Dominik Fridrich, BSc

Structural characterization of computationally designed highly stable helical oligomers with custom geometries

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieur

Master's degree programme:
Biotechnology

submitted to
Graz University of Technology

Supervisor
Univ.-Prof. Dr.rer.nat. Peter Macheroux

Co-supervisor
Mag.rer.nat. Dr.rer.nat. Gustav Oberdorfer

Institute of Biochemistry
Petersgasse 12/II, 8010 Graz

Graz, July 2020

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date, Signature

1 Abstract

Computational design of novel protein structures is a promising tool to make superior biological materials with tailor-made properties, new pharmaceuticals, or complex fine chemicals. The last decade led to an enormous progress in *de novo* protein design due to improvements of existing methods, the growth of the protein structure databases and technical advances in computational resources. Recently, a method was established for varying the parameters in the Crick coiled coil generating equations, which enables designs of completely new proteins with helical backbones. With the software suite Rosetta, low energy sequences for alternative helix supercoil arrangements were identified. This method has been successfully used to design monomeric structures, three-, four- and five-helical bundle structures using idealized coiled-coil geometries. Lately, higher complexity backbones were established, which resulted in the *de novo* design of self-assembling pore-like structures which are described in this work.

Biochemical and biophysical characterization of the designed structures, which were produced using *Escherichia coli* as expression host, was carried out. An alpha-helical circular dichroism signal was detected and thermostability up to 95°C could be observed. A binding assay for hydrophobic fluorescent ligands was established which hints at the potential of the designs to be used for downstream applications in filtering of small molecules and for nanopore sequencing. Initial crystallization conditions have been identified and improvements to the loop-geometry of one of the designs was made.

To get promising X-ray diffraction data more crystallization setups have to be carried out as well as more structural and biophysical experiments are required to examine the full potential of the pore like structures.

2 Zusammenfassung

Computergestütztes Design neuartiger Proteinstrukturen ist ein vielversprechendes Werkzeug zur Herstellung biologischer Materialien mit maßgeschneiderten Eigenschaften, die die Produktion von neuen pharmazeutischen Wirkstoffen, oder die Herstellung komplexer Feinchemikalien ermöglicht. Im letzten Jahrzehnt gab es enorme Fortschritte im Bereich des *de novo* Proteindesigns aufgrund der Verbesserung von bestehenden Computermethoden, dem Wachstum der Proteinstrukturdatenbanken und der Weiterentwicklung der zur Verfügung stehenden Rechenleistung von Computersystemen. Kürzlich wurde eine Methode zur Variation der Parameter, die in den *Crick coiled coil* Generierungsgleichungen vorkommen, etabliert. Dadurch wurde das Design völlig neuer Proteine mit alpha-helikalem Grundgerüst ermöglicht.

Unter Verwendung des Softwarepaketes Rosetta für kombinatorische Design Berechnungen wurden energiearme Sequenzen für alternative Helix-Supercoil Anordnungen identifiziert. Diese Methode wurde erfolgreich verwendet um monomere Helixbündel mit drei und vier Helices, sowie eine pentamere Fünf-Helix-Bündelstruktur mit idealisierter *coiled coil* Geometrie zu entwerfen.

Aus den *de novo* Entwürfen selbstorganisierender porenartiger Strukturen wurden vor Kurzem Protein Backbones mit hoher Komplexität erschaffen, die in dieser Arbeit vorgestellt werden. Diese Strukturen wurden unter der Verwendung von *Escherichia coli* als Expressionsorganismus produziert und biochemisch sowie biophysikalisch charakterisiert. Es wurde ein Signal für zirkulären Dichroismus, welches typisch für alpha-helikale Proteine ist, gemessen und eine thermische Stabilität bis zu 95 °C festgestellt. Weiters wurde ein Ligandenbindungstest mittels fluoreszierender Moleküle durchgeführt, der das Potenzial der Konstrukte darlegt, diese für das Downstream Processing und somit die Filterung kleinerer Moleküle oder für die DNA-Sequenzierung mittels Nanoporen zu verwenden. Erste Kristallisationsbedingungen konnten identifiziert und eine geometrische Designoptimierung einer Loopregion in einem alpha-Helix Barrel durchgeführt werden.

Um das volle Potenzial der porenartigen Strukturen zu untersuchen und zukünftig vielversprechende Röntgenbeugungsdaten zu erhalten, müssen mehr Kristalle getestet, und weitere strukturelle und biophysikalische Experimente durchgeführt werden.

3 Index

4 Introduction

Custom-made protein structures pave the way to entirely new biomolecular machines and applications. By varying the parameters in the Crick coiled coil generating equations, the procedure enables the design of completely new proteins [1]. Using the software suite Rosetta for combinatorial design calculations low energy sequences for alternative helix supercoil arrangements can be identified [2]. This method has been successfully used to design monomeric, three and four as well as a pentameric five-helix bundle structures using idealized coiled coil geometries [3]. Recently, higher complexity backbones were established, which resulted in the *de novo* design of self-assembling pore-like structures with the potential to use them as channels or transporters.

4.1 The diversity of proteins

Proteins are the foundation for essential life processes. They are nanoscale sized biological machines capable of catalysing various reactions in nature, in a way of variations which seem to be endless. In naturally occurring proteins it is possible to transform chemical energy into work or use solar energy for manufacturing complex molecules. Furthermore, they can act as natural detector for molecules or light and thus proteins are necessary in signal transduction and the connecting piece in almost any function of the living cell. Their roles are encoded in sequences of amino acids which also specify the three-dimensional structure of proteins. Through spontaneous folding these structures are self- and biologically organized, which enables them to work as multifunctional precision tools. This has shifted the focus of current protein research to the modification of natural proteins [4].

Most of protein engineering focuses on the modification of naturally occurring proteins to fine-tune or adjust their functions using techniques such as directed evolution. This contains random mutagenesis and recombination techniques, which includes cycles of generating and selecting of variants. To expand the engineering possibilities for rational design, computer-aided methods can be recruited. Today it is possible to design custom-made proteins from the ground up and try to adopt many features to address important challenges that our society faces [4].

4.2 The protein folding problem

The structure of a protein is solely determined by its amino acid sequence. Further on, the protein folds into the lowest energy state that is accessible to its amino-acid sequence, as it was originally proposed by Christian Anfinsen [5]. According to Levinthal, the native state of a protein is just a local minimum of the free energy of a provided amino acid sequence. Levinthal postulated that folding proteins from an initial random structure by examining all possible conformations would take much longer than it actually does (Levinthal paradox) [6]. A promising solution to this paradox is that the energy landscape of a specific protein is funnel shaped and at the bottom of the funnel the protein has only access to an ever-decreasing number of conformations [7]. Still, the space of possible structures is extremely large and it is difficult to search for the native state exhaustively. Computer aided sampling is usually guided by an energy function towards minimum energy states of the system and can help to identify the global minimum of the energy landscape (see Figure 1).

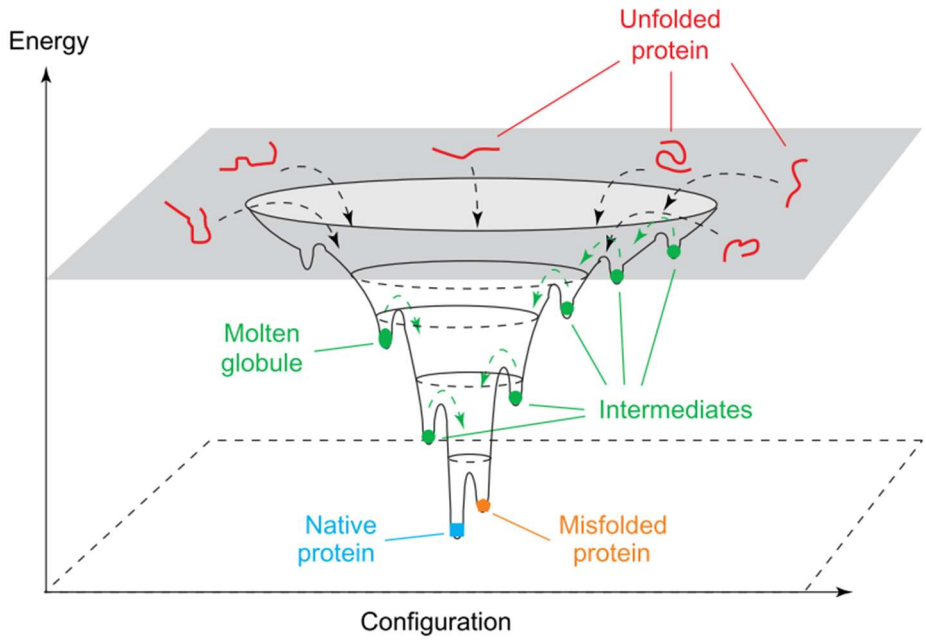

Figure 1:Schematic diagram of a protein folding energy landscape. The way of a protein to a native state. Unfolded protein folds to the native state at the bottom of the funnel. Figure is taken from Sheena E.Radford et al. [8] © Trends in Biochemical Sciences Volume 25

The burial of hydrophobic residues in the protein's core, away from the solvent, can be expected to be the driving force for protein folding. Therefore, it is important to pack the side chains of the core tightly, avoiding energetically unfavourable atomic overlaps and resulting in a minimization of the size of voids in the protein, which could be occupied by water. [9].

4.3 Protein structure prediction

Protein structure prediction can be roughly divided into two methods. First, *ab initio* structure prediction where the known amino-acid sequence acts as input and through backbone and side-chain sampling the predicted structure is produced. Second, homology modelling where the known backbone structure acts as input. In this case the side-chain sampling includes rotamers of all amino acids and generates an optimized model as output. [4]. This technique belongs to the template-based structure prediction methods. Using the Protein Data Bank (PDB), about two thirds of the todays known protein families can be modelled using templated-based methods [10,11,12].

Template-free methods generally manipulate the information from known structures; their development reflect the theoretical and technical level of protein structure prediction. The software suite Rosetta, originally developed by the lab of David Baker, is a template-free method, which assembles full-length structures based on fragments of 3-15 residues harvested from PDB structures. These fragments can be selected based on local sequence similarity and on the similarity between known and predicted secondary structure. For structure assembly, a Monte Carlo simulated annealing strategy is carried out (see Figure 2 B) [10].

Building structures based on the principles of biophysics and without relying on any previously determined structures can be defined as *ab initio* method [10]. Therefore, finding the lowest energy conformation for fixed amino acid sequences in the absence of further information about the structures of related proteins can be considered as the *ab initio* structure prediction problem. An energy function is required and a suitably accurate method for computing, which captures the interactions of all atoms in the protein with each other and the solvent. Preferably this leads to a folded protein chain with very low energy. Rosetta's energy function includes, among other terms, Van der Waals forces which favour close atomic packing, steric repulsion, electrostatic interactions and hydrogen bonds, solvation and the torsion energies of backbone and side-chain bonds [4].

4.4 Protein design

Evolution has presumably covered only a small amount of sequences that are possible for proteins. Since the number of distinct sequences for a typical protein length is 20^n (n is the number of possible amino acid residues). Naturally occurring proteins do not explore the whole sequence space because evolution proceeds by stepwise mutation and selection. Therefore, they rather cluster tightly into families and share different mutations. In contrast to evolutionary outcomes and protein engineering techniques, it should be possible to generate new proteins from scratch based on the principles of protein biophysics. The design of a protein from scratch requires methods for sampling alternative backbone and side-chain conformations by using an appropriate search algorithm and energy function [4].

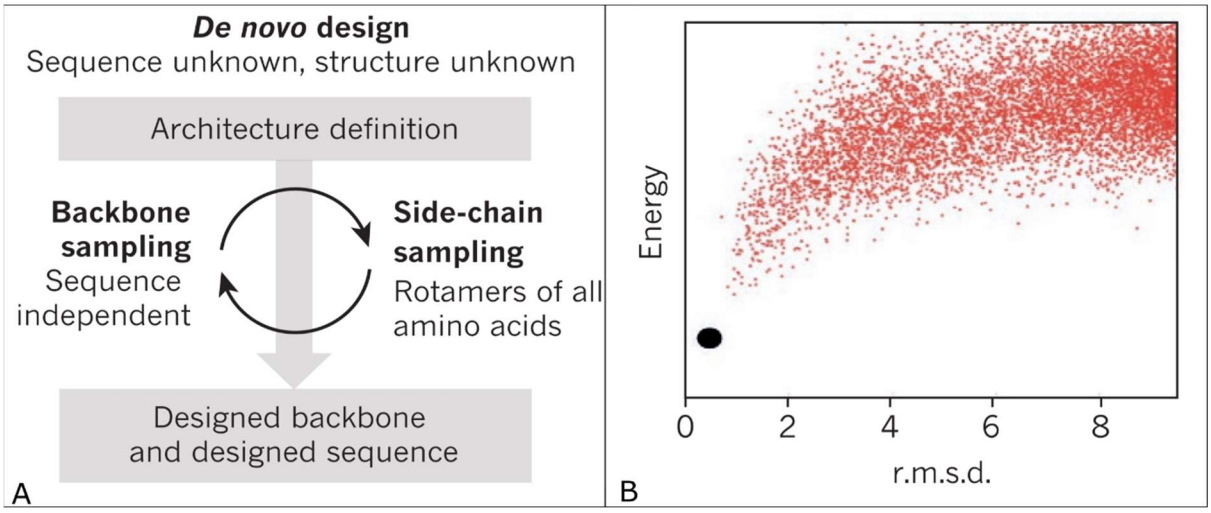

Figure 2: *De novo* design. A (In contrast to structure prediction were the sequence is known, in *de novo* design neither the sequence nor the structure is known). **B** (Energy landscape, black dot indicates the target structure, red dots represents lowest-energy structures which are plotted according to their similarity with the target) Design sequences which match on the target are selected for further characterizations. Figure is taken from P. Huang et al. [4] ©Nature.

Distinct methods are necessary for sampling alternative backbone and side-chain conformations to predict and design protein structures in combination with an adequate energy function. In contrast to protein design, the sequence of the protein is fixed in structure prediction, and the backbone structure is unknown. To overcome the protein design problem, where both the sequence and structure are unknown, the amino acid identities and the rotameric states must be sampled (see Figure 2 A) [4].

A rather new approach for protein structure prediction is to construct a learned, protein specific potential by training a neural network using deep learning techniques. In this network, backbone torsion angles and pairwise distances between residues are included. After learning and predicting many different distances this network can distribute distance information that respects covariation, local structure, and residue identities of nearby residues. These predictions can

then be combined to a protein specific potential. The central component of this method, which is called AlphaFold, is a neural network that is trained on PDB structures to predict the specific atom distances and to construct a potential that can accurately describe the shape of a protein [13].

4.4.1 Coiled coils and Helical bundles

Helical bundles were a mandatory step in the development of *de novo* protein design. The gap between simple monomeric helices and native proteins can be described through α-helical coiled coils. A left-handed coiled coil is characterized by a seven-residue geometric repeat *abcdefg* (see Figure 3). The residues *a* and *d* project towards the bundle core and are generally hydrophobic whereas *e* and *g* residues face the inter-subunit interface and are mostly polar [1,14]. Van der Waals packing between the buried residues (*a, d*) is necessary to resolve the stoichiometry and structure of coiled coils [15]. The use of flexible-backbone methods and parametric equations allowed the design of both right-handed and left-handed coiled coils [16]. By manipulation of the physicochemical and steric properties of two residues (*e, g*) 4- to 8-stranded bundles could be designed [17]. At this point proteins with the described parameters have not been found in nature and were never characterized before, which was a milestone for *de novo* protein design. The Baker Lab expanded the *de novo* protein design process by extending the use of parametric equations which made it possible to design regular bundles with a variation of geometric repeats and stoichiometry [3]. The search process for backbones that allow the formation of hydrogen-bond networks into homo- and heterodimeric coiled coils was automated [18,19]. Today, it is possible to use the known information for the design of regular coiled coils with various sizes and shapes [14], it is even possible to predict [20] and craft a folded protein structure with the online protein-folding game Foldit [21].

