-127, (1986).
63 Hodges, R., Saund, A., Chong, P., St-Pierre, S. & Reid, R. Synthetic model for two-stranded alpha-helical coiled-coils. Design, synthesis, and characterization of an 86-residue analog of tropomyosin. *J. Biol. Chem.* **256**, 1214, (1981).
64 De Crescenzo, G., Litowski, J., Hodges, R. & O'Connor-McCourt, M. Real-time monitoring of the interactions of two-stranded de novo designed coiled-coils: effect of chain length on the kinetic and thermodynamic constants of binding. *Biochemistry* **42**, 1754-1763, (2003).
65 Burton, A., Thomas, F., Agnew, C., Hudson, K., Halford, S., Brady, R. & Woolfson, D. Accessibility, Reactivity, and Selectivity of Side Chains within a Channel of de Novo Peptide Assembly. *J. Am. Chem. Soc.* **135**, 12524-12527, (2013).
66 Lai, Y.-T., Cascio, D. & Yeates, T. Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129, (2012).

67 Sinclair, J., Davies, K., Vénien-Bryan, C. & Noble, M. Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nature nanotechnology* **6**, 558-562, (2011).

68 Fletcher, J., Harniman, R., Barnes, F., Boyle, A., Collins, A., Mantell, J., Sharp, T., Antognozzi, M., Booth, P., Linden, N., Miles, M., Sessions, R., Verkade, P. & Woolfson, D. Self-assembling cages from coiled-coil peptide modules. *Science* **340**, 595-599, (2013).

69 Lai, Y.-T., King, N. & Yeates, T. Principles for designing ordered protein assemblies. *Trends Cell Biol.*, (2012).

70 Yeates, T. Nanobiotechnology: protein arrays made to order. *Nature nanotechnology* **6**, 541-542, (2011).

71 King, N., Sheffler, W., Sawaya, M., Vollmar, B., Sumida, J., André, I., Gonen, T., Yeates, T. & Baker, D. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171-1174, (2012).

72 Daggett, V. Protein folding-simulation. *Chem. Rev.* **106**, 1898-1916, (2006).

73 Biotechnology advancesKarplus, M. & Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl Acad. Sci. U.S.A.* **102**, 6679, (2005).

74 McGuffee, S. & Elcock, A. Atomically detailed simulations of concentrated protein solutions: the effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems. *J. Am. Chem. Soc.* **128**, 12098-12110, (2006).

75 Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153**, S7-S26, (2008).

76 Ritchie, D. W. Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci.* **9**, 1-15, (2008).

77 Gilson, M. K. & Zhou, H. X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21-42, (2007).

78 Goodfellow, J. & Levy, R. Theory and simulation. *Curr. Opin. Struc. Biol.* **10**, 137-138, (2000).

79 van Gunsteren, W. & Dolenc, J. Biomolecular simulation: historical picture and future perspectives. *Biochem. Soc. Trans.* **36**, 11-15, (2008).

80 Wassenaar, T., Ingólfsson, H., Priess, M., Marrink, S. & Schäfer, L. Mixing MARTINI: electrostatic coupling in hybrid atomistic-coarse-grained biomolecular simulations. *J. Phys. Chem. B* **117**, 3516-3530, (2013).

81 van Gunsteren, W. *et al.* Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed. Engl.* **45**, 4064-4092, (2006).

82 Stanton, C. L. & Houk, K. N. Benchmarking p K a Prediction Methods for Residues in Proteins. *J. Chem. Theor. Comput.* **4**, 951-966, (2008).

83 Chen, R., Li, L. & Weng, Z. ZDOCK: An initial stage protein docking algorithm. *Proteins* **52**, 80-87, (2003).

84 Ritchie, D. & Kemp, G. Protein docking using spherical polar Fourier correlations. *Proteins* **39**, 178-194, (2000).

85 Gray, J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. & Baker, D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331**, 281-299, (2003).

86 Christensen, N. & Kepp, K. Accurate stabilities of laccase mutants predicted with a modified FoldX protocol. *J. Chem. Inf. Modell.* **52**, 3028-3042, (2012).

87 Jiang, L., Althoff, E., Clemente, F., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J., Betker, J., Tanaka, F., Barbas, C., Hilvert, D., Houk, K., Stoddard, B. &

Baker, D. De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391, (2008).

88 Keskin, O., Ma, B. & Nussinov, R. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* **345**, 1281-1294, (2005).

89 Morrison, K. & Weiss, G. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **5**, 302-307, (2001).

90 Kortemme, T. & Baker, D. Computational design of protein-protein interactions. *Curr. Opin. Chem. Biol.* **8**, 91-97, (2004).

