A General, Symmetry-Based Approach for the Assembly of Proteins into Nanoscale Polyhedra

by

Aaron Sciore

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in The University of Michigan
2016

Doctoral committee:

Professor E. Neil G. Marsh, Chair
Assistant Professor Julie S. Biteen
Assistant Professor Brent R. Martin
Associate Professor Georgios Skiniotis
Professor Nils G. Walter

Dedication

This dissertation is dedicated to Dr. Thomas Morson, and to the fight against cancer.

Acknowledgements

I would like to thank everyone that made this dissertation possible. Deepest thanks to my advisor Dr. Neil G. Marsh, not only for giving me the opportunity to work in his laboratory, but also for his continued guidance over all the years working alone on a new and challenging project. I would like to thank all the Marsh lab members, past and present, who have contributed to a relaxed and productive atmosphere and contributed valuable insights over the years. I would like to thank Bekir Eser, Debasis Das, and Tad Ogorzalek for helping me understand many basic molecular biology laboratory techniques and experiment designs. I would like to thank the members of the lab that have worked on my project, undergraduate students Alexander Bader and Lawrence Chen, and graduate students Dr. Somaye Badieyan, Ajitha Sand-Cristie, and Kelsey Diffley, for their assistance with all the various odds and ends that I didn't have time to do myself. I would especially like to thank my collaborators, Dr. Neil King and Dr. William Scheffler at the University of Washington for writing the Rosetta code and running the simulations, Dr. Min Su and Prof. Georgios Skiniotis for their help in visualizing the protein cages by negative stain and cryo-TEM images, and with the refinement and fitting of the 3-d electron density reconstruction of the octahedron, Phillip Koldewey for his assistance in preparing samples for analytical ultracentrifugation and performing high-resolution data analysis by Ultrascan, and Joseph Eschweiler for his assistance in obtaining and explaining high resolution mass spectrometry data. Finally, thanks to Dr. Ben Buer and Cullen Whitmore for creating several of the cartoon representations used throughout this thesis, and to Dr. Elizabeth

Chen for illumination of the theoretical underpinnings of symmetry pairs and the various complexes they can form.

Table of Contents

List of Figures

List of Appendices

Abstract

The assembly of individual protein subunits into large-scale symmetrical structures is widespread in Nature and confers unique biological properties which have potential applications in nano-technology and medicine. While efforts to functionalize and repurpose existing protein complexes have been mainly successful, designing well-defined *de novo* protein complexes remains an unsolved problem. A major challenge in engineering *de novo* symmetrical assemblies has been to design interactions between the protein subunits so that they specifically assemble into the desired structure. Prior *de novo* protein cages have been developed with moderate success, but suffer from a lack of generalizability and require significant computational effort and screening of mutant fusion proteins. The design and optimization of a simple, generalizable approach to designing novel fusion proteins which assemble into cage-like structures will be the subject of this dissertation. We show that by genetically fusing a C_4-symmetric coiled-coil to the C-terminus of a C_3-symmetric trimeric protein via a short, flexible linker, we can assemble a well-defined 24-subunit protein cage with octahedral symmetry. The flexible nature of these assemblies alleviates the need for rigorous interface modeling, requiring only minimal computation to determine the length of the linker sequence. This is the first *de novo* designed symmetrical protein complex to incorporate a C_4 symmetry element, and we anticipate this method can be applied to a wider variety of proteins and symmetries, which may open up a new avenue of research into designer protein cages with unique, built-in functionalities.

Chapter 1

Introduction

1.1 - Natural Protein Cages

The assembly of multiple copies of a protein subunit into large, hollow, and highly symmetric complexes, referred to in this thesis as 'protein cages', is found widely throughout Nature. These natural protein cages perform a broad range of critical functions, primarily owing to the unique microenvironment of the cage interior. Access to the interior of the cage is controlled by pores of varying sizes, sharply limiting the amount of cellular machinery that can interact with the interior of the cage. Second, the residues on the interior surface of protein cages are brought into close proximity with any molecule that enters this interior, the effect of which is multiplied across every subunit in the protein cage. These features lead to behavior that would be otherwise impossible in the exterior environment.

A well-known example of the successful utilization of this microenvironment is the iron storage protein ferritin: an octahedral, 24-subunit protein cage that is highly conserved across all organisms. The interior surface of the ferritin cage contains a large number of negatively charged residues (Fig. 1.1b).[1] The negatively charged residues on the interior binds ferrous iron atoms in close proximity, catalyzing their oxidation to the ferric form which in turn serves as a nucleation site for other ferrous iron atoms, which crystallizes as iron oxide. The hollow ferritin interior can hold up to 4,500 iron atoms, which allows ferritin to participate in iron distribution

pathways in a far more effective manner than a protein containing discrete binding sites for individual metal ions.

Other natural protein cages function in a similar manner. Many viruses encapsulate their DNA or RNA in an icosahedral protein complex known as a capsid that is comprised of multiples of 60 protein subunits (Fig. 1.1c). The interior residues of capsid subunits are positively-charged, and can thus bind the viral DNA or RNA, packaging them for transmission to a new host. These capsids also exhibit another behavior that is critical for viral transmission: favorable binding interactions between adjacent capsid proteins, multiplied across the correctly assembled capsid, creates a cooperative binding effect that imbues the capsid with a high degree of stability, and protects the capsid proteins and enclosed nucleotides against environmental degradation.[2,3]

Molecular chaperones are highly conserved protein complexes that play an important role in preventing the misfolding and aggregation of cellular proteins during periods of cellular stress, such as high heat or an oxidative environment. Many molecular chaperones exist as protein cages with a variety of symmetries, not only to take advantage of cooperative binding for increased stability of the assembled protein complex in a high-stress environment, but also to generate an interior microenvironment that aids in protein refolding. The interior surface of these chaperonin protein cages is highly hydrophobic, affording an excellent binding surface for exposed hydrophobic interior residues of misfolded proteins. GroEL, a barrel-shaped 14-subunit protein cage with a 4.5 nm wide hydrophobic interior channel (Fig. 1.1e) binds a wide range of proteins, with protein fragments that are unstructured in solution exhibiting alpha helical

properties when bound to GroEL.[4] The refolding activity provided by GroEL was shown to be critical to cell function, as deletion of the GroEL gene was universally lethal to cells.[5]

The last notable function of natural protein cages involves utilizing the interior to contain reaction pathways that generate unstable intermediates. The microenvironment provides excellent temporary storage, protecting these intermediates from interacting with other parts of the cell. This behavior is exemplified by pyruvate dehydrogenase, an icosahedral complex consisting of three proteins that catalyze the conversion of pyruvate to acetyl-CoA (Fig. 1.1a). The initial reaction decarboxylates the pyruvate, generating an unstable thioester compound that is transferred quickly to a second enzyme in the complex which converts the thioester to acetyl-CoA.[6] Similar functionality can be found in primitive bacterial organelles known as bacterial microcompartments (BMCs) (Fig. 1.1f). BMCs primarily comprise hexameric shell proteins, such that they form a honeycomb pattern, interspersed with a small number of kinked, pentameric shell proteins.[7] The combination of these yields a large (40-200 nm in diameter) pseudo-icosahedral protein complex capable of encapsulating multiple enzymes in a reaction pathway to increase the efficiency of catalysis. The most well studied BMC is the carboxysome, which houses both carbonic anhydrase and RuBisCo. The carbonic anhydrase produces CO_2 from bicarbonate, which is quickly taken up by RuBisCo to further the Calvin cycle. Studies that delocalized carbonic anhydrase to the cytosol showed that significant loss of RuBisCo function occurred, suggesting that without the carboxysome, the CO_2 produced by carbonic anhydrase is too volatile and diffuses out of the cell.[8]

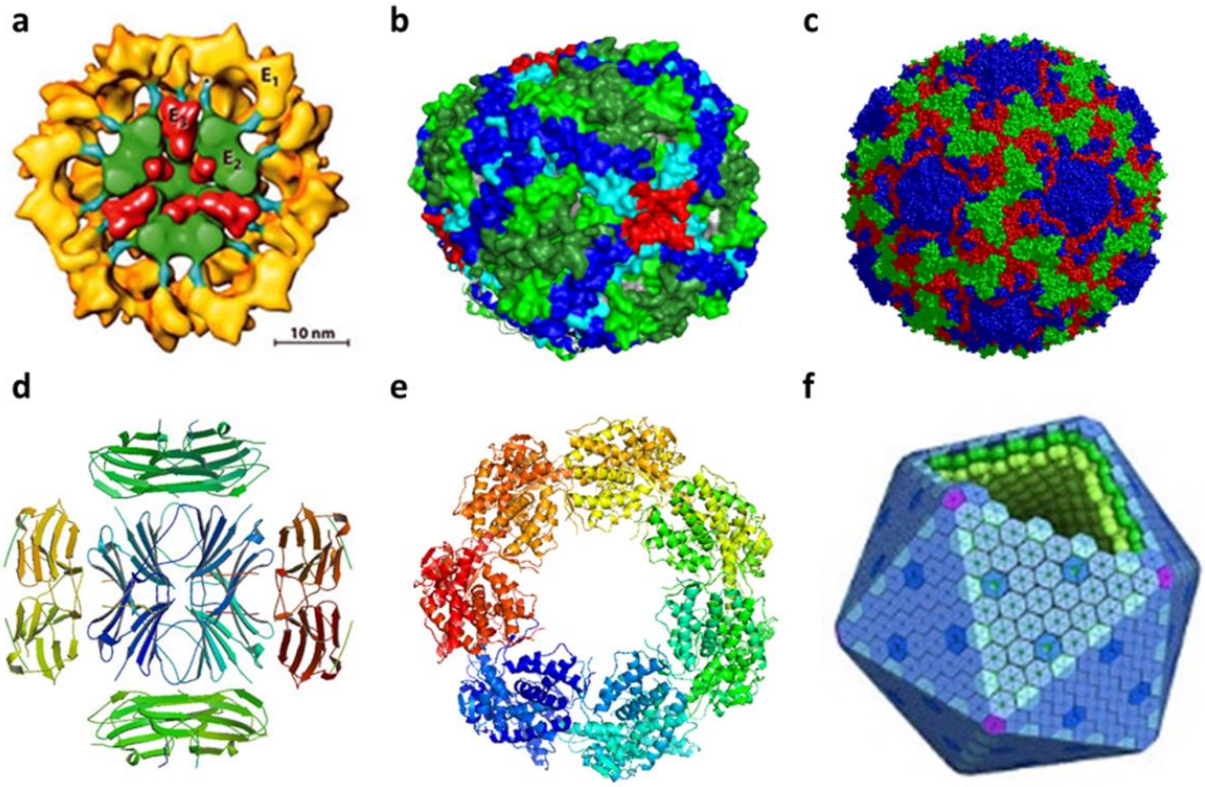


Figure 1.1. Natural protein cages**.** a) Cross-section of a 3-dimensional reconstruction of the icosahedral pyruvate dehydrogenase complex (Image taken from Ref. [9]). b) Crystal structure of octahedral bacterioferritin (PDB ID 3GVY). c) Crystal structure of icosahedral rhinovirus capsid (PDB ID 4RHV). d) Crystal structure of tetrahedral heat shock protein (PDB 2BYU). e) View along the 7-fold axis of the crystal structure of GroEL (PDB ID 1GRL). f) Cartoon of the assembly of bacterial microcompartments from hexameric and pentameric subunits (Image credit: Wikimedia).

1.2 - Functionalization of Natural Protein Cages

Naturally-occurring protein cages have been investigated for use in a diverse range of materials science and nanomedicine applications. Protein cage functionalities can be introduced at one or more of three distinct regions: the interior surface, the exterior surface, and the interfacial pores. The simplest functionalization of protein cages involves modifying a protein cage to perform tasks similar to its cellular function. Purified ferritin has been long known to react *in vitro* with excess iron salts to biomineralize iron oxide nanoparticles of a

defined size. This process was found to be quite general, with size-constrained nanoparticles of silver, platinum, palladium, cobalt oxide, and cadmium sulfide, among many others, being mineralized in ferritin's interior cavity.[10] *In vitro* biomineralization could also be replicated in heat shock proteins[11-13] and viral capsids[14] by mutating the hydrophobic or cationic interior of these proteins to metal-binding or anionic residues, with the diameters of synthesized iron oxide nanoparticles dictated by the interior diameter of these protein cages. Adding a short peptide sequence known to specifically bind an ordered assembly of CoPt to the ferritin microenvironment resulted in CoPt nanoparticles with ferromagnetic properties.[15] Protein cage-derived iron-oxide nanoparticles have also been used as a nucleation site for the synthesis of single-walled carbon nanotubes, and the diameter of these nanotubes is proportional to the size of the seed iron oxide particle.[16]

Size constrained nanoparticles, in particular iron and cobalt oxide, are of particular interest in nanoelectronics research.[17] A single layer of protein cages containing nanoparticles can be deposited onto a precoated surface in a tight, hexagonal packing pattern, with a packing density close to theoretical values.[18] In what is known as the bio-nano process, silicon wafers or other substrates are precisely patterned with hydrophobic or hydrophilic coatings, with nanoparticle-containing ferritin localizing only on the hydrophilic coat. This is then exposed to heat, burning away the ferritin to leave only the iron oxide nanoparticle, which is then reduced to yield a precisely-patterned array of metallic iron spheres.[19] This has been used to generate semiconducting logic devices on the nano scale such as thin film transistor flash memory and floating nanodot gate memory devices.[20] Parameters of the bio-nano process can be controlled

with superior precision than existing nanoelectronics technology[21], but this technology has not yet scaled to market.

The microenvironment of the protein cage interior presents an attractive target for entrapping catalytic species, essentially creating a bioreactor.[22] Palladium nanoparticles formed on the inside of ferritin cages were shown to catalytically hydrogenate olefins in solution.[23] The kinetics of this process could be controlled by varying the size of the olefin, indicating that this reaction is limited by diffusion of the olefin through ferritin's nanopores. Similarly, encapsulated gold and silver nanoparticles could catalyze the reduction of nitrophenol[24] and platinum nanoparticles synthesized inside of a tetrahedral heat shock protein could catalyze the reduction of protons to hydrogen gas.[25] In both of these reactions, the catalytic rate was increased relative to bulk metals in solution. The hollow interior of a protein cage has also sparked interest for entrapment and immobilization of enzymes, fashioning a bioreactor similar to the carboxysome. Protein cages can be assembled around enzymes in solution, trapping them, and these enzymes retain their catalytic activity.[26] However, this catalytic activity is reduced if a large number of enzymes are encapsulated in each protein cage, indicating that crowding may have a deleterious effect.[27] Using one of the larger viral capsid shells, multiple enzymes in a metabolic pathway were co-localized, but this had negligible effect on turnover rates.[28] The most promising application of protein cages as nanoreactors involves polymerization reactions. Rhodium (II) complexes that catalyze the polymerization of phenylacetylene were bound to the interior of ferritin, and polymerization was induced.[29] The polymers formed had a very narrow size distribution of 130 ± 15 monomers, which was

concentration independent. This level of precision in polymerization control may open up new avenues of research into designed smart materials.

Protein cages also hold promise for medical therapeutics. Taking advantage of a viral capsid's evolved ability to cross the cell membrane and release its genetic payload into a cell, viral capsid shell proteins are an attractive target for biocompatible functionalization. Without their viral genes, these assembled capsids are safe for medical applications and possess a sizeable interior cavity. Mimicking its natural function, the capsid interior can be loaded with therapeutic genes, and when fully assembled these genes are protected from DNAses by the capsid shell.[30] Currently, functionalized capsids are able to encapsulate plasmids as large as 17.6 kbp.[31] These viral capsids, loaded with custom DNA, may be useful for therapeutic gene delivery. For example, when an exotoxin-encoding plasmid was packaged into a viral capsid and injected into tumor cells, tumor sizes were significantly reduced both *in vitro* and *in vivo.*[32] DNA has also been encapsulated in nonviral protein cages, by mutating in positively-charged residues on the interior surface of the icosahedral lumazine synthase complex.[33]

Small molecules such as cancer therapeutics can also be encapsulated in the capsid interior[34], or covalently linked to the protein cages.[35] Drug delivery via this method takes advantage of protein cages' resilience to degradation, allowing for the timed release of therapeutics.[36] While unmodified viral capsids loaded with doxorubicin, an important anti-cancer drug, was shown to cause higher cytotoxicity in cancer cells than free doxorubicin in solution[37], these therapeutic protein cages have the drawback of depositing this toxic payload indiscriminately. Therefore, significant research has gone into decorating the capsid exterior with targeting ligands that localize the therapeutic protein cage to the cell type of interest. Viral

capsids have been decorated with both small molecules[38-40] and large biomolecules[40,41] known to bind receptors that are overexpressed by many cancer cell types, and these functionalized capsids were shown to be selectively uptaken into a range of tumor cells. Several groups have gone farther, decorating viral capsids with peptide sequences that bind specific cancer cells with high selectivity. Depending on the peptide used, functionalized capsids could selectively target Jurkat leukemia T cells[42] or human hepatocellular carcinoma cells.[43] In both cases, the protein cages functionalized with both targeting ligand and therapeutics were delivered exclusively to their target cells, inducing cell death in the majority of those cancer cells without affecting any of the control cells.

By loading the viral capsids with imaging agents, it is possible to follow the localization of protein cages within cells. For *in vitro* studies, optical imaging can be applied with excellent resolution: for example, by attaching quantum dots to the exterior of HIV viral capsids, individual capsids could be identified and tracked with single molecule imaging.[44] Optical imaging is less useful for *in vivo* applications. Viral capsids outfitted with both cancer targeting and fluorescent ligands were injected into cancerous rats, and while fluorescence was localized to the tumor cells, these cells could only be detected in mammalian tissues to a depth of 500 μm.[45]

More promising for therapeutic imaging *in vivo* is the encapsulation of positron emission tomography (PET) imaging agents or magnetic resonance imaging (MRI) contrast agents within protein cages.[46] The common PET imaging agent ^{18}F was bioconjugated into viral capsids[47] and the location of these capsids could be dynamically imaged *in vivo*.[48] Iron oxide nanoparticles, a simple and effective MRI contrast agent, could be attached to viral capsids with similar

results.[49] Incorporation of Gd^{3+} into the calcium-binding sites of the cowpea chlorotic mosaic virus led to a species with the highest relaxivity values measured to date, potentially leading to new applications with low-dose MRI contrast agents.[50] Superparamagnetic iron oxide or Gd^{3+} nanoparticles can also be incorporated into the ferritin interior, resulting in species with high relaxivity values[51]. Multiple imaging agents can also be implemented in a single protein cage, and data obtained from each technique can be correlated for increased resolution.[49]

Finally, and most promisingly, assembled viral capsids have a tendency to elicit strong immunogenic responses when introduced to a host organism.[52] This feature of viral capsids has attracted great interest in the field of vaccine development, as the viral capsid shell offers an elegant solution to an unsolved issue with current vaccine technology. Most presently-approved vaccine delivery systems exist as a compromise between effectiveness and safety concerns. Since vaccines generate the strongest immune response when multiple copies of an immunogenic epitope are presented, prior vaccine delivery systems were based around either removing the DNA from a live virus and adding the epitope of interest or adding an adjuvant such as aluminum to a vaccine displaying fewer copies of that epitope to multiply the immune response. Heterologously-expressed viral capsids offer a delivery system that doesn't have to be studiously scrubbed of viral DNA nor require careful testing to minimize the safety concerns associated with adjuvants.[53] In contrast, viral capsids are safe, biocompatible, bioavailable, and still induce a strong immune response, the only major concern being that the immune response must be modulated such that it does not produce any toxic effects to the subject.[54,55] Currently over a dozen capsid-based vaccines are approved for clinical trials or clinical use, targeting Influenza, Hepatitis A & B, HPV, and others.[56,57] More interestingly, these have been implicated

as suitable targets for therapeutic vaccinations, which create antigens for nonviral diseases, such as Alzheimer's disease and cancer. Capsid protein displaying a 9-amino acid sequence from amyloid-β proteins induced an immune response in rats, and prevented some degree of amyloid aggregation.[58] Similar therapeutic gains have been seen with arthritis and nicotine addiction with vaccines targeting cytokine receptors and nicotine receptors, respectively.[59,60] Cancer-targeting vaccines have shown promise as well, the main challenge here is to correctly target a tumor-specific antigen, as these may be similar to endogenous proteins. While initial results have been very limited in scope, capsid-based vaccines have shown therapeutic effects in the treatment of both prostate cancer and melanoma.[61,62]

1.3 - Design of *de novo* Protein Cages

Assembling protein cages *de novo* from constituent building blocks of choice provides an attractive alternative approach to re-purposing existing protein cages. The primary advantage of this approach is customizability: one can choose attachment sites that are sensitive to environmental conditions and cofactors, such that one can control cage formation and optimize it for functional purposes, and we can also control the pore and cavity size of designed protein cages. The basic requirements for making a protein cage are deceptively simple: one must have two symmetric protein domains connected at the proper dihedral angle (Fig. 1.2a). To form one of the three geometries of Euclidean solids, tetrahedral (P_{332}), octahedral (P_{432}), or icosahedral (P_{532}), any one of six symmetry pairs can be used (Fig. 1.2b), depending on the symmetry group of that Euclidean solid. Each Euclidean solid can be created by combining any two of its three symmetry operators, so an icosahedron can be formed with symmetry pairs C_3+C_2, C_5+C_2, and C_5+C_3, provided the two symmetry axes are oriented at the proper dihedral angle. In addition to

these six, any symmetry pair with a C_2 symmetry element has the potential to form prismatic geometry. This can be seen with GroEL, a 14-subunit protein cage formed from a C_7+C_2 symmetry pair.

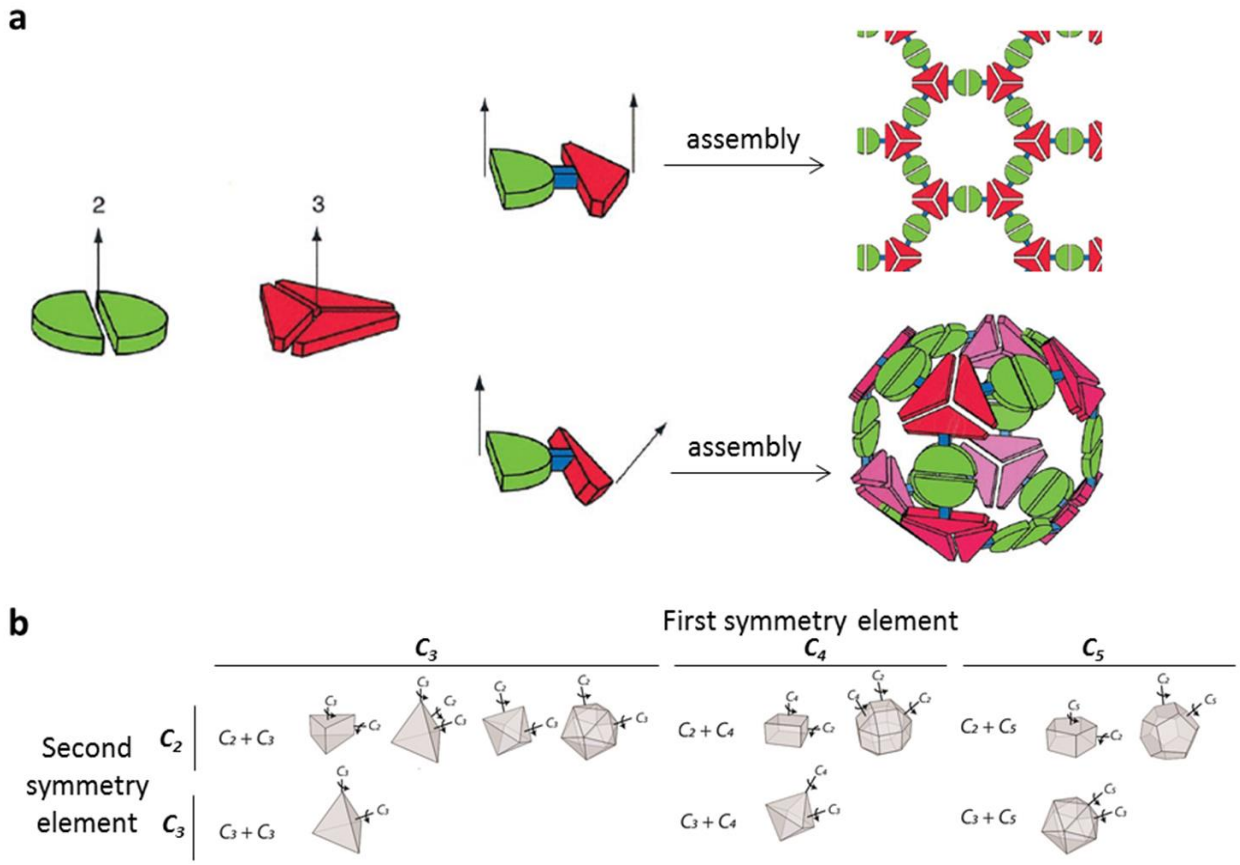


Figure 1.2. Assembly of designed fusion protein assemblies. a) Rigidly linking a homodimeric protein (green) with a homotrimeric protein (red) will result in different complexes depending on the dihedral angle imparted by the rigid linker (blue) (Image adapted from Ref 24). b) Six different symmetry pairs of fusion proteins connected at a proper dihedral angle will result in the formation of closed Euclidean solids. Additionally, any symmetry pair that includes a C_2 symmetry element can assemble into a prismatic complex (Image credit: Dr. Ben Buer).

It is important to note that creating a fusion protein with a particular symmetry pair doesn't necessarily encode exclusively its respective Euclidean solid(s), particularly if flexibility is introduced to the system. The Euclidean solids, and prismatic geometries, are simply the only geometries with regular dihedral angles. If these dihedral angles are malleable, then many

more complexes with irregular geometries can be formed (Fig. 1.3). Larger species are possible with all sets of symmetry pairs, the only requirement for a stable, closed system is that there are no unpaired symmetry elements. Thus, a C_3+C_2 symmetry pair may only form multiples of two trimers, for example, adding one trimer, with each subunit of the trimer being attached to a dimerizing unit, to a system of four trimers would yield 15 dimerizing units, making seven dimer pairs but leaving one dimerizing unit free to associate with an unpaired dimer on another complex. A C_3+C_3 symmetry pair, which is predicted to only form a tetrahedron if the two symmetry elements are rigidly connected at the correct dihedral angle, can associate into complexes consisting of any number of trimers if the two symmetry elements are flexibly attached, because each additional trimer adds three monomers of a trimerizing unit. If these two oligomerization sites are not so flexibly attached that multiple secondary oligomerization sites on a single trimer are unable to associate with each other (figure 1.4a), it is possible to envision geometrically plausible complexes with at least 4, 6, 7, or 8 trimers (a 5 trimer C_3+C_3 complex involves significant subunit torsion (figure 1.4b)). The C_3+C_4 symmetry pair can assemble into any complex with multiples of 4 trimers, shown in figure 1.3 are complexes with 8, 12, and 16 trimers that can be formed with a flexible fusion protein with a C_3+C_4 symmetry pair. The specific species formed for each symmetry pair is dependent on the range of allowed dihedral angles. For example, the hexamer of trimers of a C_3+C_3 system can form a ring with the same geometric arrangement as an octahedron but missing two opposing faces, and the 16-mer of trimers of a C_3+C_4 system is an icosahedron *sans* any four trimers that do not touch each other. If the dihedral angle of a C_3+C_3 or a C_3+C_4 system was that of an octahedron or an icosahedron, respectively, the resulting complex should specifically form the hexamer or 16-

mer of trimers, respectively. On the other hand, forming the heptamer of trimers with a C_3+C_3 system involves considerable flexibility in the range of dihedral angles.

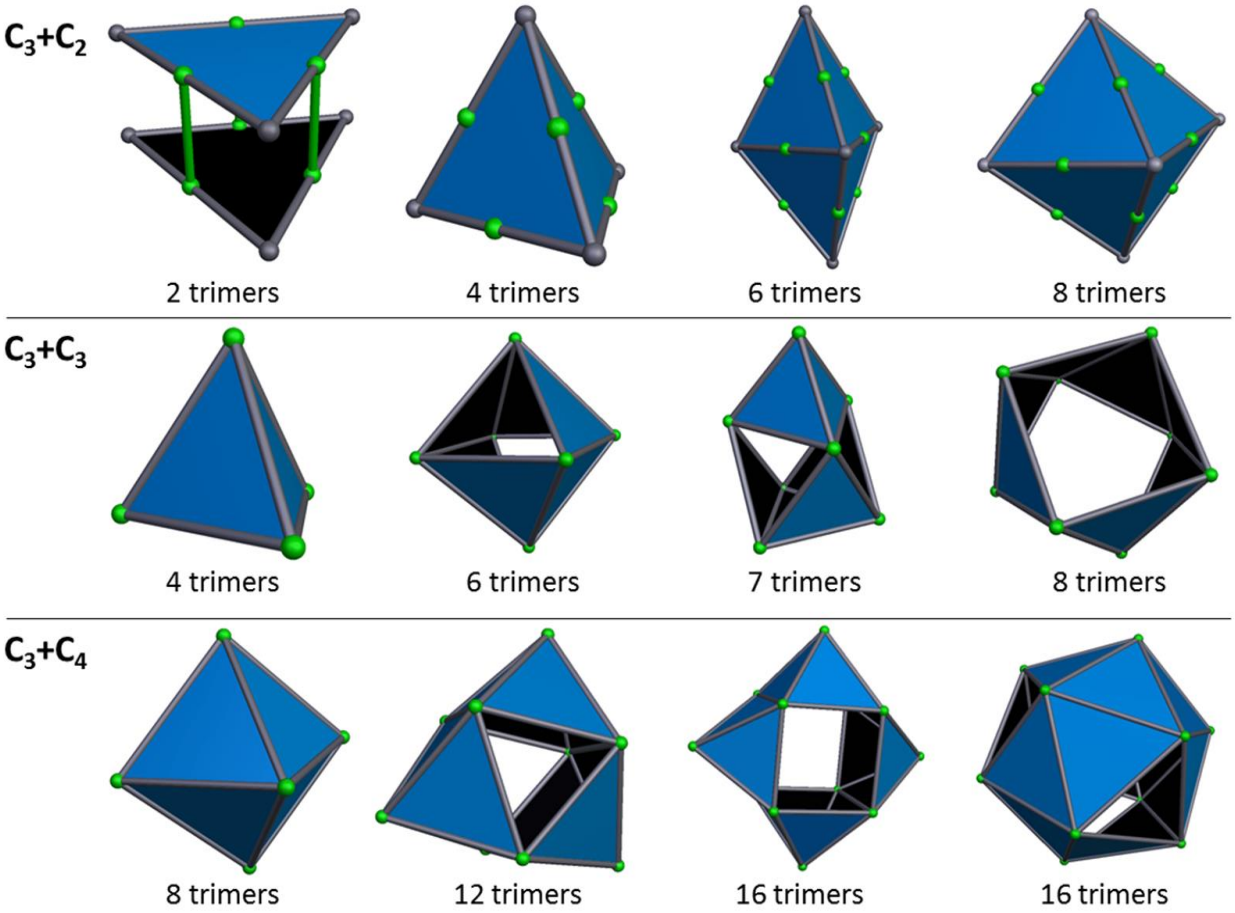


Figure 1.3. Examples of possible oligomers that can be formed with fusion proteins utilizing various symmetry pairs. Blue triangles represent a trimeric building block protein, fused with a second symmetric protein, represented as green dots at either the edges (C_2) or vertices (C_3 and C_4) of the blue triangles. These complexes may be porous and/or require a variable dihedral angle for formation, but have the proper oligomerization state at every point of attachment.