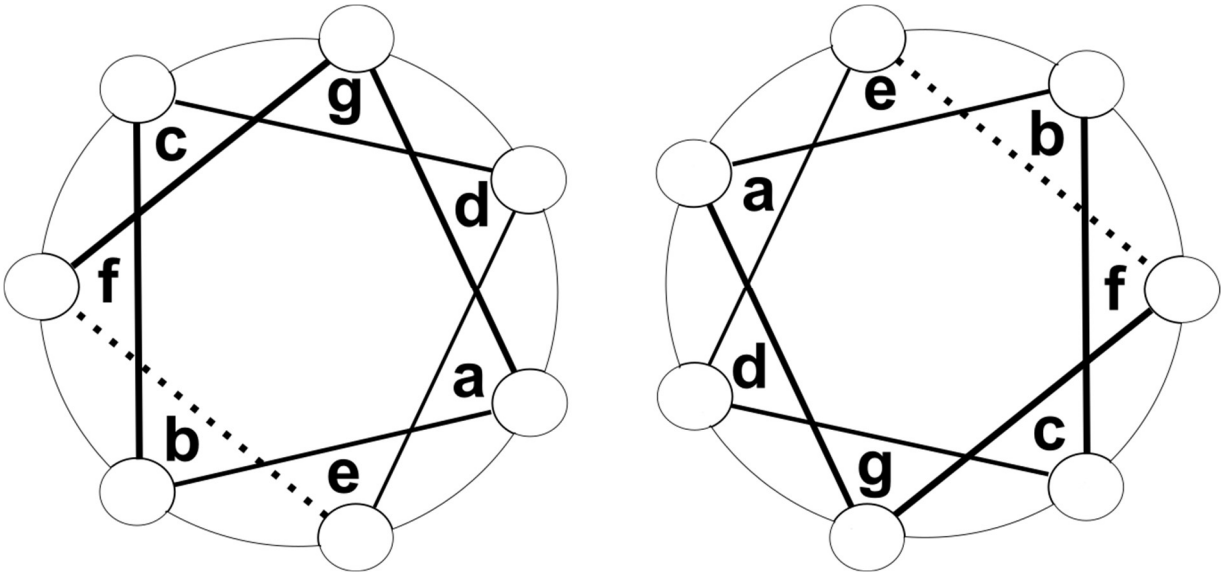

Figure 3: Heptad repeat. *abcdefg* are the possible residues of a corresponding helices. Figure is adapted from I. Korendovych et al [14] © Cambridge University Press, Quarterly Reviews of Biophysics.

4.4.2 *De novo* Design

Design of proteins from scratch requires methods for sampling backbone and side chain conformations by using a suitable energy function. This energy function must cover the physical principles for protein folding, capturing interactions of all atoms building up the protein with each other and the surrounding solvent. The calculations for this function can be performed by the Rosetta macromolecular modelling program. This software can differentiate between the unfolded, folded, and native-like conformations [22].

In *de novo* design the sequence and the structure of the backbone are unknown (see Figure 2 A) By assembling short peptide fragments or using algebraic equations to specify the geometry, the first backbones can be established parametrically. After that for each backbone a combinatorial sequence-optimization calculation is performed which is used to identify the lowest energy for the structure. To overcome the computational limits in backbone sampling, it is important to reduce the immense alternatives of backbone structures and decide to calculate only a finite number. This sequence-independent backbone construction is achieved by the requirement that the polar atoms of the backbone only make hydrogen bonds within the chain of alpha helices or beta sheets, or encounter the solvent in the exposed loop regions. This restricts the length of the possible secondary structures that are permitted [4, 23]. An additional constraint is achieved by restricting the lengths of the loops which connect α-helices and β-sheets in various packing orientations. Sequence-independent design principles have been revealed which relate the length of helices, strands and loops through simulations and analysis of protein structures. Therefore, the construction of topologies that consists of α-helices and β-sheets was made possible [16, 17]

To generate functional designs, the focus is designing ideal protein structures with α-helices and β-strands. Several steps are required to define the sequence-independent design approach for these protein structures. A topology *blueprint-file* is created which is consistent with the backbone design principles. This is necessary to specify the lengths, packing arrangement and the order of the different components. At this point, one can decide the number of α-helices and β-strands, as well as the lengths of the connecting loops for the design. This is followed by searching for protein backbones which fit the blueprint file. Once found, these are assembled from protein structure fragments using a Monte Carlo approach [4]. After the calculations with the stochastic method, rotamer optimization is used to identify a low energy amino acid sequence for each backbone. This step is followed by interchangeable cycles of backbone relaxation and additional sequence optimization to find a pair of sequence and structure that has very low energy. Finally, sequences that suit the designed structure in structure prediction calculations are tested experimentally (see Figure 2 B). [4, 15].

4.5 Nano pore sequencing

High throughput DNA sequencing methodologies have been optimized over the last years and as the technologies improve, so does the number of applications in research [18,19]. Sanger sequencing has become the sequencing standard. Itwas developed in 1977 and remains important for applications where high throughput is not required. It is still used to verify plasmid constructs or PCR products [19, 20]. The next step in the development of sequencing technologies were the second generation sequencing methods which can be grouped into two major categories sequencing by hybridization and sequencing by synthesis, whereas sequencing by synthesis is a further stage of sanger sequencing, with repeated cycles of synthesis. These methods have the advantage of running in parallel setups, thereby reducing costs.

Nanopore-sequencing is capable of analysing single molecules and belongs to the third-generation sequencing techniques. Today, science still needs cost-efficient and rapid sequencing methods. With one nanopore at 10 ms/base working speed, it would take approximately 20 years to sequence a human genome (10x coverage). To overcome the issue, parallel sequencing is required with ten thousand pores. Then it would take less than a day for a genome or, by increasing the number of pores to millions, just ten minutes. Nanopore-based instruments are capable of reads of > 100 kb and they are cheap and fast in comparison to second-generation sequencing devices [25]. The third-generation methods aim to sequence long DNA and RNA molecules by passing nucleotides through high-precision pores and measure differing currents as each molecule passes by [26].

In its simplest configuration the sequencing tool consist of a single pore within a membrane. A membrane (lipid bilayer) divides a salt solution into two sections called cis and trans. By applying a voltage across the membrane, an ion current flows through the nanopore (see Figure 4Figure 4). The primary signal to detect is the magnitude of this ion current. Therefore, molecules of interest can pass through the pore and a modulation of the ion current can be measured. The molecules are altered in the current depending on the degree to which the ion current is blocked, hence they can be distinguished by different properties of the molecule and movement through the pore [19, 21].

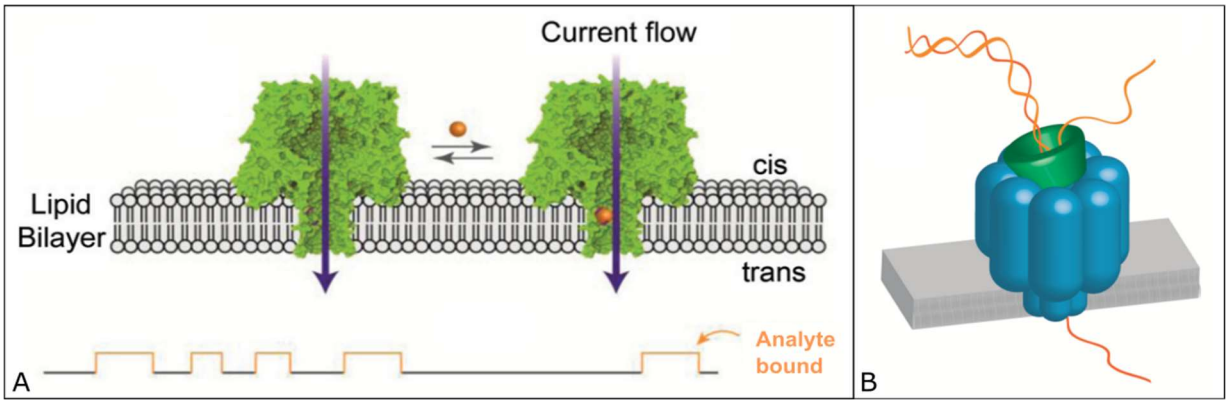

Figure 4:Schematic Nanopore. A (Ion current through an individual α-hemolysin pore is monitored. If an analyte binds within the pore a transient change in the current is detected) **B** (Nanopore strand sequencing, enzyme processes dsDNA, ssDNA is fed through individual protein pore). Both figures taken from H. Bayley et al [25] ©Clinical Chemistry.

There are two types of nanopores to distinguish biological pores, such as α-hemolysin pore (see Figure 4 A), and solid-state pores, which can be found as nanopores drilled into thin membranes of silicon nitrile [21,22,23]. By now, nanopores are capable to detect size and shape characteristics of proteins [31]. They can separate two different types of proteins [32] or polymers through different sizes and charges [33]. Moreover, DNA and RNA sequencing were achieved. For this purpose, nucleotides need to be recognized as ssDNA to pass through the nanopore. The correct threading of ssDNA into the nanopores, the identification of individual bases and passing through of DNA were already achieved (see Figure 4 B). To overcome issues regarding the reliability and precision of existing nanopores, *de novo* proteins could contribute to solve these problems. Single-stranded polynucleotides are still driven too fast through the pores for time-resolved resolution. A specifically designed pore which slows down the DNA translocation process for each base and is capable of distinguishing nucleotides with sufficient signal to noise ratio, would have an impact. To moderate the process, enzymes are used which are incorporated onto the pores. Catalytical enzymes remain an issue in terms of long-time usage. Additionally, they require a well-defined pH range for their activity [25]. In contrast, *de novo* proteins could be an ideal tool to overcome these challenges as they are remarkably stable, even at strong denaturing conditions and will not need additional enzymes to work [3]. However, substantial improvement needs still to be made for the parallelization of the system to facilitate a competitive rate of sequencing [25].

4.6 Introducing the *de novo* C3, C6, C9 set

The helix bundle structures are built out of α-helical coiled coils as the individual helices of the constructs are connected via loops. They have a molecular weight of around 9.2 kDa and share a similar backbone, which was produced by varying the parameters in Crick coiled coil generating equations. Docking of the three helix bundles and interface design should lead to self-assembling pore-like structures (see Figure 7) [3]. They share a building block of three alpha helices connected by two loops (α-helical hairpins). One of the major advantages of α-helical constructs, like helix bundles, are that they are soluble and show high thermal and chemical stability [3].

Main differences of the structures are the number of building blocks as well as the diameters of the hydrophobic cores and the total size of the constructs. Therefore, the names of the structures are derived from the number of building blocks, three building blocks for C3, six for C6 and so on. Furthermore, the building blocks are connected through different loops. The C6 and C9 protein have a kind of central accessible channel, which could be used for the recognition and binding of ligands (see Figure 5, and 7.7 Ligand Docking). The constructs share similar characteristics regarding the hydrophobic residues primarily pointing towards the interior of the helix bundle in contrast the charged residues generally at the surface of the construct.

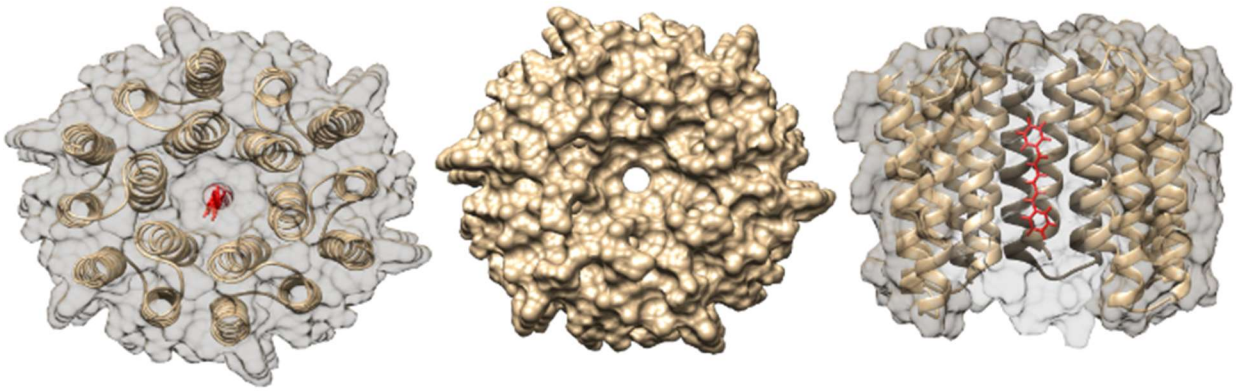

Figure 5: Possibilities for ligand-docking into 3H22 protein. Three different viewing angels of the C6 construct. The pore like structure offers the possibility to recognize or bind different ligands. As it is shown in red, the construct bind the environmentally sensitive dye (1,6-Diphenyl-1,3,5-hexatriene)

The 3H22 C3 construct has a relatively high pI due to the high amount of positively charged residues, especially Lysine. In contrast, the distribution of charges is more balanced for C6 and C9 (see Table 1).

Table 1 Data of the de novo constructs

Construct	Monomer [DA]	Total Size [kDa]	Amino acids	Extinction coefficient [M^{-1} cm^{-1}]	pI
3H22 C3	9417.4	28.25	79	11000	9.82
3H22 C6	9131.8	54.79	81	6990	6.58
3H22 C9	9271.8	83.45	79	16500	6.70

An additional feature composed of six tyrosine residues is integrated into the top of the 3H22 C6 hexamer design, building a potential selectivity filter. The monomer from this design was already successfully fused through a linker to the protein DHF119 which functions as a new anchor protein (see Figure 6) [34].

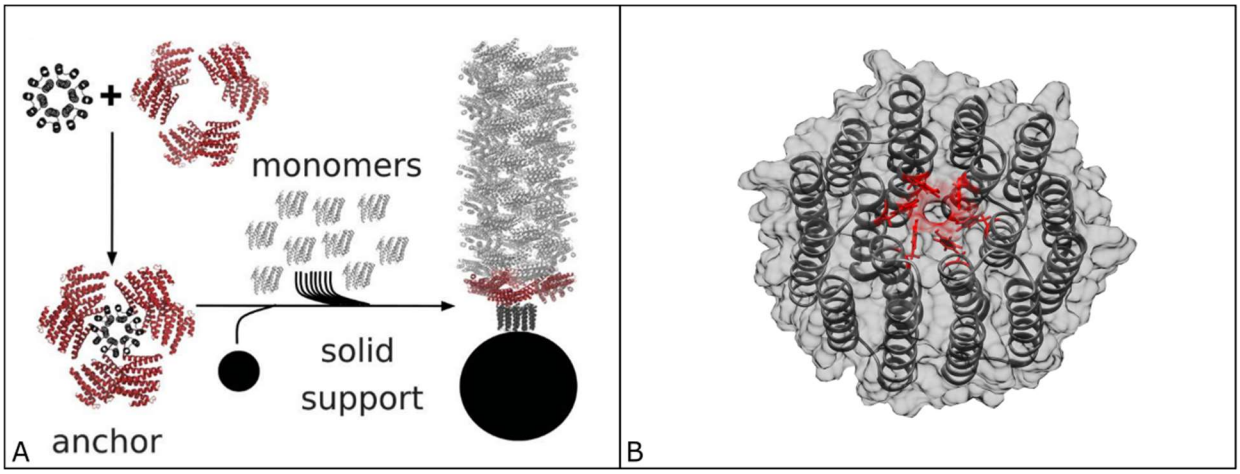

Figure 6: The construction of a fibre anchor and C6 tyrosine ring. A (The anchor protein DHF119_C6 is fused out of 3H22 monomers to the C cap of DHF119 with a GGS5 linker. The anchor is shown holding monomers which can form self-assembling helical protein filaments, **B** (stabilising tyrosine ring, indicated in red, at the top of the 3H22 C6 construct), Figure A is taken from H. Shen et al [34] ©Science.

For the calculation of the theoretical pI and molar extinction coefficient, the ExPASy ProtParam tool was used. More details about the protein sequences and specific loop residues of the constructs can be found in the Appendix (see Table 17). All constructs and mandatory design improvements were performed by Dr Gustav Oberdorfer through combinatorial design calculations. The protein sequences were back-translated into DNA, synthesized, and finally cloned into pET21-NESG vectors.