91 Thornton, J. The Hans Neurath Award lecture of The Protein Society: proteins-- a testament to physics, chemistry, and evolution. *Protein Sci.* **10**, 3-11, (2001).

92 Schreiber, G. & Fleishman, S. Computational design of protein-protein interactions. *Curr. Opin. Struc. Biol.*, (2013).

93 Der, B., Machius, M., Miley, M., Mills, J., Szyperski, T. & Kuhlman, B. Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J. Am. Chem. Soc.* **134**, 375-385, (2012).

94 Der, B., Edwards, D. & Kuhlman, B. Catalysis by a de novo zinc-mediated protein interface: implications for natural enzyme evolution and rational enzyme engineering. *Biochemistry* **51**, 3933-3940, (2012).

95 Zlotnick, A. & Mukhopadhyay, S. Virus assembly, allostery and antivirals. *Trends Microbiol.* **19**, 14-23, (2011).

96 Endres, D. & Zlotnick, A. Model-based analysis of assembly kinetics for virus capsids or other spherical polymers. *Biophys. J.* **83**, 1217-1230, (2002).

97 Keef, T., Micheletti, C. & Twarock, R. Master equation approach to the assembly of viral capsids. *J. Theor. Biol.* **242**, 713-721, (2006).

98 Elrad, O. & Hagan, M. Mechanisms of size control and polymorphism in viral capsid assembly. *Nano letters* **8**, 3850-3857, (2008).

99 Rapaport, D. Role of reversibility in viral capsid growth: a paradigm for self-assembly. *Phys. Lett.* **101**, 186101, (2008).

100 Nguyen, H., Reddy, V. & Brooks, C. Invariant polymorphism in virus capsid assembly. *J. Am. Chem. Soc.* **131**, 2606-2614, (2009).

101 Ceres, P. & Zlotnick, A. Weak Protein- Protein Interactions Are Sufficient To Drive Assembly of Hepatitis B Virus Capsids†. *Biochemistry* **41**, 11525-11531, (2002).

102 Prevelige, P., Thomas, D. & King, J. Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells. *Biophys. J.* **64**, 824-835, (1993).

103 Parent, K., Doyle, S., Anderson, E. & Teschke, C. Electrostatic interactions govern both nucleation and elongation during phage P22 procapsid assembly. *Virology* **340**, 33-45, (2005).

104 Wang, S., Chang, J. & Dokland, T. Assembly of bacteriophage P2 and P4 procapsids with internal scaffolding protein. *Virology* **348**, 133-140, (2006).

105 Stehle, T., Gamblin, S., Yan, Y. & Harrison, S. The structure of simian virus 40 refined at 3.1 A resolution. *Structure* **4**, 165-182, (1996).

106 Larson, S., Day, J., Greenwood, A. & McPherson, A. Refined structure of satellite tobacco mosaic virus at 1.8 A resolution. *J. Mol. Biol.* **277**, 37-59, (1998).

107 Keef, T., Twarock, R. & Elsawy, K. Blueprints for viral capsids in the family of polyomaviridae. *J. Theor. Biol.* **253**, 808-816, (2008).

108 Kanesashi, S.-n., Ishizu, K.-i., Kawano, M.-a., Han, S.-i., Tomita, S., Watanabe, H., Kataoka, K. & Handa, H. Simian virus 40 VP1 capsid protein forms polymorphic assemblies in vitro. *J. Gen. Virol.* **84**, 1899-1905, (2003).

109 Salunke, D., Caspar, D. & Garcea, R. Polymorphism in the assembly of polyomavirus capsid protein VP1. *Biophys. J.* **56**, 887-900, (1989).

110 Kler, S., Wang, J., Dhason, M., Oppenheim, A. & Zlotnick, A. Scaffold properties are a key determinant of the size and shape of self-assembled virus-derived particles. *ACS chem. biol.* **8**, 8, (2013).

111 Nguyen, H. & Brooks, C. Generalized structural polymorphism in self-assembled viral particles. *Nano letters* **8**, 4574-4581, (2008).

112 Sanchez, J. & Holmgren, J. Virulence factors, pathogenesis and vaccine protection in cholera and ETEC diarrhea. *Curr. Opin. Immunol.* **17**, 388-398, (2005).

113 Fisher, D. Cholera in Zimbabwe. *Ann Acad Med Singapore* **38**, 82, (2009).

114 Vanden Broeck, D., Horvath, C. & De Wolf, M. J. S. Vibrio cholerae: cholera toxin. *Int. J. Biochem. Cell Biol.* **39**, 1771-1775, (2007).