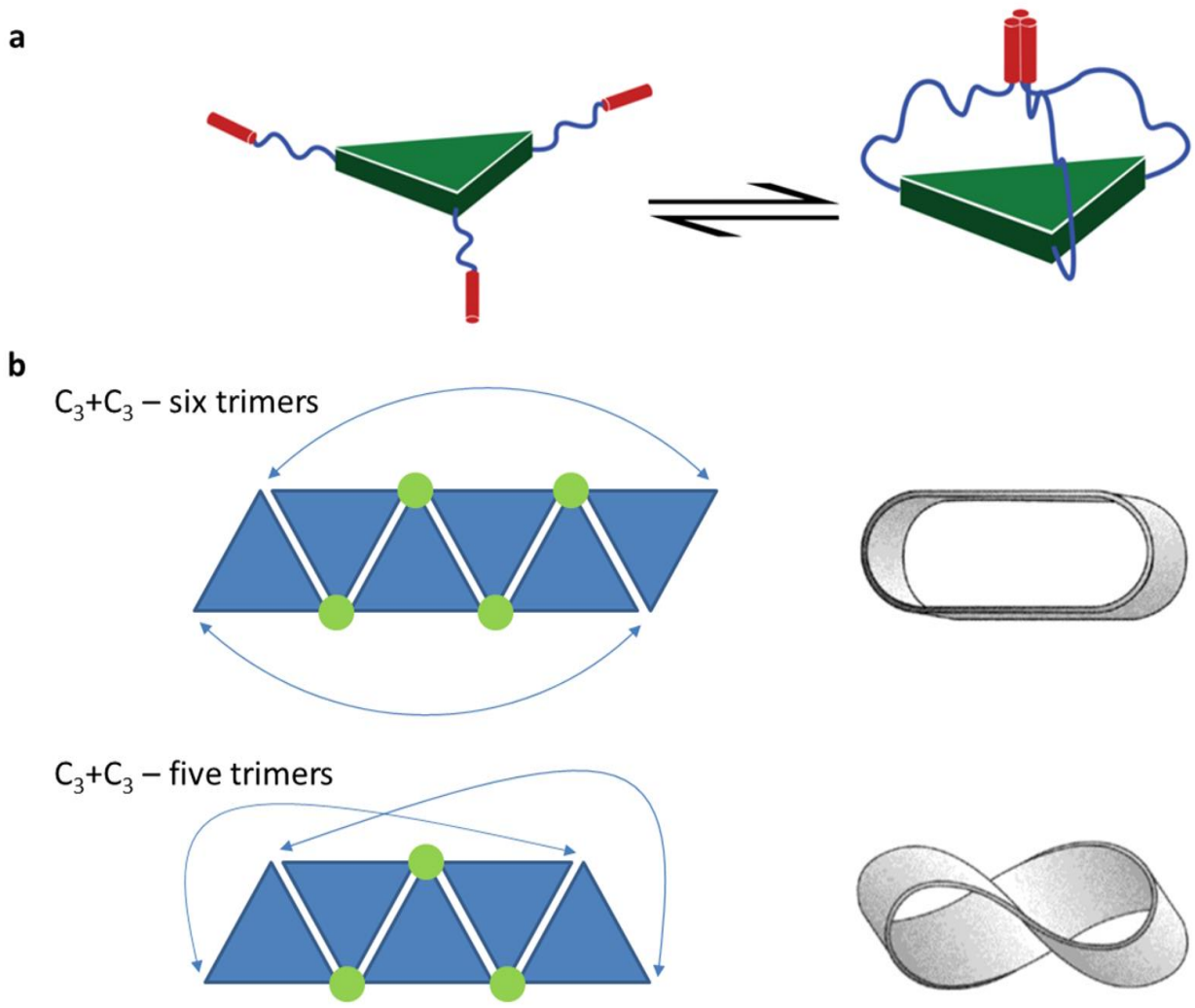


Figure 1.4. Special cases of symmetry pairs. a) A symmetry pair connected by a long enough flexible linker may self-associate at multiple oligomerization sites on the same subunit, dramatically reducing the minimum number of subunits necessary to create a closed complex. In the most drastic case shown here, a flexibly-linked C_3+C_3 symmetry pair is stable as a single trimer. b) A closed five trimer complex is possible from a C_3+C_3 symmetry pair, but requires significant subunit flexibility. Whereas a six trimer complex can be closed like a ring, a five trimer complex must be twisted like a Mobius strip to connect unpaired symmetry elements, represented here as blue arrows.

1.4 - Previously Designed *de novo* Protein Cages

The field of *de novo* designed protein cages is still in its infancy, with 2001 marking the first year that a research group described a symmetric fusion protein that oligomerized into a homogeneous, closed, supersymmetric assembly. By rigidly combining a dimeric protein

domain with a trimeric protein domain at the proper dihedral angle, Padilla and coworkers showed that this designed fusion protein self-assembled into a rigid tetrahedron that could later be crystallized (Fig. 1.6a).[63,64] Critical to this design was that both the C-terminus of the trimeric domain and the N-terminus of the dimeric coil were composed of alpha helices, so the connector of these two symmetric subunits could be a rigid alpha helix. The geometry of the species could therefore be specified by adding residues to this alpha helix connector, twisting the dihedral angle by 100° with each additional residue until it reached the necessary dihedral angle for oligomerization. When this trimeric subunit was substituted for a dimeric subunit (generating a C_2+C_2 symmetry pair), and the dihedral angle was altered to 180°, the fusion protein formed a filament. Multiple extended filaments utilizing a similar C_2+C_2 geometry have since been reported.[65,66] Interestingly, adding flexibility to one of these dimer-dimer systems with a glycine-rich linker reduces the size of the oligomers formed (Fig 1.5). While the designed fusion protein with an additional 2 or 4 flexible residues in the intersubunit linker retained a filamentous structure, the addition of 6 residues yielded a tetramer of dimers, while 8 residues yielded a trimer of dimers, and 10 residues led to a mixture of trimers of dimers and dimers of dimers.[67]

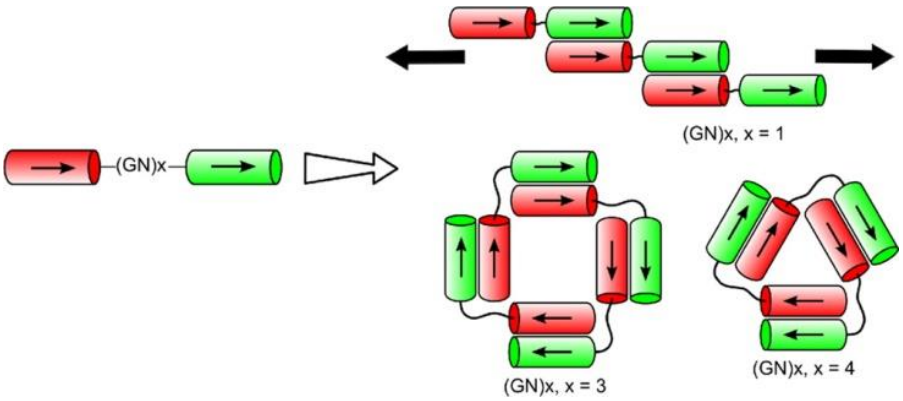


Figure 1.5. A designed C_2+C_2 symmetry paired protein assembles into filaments with a short flexible linker (top) but assembles into smaller oligomers inversely proportional to the length of the flexible linker (bottom). Image taken from Ref 27.

It is also possible to design fusion proteins that associate into lattices, but due to the extra spatial dimension that must be aligned, these are trickier to design and characterize. A fusion protein was designed with a combined three-fold and two-fold axis oriented at approximately 60° to each other, such that the resulting lattice would form a two dimensional honeycomb pattern *a la* Figure 1.2a. Instead, due to the inherent flexibility of proteins, these assemblies formed large spherical species approximately 100 nm in diameter, similar in size and shape to bacterial microcompartments.[68] Recently, due to the availability of robust computational modeling software to efficiently design orientations, a 2-dimensional square lattice was formed by combining a D_4-symmetric aldolase protein with a D_2-symmetric streptavidin protein at a 90° angle (Fig. 1.6d).[69,70] Similarly, a monomeric protein with multiple designed dimeric Zn-binding sites at axes in all three dimensions was shown to oligomerize into 2- and 3-dimensional ordered structures upon addition of zinc, depending on the conditions and ratio of zinc/protein used.[71] Finally, a crystallography-grade lattice was generated by the computational design of exterior sidechains in a homotrimeric coiled-coil, yielding a crystal with the infrequently-seen *P6* space group.[72]

The design of closed protein nanostructures, on the other hand, has been met with significantly more success. For the decade after Padilla and coworkers published their seminal 2001 paper that jump-started research into the rational design of *de novo* protein cages, there were only two notable systems of closed, spherical protein assemblies. The first was a design by the Burkhard group, which combined a pentameric coiled-coil with a trimeric coiled-coil. These elements were connected by a flexible linker sequence with a cysteine residue at each end, such that oxidative formation of the disulfide bond would lock the dihedral angle between the two coils at approximately 37°, the angle required for the formation of an icosahedron (Fig. 1.6b). At a sufficiently high concentration in oxidative conditions, this protein formed spherical complexes with the majority of species being approximately the right size for an icosahedron, though this complex was not homogeneous and was not characterized in detail.[73] Notably, however, this pseudo-icosahedral protein complex showed immunogenic potential: malaria coat proteins were added to the ends of these complexes, and when injected into rats induced an immune response that lasted for 6 months.[74]

The second design was from this laboratory, where trimeric aldolase proteins were attached through a flexible linker to one of two peptides designed to form heterodimeric antiparallel coiled-coils through complementary electrostatic interactions, one with a strong positive charge and the other with a strong negative charge (Fig. 1.6c). Solutions consisting of solely positively- or negatively-charged trimeric proteins did not oligomerize further, but when mixed together, formed a mixture of complexes with molecular weights expected from a C_3+C_2 system – the 6-subunit trigonal prism, 12-subunit tetrahedron, 18-subunit trigonal bipyramid, and 24-subunit octahedron, though these geometries could not be specifically identified.[75]

These complexes were catalytically active, indicating that protein cage formation didn't interfere with the subunit's tertiary structure. A recent study replicated this result with a flexibly-linked C_3+C_2 symmetry pair, substituting the heterodimeric coiled-coil with a homodimer and assembling these complexes *in vivo*.[76]

In 2012, the major breakthrough for designing these *de novo* protein cages came from advances in computational methods. The program Rosetta can rapidly and robustly model different docking conformations of protein-protein interfaces, using a built-in scoring function that assesses the energetic stability gained from burying hydrophobic surface residues and creating hydrogen bonds as well as the destabilizing effects of steric clashes and unfavorable Coulombic interactions. Analysis of docked protein interfaces with Rosetta has been used to predict the oligomerization state and binding surface of a self-associating protein from its crystal structure.[77,78] Additionally, new protein-protein interactions can be designed by remodeling a docked structure to add in inter-protein hydrogen bonds and hydrophobic interfaces.[79-81] King and coworkers exploited this technology to design protein cages by replicating each of the 271 proteins in the PDB that have C_3 symmetry into both tetrahedral and octahedral space, and analyzing these assemblies for steric clashes and close contacts at the inter-trimer surface (Fig. 1.6e).[82] Each trimeric protein was rotated 0.5° 240 times to sample the entire set of rotational conformations at a specific radial position that could lead to a designable trimer-trimer interface, after which the protein was translated 1 Å radially from the center of symmetry and rotational sampling continued, until the trimeric protein could be rotated 120° without touching a neighboring symmetry-generated trimer. The 20 trimers that were symmetrically docked into conformations with the largest number of surface interactions

without any steric clashes were selected for interface redesign, and favorable interactions were designed at the trimer-trimer interface. 35 potential mutants were designed from these 20 proteins with an average of 9 mutations per design, of which 24 expressed as soluble proteins and 3 oligomerized into symmetrical assemblies – one into an octahedron and two into tetrahedrons. Crystal structures of these assemblies could be determined, and were in close agreement with computational models. This approach was further extended to design constructs in which a trimeric protein was docked at the faces of a tetrahedron and either a dimeric or a trimeric protein was docked at the edges or vertices respectively. The rotational and translational space of both of these proteins was sampled, and a heteroprotein interface was designed and optimized. This led to the design and crystallographic characterization of six tetrahedral complexes with two different geometries.[83] The Rosetta framework was further exploited (Fig. 1.6f) to design a rigid linker between a trimeric and a dimeric protein in the same manner as described above by Padilla *et al*, but this design, even after the dihedral angle was optimized to form an octahedron, instead formed a mix of tetrahedrons, trigonal bipyramids, and octahedrons.[84,85]

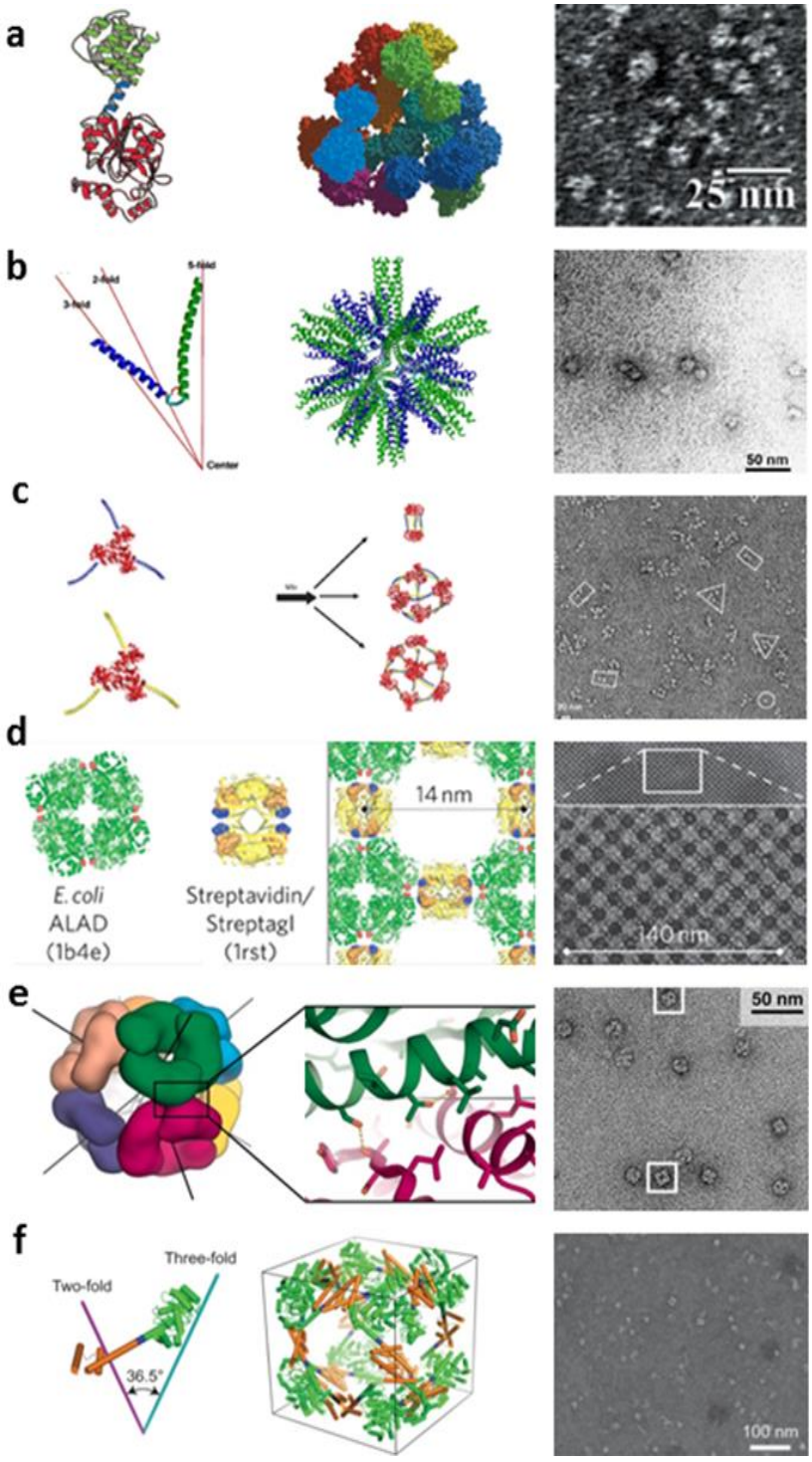


Figure 1.6. Designed protein cages (left) with their respective TEM characterizations (right). a) A rigidly assembled C_3+C_2 symmetry pair as designed by Padilla et al. assembles into complexes with tetrahedral symmetry. Images taken from reference 24. b) A trimeric coiled-coil and a pentameric coiled-coil as designed by Raman et al. is connected with a disulfide linker to yield a proper dihedral angle and oligomerize into globular complexes with the approximate weight of

an icosahedron. Images taken from reference 31. c) A C_3+C_2 heteroprotein system as designed by Patterson et al. consisting of trimers flexibly attached to either positively- or negatively-charged coiled-coils forms a variety of distinct symmetric structures when mixed together. Images taken from reference 33. d) A C_4+C_2 heteroprotein system as designed by Sinclair et al. with a rigid 180° dihedral angle assembles into a regular square lattice. Images taken from reference 29. e) A C_3 protein with a computationally designed dimeric interface by King et al. was assembled into either tetrahedra and octahedra (shown) depending on the angle of interface, with well-defined symmetry axes. Images taken from reference 35. f) A rigidly-connected C_3+C_2 symmetry pair designed by Lai et al. to connect with the dihedral angle of an octahedron instead assembles into both tetrahedra and octahedra. Images taken from reference 38.

1.5 - Project Goals

There is a sizeable and rapidly growing body of work in this nascent field of research, but there are still many symmetry pairs that have yet to be explored for their potential to form new protein cages. Of the six possible symmetry pairs that can generate Euclidean solids, the only two that have yielded robust complexes have been C_3+C_3 and C_3+C_2, and these have only assembled tetrahedral and octahedral protein complexes. This poses the question: why is it that no higher order symmetry pairs have been successfully used to assembly protein cages? Is it impossible to form these higher order structures without forming an array of misfolded complexes, larger or smaller? Most of the previously characterized cages have focused on designing rigid structures that precisely orient symmetry elements to achieve the desired geometry, but as the work by Lai et al shows, it may be difficult to separate different oligomers should the structure be insufficiently rigid. A more general solution to this problem, one that doesn't involve screening dozens of designed mutants, is desirable. Since rigid cages can only be designed with a high degree of computational work, my dissertation was focused on producing and characterizing flexibly-attached symmetry elements to determine the oligomeric

species that these form. Specifically, my project aimed to produce a designed fusion protein that has the following properties:

1) Contains two crystallographically-verified symmetry sites connected genetically by a flexible linker region.
2) Can be expressed and purified by standard biochemical techniques.
3) Assembles *in vivo* into soluble complexes, which can be further purified to yield a single, regular, symmetric species as predicted by the symmetries of the two subunits.
4) Is stable, soluble, and enzymatically active after assembly.

For the first component, we selected a trimeric esterase protein isolated from *Pseudomonis putida* (PDB ID 1ZOI), as it had several important characteristics. First, this is a relatively large protein – the side length of the trimer in the crystal structure is 7 nm, so the protein could be readily imaged by TEM. Second, this protein has strategically placed termini. The N-terminus is located towards the middle of the 'face' of the triangular protein – near the subunit inferface. This is well positioned for introducing a 6xHis tag to facilitate purification. The N-terminus is as far removed from the secondary oligomerization site as possible, minimizing the risk that the His-tag could interfere with protein cage assembly. The C-terminus of this esterase is located perpendicular to the C_3 axis, near the 'vertex' of the triangle formed by the three subunits. This is where we will attach a flexible glycine linker, followed by a designed coiled-coil as the secondary oligomerization site. These coiled-coils are going to bind together multiple bulky trimers, so it's a sensible idea to put this secondary oligomerization site in as sterically unhindered a location as possible. The coiled-coil is chosen as the secondary oligomerization site because it is small and unobtrusive, and because it is a well-characterized

system with a variety of designed homooligomers ranging from homodimers up to complexes of seven alpha helices.[86,87] These are an excellent choice to add onto a trimeric building block, as we can create 4 of the 6 symmetry pairs discussed earlier (C_3+C_2, C_3+C_3, C_3+C_4, and C_3+C_5) by choosing an appropriate coiled-coil to affix to the trimer. Since the C_3+C_2 and C_3+C_3 symmetry pairs have already been explored by other groups and are predicted to form multiple complexes with similar sizes, we selected the C_3+C_4 symmetry pair as a potential symmetry pair for the goal of forming an octahedron.

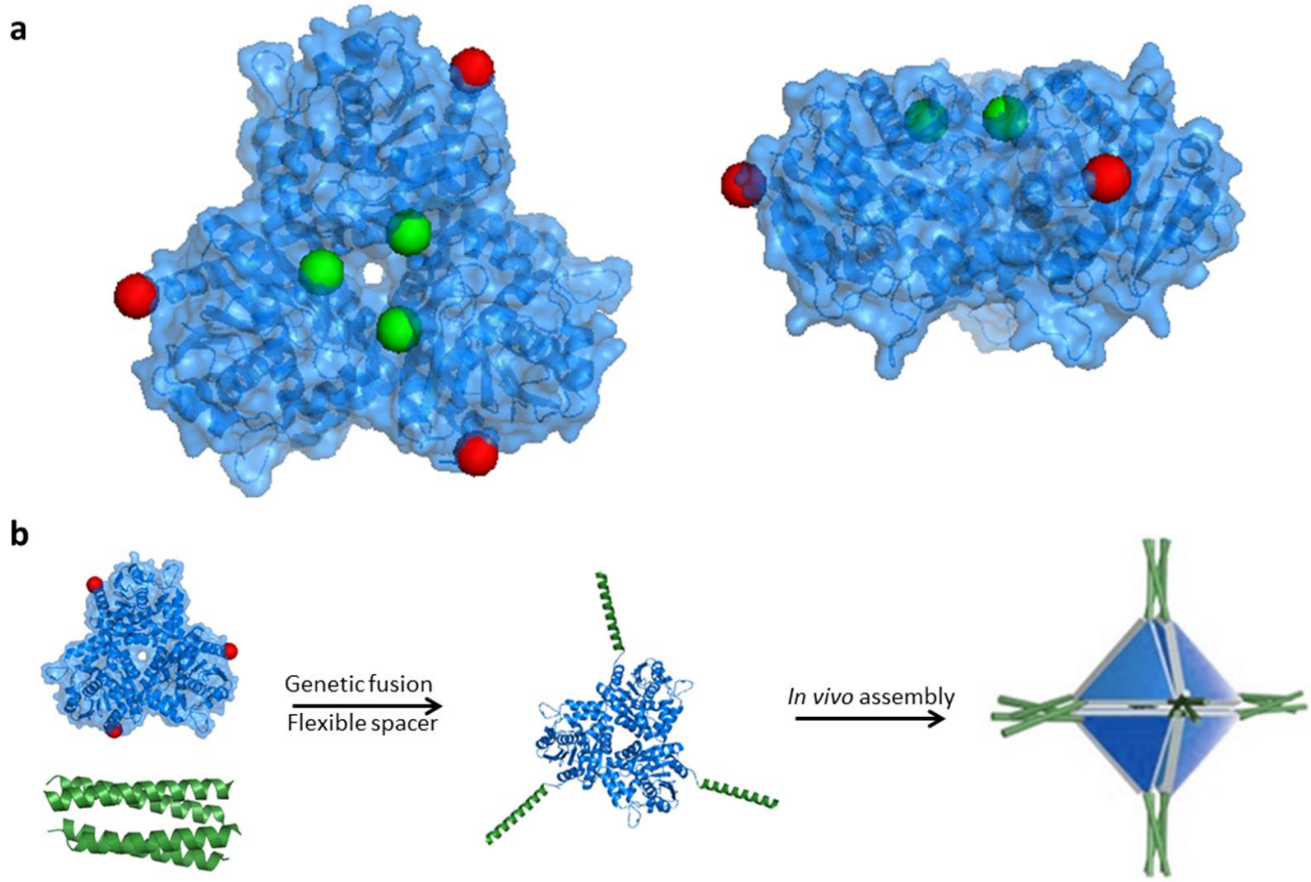


Figure 1.7. Assembly strategy for designed protein cages. a) The face (left) and edge (right) of the trimeric esterase building block we will be fusing secondary oligomerization sites to. Green spheres are the N-termini, where we will fuse a His-tag, and red spheres are the C-termini, where we will fuse a flexibly-linked coiled-coil. b) Cartoon of the assembly scheme for the octahedral protein cage, with coiled-coil linkers connecting at the C-terminus 'vertices' of trimeric proteins (Image credit: Ben Buer and Cullen Whitmore).

1.6 - References

1. Theil, E.C. Ferritin: Structure, Gene Regulation, and Cellular Function in Animals, Plants, and Microorganisms. *Annu. Rev. Biochem* **56**, 289-315 (1987).
2. Ross, P.D. et al. A Free Energy Cascade with Locks Drives Assembly and Maturation of Bacteriophage HK97 Capsid. *J. Mol. Biol.* **364**, 512-525 (2006).
3. Ceres, P., Stray, S.J. & Zlotnick, A. Hepatitis B Virus Capsid Assembly Is Enhanced by Naturally Occurring Mutation F97L. *J. Virol.* **78**, 9538-9543 (2004).
4. Landry, S.J. & Gierasch, L.M. The chaperonin GroEL binds a polypeptide in an alpha-helical conformation. *Biochemistry* **30**, 7359-62 (1991).
5. Fayet, O., Ziegelhoffer, T. & Georgopoulos, C. The groES and groEL heat shock gene products of Escherichia coli are essential for bacterial growth at all temperatures. *J. Bacteriol.* **171**, 1379-1385 (1989).
6. Perham, R.N., Duckworth, H.W. & Roberts, G.C.K. Mobility of polypeptide chain in the pyruvate dehydrogenase complex revealed by proton NMR. *Nature* **292**, 474-477 (1981).
7. Yeates, T.O., Kerfeld, C.A., Heinhorst, S., Cannon, G.C. & Shively, J.M. Protein-based organelles in bacteria: carboxysomes and related microcompartments. *Nat Rev Micro* **6**, 681-691 (2008).
8. Price, G.D. & Badger, M.R. Expression of Human Carbonic Anhydrase in the Cyanobacterium Synechococcus PCC7942 Creates a High CO(2)-Requiring Phenotype : Evidence for a Central Role for Carboxysomes in the CO(2) Concentrating Mechanism. *Plant Physiol.* **91**, 505-513 (1989).
9. Reed, L.J. A Trail of Research from Lipoic Acid to α-Keto Acid Dehydrogenase Complexes. *J. Biol. Chem.* **276**, 38329-38336 (2001).
10. Flenniken, M.L. et al. A library of protein cage architectures as nanomaterials. *Current topics in microbiology and immunology* **327**, 71-93 (2009).
11. McMillan, R.A. et al. A Self-Assembling Protein Template for Constrained Synthesis and Patterning of Nanoparticle Arrays. *J. Am. Chem. Soc.* **127**, 2800-2801 (2005).
12. Klem, M.T. et al. Bio-inspired Synthesis of Protein-Encapsulated CoPt Nanoparticles. *Adv. Funct. Mater.* **15**, 1489-1494 (2005).
13. Ishii, D. et al. Chaperonin-mediated stabilization and ATP-triggered release of semiconductor nanoparticles. *Nature* **423**, 628-632 (2003).
14. Brumfield, S. et al. Heterologous expression of the modified coat protein of Cowpea chlorotic mottle bromovirus results in the assembly of protein cages with altered architectures and function. *J. Gen. Virol.* **85**, 1049-53 (2004).
15. Klem, M.T., Young, M. & Douglas, T. Biomimetic magnetic nanoparticles. *Mater. Today* **8**, 28-37 (2005).
16. Kramer, R.M., Sowards, L.A., Pender, M.J., Stone, M.O. & Naik, R.R. Constrained iron catalysts for single-walled carbon nanotube growth. *Langmuir* **21**, 8466-70 (2005).
17. Jutz, G., van Rijn, P., Miranda, B.S. & Boeker, A. Ferritin: A versatile building block for bionanotechnology. *Chem. Rev.* **115**, 1653-1701 (2015).
18. Atsushi, M. et al. Floating Nanodot Gate Memory Devices Based on Biomineralized Inorganic Nanodot Array as a Storage Node. *Japanese Journal of Applied Physics* **45**, L1 (2006).

19. Takuro, M. et al. Direct Production of a Two-Dimensional Ordered Array of Ferritin-Nanoparticles on a Silicon Substrate. *Japanese Journal of Applied Physics* **46**, L713 (2007).
20. Kazunori, I. et al. Low-temperature Polycrystalline Silicon Thin Film Transistor Flash Memory with Ferritin. *Japanese Journal of Applied Physics* **46**, L804 (2007).
21. Kiyohito, Y. et al. Effects of Dot Density and Dot Size on Charge Injection Characteristics in Nanodot Array Produced by Protein Supramolecules. *Japanese Journal of Applied Physics* **46**, 7549 (2007).
22. Bode, S.A., Minten, I.J., Nolte, R.J.M. & Cornelissen, J.J.L.M. Reactions inside nanoscale protein cages. *Nanoscale* **3**, 2376-2389 (2011).
23. Ueno, T. et al. Size-Selective Olefin Hydrogenation by a Pd Nanocluster Provided in an Apo-Ferritin Cage. *Angew. Chem. Int. Ed.* **43**, 2527-2530 (2004).
24. Shin, Y., Dohnalkova, A. & Lin, Y. Preparation of Homogeneous Gold−Silver Alloy Nanoparticles Using the Apoferritin Cavity As a Nanoreactor. *The Journal of Physical Chemistry C* **114**, 5985-5989 (2010).
25. Varpness, Z., Peters, J.W., Young, M. & Douglas, T. Biomimetic Synthesis of a H2 Catalyst Using a Protein Cage Architecture. *Nano Lett.* **5**, 2306-2309 (2005).
26. Comellas-Aragones, M. et al. A virus-based single-enzyme nanoreactor. *Nat Nano* **2**, 635-639 (2007).
27. Minten, I.J. et al. Catalytic capsids: the art of confinement. *Chemical Science* **2**, 358-362 (2011).
28. Patterson, D.P., Schwarz, B., Waters, R.S., Gedeon, T. & Douglas, T. Encapsulation of an Enzyme Cascade within the Bacteriophage P22 Virus-Like Particle. *ACS Chem. Biol.* **9**, 359-365 (2014).
29. Abe, S. et al. Polymerization of Phenylacetylene by Rhodium Complexes within a Discrete Space of apo-Ferritin. *J. Am. Chem. Soc.* **131**, 6958-6960 (2009).
30. Štokrová, J. et al. Interactions of heterologous DNA with polyomavirus major structural protein, VP1. *FEBS Lett.* **445**, 119-125 (1999).
31. Kimchi-Sarfaty, C., Arora, M., Sandalon, Z., Oppenheim, A. & Gottesman, M.M. High Cloning Capacity of In Vitro Packaged SV40 Vectors with No SV40 Virus Sequences. *Hum. Gene Ther.* **14**, 167-177 (2003).
32. Kimchi-Sarfaty, C. et al. SV40 Pseudovirion gene delivery of a toxin to treat human adenocarcinomas in mice. *Cancer Gene Ther.* **13**, 648-657 (2006).
33. Lilavivat, S., Sardar, D., Jana, S., Thomas, G.C. & Woycechowsky, K.J. In Vivo Encapsulation of Nucleic Acids Using an Engineered Nonviral Protein Capsid. *J. Am. Chem. Soc.* **134**, 13152-13155 (2012).
34. Goldmann, C. et al. Packaging of small molecules into VP1-virus-like particles of the human polyomavirus JC virus. *J. Virol. Methods* **90**, 85-90 (2000).
35. Abbing, A. et al. Efficient Intracellular Delivery of a Protein and a Low Molecular Weight Substance via Recombinant Polyomavirus-like Particles. *J. Biol. Chem.* **279**, 27410-27421 (2004).
36. Flenniken, M.L. et al. Selective attachment and release of a chemotherapeutic agent from the interior of a protein cage architecture. *Chemical communications (Cambridge, England)*, 447-9 (2005).