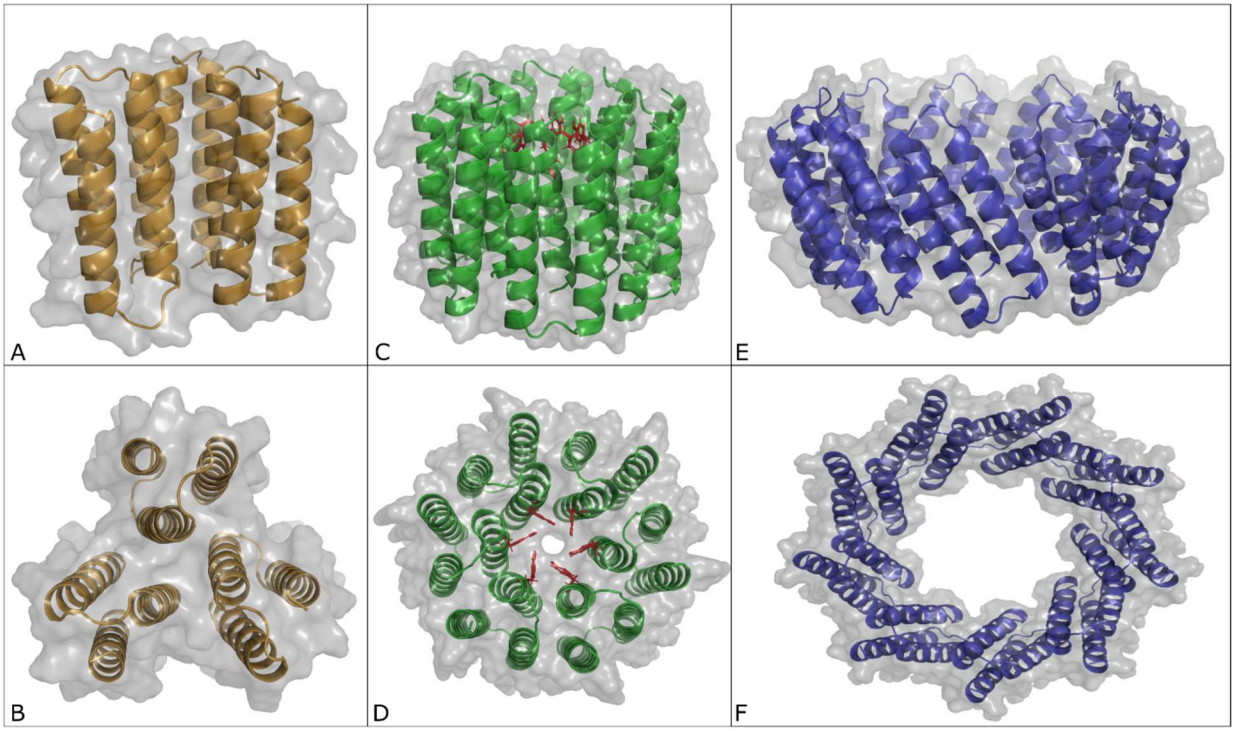

Figure 7: The de novo constructs of this thesis. A (3H22 C3 side view, three monomers self-assemble to the trimer protein) **B** (3H22 C3 top view), **C** (3H22 C6 side view, six monomers self-assemble to the hexamer protein) **D** (3H22 C6 top view, tyrosine residues indicated in red), **E** (3H22 C9 side view, nine monomers self-assemble to the nonameric protein), **F** (3H22 C9 top view)

4.7 Aim

The main focus of this thesis was the biochemical and biophysical characterization of computationally designed helical oligomers by protein expression in *Escherichia coli* followed by purification using affinity, and size exclusion chromatography. Additionally, the characterization via CD spectroscopy as well as thermal shift assay were used to confirm that the protein was fully folded. Further experiments, like fluorescence spectroscopy, crystallography and binding assays were carried out to establish a broad foundation for further research. The obtained results can be used to optimize or extend the designed structures and introduce new functions.

5 Materials

5.1 General

All chemicals and media ingredients were obtained from Roth (Karlsruhe, Germany), Sigma-Aldrich (St. Louis, MO, USA), Merck (Darmstadt, Germany), Merck Millipore (Burlington, MA, USA), Fisher Scientific (Hampton, NH, USA).

The crystallization screens were ordered from Molecular Dimensions Limited (Holland, OH, USA).

All DNA templates were provided by Dr Gustav Oberdorfer (Graz, Austria)

5.2 Culture media

Media and plates were prepared as shown below.

- LB-medium
  - 5 g/L Yeast Extract
  - 10 g/L NaCl
  - 10 g/L Tryptone
  - For Plates: 15 g/L Agar

- TB-medium
  - 12 g/L Tryptone
  - 24 g/L Yeast Extract
  - 4 mL/L Glycerol
  - 17 mM KH_2PO_4
  - 72 mM K_2HPO_4

5.3 Buffer systems

- Lysis- (Binding-) buffer: 20 mM Tris HCl, 0.5 M NaCl, 1 mM Imidazole, pH 8
- Wash buffer I: 20 mM Tris HCl, 0.5M NaCl, 30 mM Imidazole, pH 8
- Wash buffer II: 20 mM Tris HCl, 1M NaCl, 30 mM Imidazole, pH 8
- Elution buffer: 20 mM Tris HCl, 0.5M NaCl, 250 mM Imidazole, pH 8
- Crystallization buffer: 20 mM Tris HCl, 150mM NaCl, 2 mM DTT, pH 8
- Storage buffer: 20 mM NaH_2PO_4, pH 8
- Fluorescence buffer: 20 mM Tris HCl, 150mM NaCl, 2 mM DTT pH 8, 5% DMSO

5.3.1 SDS-polyacrylamide gel electrophoresis

Table 2: Instructions for the preparation of SDS-polyacrylamide gels

	Separating gel (15%)	Separating gel (17.5%)	Stacking gel (5%)
	1x	1x	1x
Acrylamide/BIS (37.5:1)	1.96 mL	2.285 mL	0.313 mL
1.5 M Tris-HCl pH 8.8	1.875 mL	1.875 mL	-
0.5 M Tris-HCl pH 6.8	-	-	0.313 mL
SDS 20%	25 µL	25 µL	12.5 µL
H_2O	1.095 mL	0.77 mL	1.845 mL
10 % $(NH4)_2S_2O_8$ (APS)	24 µL	24 µL	12.5 µL
TEMED	5 µL	5 µL	2.5 µL

- Staining solution: 75 mL/L acetic acid, 2.5 g/L Brilliant Blue R, 500 mL/L MeOH
- Destaining solution: 75 mL/L acetic acid, 200 mL/L MeOH
- Sample buffer 2x (for 100 mL): 5 mL 1 M Tris, pH 7, 25 mL 20 % SDS,20 mL glycerol, 2 mg Bromphenol Blue
- Protein standard: PageRuler™ Prestained Protein Ladder, 10 to 180 kDa (Thermo Scientific™)

5.3.2 Buffers and solutions for Western Blot

- Blotting buffer: 25 mM Tris base, 190 mM glycine, 20 % ethanol, pH 8.3
- TBS: 50 mM Tris base, 150 mM NaCl, pH 8
- TTBS: 50 mM Tris base, 150 mM NaCl, 0.2 % Triton-X-100, pH 8
- Blocking solution: 50 mM Tris base, 150 mM NaCl, 0.2 % Triton-X-100, 5 % milk powder
- Clarity Western ECL Substrate (Bio-Rad Laboratories, Inc): peroxide reagent and luminol/enhancer reagent
- Solution for detection of His-tag:
    - 1st antibody (anti-His): His-Tag XP Rabbit mAB 1:10000 dilution in TTBS with 5 % BSA
    - 2nd antibody: Anti-rabbit IgG, HRP-linked AB 1:3000 dilution in TTBS with 5 % milk powder

5.4 Fluorescence Assay Reagents

- 1,6-Diphenyl-1,3,5-hexatriene Dye: 1 mM DPH
- 6-Propionyl-2-Dimethylaminonaphthalene: 10 mM Prodan
- 1-(4-Trimethylammoniumphenyl)-6-Phenyl-1,3,5-Hexatriene-*p*-Toluenesulfonate: 10 mM TMA-DPH
- SYPRO™ Orange Protein Gel Stain (5000X concentration in DMSO)

5.5 Crystallization screen Reagents

- JCSG Screen Box 1 and 2 (Molecular Dimensions)
- Morpheus Box 1 and 2
- Midas Screen
- SG1 (ShotGun) Screen

5.6 Kits

- Plasmid purification: Wizard® Plus SV Minipreps DNA Purification Systems (Promega)
- Mutagenesis: Q5® Site-Directed Mutagenesis Kit (NEB)

5.7 Bacterial strains

- *E. coli* TOP10: plasmid amplification
- *E. coli* BL21Star (DE3): expression strain
- *E. coli* Rosetta™(DE3)pLysS: expression strain
- *E. coli* Rosetta™ 2(DE3)pLysS: expression strain
- *E. coli* BL21-CodonPlus: expression stain

5.8 Primers

Table 3: Primer for Pointmutation at sequence position 139 Threonine against Alanine.

C9-T139A	Forward Primer	5'-GCT GAA AA CAC CAC CGA TAC GAT CCG TG-3'
	Reverse Primer	5'- AGT TTC AGC AGG CGC-3'

5.9 Equipment

Autoclave	5075 MLV (Tuttnauer) V-150 (Systec)
Thermocycler:	Primus 25 advanced (Peqlab)
Protein electrophoresis equipment:	Mini-PROTEAN®3 Cell (BIO-RAD)
Centrifuges:	Avanti J-20-XP (Beckman Coulter) Avanti J-26-XP (Beckman Coulter) Eppendorf Centrifuge 5810R Heraeus Megafuge 40R (Thermo Scientific)
Sonication device	Labsonic U (B.Braun Biotech International)
FPLC:	ÄKTAexplorer 100 (GE Healthcare) Column Superdex 200 10/300 GL
Purification column:	HisTrap HP 5ml (GE Healthcare) Sephadex G-25 in PD-10 Desalting Column
Spectrophotometer:	
Circular Dichroism Spectrophotometer:	J-1500 (Jasco)
NanoDrop:	NanoDrop2000 Spectrophotometer (Thermo Scientific)
Protein Crystallization Robot:	Oryx8 (Douglas Instruments)
Microscope:	
Plate Reader:	CLARIOstar (BMG LABTECH)
Various:	Eppendorf Thermomixer comfort Amicon Ultra Centrifugal Filters 3 kDA/60 kDa (2mL, 4mL, 15mL) 96 well microtiter plates, transparent (greiner BIO-ONE) 96 well microtiter plates, white, F-bottom (greiner BIO-ONE) D-Tube™ Dialyzers (mini, midi) CrystalClear 'Duo' (Jena Bioscience)

6 Methods

6.1 Transformation

The transformation into the cells was achieved using chemical competent bacteria cells (Top10, BL21, Rosetta) for plasmid amplification as well as protein expression. A standard heat-shock protocol was applied for the preferred sample as well as a negative control. The frozen cells were thawed on ice. After that 100 µL of cells were mixed with 0.5 to 5 µL of plasmid DNA and stored on ice for 30 minutes. Later the cells were heat-shocked for 30 seconds and cooled down for 2 minutes on ice. Afterwards 150 µL SOC-medium were added for regeneration followed by incubation at 37 °C for 1 hour at 500 rpm in a shaking incubator. Finally, the cells were centrifuged for 1 minute at 12000 rpm while the supernatant was discarded. The formed pellet was then resuspended in fresh SOC-media, 150 µL of this suspension was plated on LB-agar plates, containing the type of required antibiotic.

6.2 Protein expression

Different strategies were used for protein expression.

Expression test:

This procedure was performed to get an idea which *E. coli* strain to use for protein expression. Different parameters were varied such as expression temperature, inducer concentration and expression duration to find an optimized strategy for higher protein yield. First, 2.5 mL of ONC (overnight culture) were prepared with the desired strain and antibiotic. After overnight incubation at 37 °C on a shaker (200 rpm) the ONC was used to inoculate the main culture. Therefore, 1 mL of the ONC was transferred into a flask containing 50 mL of fresh LB-medium.

- 50 mL *E. coli* BL21Star (DE3)
- 50 mL Rosetta™(DE3)pLysS
- 50 mL *B*L21-CodonPlus
- 50 mL Rosetta™ 2(DE3)pLysS

The different cultures were incubated until an OD_{600} between 0.6 and 0.8 was achieved. IPTG (Isopropyl β-d-1-thiogalactopyranoside) was added in different concentrations (0.1 M, 0.5 M, 1 M) to induce the protein expression. After 4 h of incubation on a shaker at 37 °C and 18 °C (overnight;180 rpm) the cells were harvested and prepared for an SDS gel.

Strategy 1: Protein expression 18 °C

First, 180 mL of ONC were prepared with the desired strain and antibiotics. After overnight incubation at 37 °C on a shaker (200 rpm) for 15 h, the ONC was used to inoculate the main culture. It was prepared with 10 mL of ONC into 1 L of fresh LB-medium as well as the required

antibiotics (1:1000). The main culture was incubated on a shaker (at 37 °C; 180rpm) until an OD_{600} (optical density at 600 nm) between 0.6 and 0.8 could be measured. As soon as the desired OD_{600} was reached the temperature was decreased to 18 °C and 1 mM end concentration IPTG was added to induce the protein expression. After 15 h of incubation on the shaker (at 18 °C; 180 rpm), the cells were harvested, using centrifugation (6000 rpm; 4 °C; 20 min).

Strategy 2: Protein expression 37 °C

The protein expression at 18 °C did not lead to satisfying amount of protein. For enhancing protein expression, a different strategy was chosen. First, 180 mL of ONC were prepared as already described. The main culture was inoculated with 10 mL of ONC into 1 L of fresh TB-medium, as well as the required antibiotic until OD_{600} between 0.6 – 0.8 was reached. At this point, IPTG (0.1 mM end concentration) was added to induce protein expression. After 4 h at 37 °C the cells were harvested using centrifugation (6000 rpm; 10 min). Strategy 2 was dimensioned for 18 L main culture which were produced in 18 Erlenmeyer flasks.

6.3 Cell lysis

The cells were prepared for cell lysis by washing the pellet two times with 15 mL 0.9 % NaCl solution, later the pellet was centrifuged (4600 rpm; 4 °C; 20 min). Next the supernatant was discarded, and the remaining pellet was resuspended in the lysis buffer (20 mM Tris HCl, 0.5 M NaCl, 1 mM Imidazole pH 8). Thereafter phenylmethylsulfonyl fluoride (PMSF) in a ratio of 1:10 (1 M stock) was added to the resuspension then the cells were lysed by sonication with the Labsonic U device, for two times 5 min (ice cooling the reaction tube and sonicator tip). Finally, the suspension was filled into centrifugation tubes and centrifuged (18000 rpm; 4 °C; 45 min).

6.4 Protein purification

The amount of lysed protein suspension was the decisive factor whether applied pressure or gravity columns were used. Meanwhile the Ni-NTA-columns were prepared for protein purification. Before the columns were used each one was washed and equilibrated as it is shown below.

- wash with 20 mL H_2O (or 3x column volume)
- wash with 10 mL EDTA (100 mM pH 8)
- wash with 5 mL H_2O
- add 5mL NaOH 0.5 M (closed column incubates for 10 min, then flush it)
- wash two times with 10 mL H_2O
- add 3x column volume $NiSO_4$ 100 mM
- wash with 10 mL H_2O

At this point the column can be stored in the fridge by adding column volume 20 % EtOH otherwise the column is ready to use for purification.

- Purification:
    - equilibrate the column with 10 mL lysis buffer (20 mM Tris HCl, 0.5 M NaCl, 1 mM Imidazole pH 8)
    - transfer the lysate to column (centrifuge the sample first)
    - use 20 mL of wash buffer I (20 mM Tris HCl, 0.5M NaCl, 30 mM Imidazole, pH 8)
    - use 20 mL wash buffer II (20 mM Tris HCl, 1M NaCl, 30 mM Imidazole, pH 8)
    - use elution buffer (add 1x column volume, or more until all protein is washed from the column)

Filtering the sample ensured that only clear lysate was loaded onto the equilibrated Ni-NTA column. Afterwards an SDS-PAGE analysis was performed to separate the unwanted fractions from the fractions containing the protein of interest. Depending on how the protein sample in the upcoming experiments was used one of the following buffer exchange methods were applied.

- Buffer exchange using centricons

All pooled protein fractions were concentrated to a volume of about 1 mL. Next crystallization buffer was added (20 mM Tris HCl, 150mM NaCl, 2 mM DTT pH 8) then again, the sample was concentrated to volume of 800 – 900 µL.