115 Ruddock, L., Ruston, S., Kelly, S., Price, N., Freedman, R. & Hirst, T. Kinetics of acid-mediated disassembly of the B subunit pentamer of Escherichia coli heat-labile enterotoxin. Molecular basis of pH stability. *J. Biol. Chem.* **270**, 29953-29958, (1995).

116 Fan, E., Merritt, E., Verlinde, C. & Hol, W. AB5 toxins: structures and inhibitor design. *Curr. Opin. Struc. Biol.* **10**, 680-686, (2000).

117 Merritt, E., Kuhn, P., Sarfaty, S., Erbe, J., Holmes, R. & Hol, W. The 1.25 A resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. *J. Mol. Biol.* **282**, 1043-1059, (1998).

118 Haan, L. & Hirst, T. R. Cholera toxin: A paradigm for multi-functional engagement of cellular mechanisms (Review). *Mol. Membr. Biol.* **21**, 77-92, (2004).

119 van Heyningen, S. Cholera toxin: interaction of subunits with ganglioside GM1. *Science* **183**, 656, (1974).

120 Lauer, S., Goldstein, B., Nolan, R. L. & Nolan, J. P. Analysis of cholera toxin-ganglioside interactions by flow cytometry. *Biochemistry* **41**, 1742-1751, (2002).

121 Turnbull, W., Precious, B. & Homans, S. Dissecting the cholera toxin-ganglioside GM1 interaction by isothermal titration calorimetry. *J. Am. Chem. Soc.* **126**, 1047-1054, (2004).

122 Holmner, A., Askarieh, G., Okvist, M. & Krengel, U. Blood group antigen recognition by Escherichia coli heat-labile enterotoxin. *J. Mol. Biol.* **371**, 754-764, (2007).

123 Mandal, P., Branson, T., Hayes, E., Ross, J., Gavín, J., Daranas, A. & Turnbull, W. Towards a structural basis for the relationship between blood group and the severity of El Tor cholera. *Angew. Chem., Int. Ed. Engl.* **51**, 5143-5146, (2012).

124 Hayes, E. D. & Turnbull, W. B. Monovalent and Multivalent Inhibitors of Bacterial Toxins. *Synthesis and biological applications of multivalent glycoconjugates*, 78-91, (2011).

125 De Haan, L. & Hirst, T. Cholera toxin: a paradigm for multi-functional engagement of cellular mechanisms (Review). *Mol. Membr. Biol.* **21**, 77-92, (2004).

126 Pickett, C., Twiddy, E., Coker, C. & Holmes, R. Cloning, nucleotide sequence, and hybridization studies of the type IIb heat-labile enterotoxin gene of Escherichia coli. *J. Bacteriol.* **171**, 4945-4952, (1989).

127 Reichow, S., Korotkov, K., Hol, W. & Gonen, T. Structure of the cholera toxin secretion channel in its closed state. *Nat. Struct. Mol. Biol.* **17**, 1226-1232, (2010).

128 Lencer, W. & Tsai, B. The intracellular voyage of cholera toxin: going retro. *Trends Biochem. Sci.* **28**, 639-645, (2003).

129 Mekalanos, J., Collier, R. & Romig, W. Enzymic activity of cholera toxin. II. Relationships to proteolytic processing, disulfide bond reduction, and subunit composition. *J. Biol. Chem.* **254**, 5855, (1979).

130 Wernick, N., Chinnapen, D., Cho, J. & Lencer, W. Cholera Toxin: An Intracellular Journey into the Cytosol by Way of the Endoplasmic Reticulum. *Toxins* **2**, 310-325, (2010).

131 Loregian, A., Hirst, T., Marsden, H. & Palù, G. Use of Vibrio spp. for expression of Escherichia coli enterotoxin B subunit fusion proteins: purification and characterization of a chimera containing a C-terminal fragment of DNA polymerase from herpes simplex virus type 1. *Protein Expression Purif.* **8**, 381-389, (1996).

132 Streatfield, S., Sandkvist, M., Sixma, T., Bagdasarian, M., Hol, W. & Hirst, T. Intermolecular interactions between the A and B subunits of heat-labile enterotoxin from Escherichia coli promote holotoxin assembly and stability in vivo. *Proc. Natl Acad. Sci. U.S.A.* **89**, 12140-12144, (1992).

133 Sandkvist, M., Hirst, T. & Bagdasarian, M. Alterations at the carboxyl terminus change assembly and secretion properties of the B subunit of Escherichia coli heat-labile enterotoxin. *J. Bacteriol.* **169**, 4570, (1987).