37. Aljabali, A.A.A., Shukla, S., Lomonossoff, G.P., Steinmetz, N.F. & Evans, D.J. CPMV-DOX Delivers. *Mol. Pharm.* **10**, 3-10 (2013).
38. Zhao, Q. et al. Self-assembled virus-like particles from rotavirus structural protein VP6 for targeted drug delivery. *Bioconjug Chem* **22**, 346-52 (2011).
39. Banerjee, P.S., Ostapchuk, P., Hearing, P. & Carrico, I. Chemoselective attachment of small molecule effector functionality to human adenoviruses facilitates gene delivery to cancer cells. *J. Am. Chem. Soc.* **132**, 13615-7 (2010).
40. Huang, R.K., Steinmetz, N.F., Fu, C.-Y., Manchester, M. & Johnson, J.E. Transferrin-mediated targeting of bacteriophage HK97 nanoparticles into tumor cells. *Nanomedicine (London, England)* **6**, 55-68 (2011).
41. Lockney, D.M. et al. The Red clover necrotic mosaic virus capsid as a multifunctional cell targeting plant viral nanoparticle. *Bioconjug Chem* **22**, 67-73 (2011).
42. Stephanopoulos, N., Tong, G.J., Hsiao, S.C. & Francis, M.B. Dual-surface modified virus capsids for targeted delivery of photodynamic agents to cancer cells. *ACS Nano* **4**, 6014-20 (2010).
43. Ashley, C.E. et al. Cell-Specific Delivery of Diverse Cargos by Bacteriophage MS2 Virus-like Particles. *ACS Nano* **5**, 5729-5745 (2011).
44. Joo, K.I. et al. Site-specific labeling of enveloped viruses with quantum dots for single virus tracking. *ACS Nano* **2**, 1553-62 (2008).
45. Lewis, J.D. et al. Viral nanoparticles as tools for intravital vascular imaging. *Nat Med* **12**, 354-360 (2006).
46. Cormode, D.P., Jarzyna, P.A., Mulder, W.J.M. & Fayad, Z.A. Modified natural nanoparticles as contrast agents for medical imaging. *Adv. Drug Delivery Rev.* **62**, 329-338 (2010).
47. Hooker, J.M., O'Neil, J.P., Romanini, D.W., Taylor, S.E. & Francis, M.B. Genome-free viral capsids as carriers for positron emission tomography radiolabels. *Molecular imaging and biology : MIB : the official publication of the Academy of Molecular Imaging* **10**, 182-91 (2008).
48. Flexman, J.A. et al. Magnetically targeted viral envelopes: a PET investigation of initial biodistribution. *IEEE transactions on nanobioscience* **7**, 223-32 (2008).
49. Ghosh, D. et al. M13-templated magnetic nanoparticles for targeted in vivo imaging of prostate cancer. *Nat. Nanotechnol.* **7**, 677-82 (2012).
50. Allen, M. et al. Paramagnetic viral nanoparticles as potential high-relaxivity magnetic resonance contrast agents. *Magnetic Resonance in Medicine* **54**, 807-812 (2005).
51. Uchida, M. et al. A human ferritin iron oxide nano-composite magnetic resonance contrast agent. *Magnetic Resonance in Medicine* **60**, 1073-1081 (2008).
52. Kaiser, C.R. et al. Biodistribution studies of protein cage nanoparticles demonstrate broad tissue distribution and rapid clearance in vivo. *International Journal of Nanomedicine* **2**, 715-733 (2007).
53. Kawano, M., Matsui, M. & Handa, H. SV40 virus-like particles as an effective delivery system and its application to a vaccine carrier. *Expert Review of Vaccines* **12**, 199-210 (2013).

54. Rebeaud, F. & Bachmann, M. Virus-Like Particles as Efficient Delivery Platform to Induce a Potent Immune Response. in *Innovation in Vaccinology* (ed. Baschieri, S.) 87-122 (Springer Netherlands, 2012).
55. Jennings, G.T. & Bachmann, M.F. Immunodrugs: Therapeutic VLP-Based Vaccines for Chronic Diseases. *Annu. Rev. Pharmacool. Toxicol.* **49**, 303-326 (2009).
56. Teunissen, E.A., de Raad, M. & Mastrobattista, E. Production and biomedical applications of virus-like particles derived from polyomaviruses. *J. Controlled Release* **172**, 305-321 (2013).
57. Donaldson, B. et al. Virus-Like Particles, a Versatile Subunit Vaccine Platform. in *Subunit Vaccine Delivery* (eds. Foged, C., Rades, T., Perrie, Y. & Hook, S.) 159-180 (Springer New York, 2015).
58. Wiessner, C. et al. The second-generation active Abeta immunotherapy CAD106 reduces amyloid accumulation in APP transgenic mice while minimizing potential side effects. *J. Neurosci.* **31**, 9323-31 (2011).
59. Maurer, P. et al. A therapeutic vaccine for nicotine dependence: preclinical efficacy, and Phase I safety and immunogenicity. *Eur. J. Immunol.* **35**, 2031-40 (2005).
60. Spohn, G. et al. Active immunization with IL-1 displayed on virus-like particles protects from autoimmune arthritis. *Eur. J. Immunol.* **38**, 877-87 (2008).
61. Garcia, J.A. Sipuleucel-T in patients with metastatic castration-resistant prostate cancer: an insight for oncologists. *Therapeutic Advances in Medical Oncology* **3**, 101-108 (2011).
62. Speiser, D.E. et al. Memory and effector CD8 T-cell responses after nanoparticle vaccination of melanoma patients. *Journal of immunotherapy (Hagerstown, Md. : 1997)* **33**, 848-58 (2010).
63. Lai, Y.-T., Cascio, D. & Yeates, T.O. Structure of a 16-nm Cage Designed by Using Protein Oligomers. *Science* **336**, 1129-1129 (2012).
64. Padilla, J.E., Colovos, C. & Yeates, T.O. Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2217-2221 (2001).
65. Sharp, T.H. et al. Cryo-transmission electron microscopy structure of a gigadalton peptide fiber of de novo design. *Proceedings of the National Academy of Sciences* **109**, 13266-13271 (2012).
66. Papapostolou, D. et al. Engineering nanoscale order into a designed protein fiber. *Proceedings of the National Academy of Sciences* **104**, 10853-10858 (2007).
67. Boyle, A.L. et al. Squaring the Circle in Peptide Assembly: From Fibers to Discrete Nanostructures by de Novo Design. *J. Am. Chem. Soc.* **134**, 15457-15467 (2012).
68. Fletcher, J.M. et al. Self-Assembling Cages from Coiled-Coil Peptide Modules. *Science* **340**, 595-599 (2013).
69. Ringler, P. & Schulz, G.E. Self-assembly of proteins into designed networks. *Science* **302**, 106-109 (2003).
70. Sinclair, J.C., Davies, K.M., Venien-Bryan, C. & Noble, M.E.M. Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nat Nano* **6**, 558-562 (2011).
71. Brodin, J.D. et al. Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. *Nature Chem.* **4**, 375-382 (2012).

72. Lanci, C.J. et al. Computational design of a protein crystal. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7304-7309 (2012).
73. Raman, S., Machaidze, G., Lustig, A., Aebi, U. & Burkhard, P. Structure-based design of peptides that self-assemble into regular polyhedral nanoparticles. *Nanomedicine* **2**, 95-102 (2006).
74. Kaba, S.A. et al. A Nonadjuvanted Polypeptide Nanoparticle Vaccine Confers Long-Lasting Protection against Rodent Malaria. *The Journal of Immunology* **183**, 7268-7277 (2009).
75. Patterson, D.P. et al. Characterization of a highly flexible self-assembling protein system designed to form nanocages. *Protein Sci.* **23**, 190-199 (2014).
76. Kobayashi, N. et al. Self-Assembling Nano-Architectures Created from a Protein Nano-Building Block Using an Intermolecularly Folded Dimeric de Novo Protein. *J. Am. Chem. Soc.* **137**, 11285-11293 (2015).
77. André, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proceedings of the National Academy of Sciences* **104**, 17656-17661 (2007).
78. Das, R. et al. Simultaneous prediction of protein folding and docking at high resolution. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18978-18983 (2009).
79. Huang, P.-S., Love, J.J. & Mayo, S.L. A de novo designed protein–protein interface. *Protein Sci.* **16**, 2770-2774 (2007).
80. Jha, R.K. et al. Computational Design of a PAK1 Binding Protein. *J. Mol. Biol.* **400**, 257-270 (2010).
81. Stranges, P.B., Machius, M., Miley, M.J., Tripathy, A. & Kuhlman, B. Computational design of a symmetric homodimer using β-strand assembly. *Proceedings of the National Academy of Sciences* **108**, 20562-20567 (2011).
82. King, N. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **6085**, 1171-1174 (2012).
83. King, N.P. et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103-+ (2014).
84. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* **6**, 1065-1071 (2014).
85. Lai, Y.-T., Tsai, K.-L., Sawaya, M.R., Asturias, F.J. & Yeates, T.O. Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly. *J. Am. Chem. Soc.* **135**, 7738-7743 (2013).
86. Fletcher, J.M. et al. A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth. Biol.* **1**, 240-250 (2012).
87. Thomson, A.R. et al. Computational design of water-soluble alpha-helical barrels. *Science* **346**, 485-8 (2014).

Chapter 2

Purification and Characterization of Designed Protein Constructs

All designed fusion protein constructs were analyzed by the same set of techniques under similar conditions, so in this chapter I will briefly examine the analysis techniques used to determine the size and shape of the various complexes formed from each specific fusion protein. As an example for each of these techniques, I have expressed, purified, and analyzed the trimeric esterase building block that we will be attaching secondary oligomerization sites to. The analysis of this trimeric esterase provides an excellent negative control for assaying whether a designed fusion protein oligomerizes at all, and an excellent positive control for examining the behavior of a homogeneous oligomeric complex.

2.1 - Preparation of Construct DNA

The gene encoding Oct-1 in pet28b vector was purchased (Genscript) with strategically placed restriction sites designed between the trimer and glycine linker (Kpn1), between the linker and the coil (Spe1) and after the stop codon (BamHI). All DNA modifying enzymes and reagents were purchased from New England Biolabs. A gene that encoded fragments to be double digested and subsequently ligated into double-digested Oct-1 to create the trimeric esterase and Oct-2 were purchased (IDT Technologies). All subsequent single-stranded and complementary DNA fragments with complementary overhangs to the restriction sites in Oct-1 encoding Oct-3-1, Oct-3-2, Oct-3-3, Oct-4-0, Oct-4-1, and Oct4-2 were purchased from IDT DNA

Technologies for annealing and insertion into Oct-1 DNA. Full DNA and protein sequences for all constructs discussed within this dissertation can be found in Appendix B.

The Oct-1 gene was transformed by electroporation into XL1-Blue electrocompetent *E.coli* cells and cells containing plasmids were selected by growing transformed cells on LB agar plates in the presence of kanamycin (50 mg/L) at 37 °C overnight. Single colonies were used to inoculate 5 mL of sterile LB media with kanamycin and grown at 37 °C overnight, after which they were pelleted by centrifugation at 14,000 rpm for 10 min. DNA was extracted from cell pellets using QIAprep Spin Miniprep kit and sequenced with the T7 promoter primer at the UM Sequencing Core to confirm presence of Oct-1 gene. Oct-1 DNA was double digested using SpeI and BamHI for preparation of Oct2 and double digested using KpnI and BamHI for preparation of all other constructs. Double digested Oct-1 DNA was gel purified in a 0.5% agarose gel and extracted using a QIAquick Gel Extraction Kit.

The gene encoding fragments for the trimeric esterase and for Oct-2 was transformed and purified as described above. This gene was double digested with KpnI and BamHI to create the DNA fragment that was ligated into double digested Oct-1 to create the trimeric esterase, and double digested with SpeI and BamHI to create the DNA fragment that was ligated into double digested Oct-1 to create Oct-2. Both fragments were purified on a 2% agarose gel and extracted using QIAquick Gel Extraction Kit. Reverse complementary sequences of single stranded DNA with complementary overhangs to Oct-1 that were contained coil encoding sequences for Oct-3-* and Oct-4-* were suspended in equimolar concentrations in 1x DNA ligase buffer at 95 °C, and annealed by slowly ramping down to room temperature over a period of 10 hours.

Double digested and gel purified fragments of vector containing Oct-1 (6 pmol) and DNA fragment insert (30 pmol) were mixed in the presence of T4 DNA ligase and supplied buffer system to a final volume of 20 µL. 1 µL of ligated DNA was transformed into XL1-Blue electrocompetent cells, grown up on kanamycin-containing LB agar plates overnight, and single colonies were grown up in 5 mL sterile LB media containing kanamycin overnight at 37 °C. DNA was extracted from culture pellets and sequenced as above.

50 ng of plasmid confirmed to contain genes encoding protein of interest was transformed into *E.coli* strain BL21-DE3, incubated on plates and then 5 ml overnight LB cultures as above. 700 µL of LB overnight cultures was added to 300 µL of sterile 50% glycerol solution, mixed by pipetting, flash frozen in liquid N_2, and stored at -80 °C.

2.2 - Expression and Purification of Fusion Protein Constructs

5 µL of flash frozen stock of BL21 (DE3) *E.coli* with plasmid encoding fusion protein of interest was added to 5 mL sterile LB media containing 50 mg/L kanamycin and grown up at 37 °C overnight while shaking at 220 rpm. Cultures were then added to 1 liter sterile 2xYT media with kanamycin and incubated at 37 °C while shaking at 200 rpm until the cultures reached an OD_{600} of 0.3, after which the incubation temperature was reduced to 18°C. When the 1 L culture reached an OD_{600} of 0.8-1.0, 10 µL was taken for a pre-induction sample for SDS-PAGE analysis and then induced with 10-20 mg IPTG, and protein expression proceeded overnight.

After 18 hours of protein expression, another 10 µL was taken for a post-induction sample for SDS-PAGE analysis, and the remainder was centrifuged at 4 °C and 4500 rpm for 20 min. 8 g of cell pellet was then resuspended on ice in 40 mL of cold **lysis buffer** containing 1 M urea, 300 mM NaCl, 50 mM HEPES, 50 mM imidazole, and 5% glycerol at pH 7.5, to which was

added 40 mg lysozyme, 500 units of benzonase, and one SigmaFAST protease inhibitor tablet. After the pellet was fully resuspended, it was lysed by sonication on ice and centrifuged for 30 minutes at 48,000 g and 4 °C. 10 µL of the supernatant was taken for SDS-PAGE analysis, and remaining supernatant is filtered through a 0.2 µM filter and then injected at a rate of 1 mL/min onto a 5 mL Ni-NTA agarose column pre-equilibrated at 4 °C with lysis buffer. A 10 µL sample of the flowthrough during injection was collected and stored for SDS-PAGE analysis. The 5mL Ni column was washed with lysis buffer at 1 mL/min until the OD_{280} of the flowthrough measured less than 30 mAU. Protein was eluted at 0.5 mL/min on a 25 mL gradient to 100% **elution buffer** containing 300 mM NaCl, 50 mM HEPES, 750 mM imidazole and 5% glycerol at pH 7.5. Fractions were collected every 1 mL and samples of these fractions were analyzed by SDS-PAGE to determine which fractions contained His-tagged protein and/or impurities. As a general rule, fractions 7-12 contained some His-tagged protein and some impurities, while fractions 13-25 contained pure His-tagged protein. Fractions containing mainly the protein of interest were pooled together, to which was added 500 units of benzonase to remove any transiently bound DNA, and dialyzed at 4 °C several times against **dialysis buffer** containing 100 mM NaCl, 25 mM HEPES, and 2 mM EDTA at pH 7.5.

After several days, dialyzed, Ni-purified protein was centrifuged to remove precipitate and the absorption at 280 and 260 nm was measured to determine transiently-bound DNA contamination. If the sample has an $A_{260/280}$ ratio above 1, additional benzonase is added and dialysis continues until this ratio is below 1. For all proteins except for the trimeric esterase, the protein sample is concentrated at 4 °C with an Amicon-15 100 kDa cutoff spin concentrator at 3,000 g – with the trimeric esterase a 30 kDa cutoff must be used – until a final volume is

reached of 1-1.5mL. To avoid unwanted precipitation due to concentration gradients formed during centrifugation, the sample must be mixed via pipetting every five minutes. Concentrated protein is centrifuged to remove precipitate and stored at 4 °C for further analysis and size exclusion purification.

2.3 - SDS-PAGE

For protein complexes that may oligomerize into multiple species, SDS-PAGE is a critical first step in analysis. Certain natural oligomeric proteins may co-purify with the designed fusion protein, and interfere with the size distribution analyses. The most common protein that will co-purify is the free esterase, formed if the secondary oligomerization site is proteolyzed during induction or cell lysis, and which may still have a His-tag, with which it can bind to the nickel column and co-elute. The second is GroEL, a protein with a monomeric molecular mass of 57 kDa, but an oligomeric mass of 804 kDa. This is in the range expected for octahedral complexes of the trimeric esterase, and will be co-purified with these complexes during a size exclusion preparation, if not removed during initial Ni purification. The presence of both of these contaminating proteins can be discerned by denaturing gel electrophoresis, making SDS-PAGE of Ni-affinity or size exclusion fractions a crucial tool for optimizing purification conditions. Protein samples of interest are electrophoresed on a 12% polyacrylamide gel (BIO-RAD) at 200 V in denaturing conditions. Protein ladder, purified trimeric esterase, and/or GroEL are included as standards. SEC-purified esterase trimer, with a monomeric molecular weight of 32.3 kDa, can be visualized as a single band on an SDS-PAGE gel between 25 and 37 kDa (Figure 2.1).

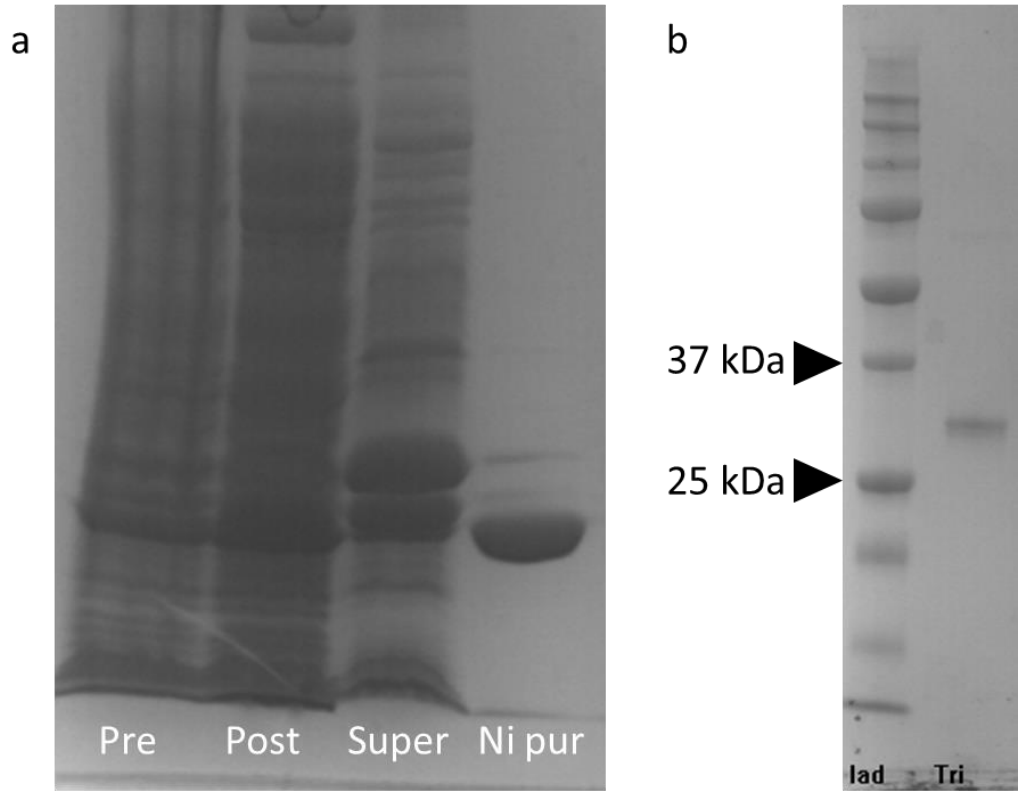


Figure 2.1. Purification of the esterase trimer assayed by SDS-PAGE. a) 10% polyacrylamide gel. Lanes 1 and 2: pre- and post-induction samples. Lane 3: supernatant after cell lysis and centrifugation. Lane 4: Ni-purified trimeric esterase. b) 12% polyacrylamide gel. Lane 1: protein ladder, lane 2: SEC-purified trimeric esterase.

2.4 - Size Exclusion Chromatography

Size Exclusion Chromatography was a quick and efficient method to examine the oligomerization states of the various protein designs, and was used on both an analytical and a preparatory scale. For both applications, we used a Superose 6 10/300 column at 4°C equilibrated with 1 column volume (25 mL) of dialysis buffer with a flow rate of 0.4 mL/min. For analytical purposes, 100 µL of clarified sample is injected onto the column, while for preparation, 500 µL of clarified sample is injected, and 0.5 ml fractions are collected starting from 5.25 ml. The void volume of this column is ~7.5 mL, and the column is calibrated by protein molecular weight standards ferritin (440 kDa) and GroEL (804 kDa), which, when

injected on the column have elution profiles with peaks around 13.8 mL and 12.0 mL, respectively. Fractions collected from preparatory size exclusion purifications can be analyzed by SDS-PAGE, native PAGE, or by rechromatography on the Superose 6 column. Fractions containing oligomeric species of interest are pooled, concentrated with spin concentrators with an appropriate molecular weight cutoff, centrifuged to remove precipitate, and then stored at 4°C for further analysis. The SEC-purified esterase trimer elutes as a single, sharp peak at 16.5mL (Figure 2.2).

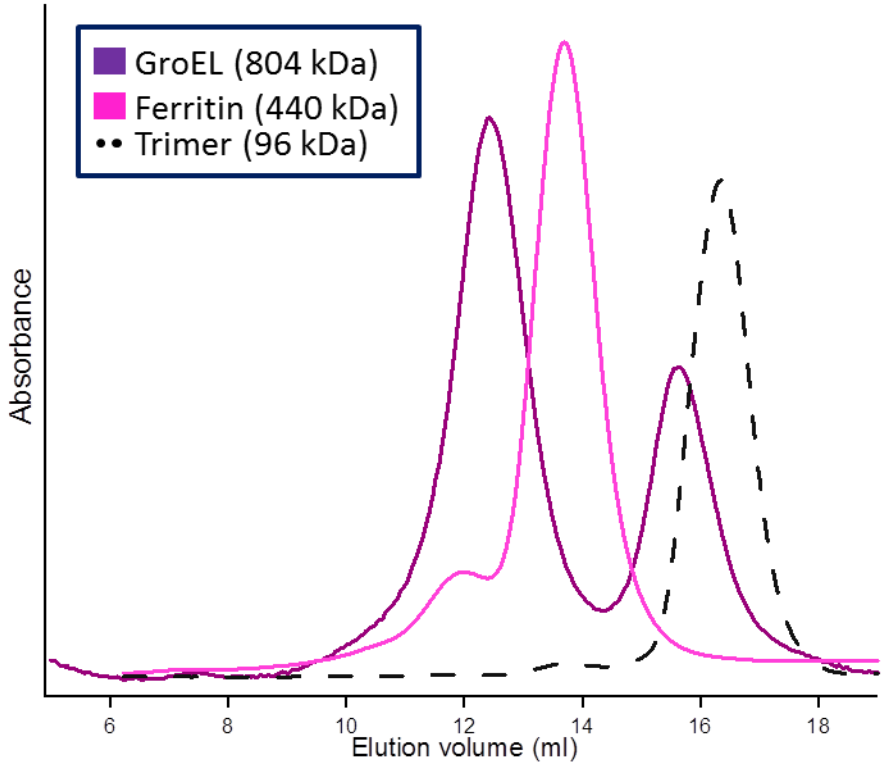


Figure 2.2. SEC molecular weight standards GroEL (804 kDa, purple) and ferritin (440 kDa, pink) elute at 12.0 mL and 13.8 mL respectively on a Superose 6 10/300 column. Unmodified trimeric esterase (96 kDa, black dashed) elutes at 16.5 mL. To reduce clutter, future SEC elution profiles of designed fusion proteins will denote the elution volumes of GroEL and ferritin standards by purple and pink vertical dashed lines.

2.5 - Native PAGE

Native PAGE proved to be an excellent technique to assay the range of oligomeric species in a sample using a very small amount of protein at a wide variety of concentrations. Only 10 µL of a protein sample at a concentration of 0.05-2 mg/ml is required to visualize bands corresponding to oligomerization states, and the resolution for this technique is significantly improved compared to size exclusion chromatography. Native PAGE electrophoresis separates species based on the size of the protein complex and its total charge; for different oligomers of the same protein building block the total charge varies minimally and so it provides an excellent tool for determining the number of different oligomers formed as well as their relative concentrations. Approximate size of these species can also be determined by comparing them to known protein standards, but as these protein standards have different shapes and charges than the protein of interest, a direct size comparison based on band mobility is impossible. Native PAGE can also be used to determine if a species is interconverting between two oligomerization states. If two species interconvert during the timeframe of the gel being run, the bands corresponding to each of those two species will be smeared towards the other. If this interconversion is rapid, it may not be possible to distinguish individual bands, instead appearing as a smear in between the two species.

Native PAGE was run with 3-8% Tris-Acetate polyacrylamide gels in Tris-Glycine buffer at pH 8.4 and 4 °C at a constant current of 70-80 V for 16-20 h. 6x native PAGE buffer (50% glycerol, 0.01% bromophenol blue) was added to all samples prior to electrophoresis. We used the two previously-described molecular weight standards ferritin and GroEL as standard

markers, and under these conditions the major band for ferritin migrates with an R_f of around 0.8 and the GroEL migrates with an R_f of around 0.5, meaning we get significant resolution in the critical area between the two, where we would expect to find the smaller oligomeric species – tetramers, hexamers, and octamers of trimers. In cases where the fusion protein oligomerizes into a species smaller than a tetramer, the native gel was run for 24 hours at 20 V, or until the dye front reaches the bottom of the gel. SEC-purified trimeric esterase runs as a single band located slightly above the dye front, and considerably below standard proteins GroEL and ferritin (Figure 2.3).

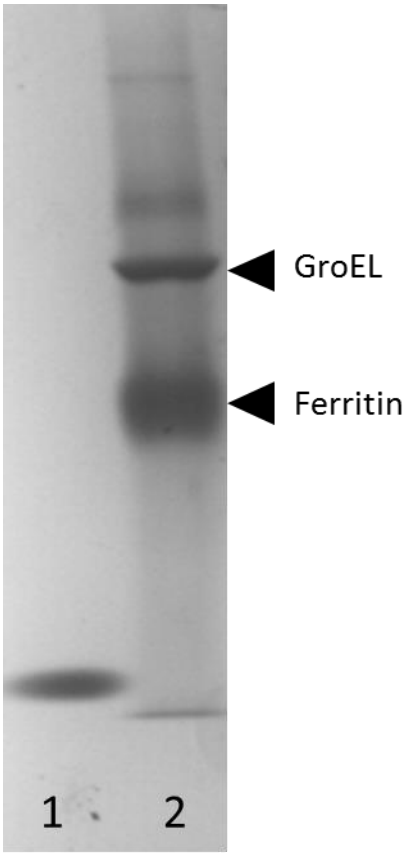


Figure 2.3. Native PAGE of trimeric esterase with standards. Lane 1 is the trimeric esterase, with a native molecular weight of 96 kDa. Lane 2 is protein standards GroEL (804 kDa, top arrow) and ferritin (440 kDa, bottom arrow).

2.6 - Analytical Ultracentrifugation

A powerful tool to determine the oligomerization state of designed fusion proteins is sediment velocity-analytical ultracentrifugation (SV-AUC), which centrifuges the sample of interest at high speeds and tracks the migration of molecules either by changes in absorption or refractive index as they sediment radially in a specialized rotor. The sedimentation of solute molecules is described by the Svedberg equation below, where u is the radial velocity of a solute molecule, ω is the angular velocity, and r is the radial position of the solute, and where M is the molar mass, v is the partial specific volume of the solute, ρ is the density of the solvent, N_A is Avogadro's number, and f is the frictional ratio. The left hand of the Svedberg equation is experimentally obtained, while the right hand must be solved by using known features of the system. The Svedberg (S) is the most common unit of measurement for sedimentation coefficients and is defined as 10^{-13} sec.[1]

$$s = \frac{u}{\omega^2 r} = \frac{M(1 - \bar{v}\rho)}{N_A f}$$

Since we are analyzing the behavior of various oligomers of the same protein, the partial specific volume will always be a constant, as will the solvent density. This leaves the relationship $s \sim M/f$, meaning that heavier species will sediment faster, and species with a high frictional coefficient will sediment slower. The frictional coefficient, defined as f/f_0, is the ratio between the observed f-value and the f-value of a perfect sphere, as defined by the equation $f_0 = 6\pi\eta R_0$, where η is the solution viscosity and R_0 is the radius of the theoretical spherical protein. Monomeric, spherical proteins generally have a frictional ratio slightly higher than 1 due to their hydration shell, while elongated or hollow protein complexes have higher frictional

ratios. The pyruvate dehydrogenase complex, for example, with its hollow interior for trapping unstable intermediates, has a frictional ratio of 2.5.[2] Highly elongated species, such as collagen or chromosomal DNA, have frictional ratios around 3-4.

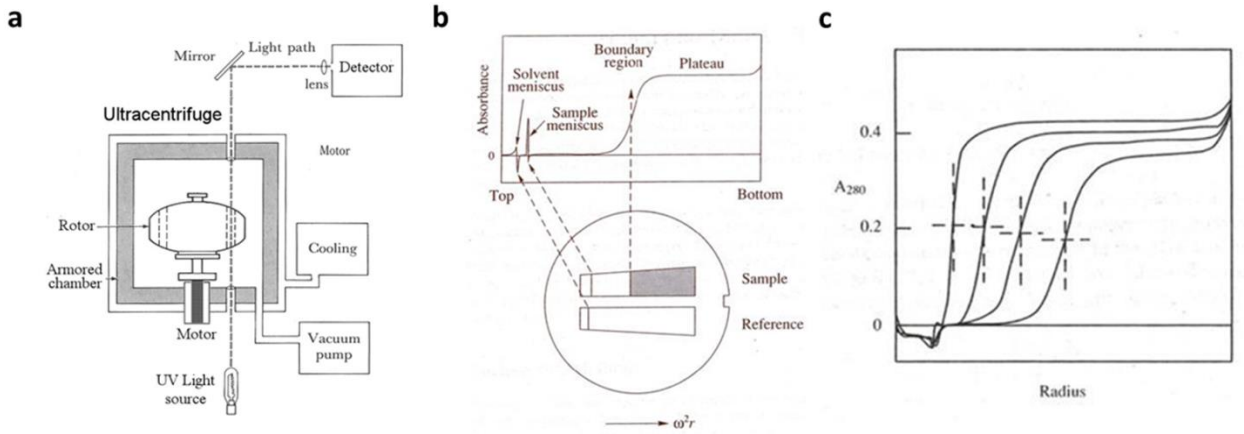


Figure 2.4. Analytical ultracentrifugation. a) Schematic of a typical ultracentrifuge. b) A single scan from an ultracentrifugation experiment. Important regions are labeled. c) Four successive scans of an ultracentrifugation cuvette. The sedimentation coefficient can be determined by measuring the velocity of the boundary midpoint. Images adapted from http://www.bioc.rice.edu/bios576/AU/AU_Page.html and https://www.mun.ca/biology/scarr/Analytical_Ultracentrifugation.html

Prior to advances in computation power, sedimentation coefficients were measured by tracking the boundary midpoint over time. Nowadays, software is able to take advantage of the wealth of data generated from a single run and model the underlying Lamm equation. The simplest, and least computationally demanding software for this task is **sedfit**, which applies a universal frictional ratio to all species in the sample, allowing a direct correlation of different sedimentation coefficients with molecular masses.[3] The frictional ratio parameter can be automatically optimized to minimize the r.m.s.d. but if multiple species in a sample have different frictional ratios, the optimized ratio will be a weighted average of all species. This is a major problem for a solute containing species with widely varying frictional ratios, as elongated

species will be reported to have artificially high molecular masses, and vice versa for globular species. Nevertheless, sedfit allows a reasonable estimation of the number, distribution, and molecular masses of the species in a sample. Using sedfit, the sedimentation profiles for the trimeric esterase and GroEL were calculated. The trimeric esterase sediments as a single species with a sedimentation coefficient of 3.8 S and a fitted frictional ratio of 1.25 (Figure 2.6). This yields a calculated molecular mass of 102 kDa, reasonably close to the expected mass of 96.9 kDa. The sedimentation profile of GroEL shows two species with sedimentation coefficients 1.8 and 14.3 S, most likely correlating to the monomer subunit and the fully assembled complex, respectively. Using a published frictional ratio for GroEL of 1.3,[4] this larger species has a calculated mass of 788 kDa, also in close agreement with the expected molecular mass of 804 kDa.