- Buffer exchange using dialysis (D-Tube™ Dialyzer)

First the dialyzer was equilibrated with H_2O for 10 minutes then the water was discarded and replaced with the concentrated protein sample. The dialyzer tube was fixed overnight in a floating rack and placed into a beaker containing exchange buffer (NaH_2PO_4 20mM pH 8) and a stir bar.

The protein was ready for fast protein liquid chromatography (FPLC).

6.5 Sodium dodecyl sulphate-polyacrylamide (SDS) gel electrophoresis

Proteins were analysed using SDS-gel electrophoresis. The samples were prepared in Eppendorf tubes. 15 µL of protein samples were mixed with 5 µL sample buffer (see 5.3.1 SDS-polyacrylamide gel electrophoresis) and incubated at 95 °C for 10-15 minutes. All the samples were centrifuged for 10 seconds and 5 µL were loaded onto the gel. In addition, 3 µL of the PageRuler™ Prestained Protein Ladder was applied to estimate the molecular weight of the expressed proteins. At the beginning the gel run with 100 V and as soon as the dye reached the separation gel the setting was changed to 180 V. Later the gel was stained with Coomassie Brilliant Blue for 10 min. Destaining of the gel was achieved while incubating the gel in destaining solution best results were achieved after overnight destaining (see 5.3.1 SDS-polyacrylamide gel electrophoresis).

6.6 Size exclusion chromatography (SEC)

The sample (max. 1 mL) was centrifuged for 5 min at 13300 rpm and prepared for fast protein liquid chromatography (FPLC). Meanwhile the sample loop was carefully flushed with crystallization buffer, though avoiding not to introduce air into the system. Followed by loading the protein sample bubble free with a 1 mL syringe on a Superdex 200 10/300 GL column. The 0.5 mL fractions were separated and collected in 15 mL Sarstedt tubes. After elution of the protein the fractions were used for an SDS-PAGE control gel. The desired FPLC fractions were pooled and used for further experiments.

6.7 Thermal shift assay

The samples were prepared in common well PCR plates with a final concentration of 5 µM of protein in crystallization buffer in a total volume of 50 µL. Moreover, 200x SYPRO Orange protein gel stain dye was added to the reaction tube and incubated for 1 h at 24 °C. In addition, a blank was prepared with crystallization buffer and dye. Fluorescence was then recorded from 20°C to 95°C

6.8 Circular Dichroism

CD measurement was consulted for tracking possible unfolding events during the heating process and to analyse the alpha helical content of the protein. Purified protein sample in NaH_2PO_4 buffer was transferred into an appropriate capillary for CD measurement. First, a test spectrum was recorded to see if a characteristic spectrum for alpha helical proteins in the range between 260 nm and 190 nm could be observed. If the concentration and the test spectrum looked good, a full CD measurement could be carried out. The alpha helical CD signal was followed at 222nm while the protein sample was exposed to a heating gradient which was between 20 to 95 °C and then reduced back to 20 °C. If a thermal unfolding event would occur, the signal should be reduced. Full spectra were recorded at 20C and 95C.

6.9 Western Blot

To verify if the expression and purification of alpha helical 3H22 C6 worked, a Western Blot analysis was performed. Different wash fractions and a small quantity of the pellet were taken during the purification of the protein. Besides, a protein with His-tag was used as a positive control for the detection. One SDS-PAGE gel was prepared for blotting and a Nitrocellulose membrane was cut to the right size. The gel, sponges and filter paper were arranged as usual for blotting in a blotting buffer filled chamber and run for one hour at 200 mA.

The samples on the membrane were stained reversibly with PonceauS for a few seconds followed by washing it with water until protein bands were visible. Afterwards the membrane was incubated for 1 h in blocking solution while shaking. Next, it was washed three time with TTBS

for 10 min. The membrane was incubated overnight with the 1st antibody (anti His; 4 °C). At the next day, it was incubated with the 2nd antibody for one hour followed by a washing step. For protein detection two Clarity Western ECL Substrates were mixed 1:1 and poured onto the membrane (for 3 min). Finally, the reaction of the substrate with the antibody-linked peroxidase could be detected using a suitable film.

6.10 Fluorescence Assay

A fluorescence assay was conducted to provide evidence if several substances like fluorescent hydrocarbons can bind or pass the protein tunnel formed by 3H22 C6 hexamer. Based on previous fluorescence measurement all dye reagents were prepared fresh in Dimethyl sulphoxide (DMSO) at the beginning of the experiment due to the high sensitivity of the measurement procedure. Afterwards three buffer-dye stocks consisting of crystallization buffer and the respective ligand (dye) were prepared.

- Prodan 10 mM
- TMA DPH 10 mM
- DPH 1 mM

A two-fold serial dilution in the range of 400 µM – 3.12 µM was made composed of purified protein and constant ligand concentration. The mixture was transferred to the 96 well microtiter plate (top/bottom optics) and incubated for 2 h in the dark while shaking. For comparison the same approach was realised for Lysozyme, Bovine serum albumin (BSA) and a blank consisting of dye and buffer. All samples were measured at 25 °C with an excitation of λ_{ex} 350 ± 20 nm and emission λ_{em} 426 ± 30 immediately after incubation as triplicates with a final volume of 100 µL per well. Binding constants (K_D) were determined by fitting the data to a single site binding model using OriginPro 8.5

$$y = \frac{B_{max} * x}{K_D + X} \qquad \text{Equation 1}$$

6.11 Crystallization

For the crystallization of the protein 3H22 C6 two different strategies were used.

Strategy 1: Manual set up Sitting Drop

CrystalClear duo Strips were used for manual crystallization and optimization. First, the strips were cleaned with pressured air and immediately sealed with crystallization tape. The tape was only removed once a screen reagent or protein was added to the well. Afterwards the reservoir was filled with 100 µL of the desired screening solution. The purified protein sample was centrifuged for 5 minutes at 13400 rpm. Hence, 1 µL of protein and 1 µL of screening solution was combined into a single drop in one of the crystallization wells. The following drop ratios were chosen.

Table 4: Manual set up Sitting Drop with varied drop ratios

Protein	Crystallization screen	
1 µL	1 µL	1:1
1 µL	0.5 µL	2:1
0.5 µL	0.5 µL	1:2

After finishing setting up the different crystallization conditions the strips were finally sealed and stored in a dark environment at 20 °C. The crystallization approaches were checked every 2 days using a microscope.

Strategy 2: Robot set up Sitting Drop

For higher throughput, a crystallization robot (Oryx8, Douglas Instruments) was used. The 96 well 3D crystal clear crystallization plates were prepared as described above. All reservoirs were filled with 30 µL of the desired screening solution. Once the robot finished pipetting the different screens the crystallization plates were sealed and stored at 4 °C in an incubator

The following drop ratios were chosen for the Midas Screen with an 3H22 C6 protein concentration of 39 mg/mL.

Table 5: Midas screen crystallization set up with robot Oryx8

	Protein	Midas screen	
Drop 1	0.25 µL	0.25 µL	1:1
Drop 2	0.6 µL	0.3 µL	2:1
Drop 3	0.25 µL	0.5 µL	1:2

The following drop ratios were chosen for the Morpheus Screen with an 3H22 C6 protein concentration of 30 mg/mL.

Table 6: Morpheus screen crystallization set up with robot Oryx8

	Protein	Morpheus screen	
Drop 1	0.25 µL	0.25 µL	1:1
Drop 2	0.6 µL	0.3 µL	2:1
Drop 3	0.25 µL	0.5 µL	1:2

The following drop ratios were chosen for the SG1 (ShotGun Screen) with an 3H22 C6 protein concentration of 35.6 mg/mL.

Table 7: SG1 screen crystallization set up with robot Oryx8

	Protein	SG1 screen	
Drop 1	0.3 µL	0.3 µL	1:1
Drop 2	0.5 µL	0.25 µL	2:1
Drop 3	0.25 µL	0.5 µL	1:2

The following drop ratios were chosen for the JCSG (Joint Center for Structural Genomics Screen) with an 3H22 C6 protein concentration of 38 mg/mL. A second JCSG crystallization set up was carried out with a protein concentration of 23.3 mg/mL.

Table 8: JCSG screen crystallization set up with robot Oryx8

	Protein	JCSG screen	
Drop 1	1 µL	1 µL	1:1
Drop 2	1 µL	0.5 µL	2:1
Drop 3	0.5 µL	1 µL	1:2

6.11.1 Manual optimization

Optimization was performed as described in Strategy 1 with JCSG B6 condition and a protein concentration of 22.1 mg/mL. The same pattern was used for a second approach with 11.05 mg/mL of 3H22 C6 protein concentration.

Table 9: Manual set up Sitting Drop JCSG screen condition B6 with varied drop ratios

	Protein	JCSG screen B6	
Drop 1	1 µL	1 µL	1:1
Drop 2	0.5 µL	0.5 µL	1:1
Drop 3	0.7 µL	0.3 µL	2.3 :1
Drop 4	0.3 µL	0.7 µL	1 :2.3

6.11.2 Seeding

Protein crystals were picked up using a loop and transferred to a new well with JCSG B6 screening solution. The crystals were streaked out in a variation of the conditions B6 (as it is shown beneath) to generate nucleation.

Table 10: Seeding of 3H22 C6 protein crystals

Total volume [µL]	JCSG screen B6	Crystallization buffer
1 µL	50 %	50 %
1 µL	70 %	30 %
1 µL	30 %	70 %

6.11.3 Harvesting

A small opening was cut into the sealed reservoir using a scalpel. Once the desired crystal was picked up the reservoir was immediately sealed again. Then, the harvested crystals were cryoprotected and stored in loops surrounded by liquid nitrogen. Afterwards, the were ready and shipped to the synchrotron.

6.12 Native Mass spectrometry

A purified protein sample was prepared for native mass spectrometry. First, the buffer was exchanged to ammoniumacetat ($C_2H_7NO_2$) buffer pH 7 using a centricon. Thereafter the sample was purified again using FPLC. The desired fraction was then dialysed. (see 6.4 Protein purification Buffer exchange using dialysis D-Tube™ Dialyzer). The protein (protein concentration 100 µMol) was shipped to Bruker and measured with kindly support by Dr. Andreas Winkler.
The samples for native MS were dissolved in 100mM HN_4Ac thereafter the samples were taken either with Captive Spray Source (750 nL/min) or with Apollo Source (3 uL/min).

6.13 Rosetta Scripts

The software suite Rosetta Scripts version 3 was applied for further structure improvements of the protein 3H22 C9. At the beginning of the experiment the unmodified 3H22 C9.pdb file was utilized as starting material. All the required files were provided by Dr Gustav Oberdorfer.
The MutateResidue mover was used to introduce the mutation at target position 46 in each of the nine chains of the protein construct. Therefore, the RosettaScripts mut_scan script was applied. The script was executed through the command line (-parser:script_vars position=A46 res=THR). Finally, a scoring script was executed for comparison between the non-mutated and mutated protein file.

6.14 Vector preparation (Point Mutation 3H22 C9)

For having more parental material an ONC was performed with *E. coli* Top 10 cells, followed by isolating the DNA using a Mini prep Kit (Wizard *Plus* SV minipreps DNA Purification System).

6.14.1 Q5 Site-Directed Mutagenesis

Later, the purified vector DNA (3H22 C9) was used for exponential amplification with mutated primers. The following reagents were assembled in a thin-walled PCR tube.

- 12.5 µL Q5 Hot Start High-Fidelity Polymerase
- 1.25 µL 10 µM Forward Primer (see 5.8 Primers)
- 1.25 µL 10 µM Reverse Primer (see 5.8 Primers)
- 1 µL Template DNA (1-25 ng/µL)
- 9 µL Nuclease-free water

The reagents were mixed completely, then transferred to a thermocycler. The applied thermocycling condition is provided in the table below.

Table 11: Thermocycling conditions for PCR

Step	Temperature [°C]	Time [s]	Cycles
Initial Denaturation	98 °C	30	1
Denaturation	98 °C	10	
Annealing	63 °C	30	25
Extension	72 °C	30	
Final Extension	72 °C	165	
Hold	4	∞	

After finishing the PCR protocol, 5 µL of the amplified plasmid was mixed with 2 µL loading dye and 8 µL water. The DNA-mixture was separated on an agarose gel.

6.15 Agarose gel

Agarose gels were used to analyse the size of the created PCR product after Q5 Site-Directed Mutagenesis. For analysis 1 % agarose gels were prepared, adding the required amount of agarose to TAE-buffer and heating the suspension until the agarose was dissolved completely. Before adding 5 µL of HD Green Plus DNA stain (Intas Science) the suspension was cooled down (30 – 40 °C). As reference DNA standard GeneRuler 1kb plus (Thermo Scientific) was used.

7 Results

7.1 Protein expression

7.1.1 Test-expression

Finding a suitable host for the protein expression was successful depending 3H22 C6 while 3H22 C9 was expressed but still needed to be modified. The protein 3H22 C3 was either expressed too weak or not at all as it is shown in Figure 8.

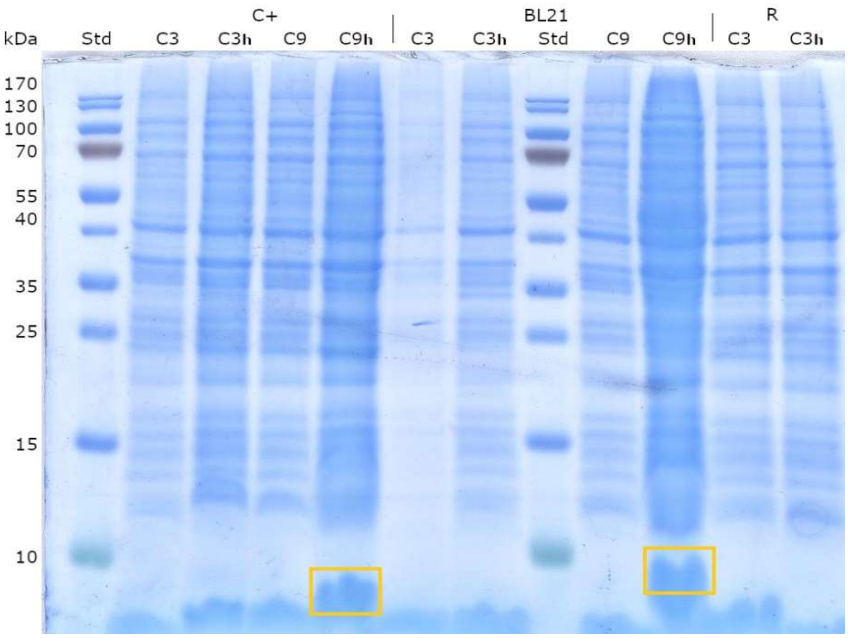

Figure 8: Test expression C3 and C9. SDS-Page analysis of samples taken from the main culture before induction (C3 or C9) and after harvesting (C3h or C9h) in comparison between *E. coli* cells Codon+, BL21 DE3 star and Rosetta cells. The first lane shows the PageRuler™ Prestained Protein Ladder, followed by the different protein samples. The monomeric state of the protein constructs is expected at around ~ 9.1 kB and is highlighted by the rectangles.

Since the expression for C6 worked out, different strains were tested (see Figure 9). The variation in IPTG concentration lead to the decision to use 0.1M of IPTG for the induction in further protein production approaches.