134 Lebens, M., Johansson, S., Osek, J., Lindblad, M. & Holmgren, J. Large-scale production of Vibrio cholerae toxin B subunit for use in oral vaccines. *Nat. Biotechnol.* **11**, 1574-1578, (1993).

135 Olsvik, O., Wahlberg, J., Petterson, B., Uhlén, M., Popovic, T., Wachsmuth, I. & Fields, P. Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains. *J. Clin. Microbiol.* **31**, 22-25, (1993).

136 Jansson, L., Angström, J., Lebens, M., Imberty, A., Varrot, A. & Teneberg, S. Carbohydrate binding specificities and crystal structure of the cholera toxin-like B-subunit from Citrobacter freundii. *Biochimie* **92**, 482-490, (2010).

137 Jobling, M. & Holmes, R. Analysis of structure and function of the B subunit of cholera toxin by the use of site directed mutagenesis. *Mol. Microbiol.* **5**, 1755-1767, (1991).

138 Merritt, E., Sarfaty, S., Chang, T., Palmer, L., Jobling, M., Holmes, R. & Hol, W. Surprising leads for a cholera toxin receptor-binding antagonist: crystallographic studies of CTB mutants. *Structure* **3**, 561-570, (1995).

139 Connell, T. D. & Holmes, R. K. Mutational analysis of the ganglioside binding activity of the type II Escherichia coli heat labile enterotoxin LT IIb. *Mol. Microbiol.* **16**, 21-31, (1995).

140 Wolf, A., Jobling, M., Saslowsky, D., Kern, E., Drake, K., Kenworthy, A., Holmes, R. & Lencer, W. Attenuated endocytosis and toxicity of a mutant cholera toxin with decreased ability to cluster ganglioside GM1 molecules. *Infect. Immun.* **76**, 1476-1484, (2008).

141 Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J. & Chao, J. C. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. Assoc. Comput. Machin.* **51**, 91-97, (2008).

142 Brooks, B. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545-1614, (2009).

143 Lazaridis, T. & Karplus, M. Effective energy function for proteins in solution. *Proteins* **35**, 133-152, (1999).

144 Schymkowitz, J., Rousseau, F., Martins, I., Ferkinghoff-Borg, J., Stricher, F. & Serrano, L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. U.S.A.* **102**, 10147-10152, (2005).

145 Guerois, R., Nielsen, J. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369-387, (2002).

146 Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382, (2005).

147 Marcello, A., Loregian, A., De Filippis, V., Fontana, A., Hirst, T. & Palu, G. Identification and characterization of an extracellular protease activity produced by the marine Vibrio sp. 60. *FEMS Microbiol. Lett.* **136**, 39-44, (1996).

148 Marcus, H., Ketley, J., Kaper, J. & Holmes, R. Effects of DNase production, plasmid size, and restriction barriers on transformation of Vibrio cholerae by electroporation and osmotic shock. *FEMS Microbiol. Lett.* **56**, 149-154, (1990).

149 Panda, D., Dasgupta, U. & Das, J. Transformation of Vibrio cholerae by plasmid DNA. *Gene* **105**, 107-111, (1991).

150 Shevchik, V., Condemine, G. & Robert-Baudouy, J. Characterization of DsbC, a periplasmic protein of Erwinia chrysanthemi and Escherichia coli with disulfide isomerase activity. *EMBO J.* **13**, 2007-2012, (1994).

151 Liu, J. & Walsh, C. Peptidyl-prolyl cis-trans-isomerase from Escherichia coli: a periplasmic homolog of cyclophilin that is not inhibited by cyclosporin A. *Proc. Natl Acad. Sci. U.S.A.* **87**, 4028-4032, (1990).

152 Dertzbaugh, M. & Cox, L. The affinity of cholera toxin for Ni2+ ion. *Protein Eng.* **11**, 577-581, (1998).

153 Jobling, M., Palmer, L., Erbe, J. & Holmes, R. Construction and characterization of versatile cloning vectors for efficient delivery of native foreign proteins to the periplasm of Escherichia coli. *Plasmid* **38**, 158-173, (1997).

154 Jobling, M. & Holmes, R. Fusion proteins containing the A2 domain of cholera toxin assemble with B polypeptides of cholera toxin to form immunoreactive and functional holotoxin-like chimeras. *Infect. Immun.* **60**, 4915-4924, (1992).

155 Kowalczykowski, S., Dixon, D., Eggleston, A., Lauder, S. & Rehrauer, W. Biochemistry of homologous recombination in Escherichia coli. *Microbiologic. Rev.* **58**, 401-465, (1994).