A more powerful tool for deconvoluting sedimentation data is **Ultrascan**, which separately calculates molecular weights and frictional coefficients for each sedimenting species.[5] This method requires significantly more computational power, and data must be submitted to supercomputer clusters for multiple rounds of Monte-Carlo analysis. Nevertheless, this method affords the highest resolution possible and gives definitive determinations of both size and shape for all species in a heterogeneous sample. Ultrascan analysis of the trimeric esterase (Figure 2.8) revealed a single, sharply-defined species with a sedimentation coefficient of 5.6 and a frictional ratio of 1.23, but a calculated molecular weight of 83.9 kDa, which is less accurate than the molecular weight calculated by sedfit. This may indicate the presence of an unmodeled monomer-trimer equilibrium, but it may also reveal an

essential drawback of relying on calculating a frictional ratio to derive molecular masses by AUC.

Analytical ultracentrifugation was performed on protein samples using a Beckman Proteome Lab XL-I analytical ultracentrifuge (Beckman Coulter, Indianapolis, IN) equipped with an AN60TI rotor. The hydrodynamic behavior of the various proteins was analyzed at a protein concentration corresponding to an OD_{280} of 0.2. 450 µL of each protein sample was loaded into pre-cooled standard sector-shaped, 2-channel Epon centerpieces with 1.2 cm path-length, and allowed to equilibrate at 6 °C for 2 h in the non-spinning rotor prior to sedimentation. Proteins were sedimented at 36,000 rpm. Absorbance data were collected at a wavelength of 280 nm. Sedimentation velocity data were initially analyzed by sedfit using a continuous c(s) distribution and a resolution of 250, fitting the frictional ratio with a simulated annealing algorithm. Selected samples of interest were further analyzed by 2-dimensional sedimentation spectrum analysis (2-DSA) using the finite element modeling module provided with the Ultrascan III software (http:/www.ultrascan.uthscsa.edu). Confidence levels for statistics were derived from 2-DSA data refinement using a genetic algorithm followed by 50 Monte Carlo simulations.

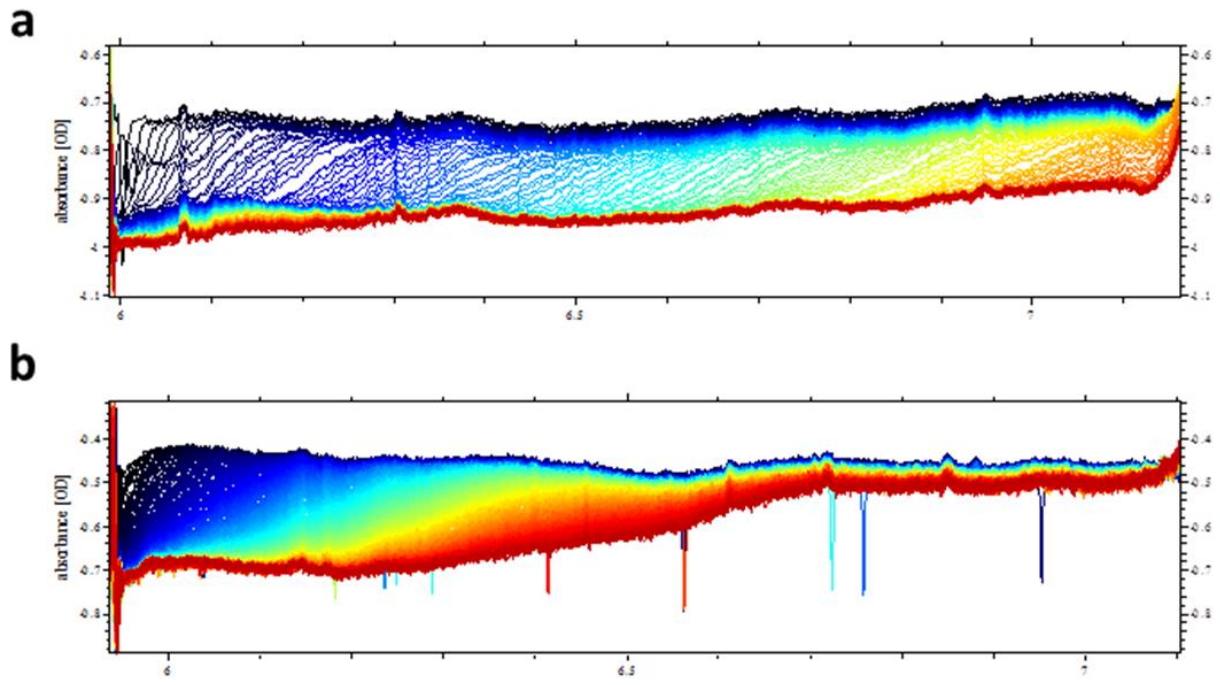


Figure 2.5. Raw AUC data for a) GroEL, and b) unmodified trimeric esterase. Violet traces represent the first scans, red traces represent the final scans.

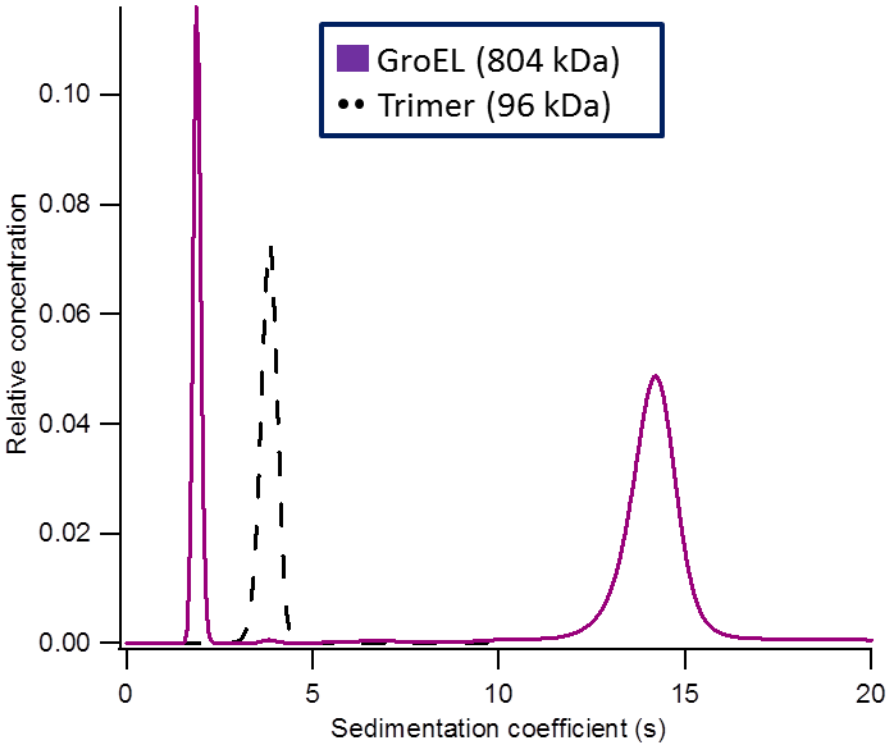
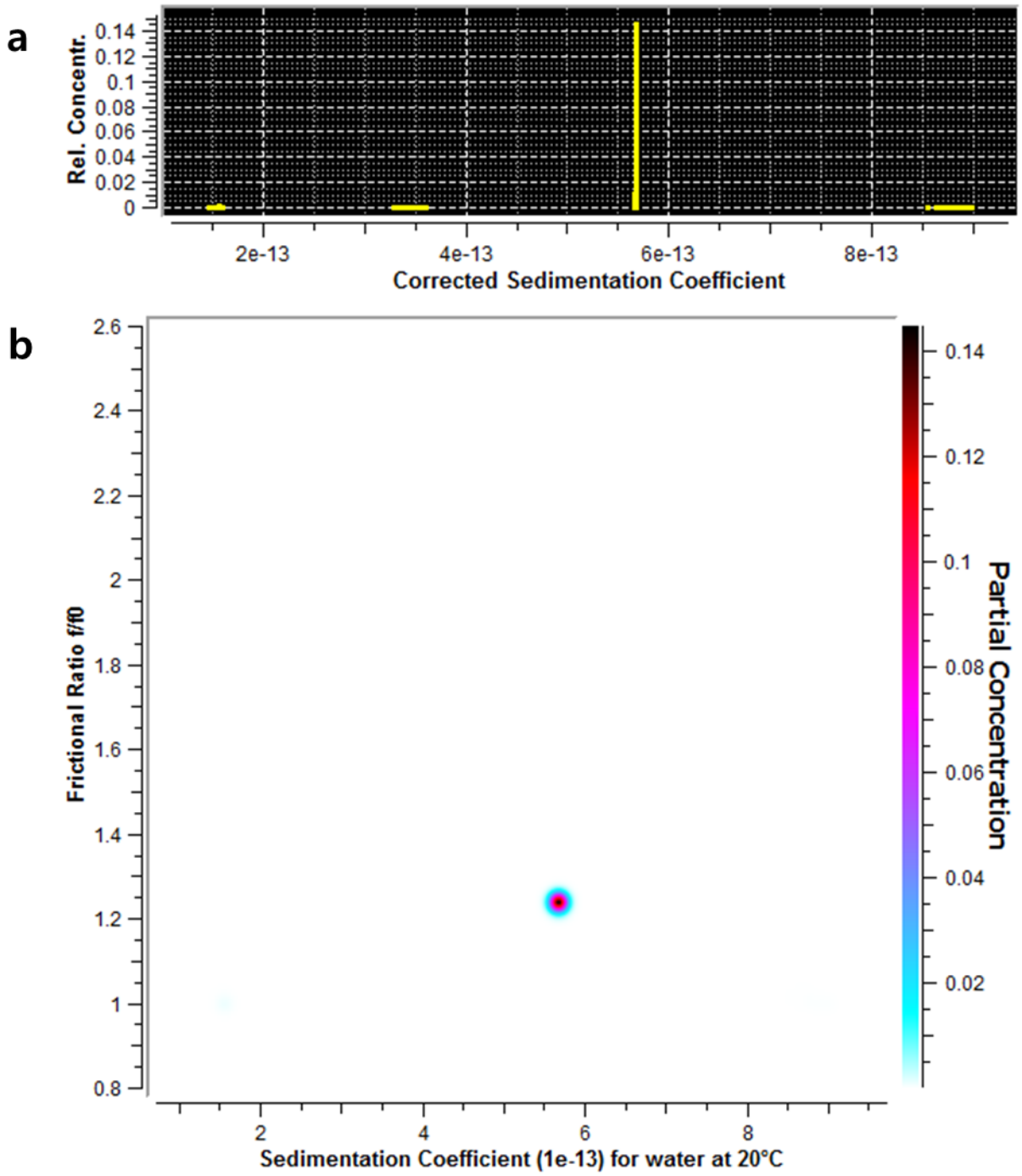


Figure 2.6. Sedimentation profiles of trimeric esterase and GroEL, analyzed by sedfit. GroEL (purple) sediments as both monomeric (57 kDa) and tetradecameric (804 kDa) species. The trimeric esterase (96 kDa, black dashed) sediments as a single species.



Tri-Est	$s_{20,w}$	M.W. (kDa)	f/f_0	Conc. %
Species 1	5.67 ± 0.004	83.9 ± 0.9	1.23 ± 0.01	95.0%

Figure 2.7. Ultrascan analysis of the trimeric esterase. The trimeric esterase could be identified as a single species by both 1D (a) and 2D (b) analytical methods, but the mass accuracy could not be improved over the lower resolution sedfit analysis.

2.7 - Transmission Electron Microscopy

Although analytical ultracentrifugation proves to be a powerful tool for determining the oligomerization states of protein assemblies, it gives only limited insight into the geometry of these species. Using Transmission electron microscopy (TEM) it is possible to directly identify the distribution of size and shape of individual oligomers in a sample, and with a high enough resolution, identify their geometry as well. Preliminary investigations of sample proteins were conducted using negative-stain TEM. Samples of interest were diluted to 0.02 mg/mL and adsorbed on a glow-discharged Formvar 400 carbon grid using conventional negative staining procedures. Imaging was performed at room temperature with a Morgagni 268(D) transmission electron microscope (FEI Company) equipped with a tungsten filament operated at an acceleration voltage of 100 kV and a mounted Orius SC200W CCD camera (Gatan Inc.). Negative stain TEM images of the trimeric esterase show a homogeneous sampling of triangular species with edge lengths of 5-7 nm, in close agreement with lengths derived from the crystal structure (Figure 2.7).

If a protein sample is deemed oligomerically homogeneous by other methods, higher resolution images can be attained with cryo-TEM , where adsorbed particles are flash frozen in liquid ethane to preserve them in their native state, as the dehydration that is a consequence of negative staining may perturb the protein cage's structure. This, as well as the small B-factors from the low temperatures maintained throughout imaging, allows significantly more regular images to be obtained, which can be selected and then averaged to generate a series of reference-free class averages that, with enough averaged particles, should represent all the unique arrangements of the protein complex that can be viewed from the top-down. The

electron density of selected class averages can be reconstructed in 3-dimensional space and the resolution improved by applying the known symmetry of the particle and by modeling in the known crystal structure of that particle or a similar complex.

For Oct-4-4, 3 µL of a protein sample concentrated to an OD_{280} of 0.5 was adsorbed onto a glow-discharged Quantifoil grid (R2/2 200 mesh) and vitrified using a Vitrobot (FEI Mark IV). The sample was imaged on a Tecnai TF20 transmission electron microscope (FEI) equipped with a field emission electron gun operated at 200 kV. Images were recorded at a magnification of 41667x on a Gatan K2 Summit camera, and binned (2 x 2 pixels) resulting in a pixel size of 4.4 Å on the specimen level. All the images were acquired using low-dose procedure to minimize radiation damages to the samples, with a defocus value in the range of 2 - 4 µm.

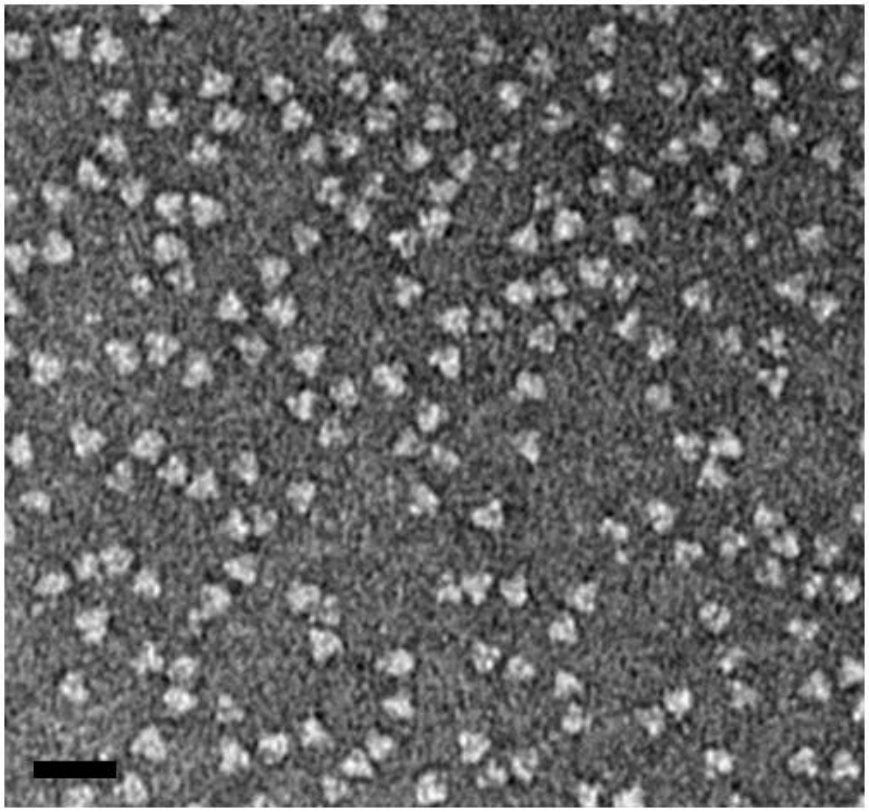


Figure 2.8. Negative stain transmission electron microscopy of trimeric esterase. Esterase is visually identifiable as possessing C_3 trimeric symmetry, with approximate side length of 7 nm. Scale bar is 20 nm.

2.8 - Ion Mobility-Mass Spectrometry

This technique can theoretically yield the highest resolution for the molecular masses of protein complexes, but is also the newest and poses its own unique challenges. Mass spectrometry has been successfully applied to native protein complexes in the gas phase for over two decades. More recently, it was discovered that a weak electric field, if applied to native protein complexes in the gas phase, induced an additional separating effect. Similar to native PAGE, these slightly positively-charged complexes pass through the electric field at a rate proportional to their collisional cross-section, which can not only be used to distinguish several species with similar masses for proteomic applications, but can also be used to probe protein

structure, dissociation, and unfolding pathways. Recently, this technique was successfully applied to other protein cages, identifying the molecular masses of a C_3+C_2 system and obtaining accurate masses for the mixture of 12-subunit tetrahedron, 18-subunit trigonal bipyramid, and 24-subunit octahedron.[6] This has interesting applications for our use, both in confirming the oligomerization state of protein complexes and in describing the shape of these complexes. The primary concern with analysis of proteins by IM-MS is that the buffer conditions must be complete free of sodium ions, which in the case of large protein complexes with many ionic groups can be difficult to achieve.

To achieve higher resolving powers, all proteins submitted to MS analysis were purified in dialysis buffer with 100 mM ammonium acetate replacing sodium chloride, buffered with acetic acid and ammonium hydroxide, then buffer exchanged into buffer containing only 100 mM ammonium acetate before ionization. When purifying large protein complexes for MS analysis, ammonium acetate was substituted for sodium chloride in all buffers used in purification: lysis, elution, and dialysis. Ammonium acetate is used in place of sodium chloride because both counterions are volatile, and as such are removed during the transition to the gas phase. Samples prepared for mass spectrometry were purified as described above and then concentrated to 40 µL. The minimum concentration of protein required for analysis is 1 µM, with higher concentrations preferred. Sample were then loaded into gold plated needles prepared in house according to [7]. Nano-electrospray-ion-mobility-TOF mass spectrometry was performed using a Synapt G2 Traveling-Wave IM-MS instrument (Waters Corp, Manchester, U.K.). Ions were generated by applying a voltage of 1.5kV between the needle and the instrument source, with further voltage drops aiding in acceleration and desolvation as ions

passed through the skimmer region of the instrument. The quadrupole region was set to RF-only mode for collection of complete mass spectra, and in some cases was tuned to isolate selected peaks for MS/MS analysis. A range of collision energies were tested for enhanced transmission and desolvation of the ions, and in some cases dissociation of the ion into its component subunits. The base values for collision energies were 20-50 V, however energies up to 150 V were utilized for dissociation experiments. The IMS region of the instrument was operated at 4mBar of nitrogen, with wave heights and wave velocities of 15 V and 150 m/s, respectively. The instrument time of flight mass analyzer was operated in sensitivity mode, and mass spectra were collected from 1000 to 15000 m/z. Data analysis was performed using the manufacturer-provided Masslynx software.

IM-MS analysis of the unmodified trimeric esterase shows one major species with m/z peaks between 4,000 and 5,000 (Figure 2.8). These calculate to a molecular mass of 97.7 kDa, the most accurate mass of all the aforementioned techniques. The positive correlation between m/z and drift time indicates that this species retains its folded structure throughout analysis.

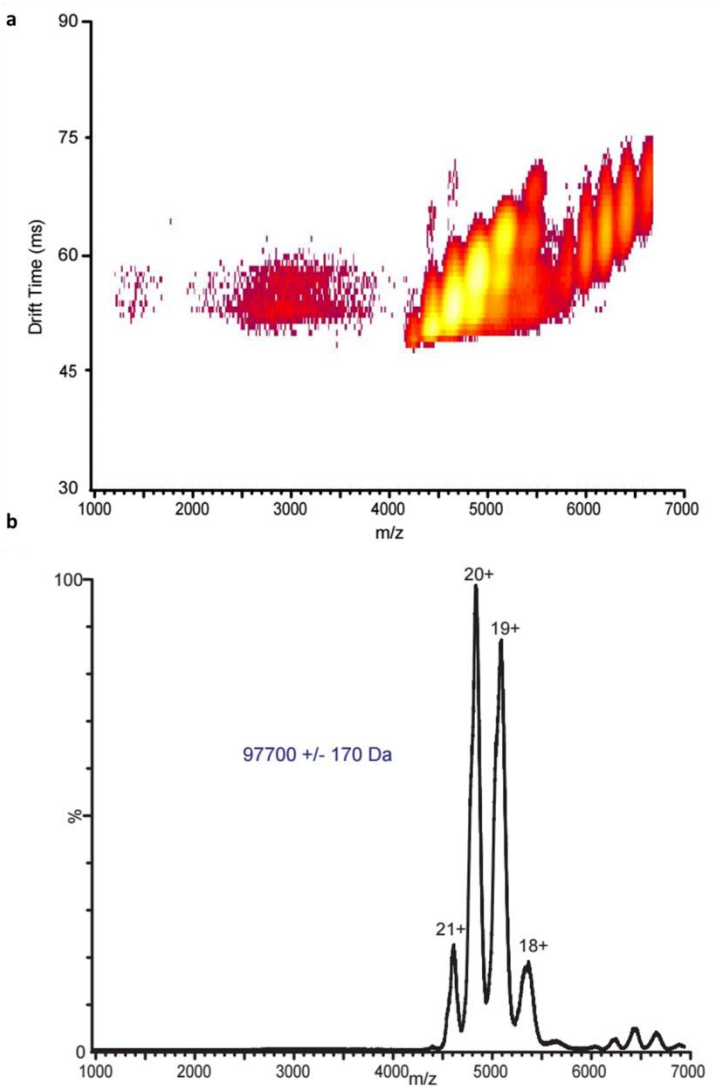


Figure 2.9. Native IM-MS of trimeric esterase. a) 2-D plot showing separation of m/z and drift time, in logarithmic scale. Species with lower charge states have a higher collisional cross section, which correlates to a higher drift time. b) 1-D plot of m/z, in linear scale. The major species has a molecular weight of 97.7 kDa, very close to the calculated molecular weight of 96.9 kDa.

2.9 - Conclusions

In this chapter I have surveyed a wide range of experimental techniques that can gauge size and/or shape of various protein complexes. These techniques differ in resolution and sample prep complexity and as such will be applied to analysis of designed fusion protein

complexes in roughly the order presented in this chapter. Using the trimeric esterase building block as a control protein, I showed that this suite of techniques as described above can accurately and complementarily identify the structure and oligomerization state of a protein complex. In this case, these techniques confirmed that the native esterase exists stably as a trimer with C_3 symmetry, and does not oligomerize further at the concentrations tested. These attributes make it an excellent building block for future investigations.

2.10 - References

1. Lebowitz, J., Lewis, M.S. & Schuck, P. Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein Sci.* **11**, 2067-79 (2002).
2. Bosma, H.J., De Kok, A., Van Markwijk, B.W. & Veeger, C. The size of the pyruvate dehydrogenase complex of Azotobacter vinelandii. *Eur. J. Biochem.* **140**, 273-280 (1984).
3. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606-19 (2000).
4. Behlke, J., Ristau, O. & Schönfeld, H.-J. Nucleotide-Dependent Complex Formation between the Escherichia coli Chaperonins GroEL and GroES Studied under Equilibrium Conditions. *Biochemistry* **36**, 5149-5156 (1997).
5. Demeler, B. UltraScan - A comprehensive data analysis software package for analytical ultracentrifugation experiments. in *Analytical Ultracentrifugation: Techniques and Methods* 210-230 (2005).
6. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* **6**, 1065-1071 (2014).
7. Ruotolo, B.T., Benesch, J.L.P., Sandercock, A.M., Hyung, S.-J. & Robinson, C.V. Ion mobility-mass spectrometry analysis of large protein complexes. *Nat. Protocols* **3**, 1139-1152 (2008).

Chapter 3

Initial attempts at protein cage design

In keeping with a design philosophy that constrains as few parameters as possible, the initial fusion protein was designed using only basic geometric considerations. A previous student working on this project used a dimeric, antiparallel coiled-coil connected to the trimeric building block with a short flexible linker to assemble a mixture of cage-like structures.[1] For our system, we plan to use tetrameric, rather than dimeric coils, as well as coils oriented in a parallel, rather than antiparallel orientation, both of which orient the attached esterase subunits in close proximity to each other. We therefore ensured that steric hindrance between symmetric subunits would not be a concern by attaching a 12-residue linker containing 10 glycines to the C-terminus of the trimeric esterase. This was followed by a 7-heptad coiled-coil sequence that contained phenylalanine residues at all but one of the *a* and *d* positions on the coiled-coil heptad repeat, and which had previously been crystallographically verified to associate into a parallel tetramer.

3.1 - Oligomerization of Coiled-Coils

Since this project employs coiled-coil motifs as a fundamental symmetric building block, it is worthwhile to engage in a brief overview of the theory behind coiled-coil assembly and prediction of their oligomerization states, as this plays a critical part in this project as well as for future research aims and protein cage functionalization. The alpha helix is the most common

protein secondary structure and is highly regular, with every residue twisting the helix by 100°. This means that every seven residues the helix will have completed slightly under two turns, and the sidechain of the eighth residue will be oriented in the same direction as the first residue. Therefore, a common way to represent the packing of alpha helix sequences is by breaking down the sequence into heptad repeats, denoted by letters *a*-*g*. When two or more alpha helices are placed next to each other, creating a coiled-coil, residues at the *a* and *d* positions of each alpha helix will be in closest proximity to each other, as shown in Figure 3.1. These residues are usually hydrophobic, such that they pack with a 'knobs into holes' design, although polar contacts and salt bridges have also been observed at these interior positions. Consequently, the other five residues in the heptad repeat tend to be hydrophilic to minimize disruption of this hydrophobic pocket. Also of note are the *e* and *g* positions (and in higher oligomerization states, the *b* and *c* positions) in the heptad repeat, which, although usually hydrophilic, interact with *e* and *g* residues on neighboring alpha helices, allowing for Coulombic interactions and hydrogen bonding to further determine specificity and increase the enthalpy of assembly.[2]

While this associative behavior has been known to exist in coiled-coils since the 1950's[3] and the heptad repeat pattern was discovered in 1972[4], the relationship between the hydrophobic packing region and the oligomerization state was not elucidated until 1993, when Harbury and coworkers discovered that the organization of the hydrophobic residues leucine, isoleucine, and valine in the *a* and *d* positions heavily influenced the oligomerization state of the coiled-coil.[5] Particularly, they found that a dimeric coil could be specified by using residues of isoleucine and leucine at the *a* and *d* positions respectively, a trimeric coil could be specified

with *a,d* = I,I, and a tetrameric coil could be specified with *a,d* = L,I. Higher oligomerization states can also be directed by mutating in hydrophobic residues at the *e* and *g* positions.[2]

Furthermore, orientation of the coiled-coil systems could be specified by controlling the identity of the *e* and *g* positions.[6] In a parallel orientation, *e* residues pack next to neighboring *g* residues, while in an antiparallel orientation, *e* residues pack next to neighboring *e* residues, and *g* residues pack next to neighboring *g* residues. Therefore, to specify a parallel orientation with a dimeric coil, one simply makes the *e* and the *g* residues two amino acids with opposing charges, and to specify an antiparallel orientation, one makes two separate helices, one with all negatively-charged residues at the *e* and *g* positions, and the other with all positively-charged residues at those positions.[7] It is also possible to specify a homodimeric antiparallel coiled-coil by making the *e* and *g* positions on the N-terminal half of the coil negatively-charged, but positively-charged on the C-terminal half of the coil.[8]

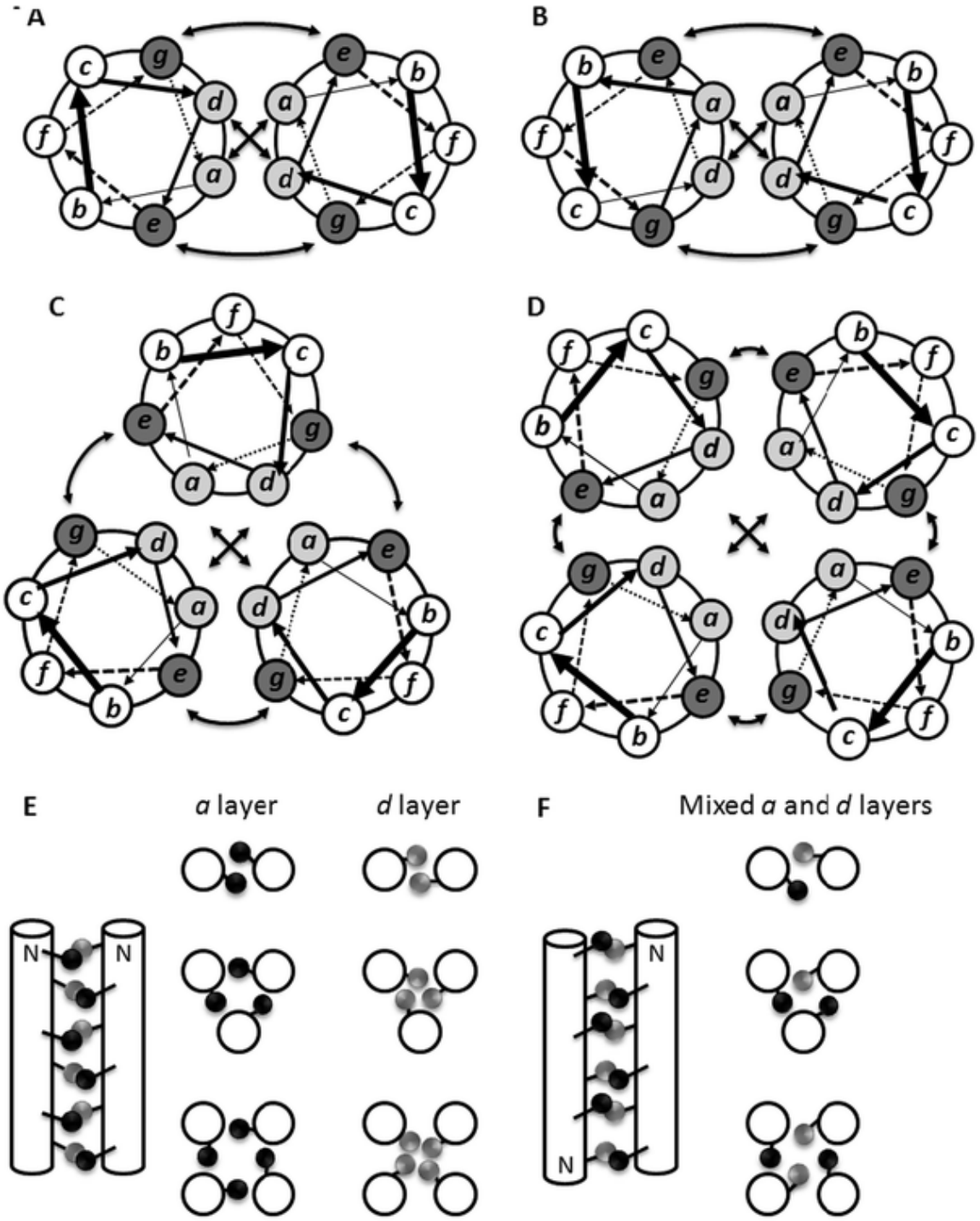


Figure 3.1. Spatial placement of the members of the heptad repeat in different coiled-coil systems, taken from Ref. 8. A) A parallel dimeric coiled-coil. B) An antiparallel dimeric coiled-coil. C) A parallel trimeric coiled-coil. D) A parallel tetrameric coiled-coil. E) Parallel coiled-coils pack all a and all d residues next to a and d residues on neighboring alpha helices. F) Antiparallel coiled-coils pack a residues next to d residues on neighboring alpha helices.