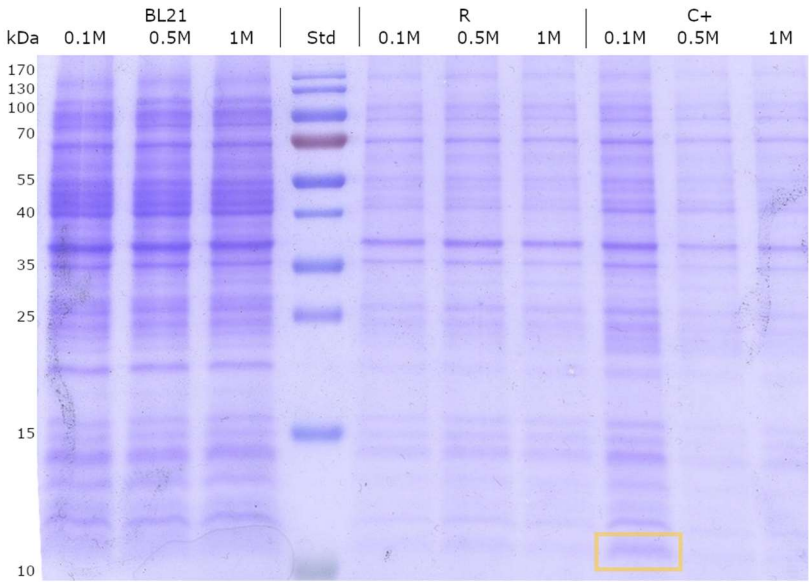

Figure 9: Test expression C6 with different IPTG concentrations. SDS Page analysis of samples after harvesting while the IPTG concentration was varied for induction. BL21(three different BL21 cultures expressing 3H22 C6 with IPTG concentrations of 0.1 M, 0.5 M and 1 M), R (three different Rosetta™(DE3)pLysS cultures expressing 3H22 C6 with IPTG concentrations of 0.1 M, 0.5 M and 1 M, C+ (three different BL21-CodonPlus cultures expressing 3H22 C6 with IPTG concentrations of 0.1 M, 0.5 M and 1 M)

Through all protein expression approaches it turned out that *E. coli* Codon+ cells worked best for the production of 3H22 C6 protein (see 6.2 Protein expression Strategy: Protein expression 37 °C). At this point the work was mainly continued with 3H22 C6, since it was not possible to produce the required amount of the other proteins for further experiments. Nevertheless, which type of expression strain was used the C3 protein was either expressed in too low amount or not at all. Throughout the laboratory work 12 batches of the C6 protein were produced in total. Across all these batches optimizations were performed for increasing the amount of protein.

7.1.2 Protein Purification

The purification was performed as described in section 6.4 Protein purification using a HisTrap HP 5 ml Ni-NTA column, for the reason that the expressed protein contained an N-terminal His_6-Tag.The collected protein fractions were analysed using SDS-PAGE. One of the typical SDS gels obtained in those purification approaches is shown in Figure 10.

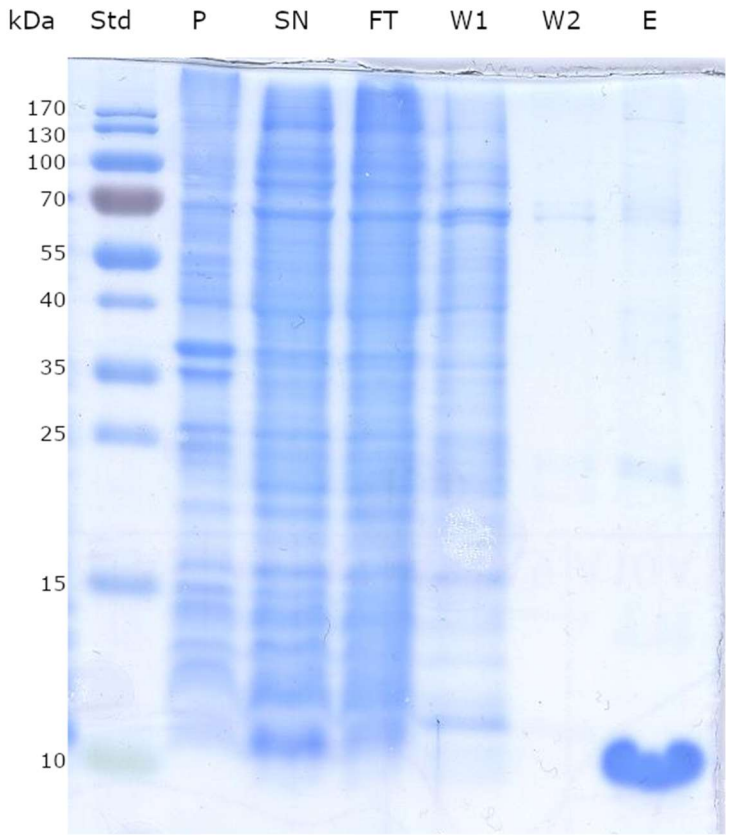

Figure 10: SDS-Page Analysis of protein fractions during 3H22 C6 purification. Std (PageRuler™ Prestained Protein Ladder), P (Pellet), SN (Supernatant), FT (flowthrough), W1 (wash fraction 1), W2 (wash fraction 2), E (elution fraction)

To get a purer protein sample the elution fractions were additionally prepared for SEC which led to the upcoming results.

7.1.3 SEC – and optimization 3H22 C6 production

The FPLC was a necessary method for the determination of the protein size since the protein tended to form different oligomeric states in various amounts. This was avoided through optimization as it is shown in Figure 11. As it turned out from all the tested *E. coli* strains the BL21 CodonPlus cells worked best for the production since it produced more than four times the amount of its predecessor the BL21 Star cells. The retention time of the protein is subject to a light shift to the right and seemed too high but as it is shown in the results of native mass spectroscopy experiments (see 7.8 Native Mass spectrometry of 3H22 C6) the protein started to eluate between 14.8 mL and 17.8 mL. As the desired protein is a Hexamer with a theoretical size of 54.7 kDa.

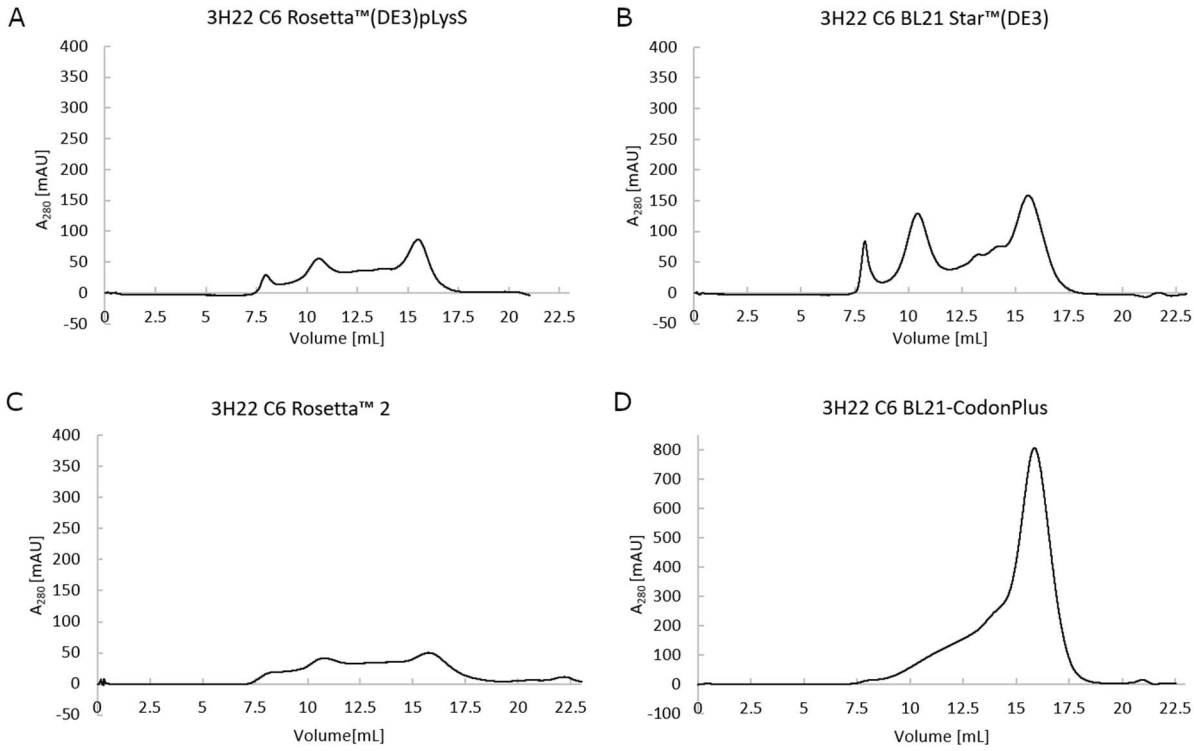

Figure 11: Comparison of 3H22 C6 protein production in different *E. coli* strains. A (Batch 1 3H22 6 expressed in Rosetta™(DE3)pLysS, B (Batch 4 3H22 C6 expressed in BL21 Star™), C (Batch 4 3H22 C6 expressed in Rosetta™ 2), D (Batch 11 3H22 C6 expressed in BL21-CodonPlus).

All the batches shown in Figure 11 (A, B, C) were performed with 3 L LB-medium main culture except for case D. The switch over to TB-medium increased the amount of produced protein a lot (see Table 12). To get enough protein for crystallization experiments the purified fractions of case A and B were pooled together. Regarding case D, the amount of collected fractions were split into three different samples whereas each sample was concentrated again. One major difference of Case D to the others was that protein expression was performed during 37 °C as it is described in 6.2 Protein expression.

Table 12:Comparison values of 3H22 C6 protein production

Case	Medium [L]	A_{280} [mAU]	Protein conc. [mg/mL]	Protein[µL]
A	3 L LB	86.46	25.02	100
B	3 L LB	158.66		
C	3 L LB	50.15	6.31	100
D	1 L TB	806.65	22.81	1350

As it was very important to have pure protein samples for crystallization setups, a second round of SDS-Page analysis (see Figure 12) and SEC were performed. Towards most purifications it

was a match between not only having enough protein but pure protein in the right oligomerization state, since usually only one or two fractions could be used for a second FPLC run.

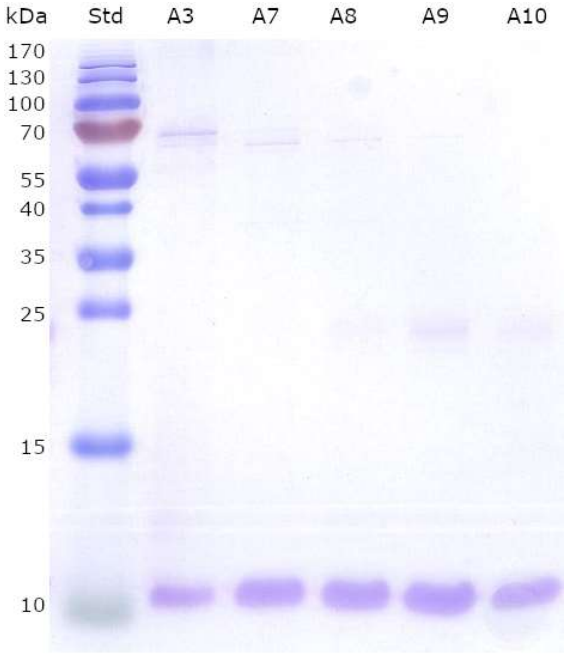

Figure 12: SDS-Page Analysis of FPLC fractions 3H22 C6. Std (PageRuler™ Prestained Protein Ladder), A3 – A10 (protein fractions after FPLC), purified protein sample near the 10 kDa mark. Fractions A8, A9, A10 were used for the second SEC run.

7.2 Crystallization and optimization

The most promising protein crystallization condition for 3H22 C6 was the B6 condition of the JCSG screen (see Figure 14).

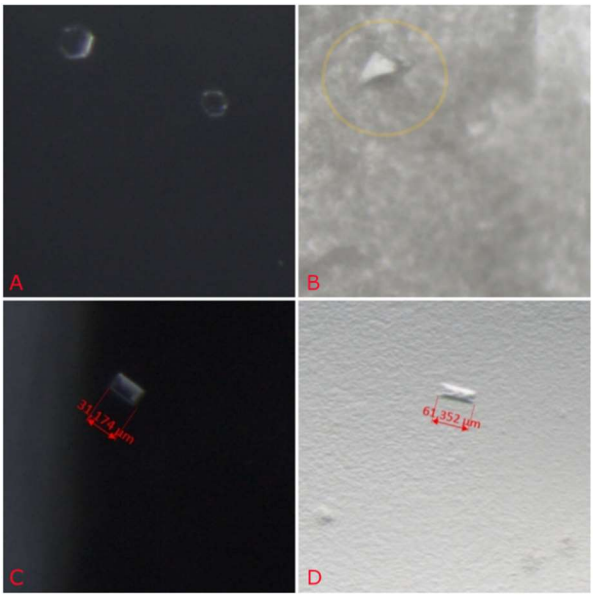

Figure 13: Comparison of four different crystals JCSG screen. A hexagonal shaped crystal (H12 drop 1), **B** octahedral shaped crystal (A8 drop 3), **C** cuboidal shaped crystal(F4 Drop2), **D** rectangular shaped crystal(H12 drop 3) all crystals were developed out of a protein sample with a concentration of 38 mg/mL.

Unfortunately, no data could be observed after grid scan for the crystals seen in picture A, B and D above. In one orientation the crystal from picture C showed reflections with low resolution however no further data could be observed.

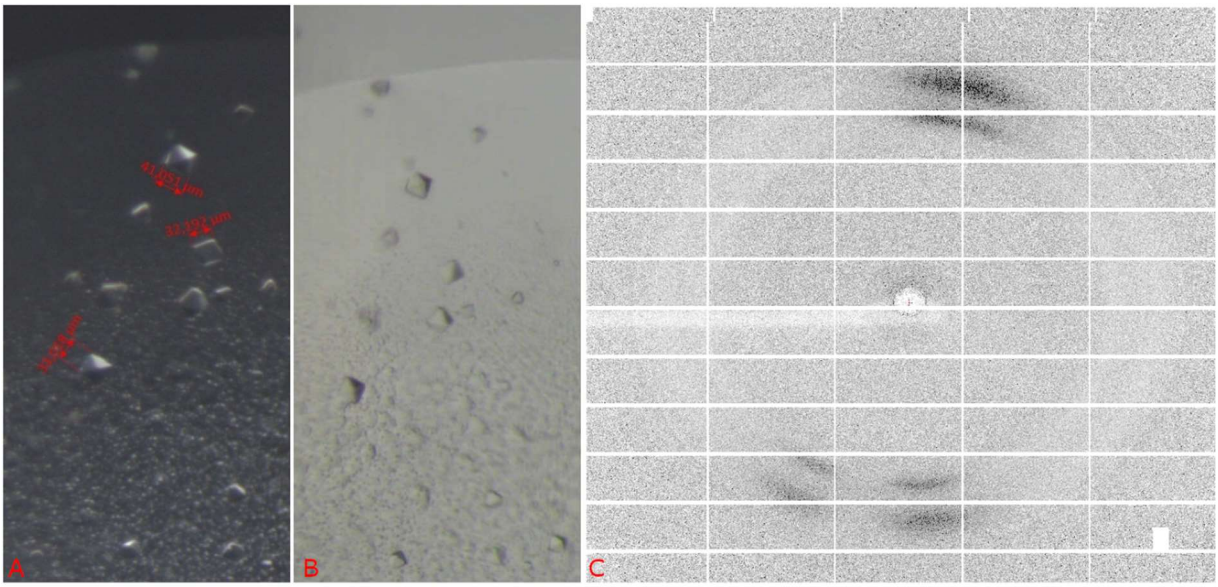

Figure 14:3H22 C6 protein crystals. A (JCSG B6 condition formed octahedral shaped crystals, with length measurement), **B** (Picture A from a different viewing angel and illumination), **C** (Picture of the diffraction of C6 crystal ~ 20 – 30 Å). The crystals were developed out of a protein sample with a concentration of 23.3 mg/mL

The picture shown above indicates octahedral shaped crystals nevertheless very weak protein diffraction was the result for most of them except for the one which showed diffraction in the range of 20 – 30 Å (see Figure 14 C).

7.3 Western Blot

To ensure that the protein which crystallized was the 3H22 C6 a Western Blot was performed with 5 µL of the protein sample.

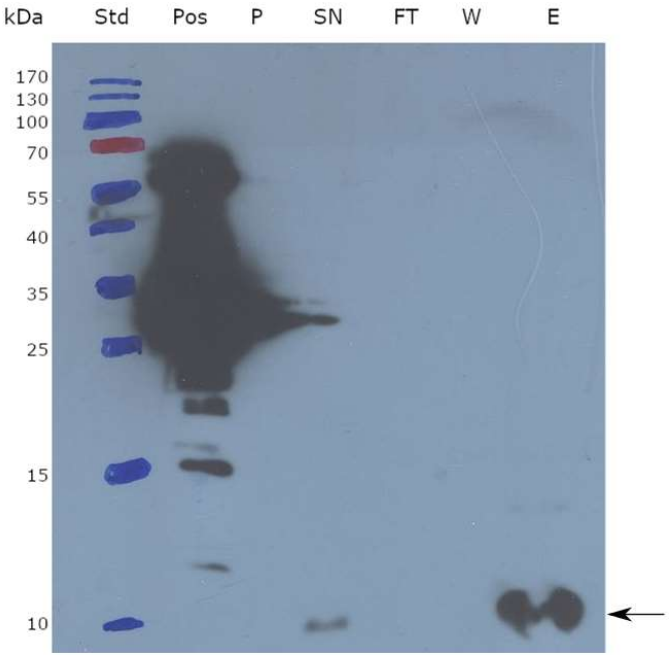

Figure 15: Western Blot of 3H22 C6. Pos (YCND as positive control), P (pellet), SN (supernatant), FT (flow through), W (wash fraction), E (elution fraction). Expected fraction was the arrow indicates due to the His-tag of the protein construct.