156 Schagen, F., Rademaker, H., Rabelink, M., van Ormondt, H., Fallaux, F., van der Eb, A. & Hoeben, R. Ammonium sulphate precipitation of recombinant adenovirus from culture medium: an easy method to increase the total virus yield. *Gene Ther.* **7**, 1570-1574, (2000).

157 Bhakuni, V., Xie, D. & Freire, E. Thermodynamic identification of stable folding intermediates in the B-subunit of cholera toxin. *Biochemistry* **30**, 5055-5060, (1991).

158 Ganesh, C., Shah, A., Swaminathan, C., Surolia, A. & Varadarajan, R. Thermodynamic characterization of the reversible, two-state unfolding of maltose binding protein, a large two-domain protein. *Biochemistry* **36**, 5020-5028, (1997).

159 Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I. & Le Mercier, P. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**, 82, (2011).

160 Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. & Players, F. Predicting protein structures with a multiplayer online game. *Nature* **466**, 756-760, (2010).

161 Harbury, P., Kim, P. & Alber, T. Crystal structure of an isoleucine-zipper trimer. *Nature* **371**, 80-83, (1994).

162 Harbury, P., Zhang, T., Kim, P. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401-1407, (1993).

163 Nautiyal, S. & Alber, T. Crystal structure of a designed, thermostable, heterotrimeric coiled coil. *Protein Sci.* **8**, 84-90, (1999).

164 Ogihara, N., Weiss, M., Degrado, W. & Eisenberg, D. The crystal structure of the designed trimeric coiled coil coil-VaLd: implications for engineering crystals and supramolecular assemblies. *Protein Sci.* **6**, 80-88, (1997).

165 Wong, S., Baron, R. & McCammon, J. Hot-spot residues at the E9/Im9 interface help binding via different mechanisms. *Biopolymers* **89**, 916-920, (2008).

166 Meenan, N., Sharma, A., Fleishman, S., Macdonald, C., Morel, B., Boetzel, R., Moore, G., Baker, D. & Kleanthous, C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc. Natl Acad. Sci. U.S.A.* **107**, 10080-10085, (2010).

167 Keeble, A., Kirkpatrick, N., Shimizu, S. & Kleanthous, C. Calorimetric dissection of colicin DNase--immunity protein complex specificity. *Biochemistry* **45**, 3243-3254, (2006).

168 Li, W., Keeble, A., Giffard, C., James, R., Moore, G. & Kleanthous, C. Highly discriminating protein-protein interaction specificities in the context of a conserved binding energy hotspot. *J. Mol. Biol.* **337**, 743-759, (2004).

169 Wallis, R., Leung, K., Osborne, M., James, R., Moore, G. & Kleanthous, C. Specificity in protein-protein recognition: conserved Im9 residues are the major determinants of stability in the colicin E9 DNase-Im9 complex. *Biochemistry* **37**, 476-485, (1998).

170 Chen, H. N. & Woycechowsky, K. J. Conversion of a Dodecahedral Protein Capsid into Pentamers via Minimal Point Mutations. *Biochemistry*, (2012).

171 Gibson, D., Young, L., Chuang, R.-Y., Venter, J., Hutchison, C. & Smith, H. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**, 343-345, (2009).

172 Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PloS one* **3**, (2008).

173 Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS one* **4**, (2009).

174 Schlegel, S., Rujas, E., Ytterberg, A., Zubarev, R., Luirink, J. & de Gier, J.-W. Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels. *Microb. Cell Fact.* **12**, 24, (2013).

175 O'Shea, E., Lumb, K. & Kim, P. Peptide 'Velcro': design of a heterodimeric coiled coil. *Current biology : CB* **3**, 658-667, (1993).

176 Krylov, D., Barchi, J. & Vinson, C. Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids. *J. Mol. Biol.* **279**, 959-972, (1998).

177 Burkhard, P., Stetefeld, J. & Strelkov, S. Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol.* **11**, 82-88, (2001).

178 Fejer, S., Chakrabarti, D. & Wales, D. Emergent complexity from simple anisotropic building blocks: shells, tubes, and spirals. *ACS nano.* **4**, 219-228, (2010).

179 Seebeck, F., Woycechowsky, K., Zhuang, W., Rabe, J. & Hilvert, D. A simple tagging system for protein encapsulation. *J. Am. Chem. Soc.* **128**, 4516-4517, (2006).

180 Fukuta, S., Magnani, J., Twiddy, E., Holmes, R. & Ginsburg, V. Comparison of the carbohydrate-binding specificities of cholera toxin and Escherichia coli heat-labile enterotoxins LTh-I, LT-IIa, and LT-IIb. *Infect. Immun.* **56**, 1748-1753, (1988).