Considerable research since then has gone into determining the oligomerization states of various natural and *de novo* designed coiled-coil systems,[9] and for the purposes of this project, any parallel tetrameric coiled-coil would suffice. We therefore initially chose a *de novo*-designed coiled-coil tetramer based upon a membrane-bound trimeric coiled-coil system found in *E.coli*. Liu et al. found that if the *a* and *d* interior residues of this trimeric coil were replaced entirely by phenylalanine residues, the coiled-coil would form a pentamer with a spiral packing motif (Figure 3.2).[10] If, however, a central phenylalanine residue is mutated into a methionine, this spiral motif is disrupted and instead oligomerizes as a tetrameric coiled-coil with a more classical knobs-into-holes arrangement. It is this tetrameric coiled-coil that was added into Oct-1 (Figure 3.3). This coiled-coil has one other unusual property: its thermal midpoint of unfolding (T_m), at 54 °C, is much lower than can be expected from a seven-heptad coil. For a comparison, the four-heptad tetramer coiled-coil described by Harbury has a T_m of over 100 °C. This low T_m indicates that the coiled-coil system is not tightly associated and may be in equilibrium with the monomer, which could be either a good or bad thing for the successful oligomerization of an octahedron. A loosely-associating coil would be good if these oligomerizing proteins have a tendency to become kinetically trapped into higher order oligomers, as it would slowly re-associate into the entropically preferable octahedral complex. If the coil is too loose, however, it could lead to the octahedron breaking apart, and if enough proto-octahedron complexes assemble together, it can lead to aggregation rendering the complexes insoluble, meaning we should see this complex slowly falling out of solution over time.

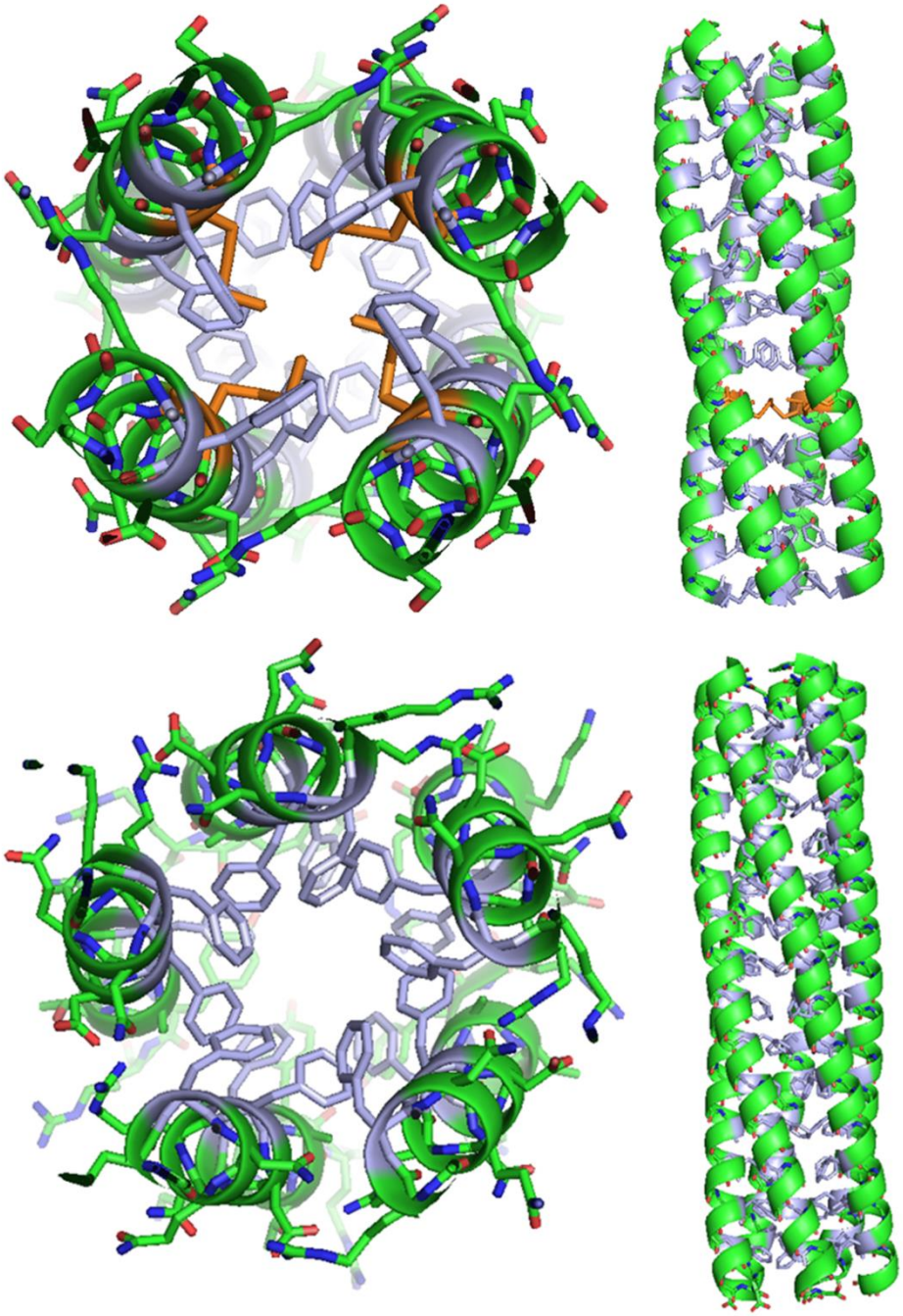


Figure 3.2. Crystal structures of the tetrameric coiled-coil motif inserted into Oct-1 and the pentameric coiled-coil that it is based on. The pentameric coil (bottom, PDB ID 2GUS) has phenylalanines at every *a* and *d* position and packs with a spiral motif, whereas the tetrameric coil (top, PDB ID 2GUV) has a single methionine at the *d* position that breaks this spiral motif in favor of a knobs-into-holes arrangement.

Figure 3.3. Schematic of the design of fusion protein Oct-1. The 6xHis tag is at the N-terminus. Residues at the *a* and *d* positions on the coiled-coil are bolded.

3.2 - Characterization of Oct-1

Oct-1 was successfully transformed, expressed, and purified as described in Chapter 2. Oct-1 yielded around 25 mg/L protein after Ni purification with modest (~90%) purity, and could be size exclusion purified to remove all other visible protein contaminants (Figure 3.4).

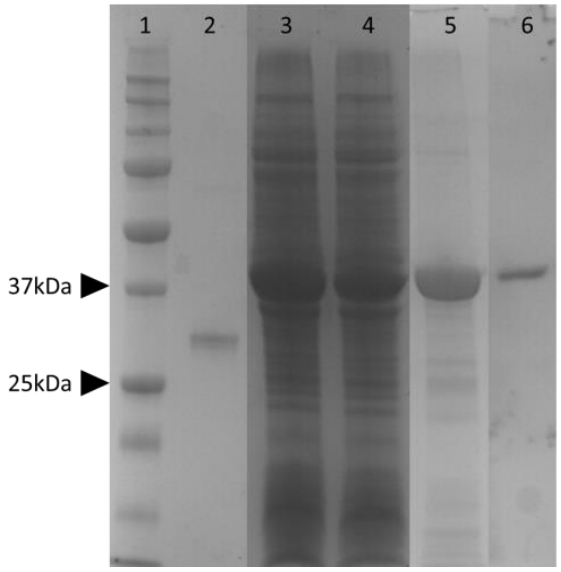


Figure 3.4. SDS-PAGE purification of Oct-1. Lane 1: protein standards ladder. Lane 2: trimeric esterase. Lane 3: supernatant. Lane 4: flowthrough. Lane 5: Ni-purified Oct-1. Lane 6: SEC-purified Oct-1.

3.2.1 - Size Exclusion Chromatography of Oct-1

Size exclusion profiles of Oct-1 one day after Ni purification shows the protein forms a broad range of oligomers larger than the unmodified esterase trimer. Sizeable amounts of

these oligomers appear in the void volume, indicating that these are complexes that are too large to interact with the column matrix (>5 MDa, for the Superose 6 beads). The profile of the same sample after 1 week in 4 °C shows a similarly broad range of species, but all with a much lower absorbance. After two weeks, however, the size exclusion profile shows something markedly different: nothing soluble remains in the void volume, and a single, broad peak is left at 15.4 ml – slightly larger than the absorbance of the trimeric esterase (Figure 3.5).

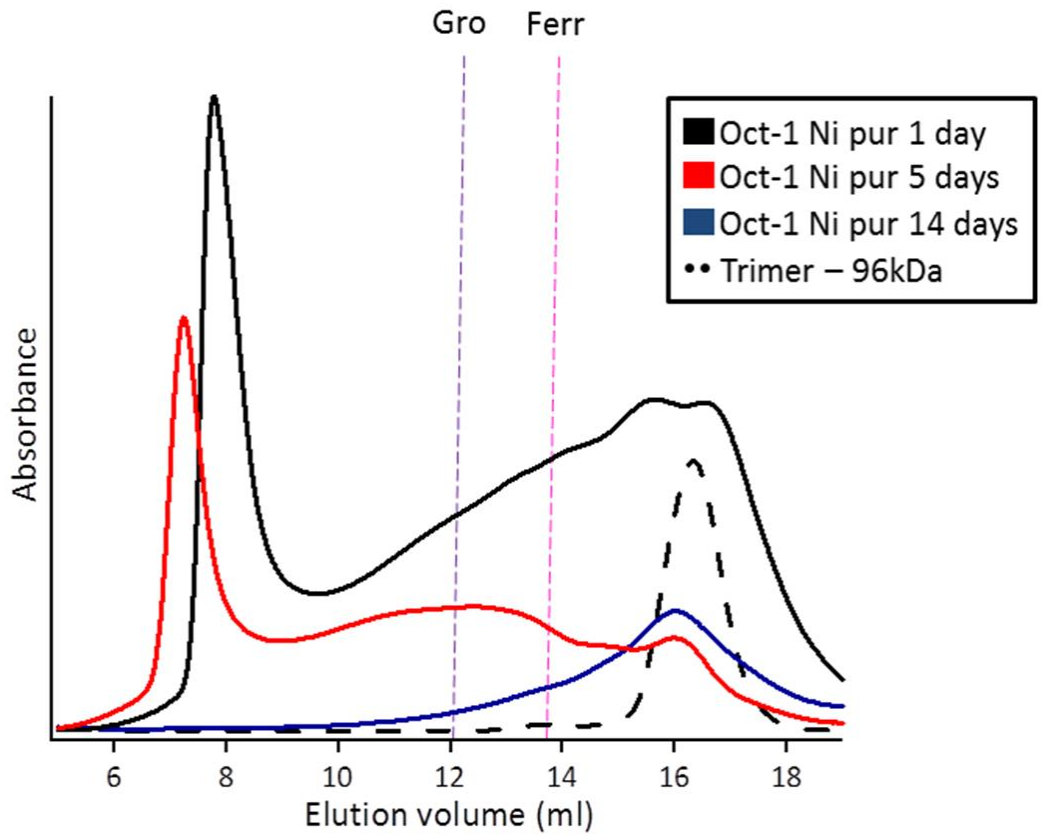


Figure 3.5. Size exclusion profiles of Ni-purified Oct-1 after 1 day of storage at 4 °C (black), 5 days (red), and 2 weeks (blue). Elution profile of trimeric esterase is indicated as black dashed line. Elution volumes of standard proteins GroEL (purple, 804 kDa) and ferritin (pink, 440 kDa) are denoted by vertical dashed lines.

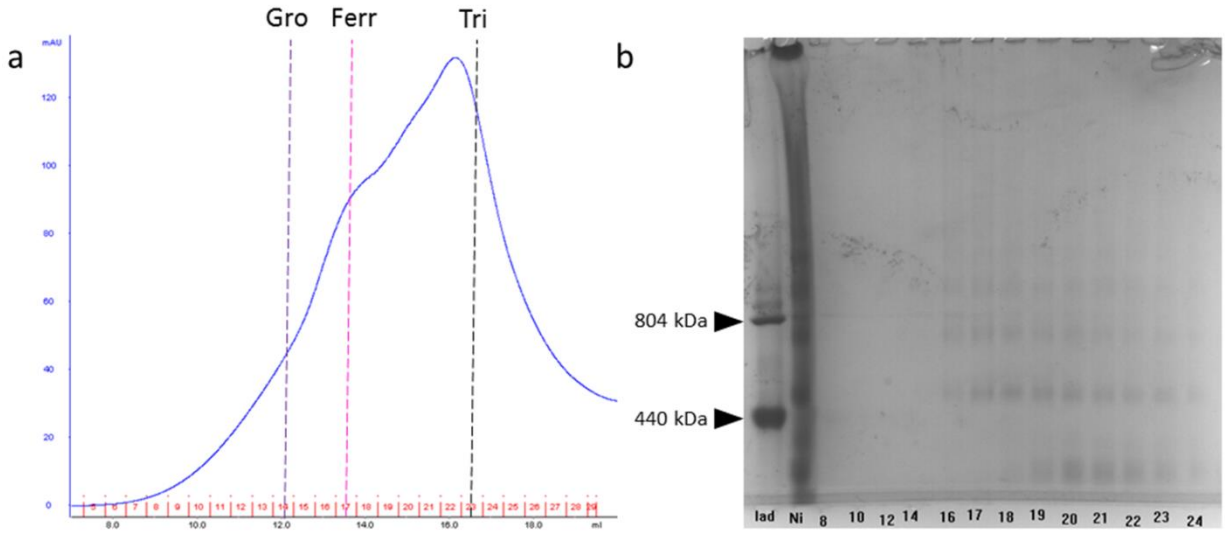


Figure 3.6. SEC purification of Oct-1. a) Elution profile of Oct-1 with fractions at bottom. b) Native PAGE of Ni-purified Oct-1 with corresponding purification fractions from (a). Ladder lane consists of GroEL (top band) and ferritin (bottom band).

This Ni-purified Oct-1 sample was then concentrated and purified using a preparatory Superose 6 column, and fractions were collected (Figure 3.6). Native PAGE of concentrated, Ni-purified Oct-1 shows at least 8 distinct bands with a wide range of molecular weights and smearing in between them, indicating that there is interconversion between different species during electrophoresis. This could be from gaining or losing oligomers, or it could be from interconversion from a compact to an extended state. Native PAGE analysis of column fractions taken from the SEC purification of Oct-1 and stored at 4 °C for several hours shows that a broad range of bands can be observed in all fractions, indicating that Oct-1 rapidly interconverts between different species.

As indicated in figure 3.6, fractions 18-22 were pooled and analyzed by both SEC and AUC. The size exclusion profile of this Oct-1 preparation has an expected broad band with a peak at 15.6 ml, near where the trimeric esterase elutes (Figure 3.7). SEC analysis of this same

fraction after 2 weeks at 4 °C shows a similar, slightly broadened elution profile indicating that this distribution of oligomers is fairly stable in solution.

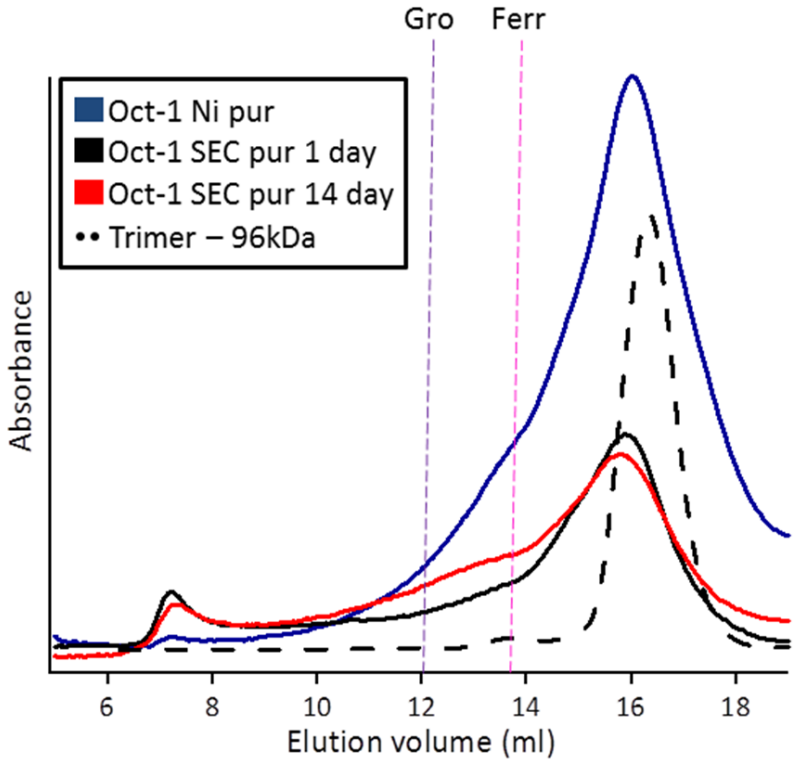


Figure 3.7. Size exclusion elution profiles of Ni-purified Oct-1 (blue), SEC-purified Oct-1 after 1 day of storage at 4 °C (black), and SEC-purified Oct-1 after 2 weeks of storage at 4 °C (red). In comparison to Figure 3.5, SEC-purified Oct-1 appears to be much more stable. Trimeric esterase is indicated as black dashed line.

3.2.2 - Analytical Ultracentrifugation of Oct-1

Analysis of the sedimentation velocity analytical ultracentrifugation profile of SEC-purified Oct-1 depicts a mixture of sedimenting species (Figure 3.9). Using the program 'sedfit' to fit the sedimentation data, we can resolve three major species with sedimentation coefficients of 4.5, 7.1, and 10.7 S. Using the fitted frictional ratio of 1.48, these species correspond roughly to molecular masses of 177, 349, and 642 kDa, which represent complexes of approximately two (225 kDa), four (450 kDa), and six (675 kDa) trimers of Oct-1 respectively.

Several smaller peaks corresponding to larger complexes can also be observed. More concerningly, there is a small side peak corresponding roughly to the s-value of the trimer, indicating that the coil on Oct-1 does in fact associate weakly and spontaneously dissociates. Obviously, this is a poor attribute when one is trying to design a system of associating trimers.

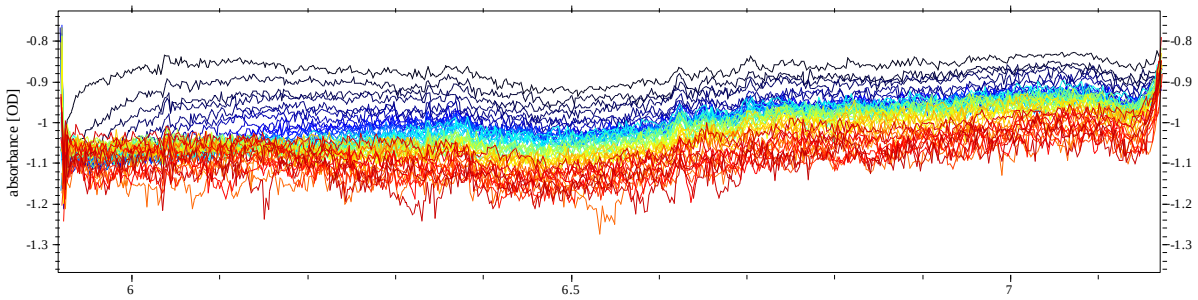


Figure 3.8. Raw sedimentation velocity data for Oct-1. Violet traces represent the first scan, Red traces represent the final scan.

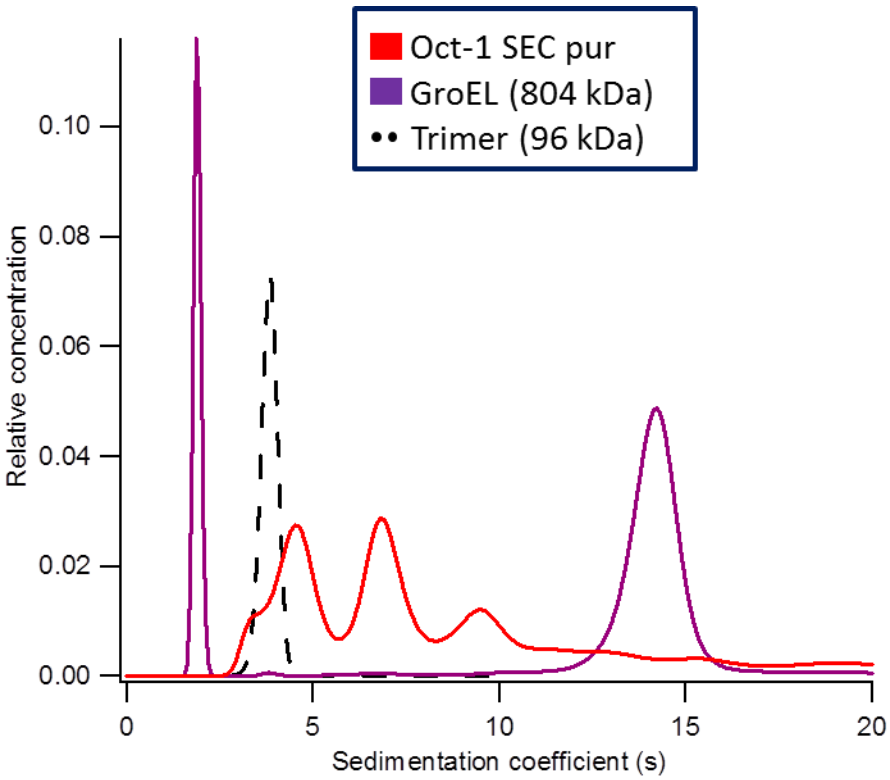


Figure 3.9. Sedimentation profile of Oct-1 after SEC purification (red). Sedimentation profiles of standard proteins GroEL (804 kDa, purple) and esterase trimer (96 kDa, black dashed) are included.

3.2.3 - Transmission Electron Microscopy of Oct-1

Negative-stain TEM images of SEC-purified (F18-F22) Oct-1 show a variety of small oligomers of trimers with no discernable pattern to oligomerization or geometry (Figure 3.10a). TEM images of the void volume fraction of Oct-1 purified in salt- and urea-free buffers show much larger species, some with the proper geometry of an octahedron and roughly 8 trimers, but most particles are larger than expected for an octahedral cage (b).

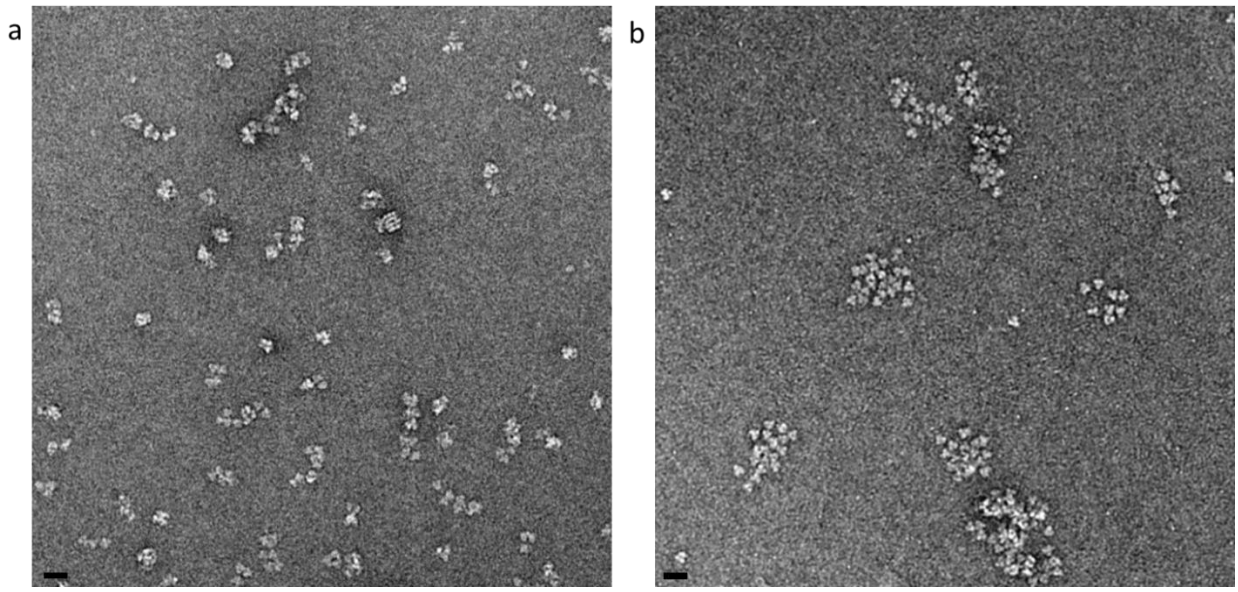


Figure 3.10. Negative stain TEM of Oct-1 after purification in 100 mM NaCl (a), and 0mM NaCl (b). Scale bars are 20 nm.

3.3 - Insertion of a More Stable Coiled-Coil and Design of Oct-2

Oct-1 seems to exist in a multitude of interconverting oligomeric states, many of which are unstable over time. It is a reasonable assumption that the weakly-associating tetrameric coil was the cause of this behavior. To test this hypothesis, we replaced the coiled-coil on Oct-1 with one based on the tetrameric coiled-coil sequences designed by Fletcher et al, which has a thermal midpoint of unfolding of over 100 °C.[11] Inadvertently, an error was introduced into this coiled-coil design that was not realized until a thorough characterization was undertaken. Instead of the canonical *a,d* = L,I interior hydrophobic packing sequence that yields a parallel tetrameric coiled-coil, we inserted a sequence with *a,d* = L,L, known to form an antiparallel trimeric coiled-coil[12], but with single residue modifications can be turned into either a dimeric or a trimeric parallel coiled-coil, suggesting oligomerization is highly context dependent.[5,13,14] Regardless of whether this coil oligomerizes into a dimer, a trimer, or possibly both, having a strongly-bound coiled-coil system should mean the protein building blocks, when formed into

stable oligomers, will be less likely to dissociate and become insoluble aggregates over time. However, because of the long length of their linkers compared to the length of the base esterase trimer, there exists the possibility that these coils may associate with other coils on the same trimer, leading to the creation of a stable monomer or other small aggregates. The second generation of this fusion protein system, consisting of the trimeric esterase connected via a 12 residue linker to a 4 heptad *a,d* = L,L coiled-coil, was termed Oct-2.

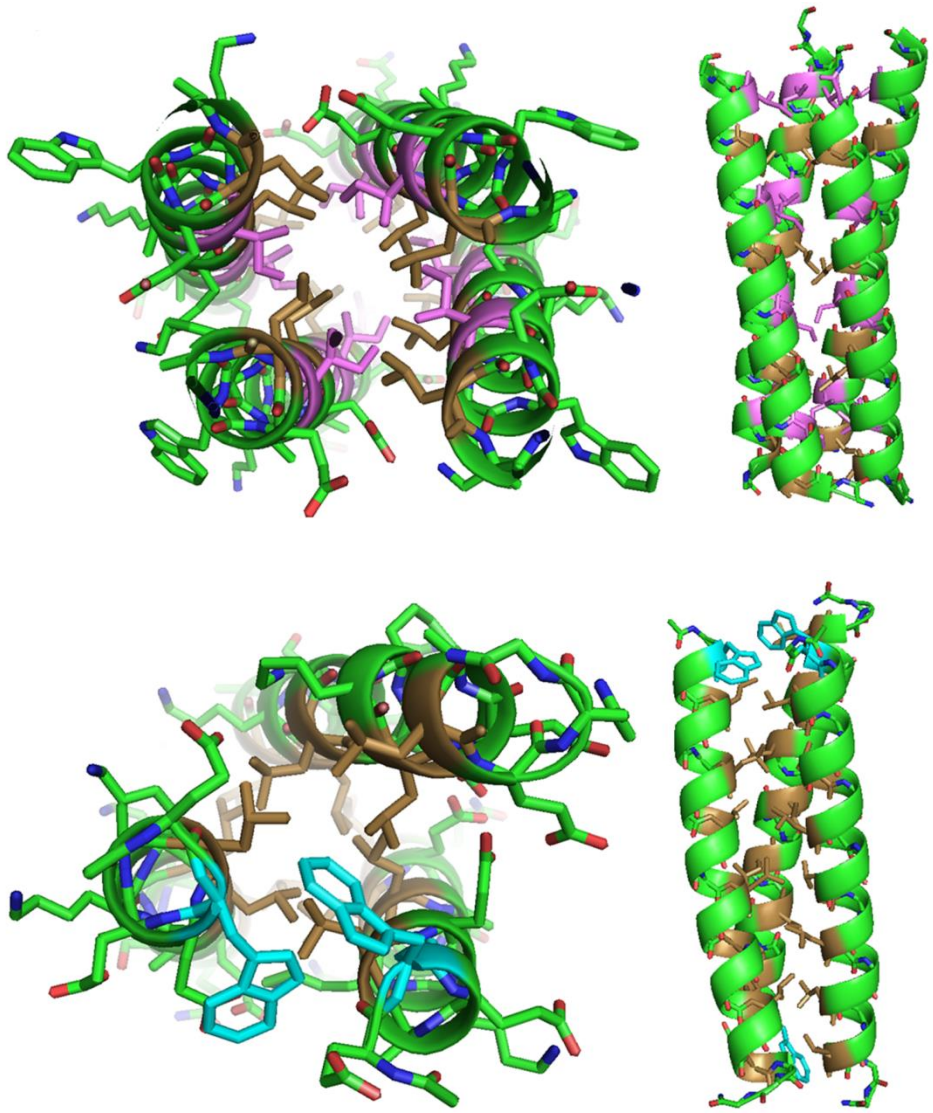


Figure 3.11. Crystal structures of the parallel, tetrameric motif that was supposed to be inserted to replace the coil in Oct-1, and the antiparallel, trimeric motif that was inserted into Oct-2. The tetrameric motif (top, PDB ID 3R4A) has leucine residues (brown) at the a positions and isoleucine residues (purple) at the d positions, while the trimeric coil (bottom, PDB ID 1COS) has leucine residues at both a and d positions. Tryptophan residues (teal) at the N-terminus of each strand of the trimeric coil are highlighted to illustrate the antiparallel orientation of the coiled-coil.