Unfortunately, too much of YCND positive control was loaded onto the gel but the result shows that there is no protein left in the pellet and only a small amount was found in the supernatant compared to the eluate fraction. Most of the protein was found in the eluate which is indicated by the arrow (see Figure 15). The result confirms that the eluted fraction was an expressed monomeric state of the 3H22 C6 construct with a His-tag and that the purification of the pellet worked as intended.

7.4 Thermal shift assay

To prove thermal stability of 3H22 C6 a thermal shift assay was performed to imply if unfolding events happened during the heating process.

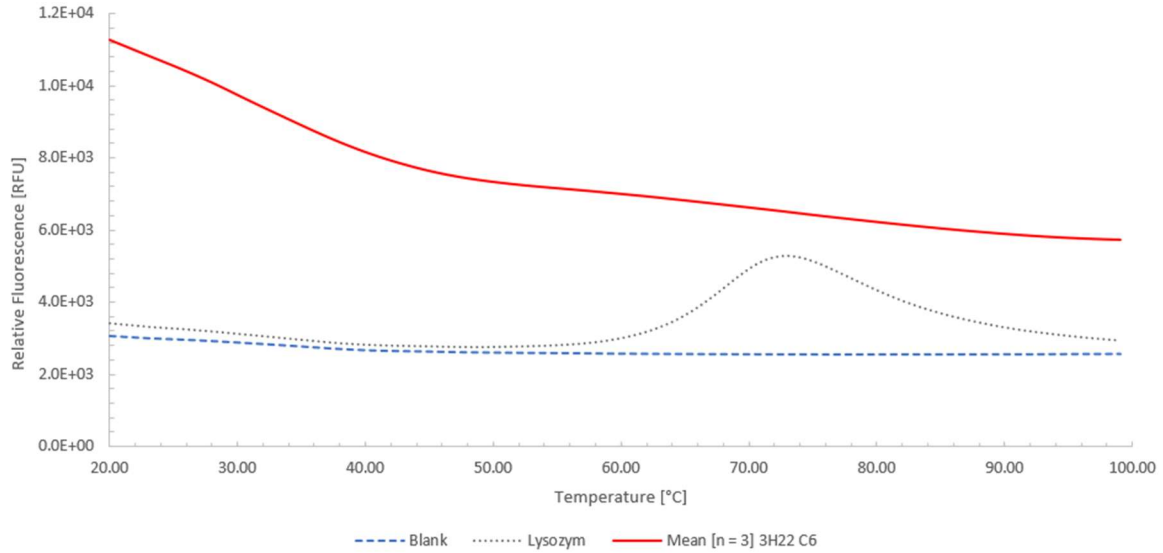

Figure 16: Thermal shift assay of 3H22 C6. (5 µM Protein in 20 mM Tris HCl, 150mM NaCl, 2 mM DTT pH 8), Blank (Buffer with SYPRO Orange Dye), Lysozyme for unfolding comparison.

Since the 3H22 C6 Protein has a hydrophobic core and the SYPRO Orange dye can possibly access the core of the protein from one direction (see Figure 17) the signal is significantly higher at the beginning of the experiment hence by increasing the temperature a self-quenching effect seems to happen (see Figure 16). The unfolding of Lysozyme starts at 60 °C and was used as a reference sample. The SYPRO Orange Dye can bind non-specifically to hydrophobic surfaces as it is the case for 3H22 C6, moreover water strongly quenches its fluorescence [35]. Once the protein unfolds, the exposed hydrophobic surfaces bind the dye, which should have shown an increase in fluorescence. Anyhow this was not the case for 3H22 C6 which demonstrates that the protein did not show any unfolding events.

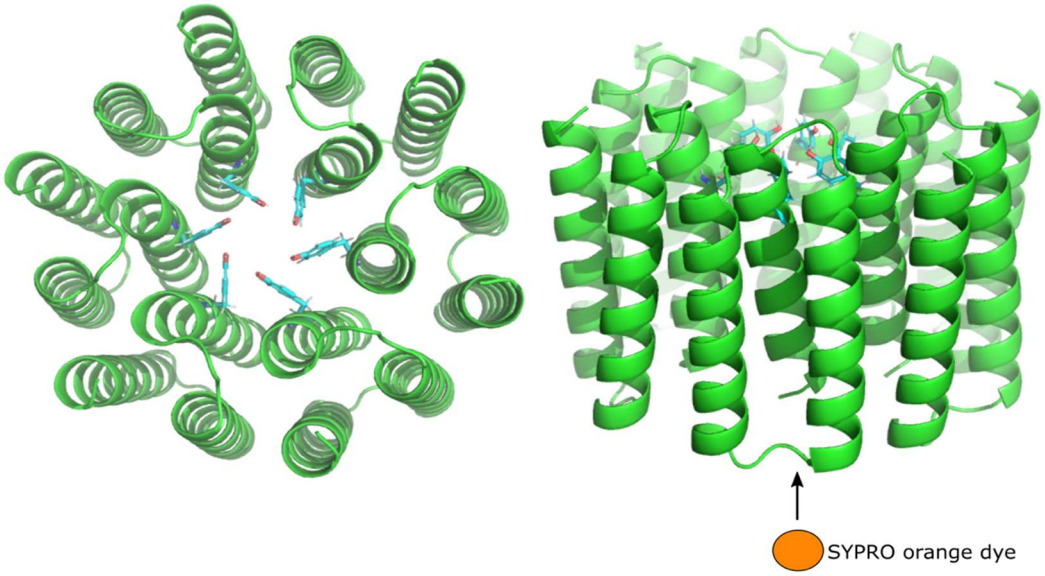

Figure 17: 3H22 C6 SYPRO Orange enters from one side. The SYPRO Orange dye can only enter at the opposite site of Tyrosine residues which are indicated in blue.

7.5 Circular Dichroism

The CD spectrum was recorded to support the result (see Figure 16) of the thermal shift assay using 3H22 C6 dissolved in 20 mM NaH2PO4 at pH 8. Usually, the secondary structures of α-helical proteins have a positive peak at 193 nm and two negative peaks at 222 nm and 208 nm [36].

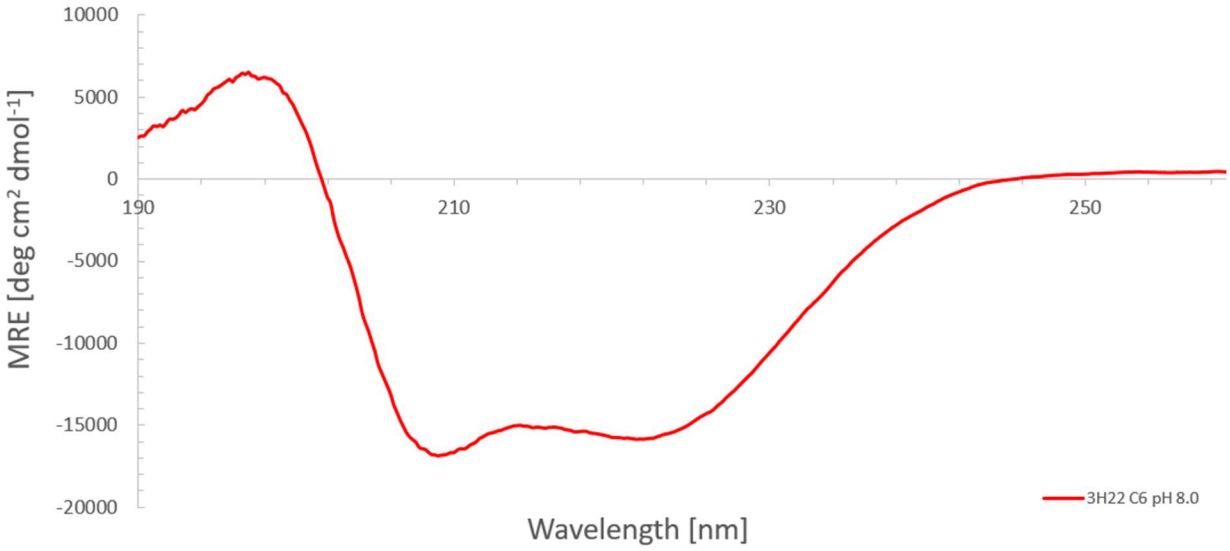

Figure 18:CD Spectra 3H22 C6. The mean residue ellipticity (MRE) of 3H22 C6 (pH 8.0)

The result indicates an α-helical fold (see Figure 18). The protein showed the most positive peak at 196 nm and two negative peaks at 209 nm and 221 nm.

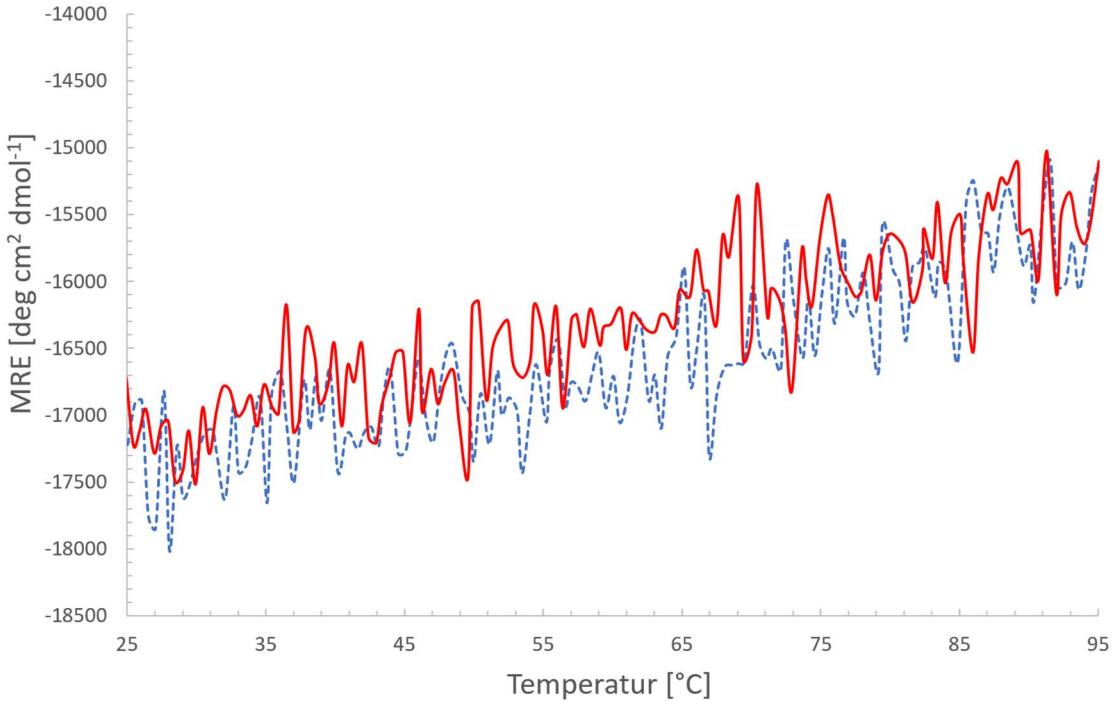

Figure 19: Temperature scan 3H22 C6. The Temp-scan monitored from 20 to 95 °C (up) and 95 to 20 °C (down).

No unfolding event could be observed while measuring the sample at 209 nm during up and down scan.

7.6 Ligand Binding Fluorescence Measurements

Fluorescence spectroscopy was used to investigate if hydrophobic, flat achiral molecules like DPH and TMA-DPH can bind in the hydrophobic environment of the 3H22 C6 core and thus by increasing the protein concentration also the fluorescence should increase until saturation is reached. Binding was quantified by increasing the amount of protein into fixed, 1µM concentration of ligands (see Figure 20)

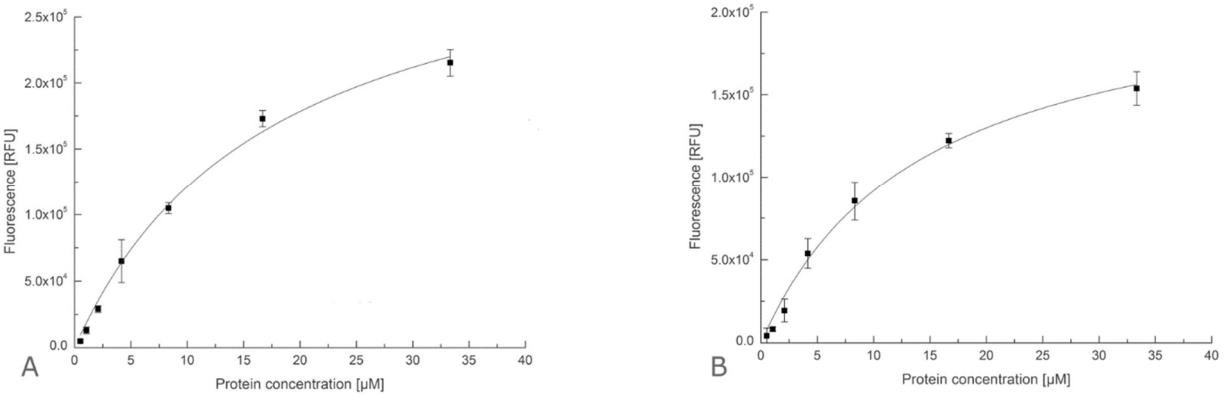

Figure 20: Saturation binding of 3H22 C6 to different ligands. The binding was recorded at λex 350 ± 20 nm and λem 426 ± 30. **A** (DPH), **B** (TMA-DPH). Conditions: protein concentration 3 - 40 µM mixed with the respective ligand.

The two dyes (see Figure 20) are environmentally sensitive. A dye like DPH fluoresces (λ_{max}= 455 nm) only when it is in hydrophobic environments, as negative control for background binding lysozyme was used as it doesn't have any channels or pores [37].To calculate the dissociation constants (K_D) which are shown in Table 13 the resulting saturation binding curves were fitted to a single-site binding model.

Table 13: Dissociation constants of small ligands binding to 3H22 C6

	K_D [µM]
DPH	17.8 ± 2.6
TMA-DPH	14.4 ± 2.4
Prodan	-

As it is shown in Figure 21 the protein accommodated TMA-DPH in the channel, whereas Prodan did not fit into it. No dissociation constant could be observed for Prodan.

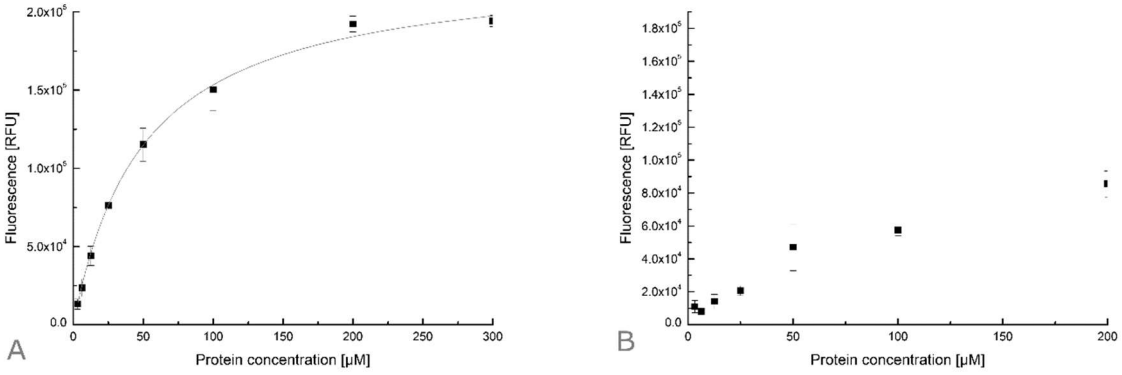

Figure 21: Saturation binding of TMA-DPH and Prodan. A (TMA-DPH), **B** (Prodan), the fluorescence measurement was recorded at 350 ± 20 nm.