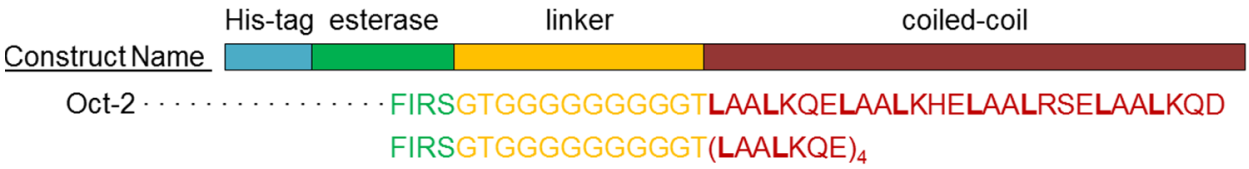


Figure 3.12. Schematic of the design of fusion protein Oct-2. The 12 residue linker was retained, while the coiled-coil in Oct-1 was swapped for a coiled-coil with leucine residues at the *a* and *d* positions. The amino acid sequence encoding the coiled-coil contains minor mutations at the e, f, and g positions to accommodate restriction sites added to the DNA, but will henceforth be abbreviated to the heptad repeat.

3.3.1 - Characterization of Oct-2

Plasmid containing Oct-2 was successfully transformed into competent *E.coli* cells and expressed by standard methods. Oct-2 expressed as a soluble protein in yields of around 10 mg/L after Ni purification, and a protein of a similar size to Oct-1 could be identified as one of the major bands in SDS-PAGE (Figure 3.13). When this impure protein is analyzed by size exclusion chromatography, the resulting elution profile is unexpected: the major peaks all elute significantly later than the unmodified esterase trimer (Figure 3.14). This implies that if this impure protein mixture contains Oct-2 at all, then the vast majority of Oct-2 doesn't even oligomerize into a trimer. There is a very small, broad peak that appears to elute at a similar volume to the unmodified esterase trimer, but this fraction could not be purified in sufficient quantity to analyze. Native PAGE of Ni-purified Oct-2 shows a single band of protein at the dye front, and several faint bands around standard proteins ferritin and GroEL (Figure 3.15). Transmission electron microscopy of nickel-purified Oct-2 confirmed the presence of small assemblies consisting of 2-5 trimers, but these seem to associate in an arbitrary manner with no discernable geometry (Figure 3.16). As with Oct-1, there are also a number of unassociated

trimeric species present, but from a raw particle count from the TEM micrographs, the relative concentration of unassociated trimeric species in Oct-2 appears much higher.

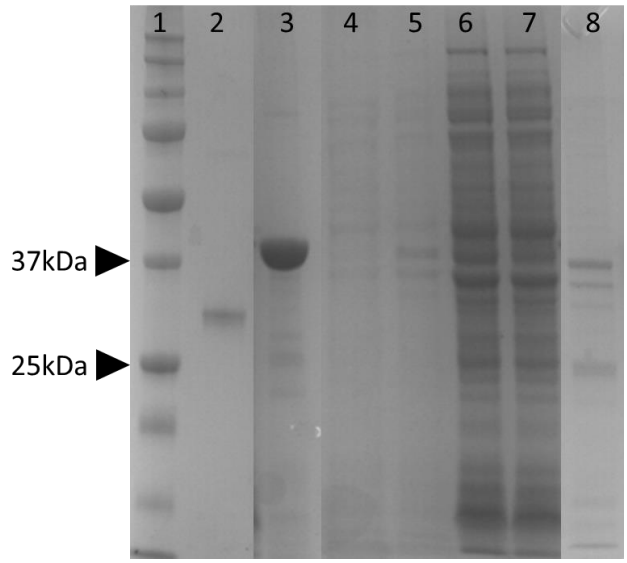


Figure 3.13. SDS-PAGE of Oct-2 purification. Lane 1: ladder. Lane 2: trimeric esterase. Lane 3: Oct-1. Lanes 4 and 5: Oct-2 before and after induction. Lanes 6 and 7: Oct-2 supernatant after lysis and flowthrough after injection on Ni column. Lane 8: concentrated, Ni-purified Oct-2.

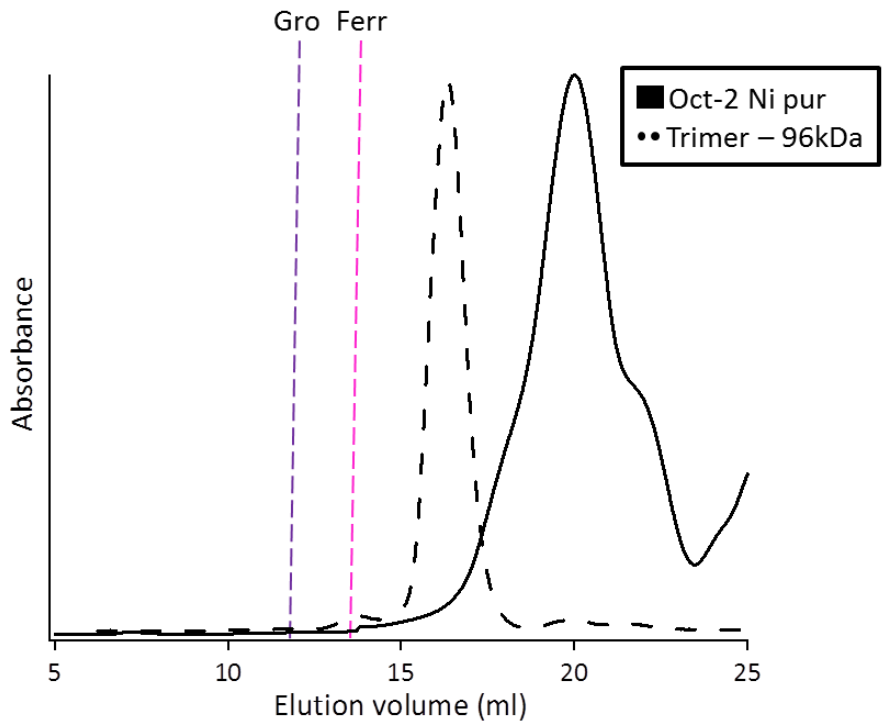


Figure 3.14. Size exclusion chromatography of Oct-2. Ni-purified Oct-2 (solid black) elutes noticeably later than the unmodified trimeric esterase (black dashed). Molecular weight standards GroEL (804 kDa) and ferritin (440 kDa) are purple and pink vertical lines, respectively.

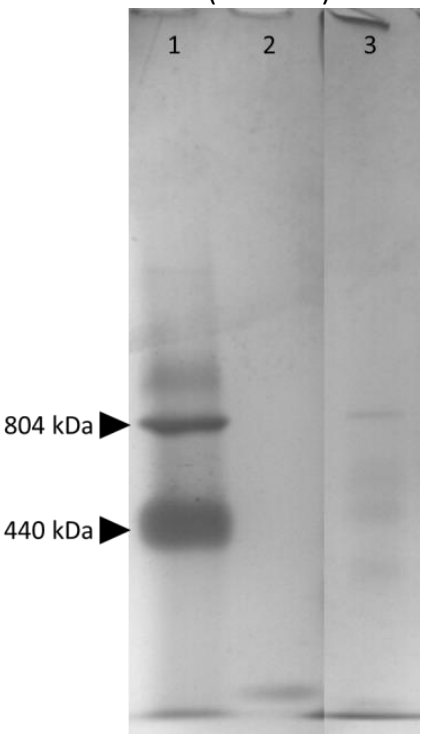


Figure 3.15. Native PAGE of Oct-2. Lane 1: protein standards ferritin and GroEL. Lane 2: unmodified esterase trimer. Lane 3: Ni-purified Oct-2.

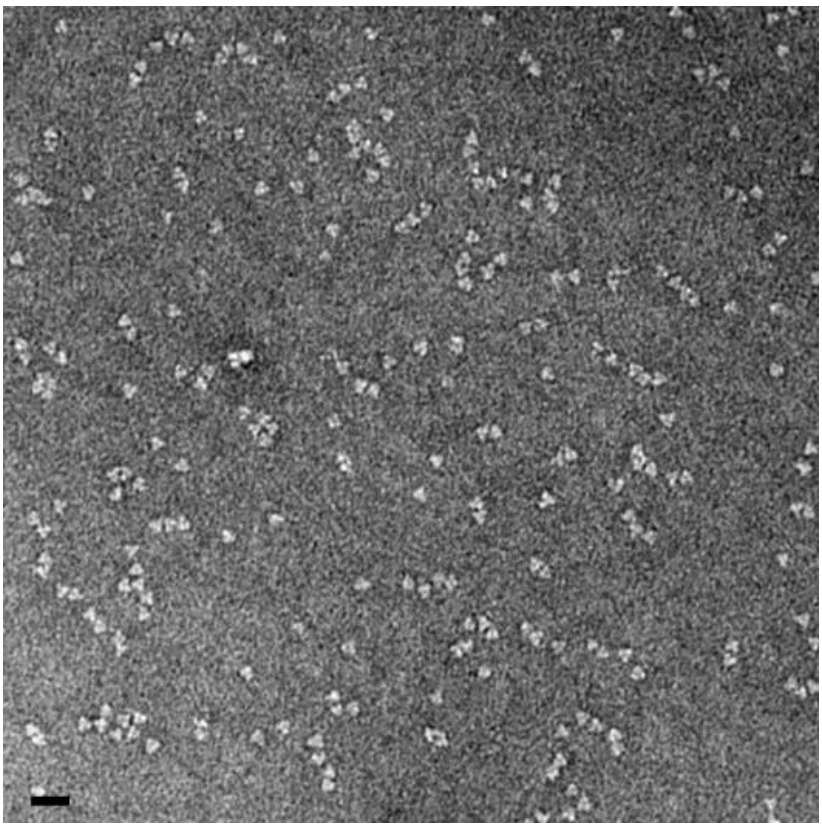


Figure 3.16. Negative stain TEM of Oct-2 shows a heterogeneous mixture of small complexes and unassociated trimers. Scale bar is 20nm.

3.4 - Conclusions

The ultimate goal of this project is to design protein constructs that efficiently assemble into larger, regular complexes in a symmetry-controlled manner. Using the series of characterization techniques described in Chapter 2, I have analyzed two designed fusion proteins. These fusion proteins comprise three parts: the trimeric esterase building block, the flexible linker, and the oligomerizing coiled-coil. The first construct, Oct-1, had a 12-residue linker attaching the trimeric esterase to a weakly-associating tetrameric coiled-coil. Oct-1 purified in fairly high yields as a broad range of species but over a period of weeks re-associated into both aggregate and various small oligomers, with the majority of stable species having

masses corresponding to 2, 4, and 6 trimers. The second construct, Oct-2, retains this long glycine-rich linker but had a more strongly-associating coiled-coil that could plausibly oligomerize into either a dimer or a trimer, depending on environmental conditions. This construct could not be purified in either high yields or high purity, but small, arbitrarily-associating complexes of trimers were visible in TEM images.

The conclusion from the characterization of these two initial constructs is that neither of them oligomerized as intended, regardless of the symmetries of their appended coiled-coil. This indicates that the problem doesn't lie with the identity of the coiled-coil, but rather with one of the other two parts of the fusion protein. Our focus, therefore, turned toward optimization of the flexible linker sequence.

3.5 - References

1. Patterson, D.P. et al. Characterization of a highly flexible self-assembling protein system designed to form nanocages. *Protein Sci.* **23**, 190-199 (2014).
2. Walshaw, J. & Woolfson, D.N. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.* **307**, 1427-50 (2001).
3. Watson, J.D. & Crick, F.H.C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738 (1953).
4. Sodek, J., Hodges, R.S., Smillie, L.B. & Jurasek, L. Amino-Acid Sequence of Rabbit Skeletal Tropomyosin and Its Coiled-Coil Structure. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 3800-3804 (1972).
5. Harbury, P.B., Zhang, T., Kim, P.S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401-7 (1993).
6. Betz, S.F., Bryson, J.W. & DeGrado, W.F. Native-like and structurally characterized designed α-helical bundles. *Curr. Opin. Struct. Biol.* **5**, 457-463 (1995).
7. Myszka, D.G. & Chaiken, I.M. Design and Characterization of an Intramolecular Antiparallel Coiled Coil Peptide. *Biochemistry* **33**, 2363-2372 (1994).
8. Gurnon, D.G., Whitaker, J.A. & Oakley, M.G. Design and Characterization of a Homodimeric Antiparallel Coiled Coil. *J. Am. Chem. Soc.* **125**, 7518-7519 (2003).
9. Apostolovic, B., Danial, M. & Klok, H.-A. Coiled coils: attractive protein folding motifs for the fabrication of self-assembled, responsive and bioactive materials. *Chem. Soc. Rev.* **39**, 3541-3575 (2010).

10. Liu, J., Zheng, Q., Deng, Y., Kallenbach, N.R. & Lu, M. Conformational Transition between Four and Five-stranded Phenylalanine Zippers Determined by a Local Packing Interaction. *J. Mol. Biol.* **361**, 168-179 (2006).
11. Fletcher, J.M. et al. A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth. Biol.* **1**, 240-250 (2012).
12. Lovejoy, B. et al. Crystal structure of a synthetic triple-stranded alpha-helical bundle. *Science* **259**, 1288-1293 (1993).
13. Betz, S., Fairman, R., O'Neil, K., Lear, J. & Degrado, W. Design of Two-Stranded and Three-Stranded Coiled-Coil Peptides. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **348**, 81-88 (1995).
14. DÜRR, E. & BOSSHARD, H.R. Folding of a three-stranded coiled coil. *PRS* **9**, 1410-1415 (2000).

Chapter 4

Optimization of Linker Length and Purification of Oct-3

4.1 - Optimization of Linker Length

In the previous chapter I examined two fusion protein constructs consisting of a trimeric esterase bound by a 12 residue glycine-rich linker to one of two coiled-coils: a weakly-associating tetrameric coiled coil and a strongly-associating coiled-coil that may be either dimeric or trimeric. Both of these constructs formed a heterogeneous mixture of complexes, the majority of which were smaller than an octahedron. This suggests the possibility that multiple coils on a single trimer may be associating with each other, which would dramatically lower the number of trimers necessary to create a closed system in which every coil oligomerizes correctly. Therefore, we undertook an effort to minimize the glycine linker length and observe the effects on oligomerization.

A rough calculation indicates that a twelve residue linker is sufficiently long to bridge the gap between two C-termini of the esterase (Figure 4.1). To prevent trimer self-association, the length of the linker needs to be less than half of the distance between two C-termini. While it would be possible to systematically decrease the linker length until a suitable length is determined, it is far easier, and in line with the this project's goals of designing as few constructs as possible, to approach this problem computationally and model the approximate minimal distance it would take to bridge the two symmetric subunits when assembled into an octahedron. As noted in Chapter 1, this type of modeling has already been applied with

significant success by the Yeates and Baker labs to generate crystallographically-verified complexes of tetrahedrons and octahedrons.[1-3]

We collaborated with Neil King in the Baker lab to use Rosetta to generate a model containing the trimeric and tetrameric subunits replicated at their respective locations in octahedral space, similar in design to Ref. 1. Crystal structures of these subunits were each independently rotated and radially translated by 0.5° and 1 Å, discarding all models with steric clashes, as defined by having 2 backbone atoms from separate subunits within 4 Å of one another (Figure 4.2). We then collected the distance between the C-terminus of the trimeric esterase and the N-terminus of the tetrameric coiled-coil of every model without steric clashes, and the 20 models with the lowest interterminus distance were saved as coordinates files. The interterminus distances of these models ranged from 9.1 to 17 Å. As an amino acid generally spans around 3.5 Å, this implies the gap between subunits could be bridged by a minimum of 3 residues.

4.2 - Design of Oct-3 Constructs

Based on this modeling, we designed three new fusion proteins constructs of the trimeric esterase fused with the coil in Oct-2 by three different lengths of linkers consisting of 3, 4, and 5 residues long. Although the residue that would add the most degrees of freedom to the fusion protein is glycine, the restriction site that was added to Oct-1 to allow for genetic modification encodes for a glycine and a threonine. Therefore, the three linkers we created had amino acid sequences: Esterase-GTG-Coil, Esterase-GTGG-Coil, and Esterase-GTGGG-Coil, which we termed Oct-3-3, Oct-3-4, and Oct-3-5, the second number referring to the length of the glycine-rich linker (Figure 4.3). While linkers with a threonine in them will have lower flexibility

due to the restricted range of Ramachandran angles compared to glycine, this may prove not to matter too much in practice. A more central concern is that the linker disrupts the alpha helices at the C-terminus of the esterase and the N-terminus of the coil, which would lower the effective length of adjacent residues in the linker. Since there are glycines bracketing the threonine in these designed linkers, the threonine's presence should not be a confounding factor. Optimizing this linker sequence may be a potential avenue of further research.

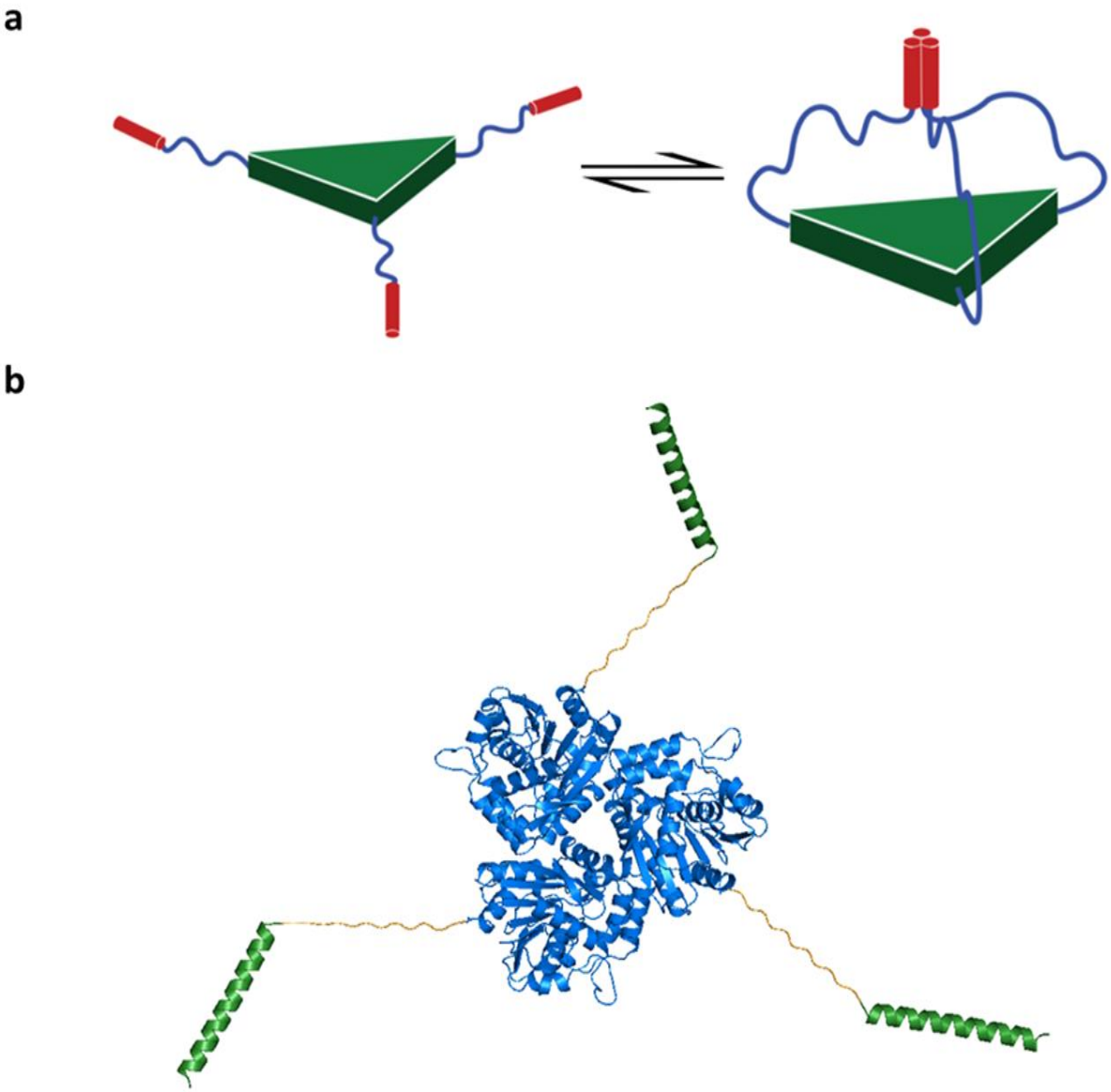


Figure 4.1. Potential problems with a long flexible linker. a) A C_3+C_3 symmetry pair with a long enough linker can stably self-assemble into a single trimer. b) A model of a single trimeric subunit of Oct-1. The length of the 12 residue linker from Oct-1 and Oct-2 is roughly equivalent to the length from one C-terminus on an esterase subunit to another, indicating that some degree of trimer self-association is likely occurring in previous constructs.

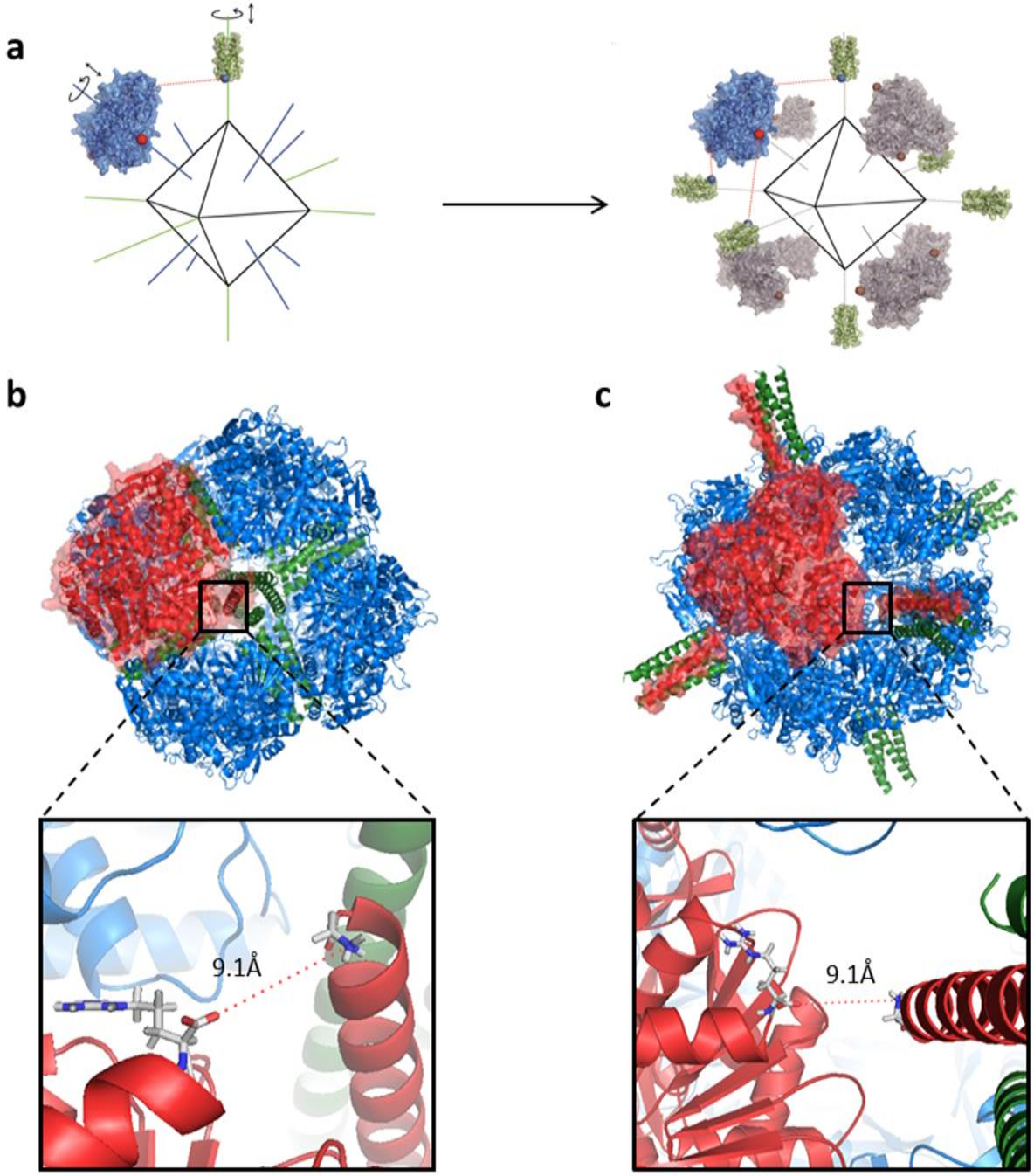


Figure 4.2. Symmetry-constrained model of minimum interterminus distances. a) Crystal structures of the trimeric esterase (blue) and tetrameric coil (green) were oriented along their individual symmetry axes. Each symmetry element was independently rotated and then radially translated, after which they were replicated to octahedral space. Models without steric hindrance had the interterminus distances measured. Two models, one with the coils on the inside (b) and one with the coils on the outside (c) both had the minimum interterminus distance of 9.1 Å.

However, the coil implemented in Oct-3 is not a tetramer-forming coil but rather a coil that may form either a dimer or a trimer, or potentially both. This means there will be fewer trimers crowded around a single coiled-coil, so we can anticipate that the minimum linker length to prevent steric effects will be shorter than the model predicted. Since the C-terminus of the esterase is located perpendicular to its C_3 axis, structures with a larger dihedral angle such as tetrahedrons and trigonal prisms may still be sterically precluded from forming.

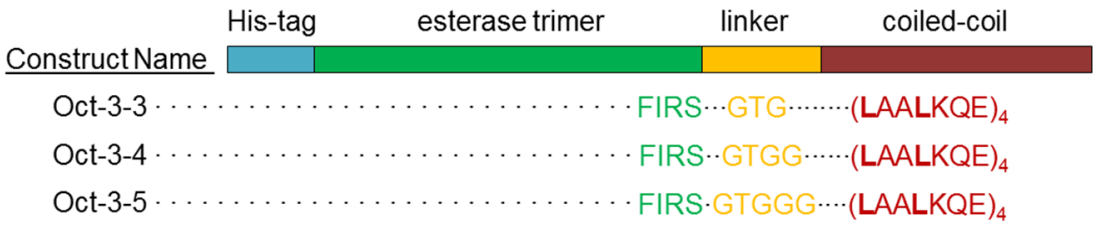


Figure 4.3. Design of the three Oct-3 fusion proteins. The 6xHis-tag is located at the N-terminus. Bolded residues represent oligomerization-determining a and d positions.

4.3 - Purification of Oct-3 Constructs

All three Oct-3 proteins could be transformed and expressed by methods previously discussed. Ni-affinity purification yielded 1-3 mg of impure protein for each construct, and all other protein contaminants detectable by SDS-PAGE could be removed by size exclusion chromatography (Figure 4.4). Size exclusion chromatography of the Ni-purified constructs showed that all three constructs had well-defined peaks eluted near molecular weight standards GroEL and ferritin, indicating that these constructs are forming small complexes of trimers. While there are also peaks in the Ni-purified samples that correspond to smaller complexes and larger complexes, these peaks are minor and distinct from the peaks of interest, and can be removed by SEC purification. After SEC purification, all Oct-3 complexes are stable in solution at 4 °C for months, and purified complexes do not re-oligomerize into larger or smaller

species. This is in direct contrast to the behavior exhibited by Oct-1, where a SEC column fraction re-oligomerizes into a broad range of complexes just hours after purification.

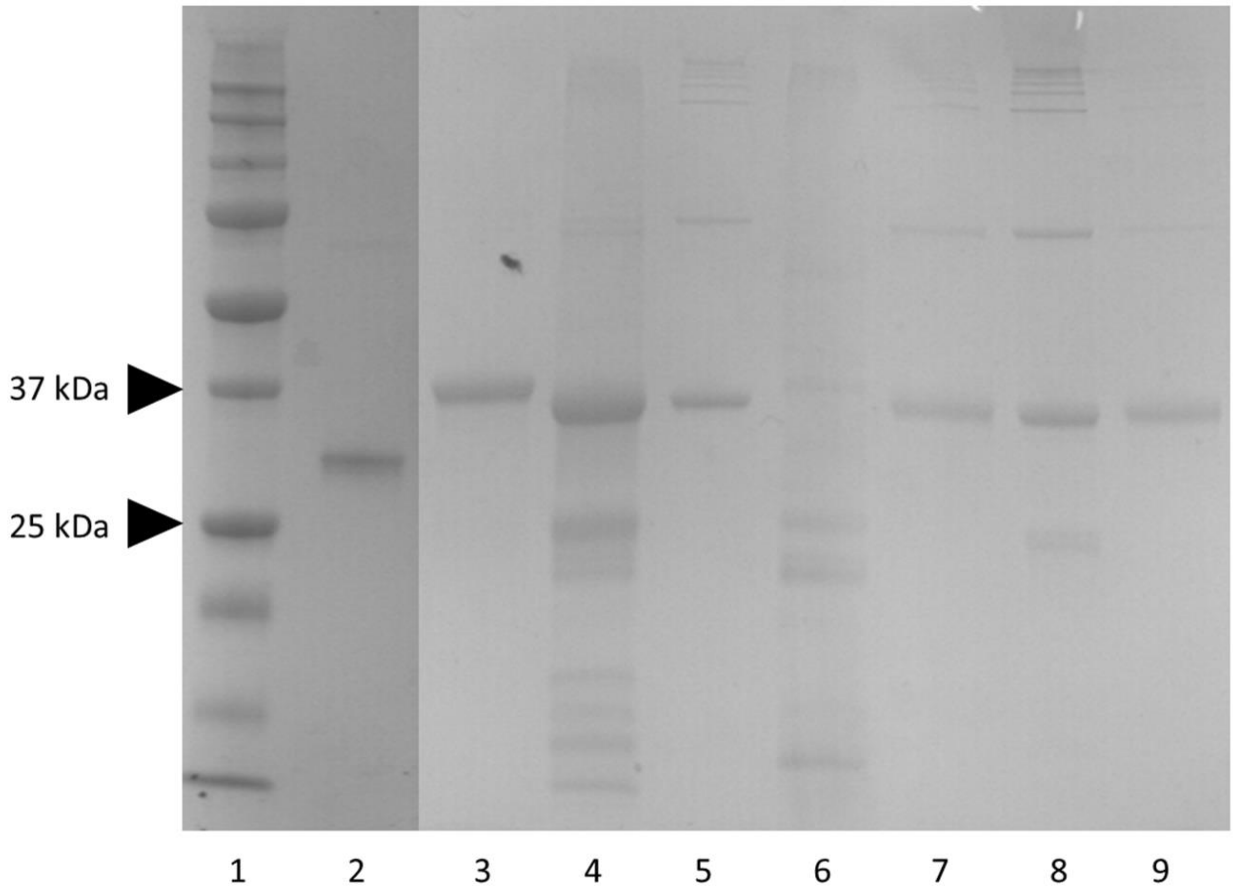


Figure 4.4. SDS-PAGE analysis of Oct-3 constructs. Lane 1: ladder. Lane 2: trimeric esterase. Lane 3: Oct-1 after SEC purification. Lanes 4, 5: Oct-3-3 after Ni and SEC purification, respectively. Lanes 6, 7: Oct-3-4 after Ni and SEC purification, respectively. Lanes 8, 9: Oct-3-5 after Ni and SEC purification, respectively.

4.4 - Size Exclusion Chromatography and Native PAGE of Oct-3 Constructs

Size exclusion chromatography analysis of SEC-purified Oct-3 complexes shows that Oct-3-4 and Oct-3-5 have very similar elution profiles, both with fairly sharp peaks and elution volumes of 12.7 and 12.5 ml, respectively (Figure 4.8). SEC-purified Oct-3-3 elutes with a comparatively broader peak with an elution volume of 11.85 ml and a shoulder peak at 13.5 ml.

None of these constructs exhibit absorbance in the range expected for a free trimer, which is a promising sign.

Native PAGE offers a higher resolution window into the relative concentration and sizes of each of the complexes formed by the three Oct-3 constructs. Correlative with the results from size exclusion chromatography, Oct-3-4 and Oct-3-5 appear to comprise the same three distinct complexes that are barely resolved (Figure 4.9). The relative concentration of these three bands differs between constructs, however. In Oct-3-4, the band correlating to the smallest species appears to be the highest relative concentration, whereas in Oct-3-5 the reverse is apparent. The native PAGE of Oct-3-3, on the other hand, shows a much broader distribution of complex size. The lowest band in Oct-3-3, which migrates with the same R_f as the lowest band in Oct-3-4 and Oct-3-5, also appears to be in the highest concentration. This is interesting, because the SEC elution profile of Oct-3-3 indicates that in sum the larger species are more prevalent. The lack of smearing in all three constructs additionally shows that there is very little interconversion between different complexes, confirming that these species are well-formed and thermodynamically stable.

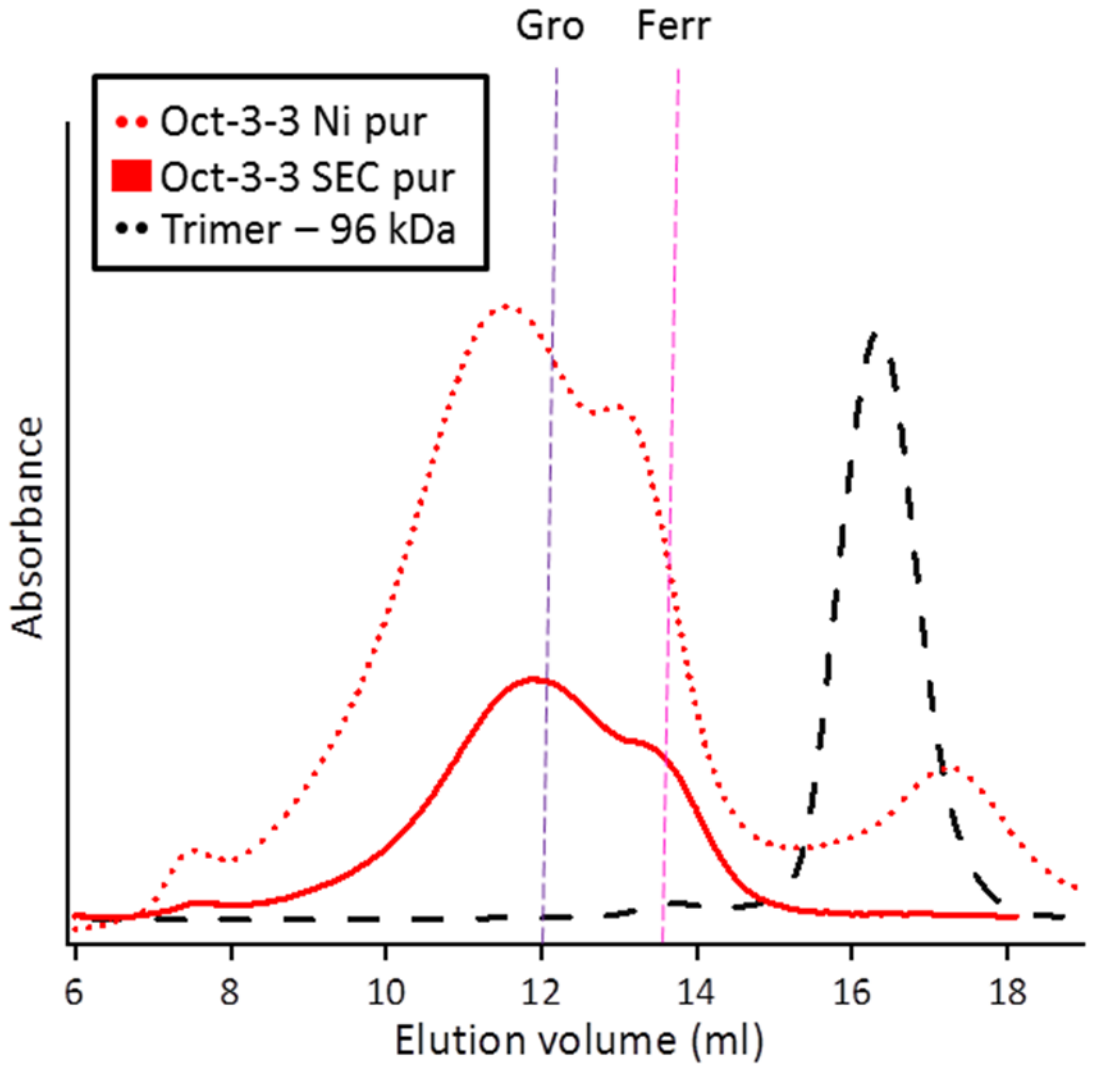


Figure 4.5. Size exclusion profiles of Ni- and SEC-purified Oct-3-3. The two side peaks present in Ni-purified Oct-3-3 are removed after size exclusion purification. Standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

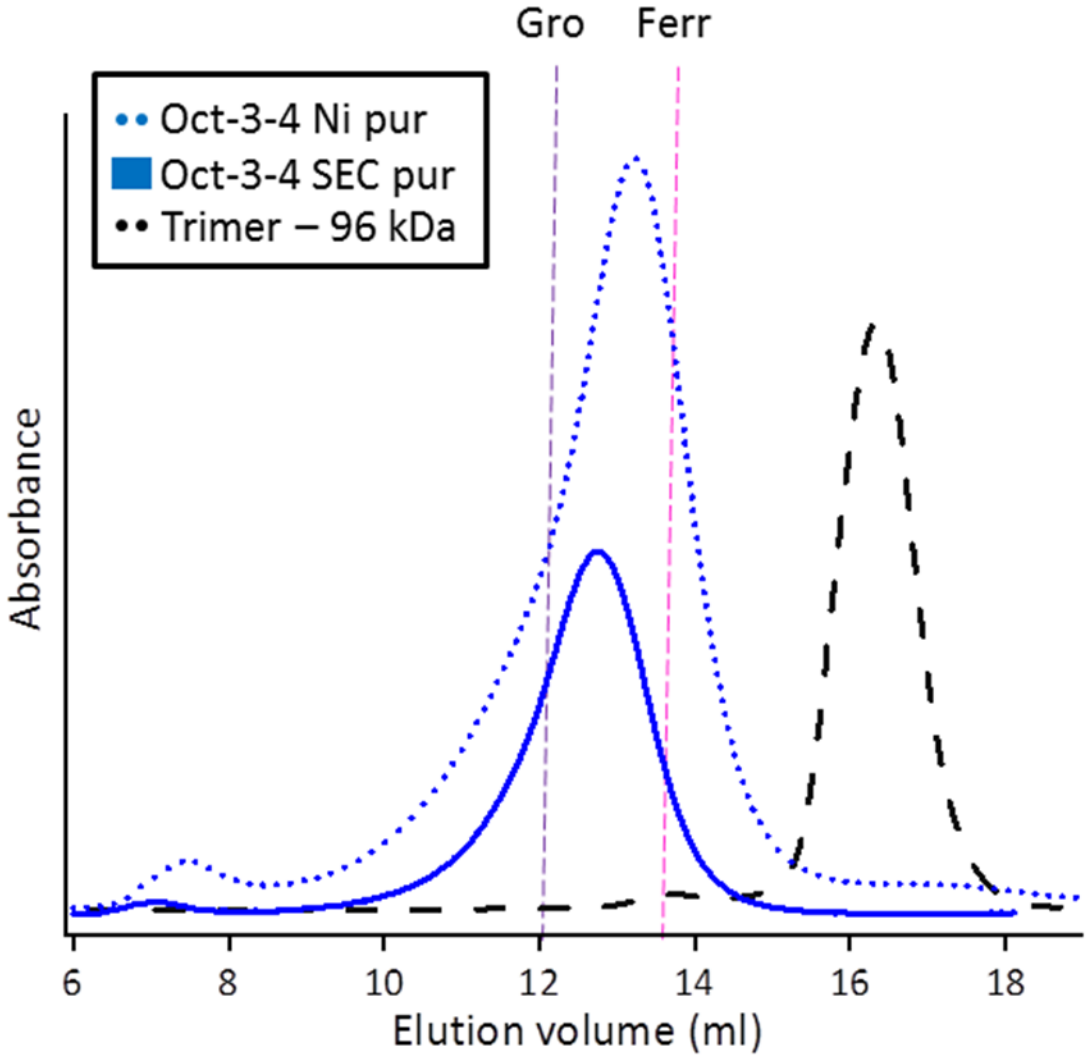


Figure 4.6. Size exclusion profiles of Ni- and SEC-purified Oct-3-4. The two side peaks present in Ni-purified Oct-3-4 are removed after size exclusion purification. Standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

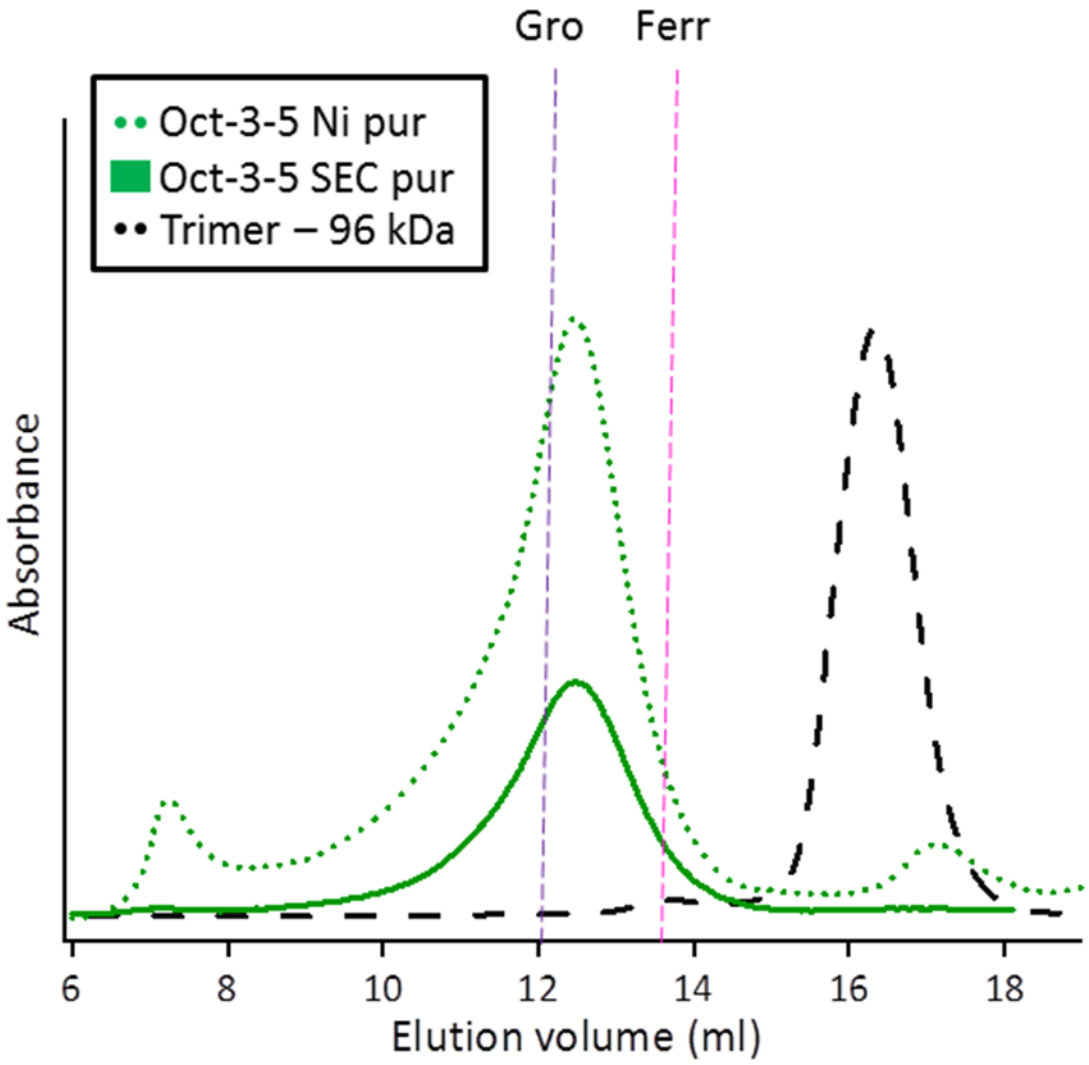


Figure 4.7. Size exclusion profiles of Ni- and SEC-purified Oct-3-5. The two side peaks present in Ni-purified Oct-3-5 are removed after size exclusion purification. Standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

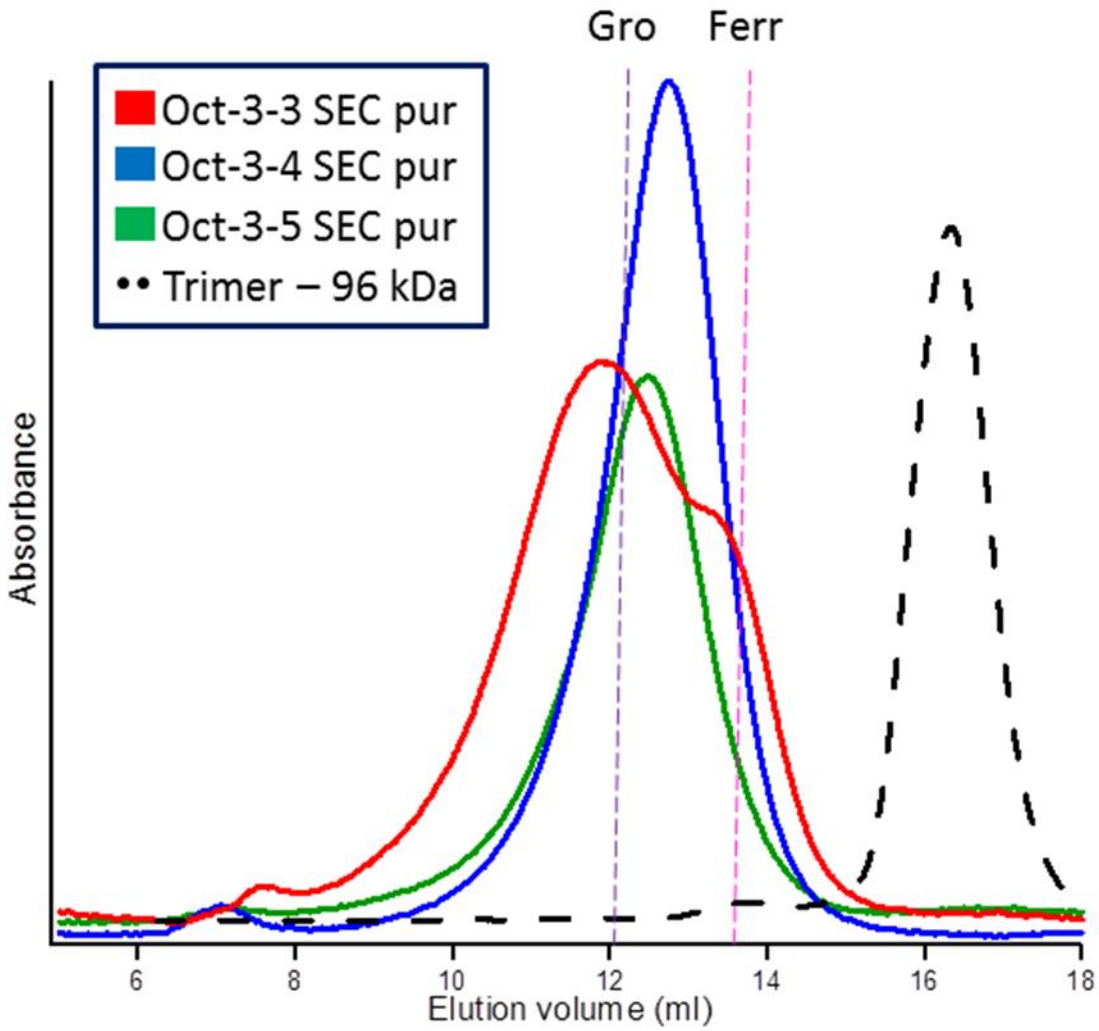


Figure 4.8. SEC profiles of Oct-3 constructs after SEC purification. Oct-3-4 (blue) and Oct-3-5 (green) assemble into complexes with a similar size. Oct-3-3 (red) assembles into species with a wider range of sizes. None of these constructs form complexes with sizes close to the unmodified trimeric esterase (black dashed). Elution volumes of standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

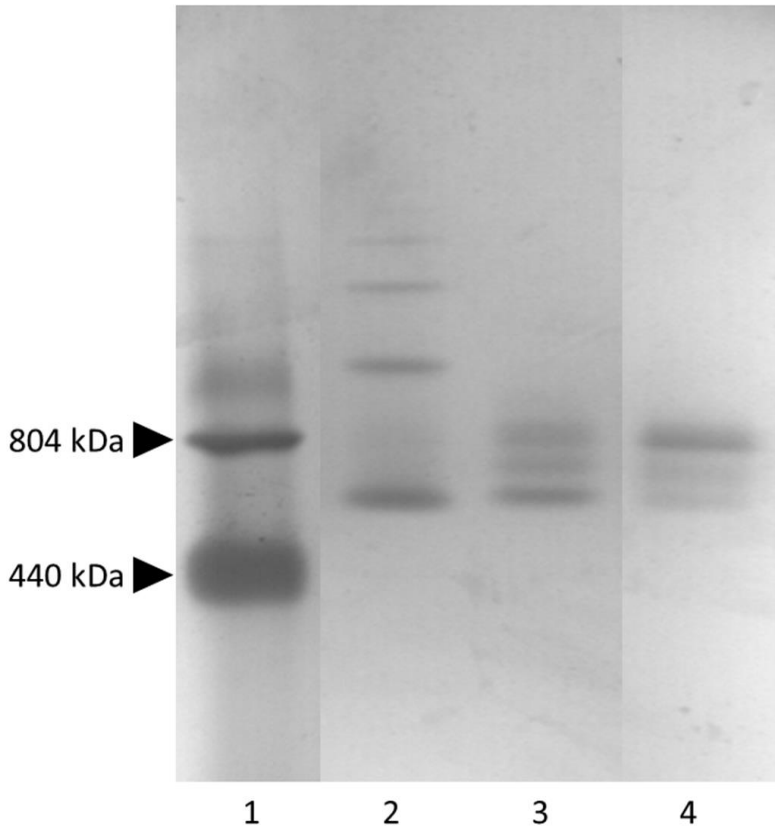


Figure 4.9. Native PAGE of SEC-purified Oct-3 complexes. Lane 1: Ladder comprised of standard proteins GroEL (top band) and ferritin (bottom). Lanes 2, 3, and 4: SEC-purified Oct-3-3, Oct-3-4, and Oct-3-5, respectively. All three constructs show multiple complexes in the region expected for a tetrahedron or an octahedron, with Oct-3-4 and Oct-3-5 behaving similarly, and Oct-3-3 having a comparatively broader range of complexes formed.

4.5 - Analytical Ultracentrifugation of Oct-3 Constructs

For a closer look at the individual complexes formed from each construct, SEC-purified samples of Oct-3 complexes were analyzed by analytical ultracentrifugation. Sedfit analysis of SV-AUC data collected from SEC-purified Oct-3 constructs revealed a distribution of complexes that correlated well with native PAGE (Figure 4.11). Oct-3-3 exists as an assortment of complexes with a broad range of s-values from 9.8 to 25, with the most prevalent species also being the smallest. Using a fitted frictional ratio of 1.29, this species at 9.8 *S* has a molecular

weight of 476 kDa, around the expected mass of a tetrahedron. A peak at 9.8 *S* is also found to a lesser degree in Oct-3-4 and possibly Oct-3-5, matching with results from the native gel. In Oct-3-4 there appear to be three major species with similar concentrations and a tight distribution of s-values from 9.8 to 12.8. This species with an s-value of 12.8 is also the only major peak in Oct-3-5. Fitting the frictional ratio for the Oct-3-5 dataset yields a frictional ratio of 1.5; using this value we can calculate a molecular weight of 860 kDa for the major species at 12.8 *S*, very close to the expected molecular mass of an octahedron. However, it should be noted that these are fitted according to a least-squares algorithm, and may not reflect true molecular weights. For example, if the frictional ratio is 1.3, the species with an *s*-value of 12.8 would have a molecular weight of 675 kDa, or approximately six trimers. In theory, it follows that if a globular species is hollow, like is expected from a well-formed symmetrical protein cage, the size of the hollow interior should grow faster than the molecular mass. As such, it makes sense that larger well-formed protein cages would have higher frictional ratios. We can use this relation to identify large complexes with a low (near unity) frictional ratio as being misfolded and collapsed, but it is of little use to validate the frictional ratios and molecular masses of specific peaks.

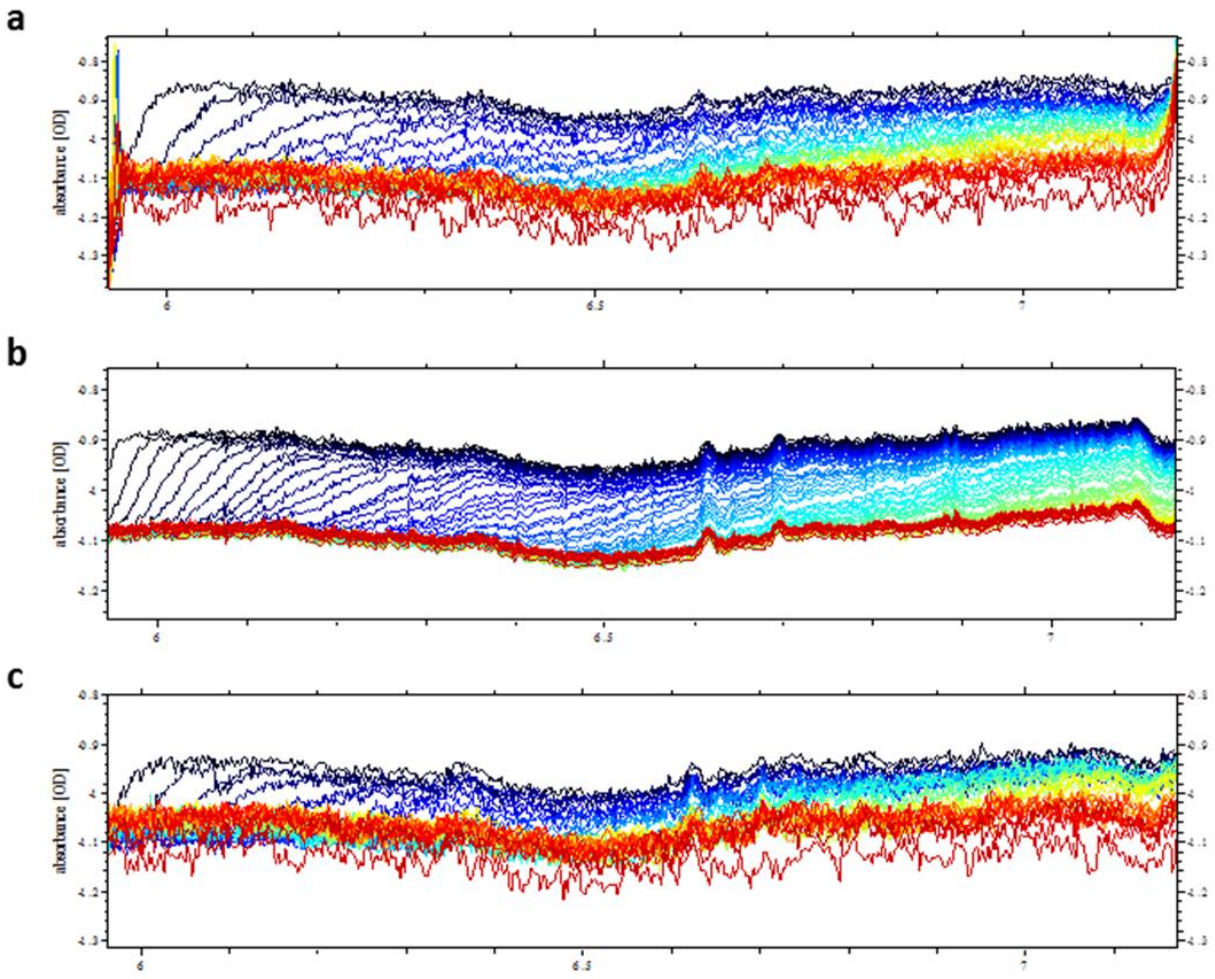


Figure 4.10. Raw sedimentation velocity-AUC data for Oct-3-3 (a), Oct-3-4 (b), and Oct-3-5 (c). Violet traces represent the first scan, Red traces represent the final scan.

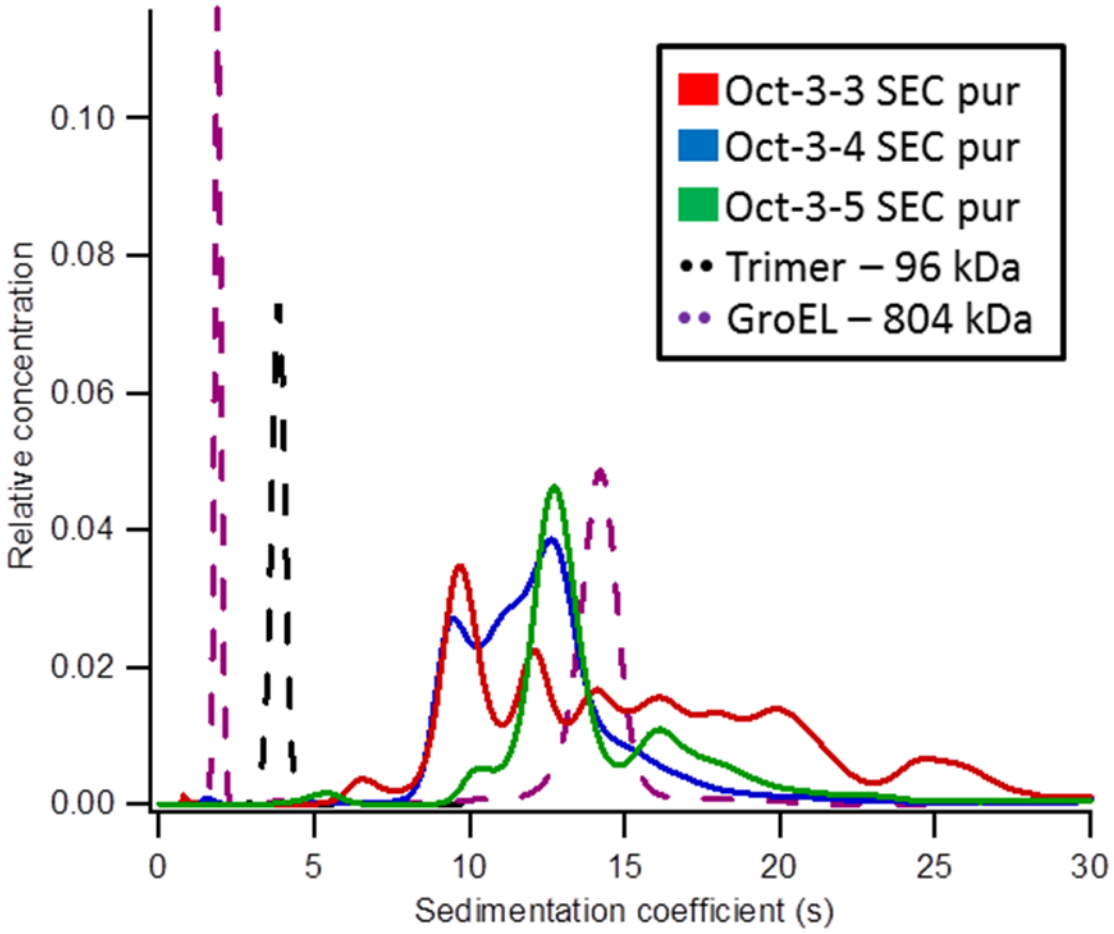
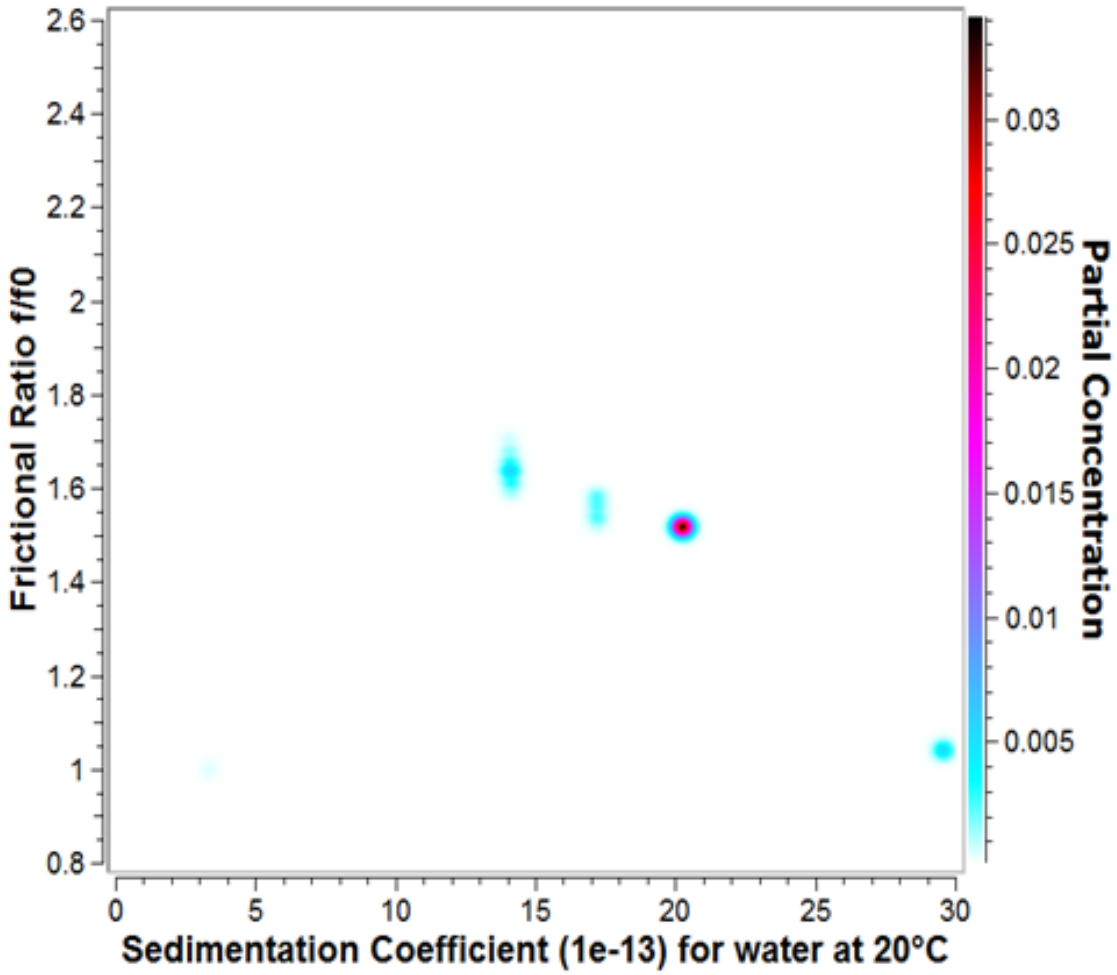


Figure 4.11. Sedimentation traces of Oct-3 constructs analyzed with sedfit**.** Oct-3-3 (red), Oct-3-4 (blue), and Oct-3-5 (green) each have multiple peaks with a distribution that correlates to native PAGE analysis. The major peaks for all three fusion proteins are between the s-values of the unmodified trimer (black dashed) and GroEL (purple dashed).