The results show that more fluorescence can be detected the higher the concentration of the protein is until it is saturated. Furthermore, this indicated the correct formation of six monomeric building blocks to a hexameric shaped protein with a hydrophobic core.

7.7 Ligand Docking

The already mentioned ligands DPH and TMA-DPH were chosen for the docking procedure (see Figure 22). Next, the binding modes were scored using their FullFitness (kcal/mol) and clustered. This clusters of complexes were confronted to an accurate and slower scoring scheme, the FullFitness, and then the ranks of their centers were updated [38]. Afterwards the clusters were ranked according to the average FullFitness of their elements. This led to 85 clusters which contained 255 different binding modes for one molecule DPH. Depending the second ligand, 46 clusters were found containing 256 different binding modes for one molecule TMA DPH. The two binding modes with highest score are shown in Table 14.

Table 14: Calculated binding modes for two different ligands into 3H22 C6 using SwissDock

Ligand	FullFitness [kcal/mol]	Estimated $\Delta G_{elec;solv}$ [kcal/mol]	Binding modes
DPH	-2118.72	-7.04	255
TMA-DPH	-2140.27	-8.40	256

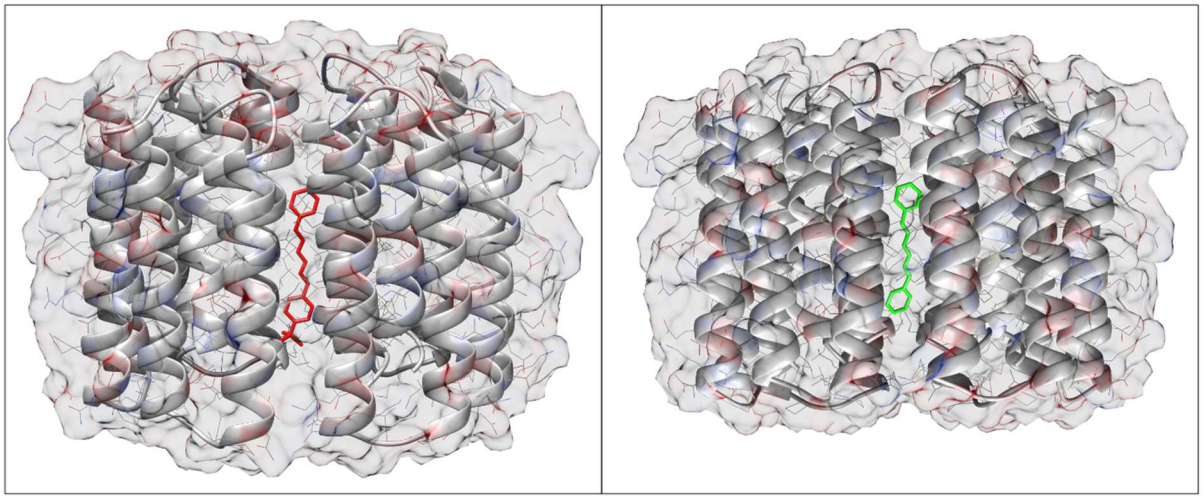

Figure 22: Binding mode for two different ligands. On the left side one molecule of TMA-DPH (indicated in red) is bound into the hydrophobic lumen of 3H22 C6. On the right side one molecule of DPH (indicated in green) is bound.

7.8 Native Mass spectrometry of 3H22 C6

For further confirmation that six monomers can form an hexameric protein complex (3H22 C6) mass spectroscopy (MS) experiments were performed. The data was kindly provided by BRUKER and Dr Andreas Winkler. Furthermore, the method was chosen to overcome the problem ensuring that the right protein fractions were taken from the purification procedure to get enough of pure protein in the desired oligomerization state for characterization and crystallization experiments.

With liquid chromatography-mass spectrometry (LC-MS) a monomer with approximately 9125 Da could be observed (see Figure 23). This result fits to the molecular weight of 9131 Da which was observed with ProtParam. Additionally, it indicates that the samples which were used for other experiments like the Western Blot (see chapter 7.3) were the protein in the desired oligomerization state.

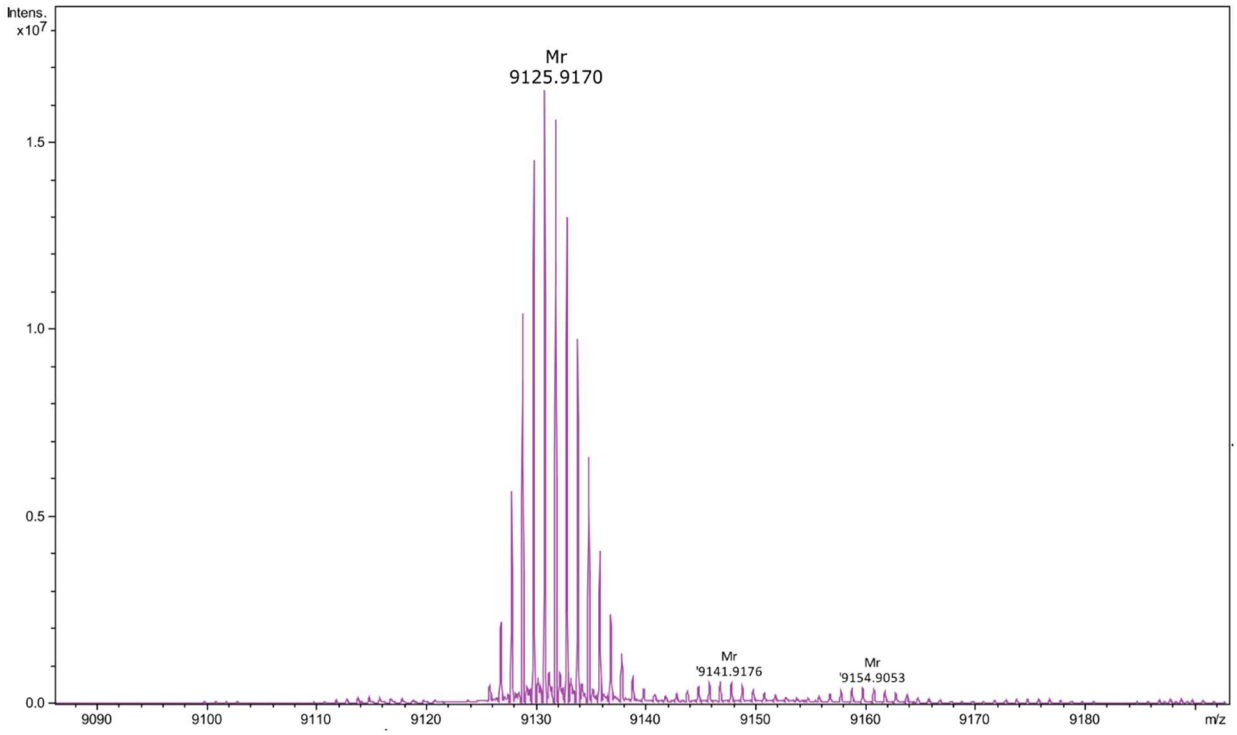

Figure 23: LC-MS Deconvoluted Monomer: A monomer with 9125 Da could be observed of a 5 uL injected protein sample.

Fortunately, the maximal entropy deconvolution of the charge envelope revealed a protein complex of 55.7 kDa (52.3 and 59.1 kDa) which is shown in Figure 24. This additionally validates that a hexameric conformation is possible as it was designed to be.

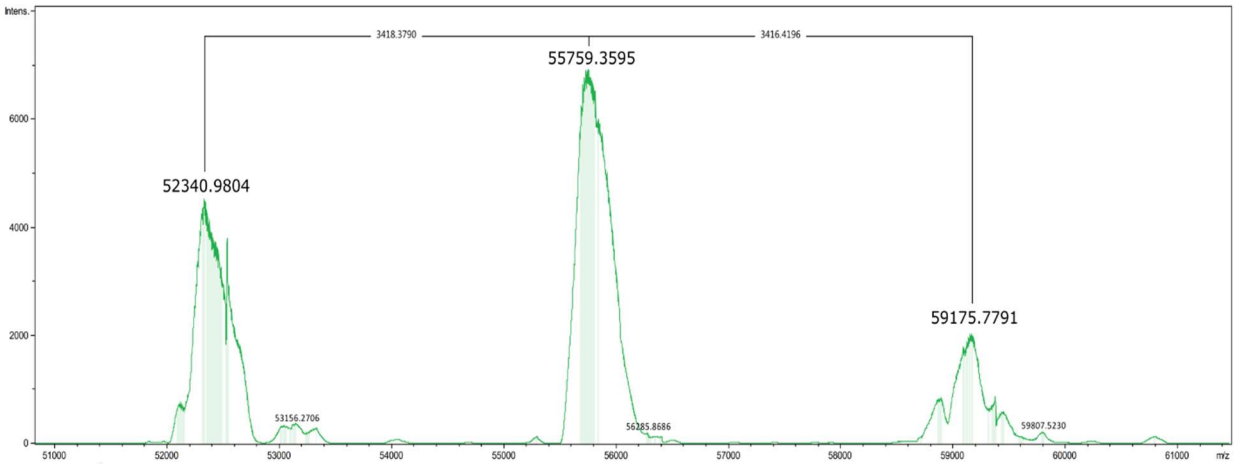

Figure 24:Protein complex of 55.7 kDa. Maximum Entropy (MaxEnt) Deconvolution of an electrospray ionization (ESI) sample (Average spectrum 2.9 min acquisition). Distribution of masses of a hexameric complex of approximately 55 kDa by charge could be observed through the “Charge State Ruler Tool” of BRUKER.

Other multimeric states, which were also present, possibly represent artifacts. By increasing harsher conditions for the desolvation, a preferential loss of lower charged states of the hexamer and diminished occurrence of possible 12-meric and18-meric aggregates as potential artifacts were induced. This result was accompanied by increasing occurrence of a 2283.9 m/z fragment (see Figure 25).

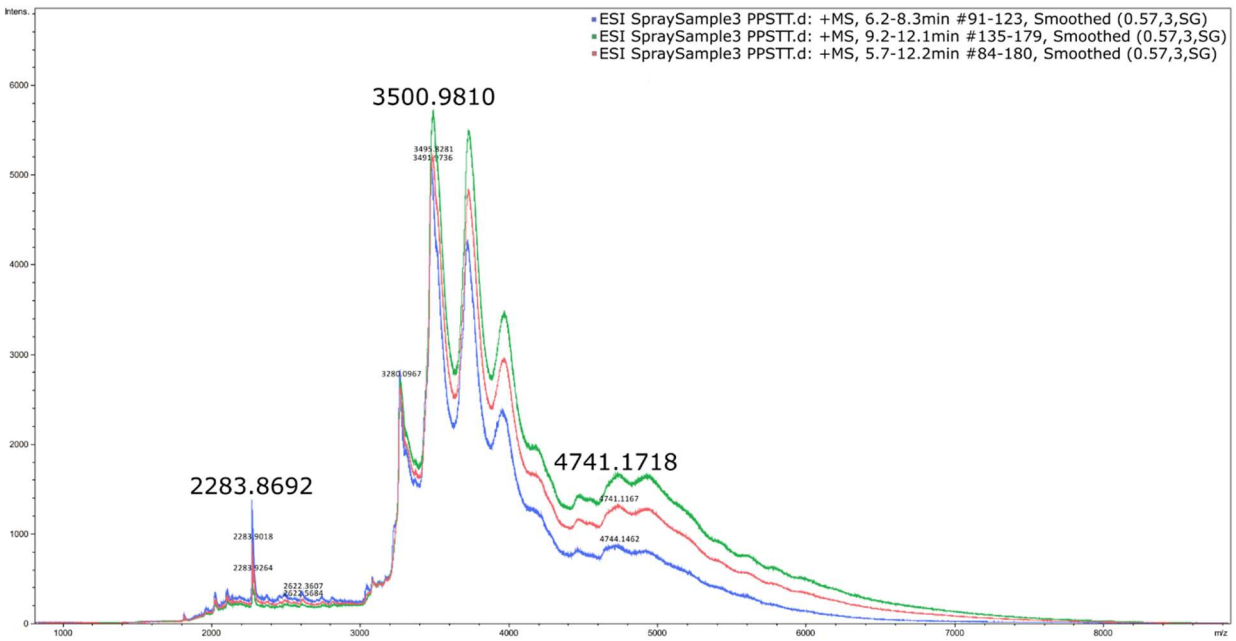

Figure 25: Effects of enhanced in-source collision-induced dissociation (ISCID) and collision energy. Blue (ESI spray sample 6.2-8.3min), green (ESI spray sample 9.2-12.1min), red (ESI spray sample 5.7-12.2 min)

7.9 3H22 C9 refinement

The original C9 design was expressed and purified. The data showed results for a construct in the range of an octameric respectively nonameric structure. It was successfully crystallized but unfortunately did not form as it was supposed to form, it crystallized as a kind of stable dimer of tetramers. The redesign of the 3H22 C9 protein was performed to change the position of a single helix forming alanine residue to a threonine. This was not only accomplished to get a more stable loop region but rather to get an alpha helical capping residue in the design. In contrast to alanine the residue forms an additional hydrogen bond with the backbone and should further improve the stability and lead to the correct formation of the desired structure (see Figure 26)

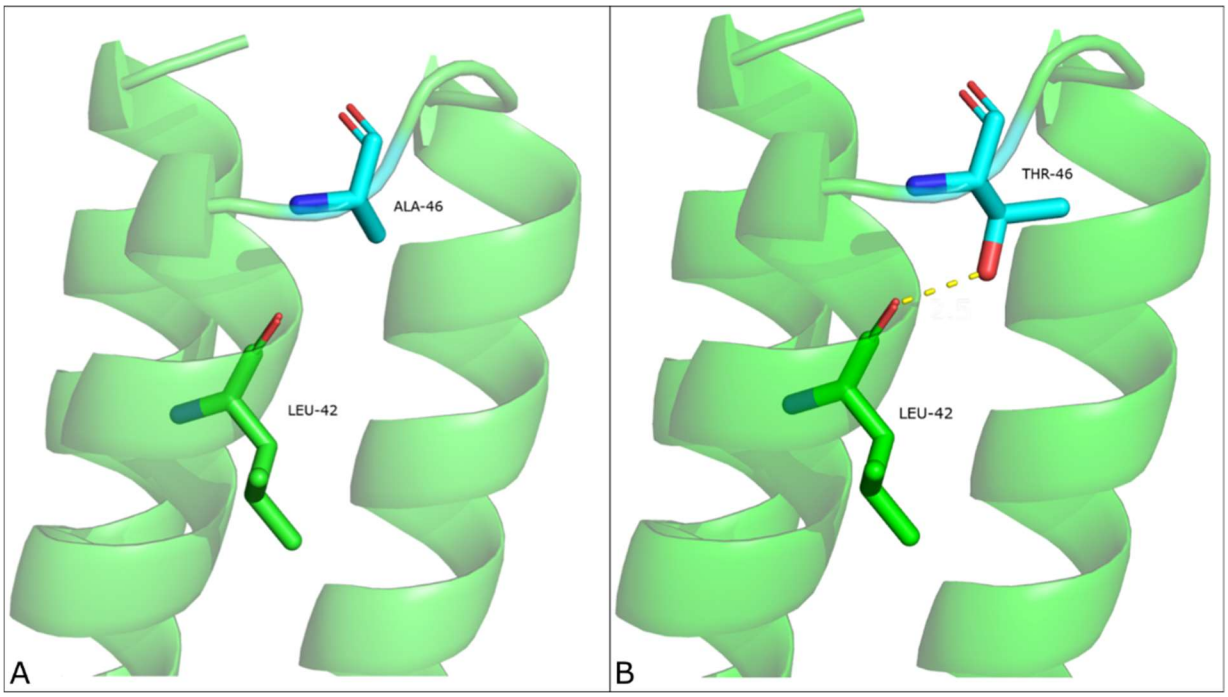

Figure 26: Refinement of 3H22 C9. The figure shows the loop regions of one of the nine chains of the protein construct. **A** (original C9 design with ALA-46), **B** (C9 refinement structure with THR-46 residue and an additional hydrogen bond between LEU-42 and THR-46)

To introduce the mutation into the design the *MutateResidue* method of the RosettaScripts Mover was used. Threonine was integrated into each chain at *target position 46* the resulting structure was scored after relaxation (see Table 15).

Table 15: Total score comparison of original C9 construct and the refinement.

Pose	Total score [REU]
3H22_C9	-1387.714
3H22_C9_refinement	-1535.262

8 Discussion

Only one of the three constructs was successfully expressed in *E. coli*, however, comprehensive optimization for the expression of 3H22 C6 could be established. The main difference between the constructs were the total number of chains. The C3 construct was designed to form only three chains. It was not possible to express enough C3 for further experiments. Not even a sensitive technique like the Western blot led to a signal, the problem may lie on a harmful effect that blocked or inhibited the heterologous protein expression [39]. There was always a slow growth pattern thus the issue could have been gene toxicity. In every expression approach OD_{600} of 0.6 - 0.8 was reached after 5 to 7 h regardless which *E. coli* strain was used (see Figure 8).This could explain why no protein was expressed and a toxicity may have caused a significant decline in bacterial growth, which dramatically decreased expression capabilities [40]. In the first expression test it was possible to express 3H22 C9, but still a point mutation was required (see Figure 8). Since the expression of 3H22 C6 worked well, it was decided to concentrate on the biochemical and biophysical characterization as well as crystallization of this protein. At the beginning of the laboratory work it was incredibly hard to get enough and especially pure 3H22 C6. The general workflow including transformation, protein expression and purification was established and optimized as it is described in Strategy 2: Protein expression 37 °C (see 6.2 Protein expression). Expression strategy 2 resulted in the highest amount of protein, however, more oligomeric states were observed during the last step of purification (SEC) in comparison to strategy 1 (expression at 18 °C over night).. , This was a compromise between having low protein-quantities or higher ones with various oligomerization states. The switch to more nutrient rich TB-Media brought a significant impact in protein production (see Table 16). Due to the fact that more biomass was produced, and thus more cells were available for expression, the overall yield improved. To avoid protein samples with undesired oligomeric states, more expression temperatures should have been considered, especially lower temperatures than 37 °C. Moreover, a different temperature could improve protein folding and potentially reduce the amount of time which was used for additional purification steps.

Table 16: Comparison of two *E. coli* strains in 3H22 C6 production.12 L batch at the very beginning of the thesis versus the last 3 L batch at the end.

Production strain	[mg protein] / [g wet cell pellet]
E. coli BL21Star (DE3) – 12 L	0.60
E. coli BL21-CodonPlus – 3 L	1.18

One thought we had was that due to the relatively high expression temperature, protein folding was impaired. Curiously, after performing an SDS gel analysis, most of the sample was always identified as the desired protein (monomer) with a few impurities. To overcome the problem with

the different states, some experiments were performed where it was tried to shift the oligomers to the desired conformation. First of all, a second purification step via FPLC was completed, followed by either heating up the purified protein sample, storing it in the fridge, or drastically changing buffer conditions. Unfortunately, we did not have any success. Therefore, there was always a lack of protein for crystallization and characterization experiments.

The thermal shift assay showed a high thermal stability for the 3H22 C6 construct. Since the C6 has six tyrosine residues which form a ring at one end of the protein (see Figure 17), it would have been interesting to see if the C3 and C9 construct would have behaved similar lacking this ring, or if the tyrosine ring had a significant impact on stability. For future experiments, it could be interesting to additionally measure chemical denaturation with guanidine hydrochloride or other denaturing compounds.

Moreover, the results of thermal stability could be confirmed in a CD thermoscan during heating of the protein to 95 °C followed by cooling down to the initial temperature of 25°C (see Figure 19). The CD spectra of the C6 construct showed an α-helical signal (see Figure 18) that compares well to known values for the secondary structures of α-helical proteins. Characteristics include a positive peak at 193 nm and two negative peaks at 222 nm and 208 nm [36]. Similar CD spectra should be expected for C3 and C9. A further spectroscopic analysis for secondary structure determination could have been NMR chemical shifts. However, the CD measurement provided a rapid solution for the analysis of the secondary structure instead of assigning the chemical shifts to particular residues, which would have made the determination considerably slower [41].

To ensure that the purified protein sample which was later used for different crystallization setups was really the desired one, a Western blot was performed. At this state of research, the different fractions of the FPLC could not be associated to different oligomerization states, they were only assumptions. The Western blot was used to corroborate previous results and assumptions for our C6 protein. No signal was observed for C3, hence no successful expression for this construct could be confirmed.

Protein production of the monomeric building blocks is one thing, but the fact that they form a hexamer had to be proven. There was the idea of a fluorescence assay, since a ligand like DPH fluoresces (λ_{max} =455 nm) only when it is in hydrophobic environment. This region can be found if six building blocks self-assemble to the designed structure of 3H22 C6. In case of specific interaction between the ligand and the protein, a concentration-dependent saturation should occur (see Figure 20). To test this hypothesis, three fluorescent ligands of different sizes were

chosen: DPH, TMA-DPH and Prodan. Lysozyme was used as negative control because it does not have any channel or pores for specific binding. The resulting saturation binding curves were fit to a single-site binding model to return dissociation constants (see Table 13). Since the core of the construct is composed of hydrophobic residues and the ligands were bound (see 7.7 Ligand Docking), the experiment could be considered a success.

Coincidentally, there was the option to use native mass spectrometry for further protein analysis. The MS result finally confirmed that there is a monomer with approximately 9125 Da where six of them can build up the hexamer, which was detected as well with 55.7 kDa. Furthermore, there is finally a reason for the oligomerization states because it seems that the protein can stack to 12-meric and 18-meric aggregates (see Figure 25). In which exact form these aggregates can appear requires more investigation. Perhaps one hexamer stacked over another one or they dformed a chain-like structure.

The first crystallization setups were performed by hand and for most conditions the protein precipitated immediately. However, a barely visible crystal was grown in one of them but was too small to harvest. Previous work showed that the 3H22 C6 does not crystallize until a protein concentration of 20 mg/µL or higher was used. Therefore, most setups were performed using the Oryx8 robot, which led to the promising result of octahedral shaped crystals (see Figure 14). Unfortunately, no recorded diffraction data was appropriate enough to solve the crystal structure.

9 Conclusion

The aim was the biochemical and biophysical characterization of computationally designed helical oligomers by protein expression in *E. coli* and purification using affinity, and size exclusion chromatography. Additionally, the characterization via CD spectroscopy as well as thermal shift assays was used to confirm that the protein is fully folded. Finally, we could obtain crystallization conditions and managed to improve the loop geometry of one particular design, 3H22C9. Due to the fact that there was the opportunity for fluorescence measurements and ligand docking, the originally planned characterization experiments could be extended. The goal was primarily reached for 3H22 C6 as it was expressed and characterized successfully. Regarding the crystal structure, future work needs to be done for growing bigger and especially more crystals to get useable diffraction data. Unfortunately, the 3H22 C3 protein could not be characterized because no protein expression was observed. Remodelling and refinement of one of the connecting loops of 3H22 C9 was achieved using Rosetta by introducing an alpha helical capping residue. This still requires experimental confirmation.

For the future, promising objectives need to be studied by computationally aided methods or through additional ligand docking experiments. However, more research is still required to confirm the crystal structures of the full *de novo* set of C3, C6, C9 constructs.

10 Appendix

10.1 Sequences

Table 17 Sequences of *de novo* designed protein loops

Construct	Loop	Amino acid code
3H22 C9	Loop 1	KNTTDT
	Loop 2	HGVHE
3H22 C3	Loop 1	NGVTQ
	Loop 2	NSDKT
3H22 C6	Loop 1	NSDNT
	Loop 2	NGVPT

3H22 C3:

```
        10          20          30          40          50          60
MTWELIIKLI KEANKLLEKL KKNGVTQTII MEVWMRMIIL VMLLENSDKT KEAAEKMLKK

        70
MKELFKKAKG SGWHHHHHH
```

3H22 C6:

```
    10          20          30          40          50          60
MTETLIRLLE ELARVLLEIL KQNGVPTNVI EAVRKAMEIL LKMLKNSDNT AEAAAYMAIA

        70          80
MILLLILAKG SGWLEHHHHH H
```

3H22 C9:

```
        10          20          30          40          50          60
MTEEEIKKLE KLARLLLEAL KEHGVHEAII RAVEFLMRLL LKLLKNATDT IRAAAEMLKE

        70
MWQAFEDARG SGWHHHHHH
```

11 List of Sources

[1] F. H. C. Crick, “The Fourier transform of a coiled-coil,” *Acta Crystallogr.*, vol. 6, no. 8, pp. 685–689, Sep. 1953.

[2] A. Leaver-Fay *et al.*, “Rosetta3: An object-oriented software suite for the simulation and design of macromolecules,” in *Methods in Enzymology*, vol. 487, no. C, Academic Press Inc., 2011, pp. 545–574.

[3] P. S. Huang *et al.*, “High thermodynamic stability of parametrically designed helical bundles,” *Science*, vol. 346, no. 6208, pp. 481–485, Oct. 2014.

[4] P.-S. Huang, S. E. Boyken, and D. Baker, “The coming of age of *de novo* protein design,” *Nature*, vol. 537, no. 7620, pp. 320–327, Sep. 2016.

[5] C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen, “The Genetic Control of Tertiary Protein Structure: Studies With Model Systems,” *Cold Spring Harb. Symp. Quant. Biol.*, vol. 28, no. 0, pp. 439–449, Jan. 1963.

[6] L. Cruzeiro and L. Degrève, “Exploring the Levinthal limit in protein folding,” *J. Biol. Phys.*, vol. 43, no. 1, pp. 15–30, Mar. 2017.

[7] K. A. Dill and H. S. Chan, “From levinthal to pathways to funnels,” *Nature Structural Biology*, vol. 4, no. 1. pp. 10–19, Jan-1997.

[8] S. E. Radford, “Protein folding: Progress made and promises ahead,” *Trends in Biochemical Sciences*, vol. 25, no. 12. Elsevier Current Trends, pp. 611–618, 01-Dec-2000.

[9] P. J. Fleming and G. D. Rose, “Do all backbone polar groups in proteins form hydrogen bonds?,” *Protein Sci.*, vol. 14, no. 7, pp. 1911–1917, Jul. 2005.

[10] H. Deng, Y. Jia, and Y. Zhang, “Protein structure prediction,” *Int. J. Mod. Phys. B*, vol. 32, no. 18, Jul. 2018.

[11] S. Ovchinnikov *et al.*, “Protein structure determination using metagenome sequence data,” *Science*, vol. 355, no. 6322, pp. 294–298, Jan. 2017.

[12] “The Protein Data Bank.,” *Methods Biochem. Anal.*, vol. 44, pp. 181–198, 2003.

[13] A. W. Senior *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, pp. 706–710, Jan. 2020.

[14] I. V. Korendovych and W. F. DeGrado, “*De novo* protein design, a retrospective,” *Quarterly Reviews of Biophysics*, vol. 53. Cambridge University Press, 2020.

[15] P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber, “A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants,” *Science*, vol. 262, no. 5138, pp. 1401–1407, Nov. 1993.

[16] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim, “High-resolution protein design with backbone freedom,” *Science*, vol. 282, no. 5393, pp. 1462–1467, Nov. 1998.

[17] A. R. Thomson *et al.*, “Computational design of water-soluble α-helical barrels,” *Science*, vol. 346, no. 6208, pp. 485–488, Oct. 2014.

[18] S. E. Boyken *et al.*, “*De novo* design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity,” *Science*, vol. 352, no. 6286, pp. 680–687, May 2016.

[19] Z. Chen *et al.*, “Programmable design of orthogonal protein heterodimers,” *Nature*, vol. 565, no. 7737. Nature Publishing Group, pp. 106–111, 03-Jan-2019.

[20] S. Cooper *et al.*, “Predicting protein structures with a multiplayer online game,” *Nature*, vol. 466, no. 7307, pp. 756–760, Aug. 2010.

[21] B. Koepnick *et al.*, “*De novo* protein design by citizen scientists,” *Nature*, vol. 570, no. 7761. Nature Publishing Group, pp. 390–394, 20-Jun-2019.

[22] R. F. Alford *et al.*, “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design,” *J. Chem. Theory Comput.*, vol. 13, no. 6, pp. 3031–3048, Jun. 2017.

[23] Y. R. Lin *et al.*, “Control over overall shape and size in *de novo* designed proteins,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 40, pp. E5478–E5485, Oct. 2015.

[24] N. Koga *et al.*, “Principles for designing ideal protein structures,” *Nature*, vol. 491, no. 7423, pp. 222–227, Nov. 2012.

[25] H. Bayley, “Nanopore sequencing: From imagination to reality,” *Clin. Chem.*, vol. 61, no. 1, pp. 25–31, Jan. 2015.

[26] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, “Overview of Next-Generation Sequencing Technologies,” *Curr. Protoc. Mol. Biol.*, vol. 122, no. 1, p. e59, 2018.

[27] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, 1977.

[28] A. H. Laszlo, I. M. Derrington, and J. H. Gundlach, “MspA nanopore as a single-molecule tool: From sequencing to SPRNT,” *Methods*, vol. 105. Academic Press Inc., pp. 75–89, 01-Aug-2016.

[29] M. Wanunu, T. Dadosh, V. Ray, J. Jin, L. McReynolds, and M. Drndić, “Rapid electronic detection of probe-specific microRNAs using thin nanopore sensors,” *Nat. Nanotechnol.*, vol. 5, no. 11, pp. 807–814, Nov. 2010.

[30] A. J. Storm, J. H. Chen, X. S. Ling, H. W. Zandbergen, and C. Dekker, “Fabrication of solid-state nanopores with single-nanometre precision,” *Nature Materials*, vol. 2, no. 8. European Association for Cardio-Thoracic Surgery, pp. 537–540, 2003.

[31] E. C. Yusko *et al.*, “Controlling protein translocation through nanopores with bio-inspired fluid walls,” *Nat. Nanotechnol.*, vol. 6, no. 4, pp. 253–260, Feb. 2011.

[32] D. Fologea, B. Ledden, D. S. McNabb, and J. Li, “Electrical characterization of protein molecules by a solid-state nanopore,” *Appl. Phys. Lett.*, vol. 91, no. 5, p. 053901, Jul. 2007.

[33] J. E. Reiner, J. J. Kasianowicz, B. J. Nablo, and J. W. F. Robertson, “Theory for polymer analysis using nanopore-based single-molecule mass spectrometry,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 27, pp. 12080–12085, Jul. 2010.

[34] H. Shen *et al.*, “*De novo* design of self-assembling helical protein filaments,” *Science*, vol. 362, no. 6415, pp. 705–709, Nov. 2018.

[35] M. Zhou, Q. Li, W. Kong, and R. Wang, “Experimental methods used for identifying small-molecule inhibitors of protein-protein interaction,” in *Targeting Protein-Protein Interactions by Small Molecules*, Springer Singapore, 2018, pp. 95–133.

[36] N. J. Greenfield, “Using circular dichroism spectra to estimate protein secondary structure,” *Nat. Protoc.*, vol. 1, no. 6, pp. 2876–2890, Jan. 2007.

[37] F. Thomas *et al.*, “*De Novo*-Designed α-Helical Barrels as Receptors for Small Molecules,” *ACS Synth. Biol.*, vol. 7, no. 7, pp. 1808–1816, Jul. 2018.

[38] A. Grosdidier, V. Zoete, and O. Michielin, “EADock: Docking of small molecules into protein active sites with a multiobjective evolutionary optimization,” *Proteins Struct. Funct. Genet.*, vol. 67, no. 4, pp. 1010–1025, Mar. 2007.

[39] L. Dumon-Seignovert, G. Cariot, and L. Vuillard, “The toxicity of recombinant proteins in Escherichia coli: A comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3),” *Protein Expr. Purif.*, vol. 37, no. 1, pp. 203–206, Sep. 2004.

[40] F. Saida, M. Uzan, B. Odaert, and F. Bontems, “Expression of Highly Toxic Genes in E. coli: Special Strategies and Genetic Tools,” *Curr. Protein Pept. Sci.*, vol. 7, no. 1, pp. 47–56, Feb. 2006.

[41] J. T. Pelton and L. R. McLean, “Spectroscopic methods for analysis of protein secondary structure,” *Analytical Biochemistry*, vol. 277, no. 2. Academic Press Inc., pp. 167–176, 15-Jan-2000.