4.6 - 2-Dimensional Sedimentation Analysis of Oct-3-4

In an attempt to gain a fuller understanding of the characteristics of specific complexes formed by these three constructs, we analyzed the sedimentation data from the Oct-3-4 construct with Ultrascan. Analysis of the data from Oct-3-4 by Ultrascan dropped the r.m.s.d. of the fit from 0.12 to 0.027, a dramatic increase in resolution (Figure 4.12). Five distinct species

were isolated in total, three major and two minor species. The two minor species have molecular weights of 281 and 786 kDa, and in total account for a tenth of the total concentration. Both of these species have frictional ratios near 1, indicating that these species are collapsed and do not maintain a hollow interior. The three major species exist in similar concentrations and have $s_{20,w}$-values of 14.08, 17.1, and 20.02, with frictional ratios between 1.63 and 1.51, indicating that these species are either elongated or have a hollow interior. These three species calculate to molecular weights of 510, 633, and 783 kDa, fairly close to the molecular masses expected from species of five (537 kDa), six (645 kDa), and seven (752 kDa) trimers. Conspicuously absent from this analysis is the 4-trimer tetrahedron, which likely indicates that none of these linkers are sufficiently long to overcome the steric hindrance of forming a tetrahedron. The presence of a 5-trimer species is also a curiosity, as even with a C_3+C_3 symmetry pair, this species requires extremely flexible geometry to place every alpha helix into a trimeric coiled-coil (see Figure 1.4b). This points to the **a,d** = L,L alpha helix used in Oct-3 being either stable as an unbound monomer or stable as a dimeric coiled-coil, while primarily oligomerizing as a trimeric coiled-coil. Both cases are viewed as possible outcomes, but due to the amphiphilic nature of the coil, it is more likely that this coil dimerizes than exists as a random-coil monomer.



Oct-3-4	$s_{20,w}$	M.W. (kDa)	f/f_0	Conc. %
Species 1	3.31 ± 0.08	281 ± 14	1.00 ± 0.01	3.7%
Species 2	14.08 ± 0.04	510 ± 23	1.63 ± 0.05	23.5%
Species 3	17.1 ± 0.04	633 ± 19	1.55 ± 0.03	35.5%
Species 4	20.02 ± 0.03	783 ± 10	1.51 ± 0.02	30.6%
Species 5	29.5 ± 0.07	786 ± 16	1.04 ± 0.02	6.4%

Figure 4.12. 2D-sedimentation analysis of Oct-3-4 by Ultrascan. Three major and two minor species were isolated, the characteristics of which are tabulated above. Species 2, 3, and 4 correspond to well-formed, hollow complexes with molecular masses consistent with those predicted for 5, 6, and 7 trimers, while species 5 corresponds to a complex of 7 trimers that is misformed and lacks a hollow interior.

4.7 - Transmission Electron Microscopy of Oct-3-4

Finally, negative stain TEM images were collected for SEC-purified Oct-3-4. These showed a heterogeneous mixture of mostly globular structures around 15-25 nm in diameter (Figure 4.13). In some of these structures, there is an apparent C_3 axis, while in other structures, single trimers appear to be attached to the complex only at the vertex of a single trimer, with the other two oligomerizing vertices are exposed to solution. The presence of these exposed single trimers may indicate that the coil attached to Oct-3 is stable as a monomer in solution, but could also be caused by the protein cages collapsing upon exposure to negative stain.

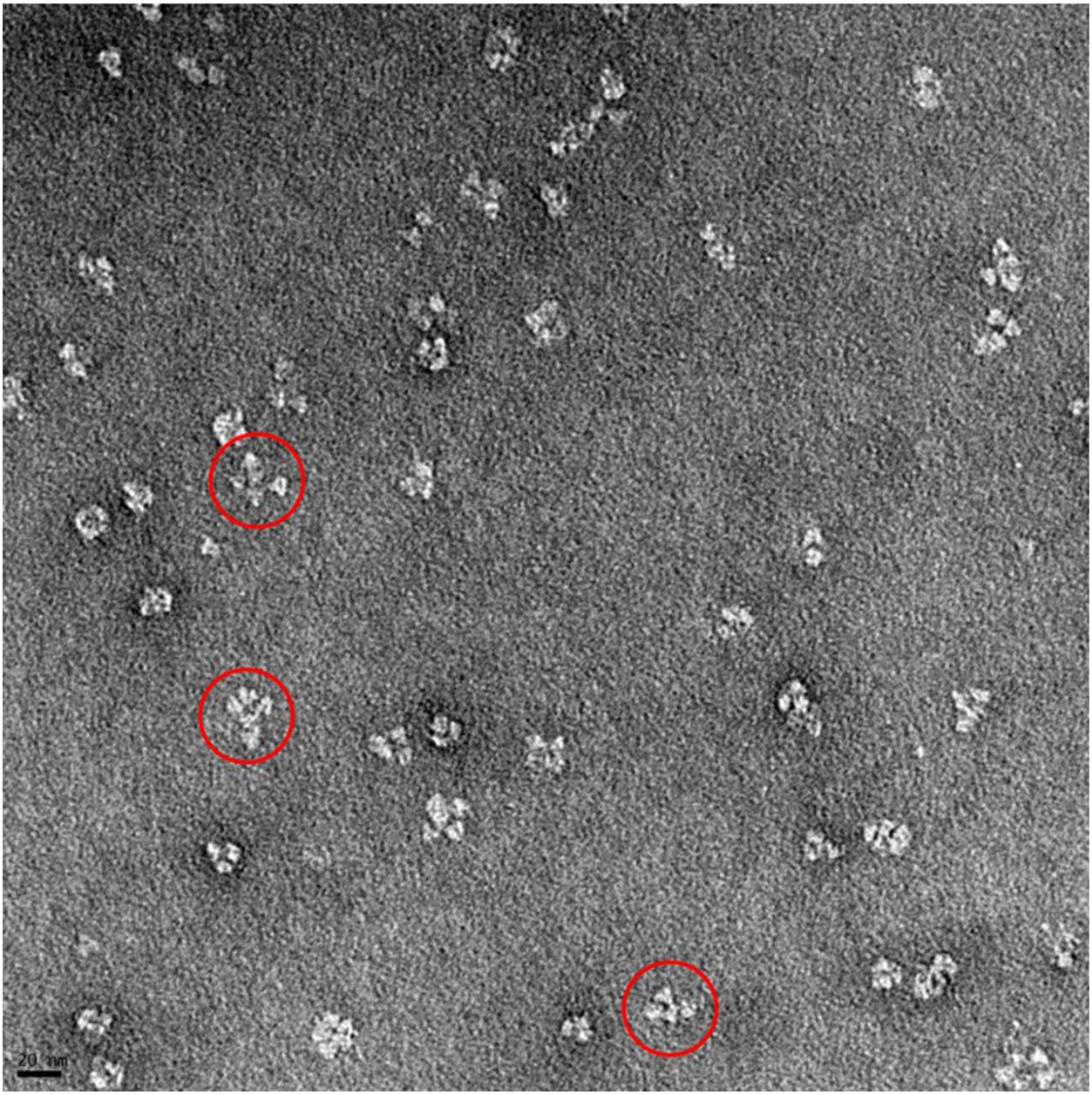


Figure 4.13. Transmission electron micrographs of Oct-3-4. Observed particles have sizes ranging from 15-25 nm in diameter and are mainly globular. Geometry on most species is indistinct, but with several cases a C_3 symmetry axis is apparent. Several complexes (red circles) appear to have an isolated trimer attached at only one or two vertices. Scale bar is 20 nm.

4.8 - Conclusions

Following up on the work done in Chapter 3 where two trimer-coil systems with 15 residue linkers were shown to inadequately oligomerize, I designed in collaboration a modeling

algorithm of the octahedron that minimized the distances between the termini of the two oligomerizing groups without steric clashes. This model yielded a minimum interterminus distance of 9.1 Å, which was used to design fusion proteins with 3, 4, and 5 residues in their linkers. These fusion proteins all oligomerized into multiple complexes in the size range expected for an octahedron, and while aggregates were present, there were no detectable free trimers in any of the samples. Complexes in the appropriate size range could be isolated by size exclusion chromatography, and after purification each fusion protein contained at least 3 distinct species. Analytical ultracentrifugation identified the smallest complex formed by all three fusion proteins as a species consisting of 5 trimers, which theoretically requires a highly flexible system. Since this system is designed with low flexibility, this points to the possibility of the coiled-coil in Oct-3 oligomerizing into multiple oligomerization states. While this may be worth looking into in more detail, such a line of research is outside of the scope of this project, and so the next chapter will explore the effects of replacing the *a,d* = L,L coil with the crystallographically verified tetrameric coil where the *a* and *d* positions on the heptad repeat are comprised of leucine and isoleucine, respectively.

4.9 - References

1. King, N.P. et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103-+ (2014).
2. King, N. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **6085**, 1171-1174 (2012).
3. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* **6**, 1065-1071 (2014).

Chapter 5

Purification of Oct-4 and Visualization of the Octahedron

5.1 - Design of Oct-4 Constructs

Following the realization that instead of the intended, crystallographically verified tetrameric coiled-coil, we had accidentally attached to the trimeric esterase a coiled-coil that may exist as a mixture of dimers and trimers, we redesigned the DNA insert to replace all four leucine residues at the *d* positions in the coiled-coil heptad with isoleucines. We designed three constructs, with 2, 3, and 4 residues in the linker, as it had occurred to us that it would be interesting to investigate the effect of making a fusion protein construct with a linker that was shorter than the predicted minimum inter-terminus distance. These constructs had the linker sequences: Esterase-GT-Coil, Esterase-GTG-Coil, and Esterase-GTGG-Coil, and were named Oct-4-2, Oct-4-3, and Oct-4-4, respectively (Figure 5.1). As noted in section 4.1, the linker sequence of Oct-4-2 may have an intrinsic problem: there is no glycine separating the threonine residue in the linker from the coiled-coil, so this threonine residue may adopt an alpha helical conformation. This would reduce both the effective length of the linker and potentially the range of allowable dihedral angles between the two oligomerizing subunits. However, Oct-4-2 is *designed* to not oligomerize correctly, as the two residues in its linker would not be long enough in any conformation to bridge the theoretical minimum inter-terminus distance. Therefore, regardless of the degree of lost flexibility in the threonine, Oct-4-2 should be expected to oligomerize in an irregular pattern.

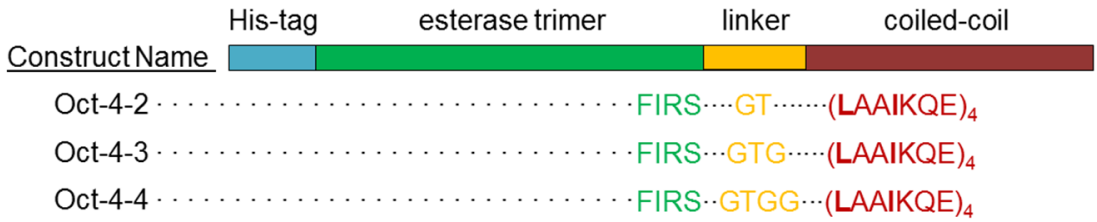


Figure 5.1. Design of the three Oct-4 fusion protein constructs. The 6xHis-tag is located at the N-terminus. Bolded residues represent oligomerization-determining *a* and *d* positions.

5.2 - Expression and Purification of Oct-4 Constructs

All three Oct-4 constructs could be transformed and expressed by methods described in Chapter 2. After lysis and centrifugation, Oct-4-2 and Oct-4-4 were found primarily in the insoluble fraction, but 1-3 mg/L of soluble fusion protein could be isolated after Ni-affinity purification. In contrast, only a miniscule amount of Oct-4-3 was present in the cell supernatant after lysis and centrifugation, and yields from Ni-affinity purification were not high enough to further purify by size exclusion. This is an unexpected result, and currently there is no definitive explanation for this behavior. Further purification of Oct-4-2 and Oct-4-4 by size exclusion chromatography removed all remaining protein impurities detectable by SDS-PAGE (Figure 5.2).

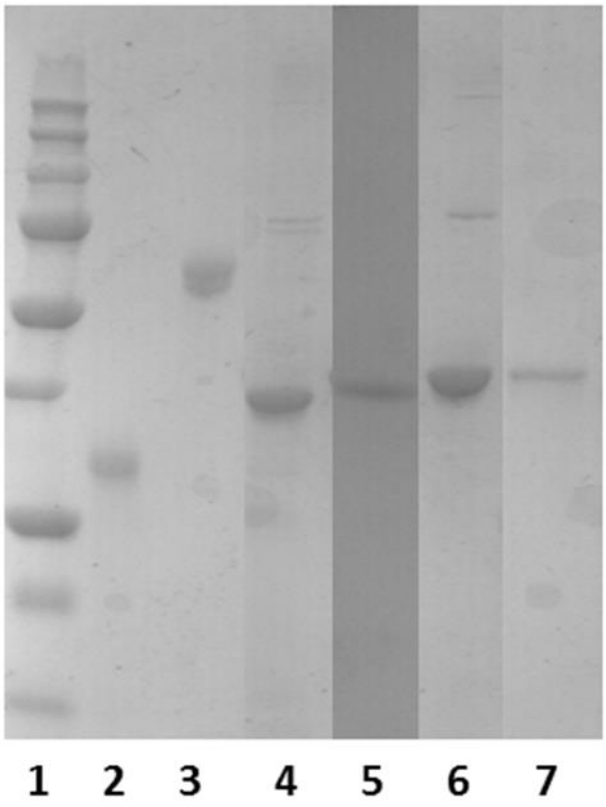


Figure 5.2. SDS-PAGE of Oct-4 constructs. Lane 1: ladder. Lane 2: trimeric esterase. Lane 3: GroEL. Lane 4: Oct-4-2 after Ni purification. Lane 5: Oct-4-3 after Ni purification. Lane 6: Oct-4-4 after Ni purification. Lane 7: Oct-4-4 after SEC purification.

5.3 - Size Exclusion Chromatography and Native PAGE of Oct-4 Constructs

After nickel affinity purification, the SEC elution profiles of the three Oct-4 constructs display features similar to the elution profiles of the Oct-3 constructs: a peak in roughly the same region as the Oct-3 constructs and molecular weight standards GroEL and ferritin, bracketed by a peak in the void volume and a peak that elutes later than the unmodified trimeric esterase (Figure 5.3). Size exclusion purification of both Oct-4-2 and Oct-4-4 was able to remove both of these contaminant peaks. Oct-4-3 and Oct-4-4 have similar elution profiles and run identically on a native PAGE, the only difference being the reduced yield of Oct-4-3. This elution profile is characterized by a sharp peak at 12.1 ml, with the front of the peak being

broader than the tail (Figure 5.5). This is promising, as it indicates a sharp lower bound for the size of the complexes formed. It is also the sharpest peak of any of the constructs purified thus far. Oct-4-2, on the other hand, has a broader peak centered around 10.5-11 ml (Figure 5.4). Both sides of this peak are equally broad, indicating that Oct-4-2 forms a range of complexes with minor species that are both smaller and larger than the major species. Native PAGE of the Oct-4 complexes confirms these observations (Figure 5.6). Oct-4-2 electrophoreses as at least five distinct major bands with a number of minor bands also present, both smaller and larger than the major bands. All of these bands, excepting a few minor species, have a much lower R_f than standard protein GroEL. Oct-4-3 and Oct-4-4 have electrophoresis profiles similar to Oct-3-3: the band with the highest intensity is also the smallest species on the gel, with 4-5 minor bands corresponding to larger species. The major species in Oct-4-3 and Oct-4-4 has a slightly lower R_f than GroEL, which is also promising as the octahedron is expected to be slightly heavier than GroEL. Since the major species in Oct-4-4 is also the smallest, we attempted to selectively purify this species by collecting and combining fractions from the tail end of the size exclusion peak.

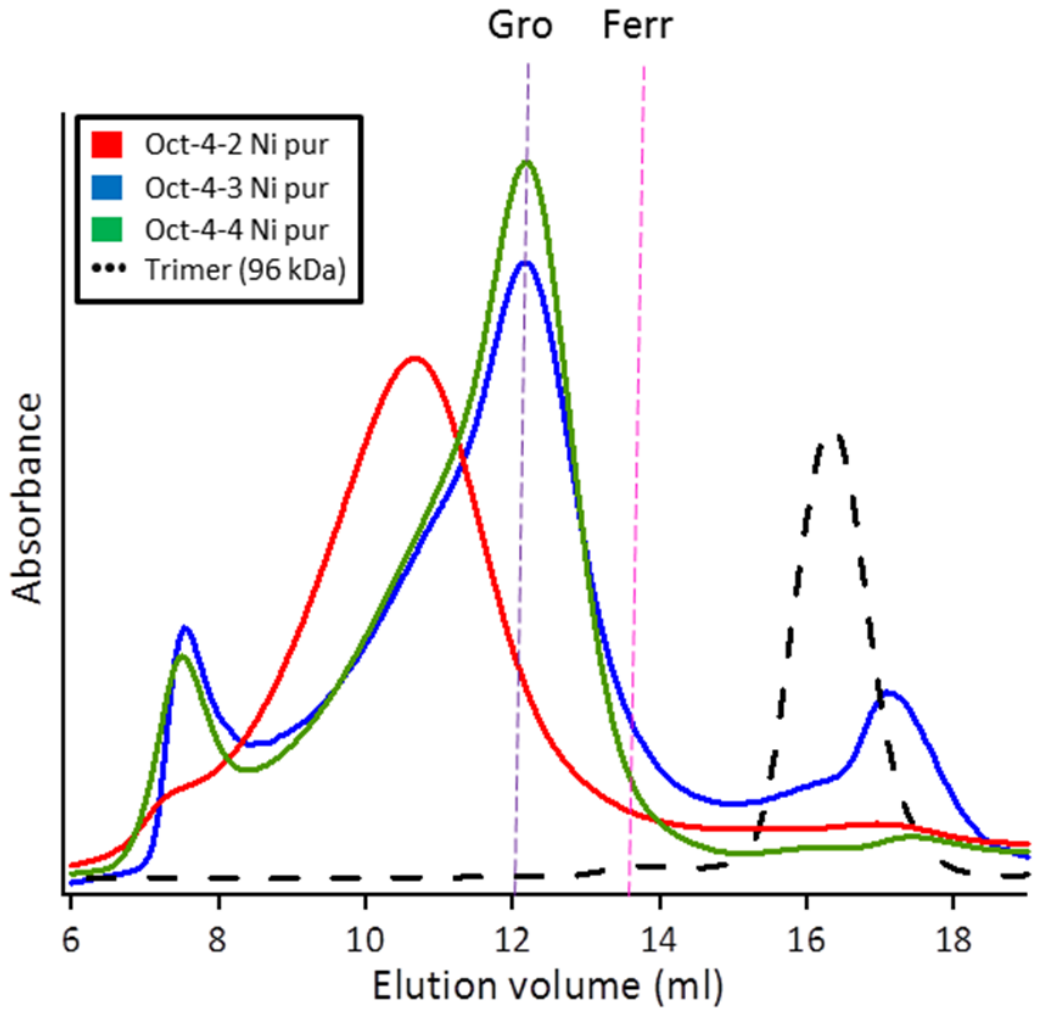


Figure 5.3. Elution profiles of Ni-purified Oct-4 constructs. Oct-4-2 (red) elutes as a broader range of larger species, while Oct-4-3 (blue) and Oct-4-4 (green) have similar elution profiles. None of the Oct-4 constructs show a peak at the elution volume of the unmodified trimeric esterase (black dashed). Elution volumes of standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

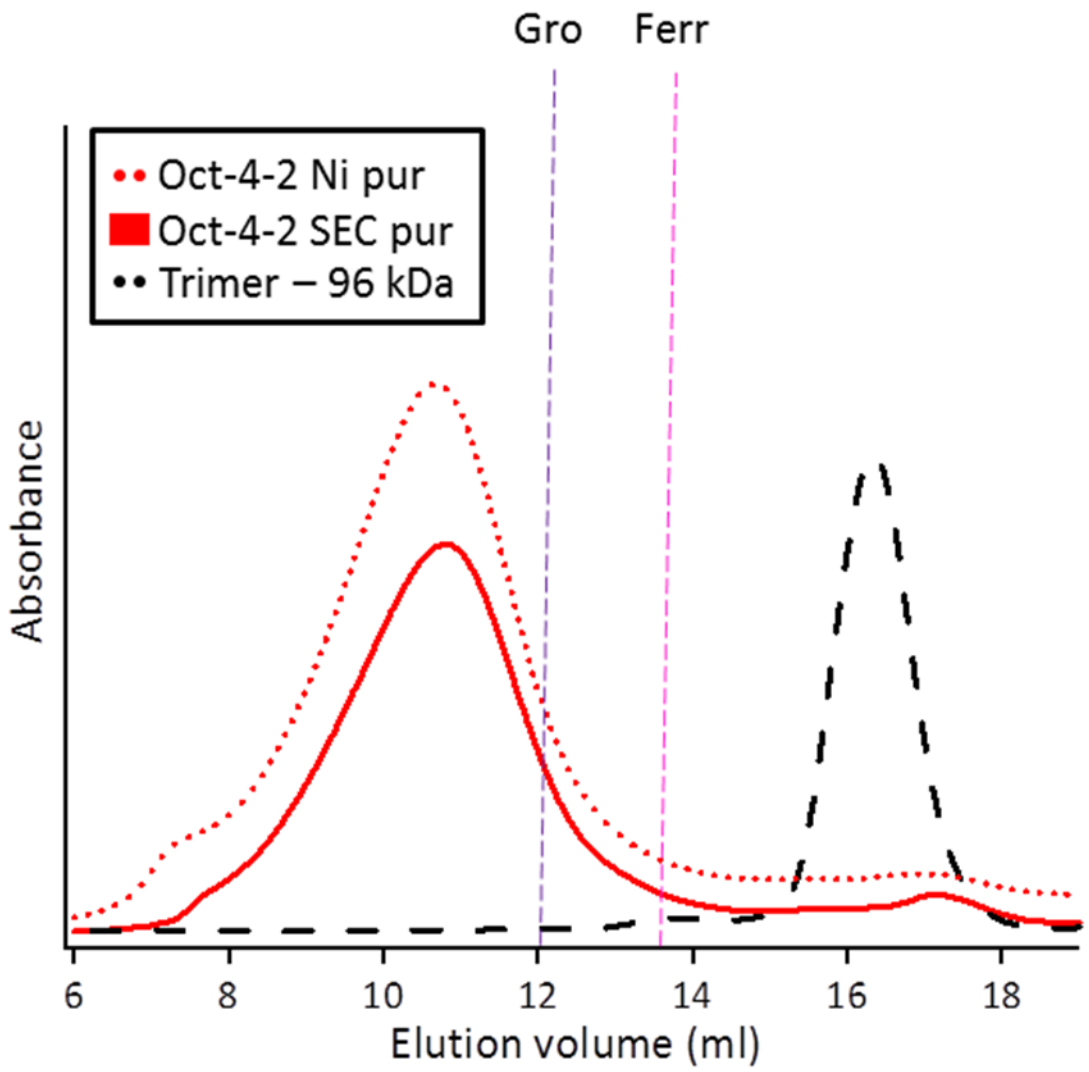


Figure 5.4. Elution profile of Oct-4-2 after Ni purification (dotted red) and SEC purification (solid red). The two side peaks present in Ni-purified Oct-4-2 are removed after size exclusion purification. Elution volumes of standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

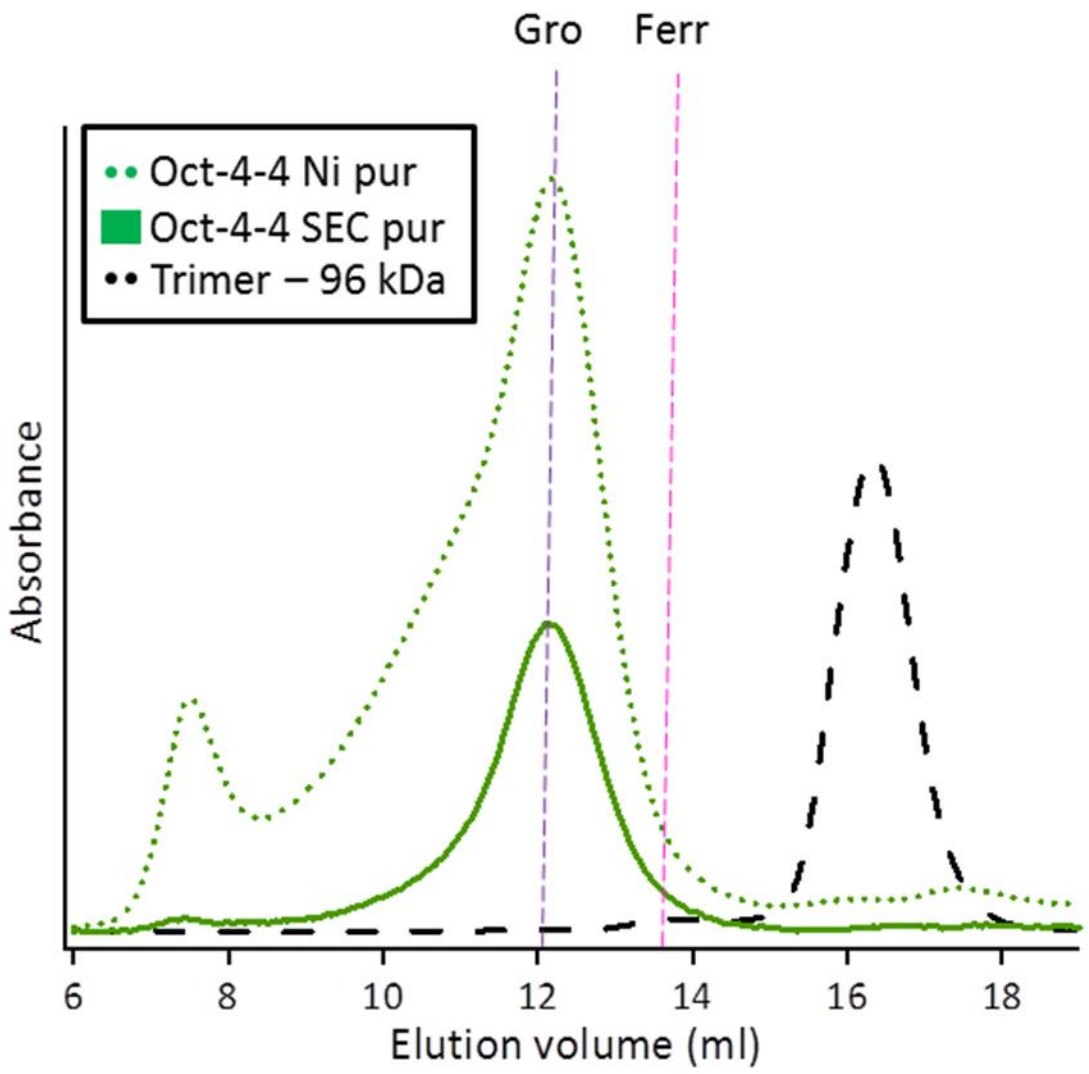


Figure 5.5. Elution profile of Oct-4-4 after Ni purification (dotted green) and SEC purification (solid green). The two side peaks present in Ni-purified Oct-4-4 are removed after size exclusion purification. Elution volumes of standard proteins GroEL (804 kDa, purple dashed) and ferritin (440 kDa, pink dashed) are marked.

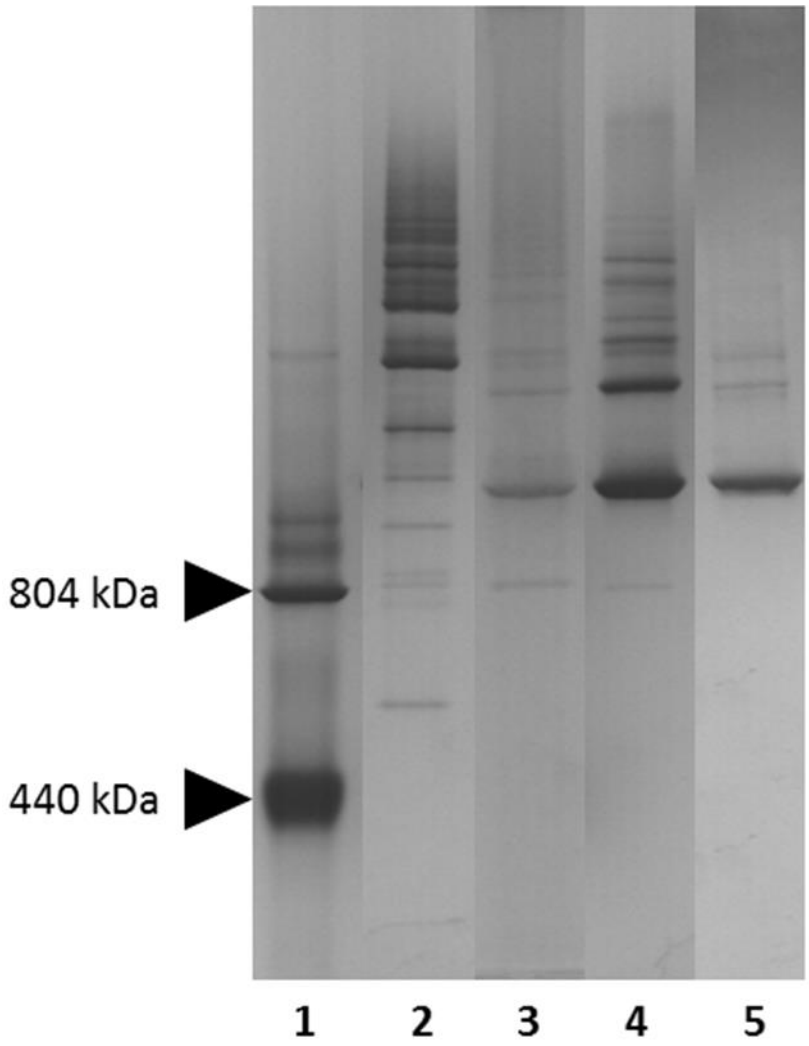


Figure 5.6. Native PAGE of Oct-4 constructs. Lane 1: standard proteins GroEL (top, 804 kDa) and ferritin (bottom, 440 kDa). Lane 2: Ni-purified Oct-4-2. Lane 3: Ni-puriifed Oct-4-3. Lane 4: Ni-purified Oct-4-4. Lane 5: SEC-purified Oct-4-4. Careful SEC purification can remove the majority of oligomeric impurities.

Ni-purified Oct-4-2 and Oct-4-4 were chromatographed on the Superose 6 size exclusion column and fractions were collected every 0.5 ml. These fractions were then analyzed on both the size exclusion column[1] and native PAGE the following day. As indicated in Figure 5.8, SEC fractions of Oct-4-2 rechromatographed as broad peaks roughly centered around the fraction

[1] Note: The rechromatography experiments described in the next two paragraphs were done on a Superose 6 column with a larger bed volume, thus the elution volume for all samples is slightly higher than previously described.

they were taken from: fraction 9, taken from 9.25-9.75 ml, has a peak centered at 10.1 ml; fraction 12, taken from 10.75-11.25 ml, has a peak centered at 11.0ml; fraction 15, taken from 12.25-12.75 ml, has a peak centered around 11.5 ml. Native PAGE of these SEC fractions confirms that multiple species were found in all fractions, although some degree of oligomeric purification can be attained – the smallest complexes are not found in the earliest fractions, and the largest complexes are not found in the latest fractions (Figure 5.8b). These complexes are only somewhat stable – after a day, the peak of fraction 9 has a higher elution volume than the fraction it was taken from, while the peak of fraction 15 has a lower elution volume than the fraction it was taken from. This indicates that the larger complexes isolated in fraction 9 and the smaller complexes isolated in fraction 15 are re-equilibrating to the broad range of complexes seen in the Ni-purified Oct-4-2, in a manner analogous to the behavior seen in SEC-purified Oct-1 (see Figure 3.7), although on a slower timescale. Taken together, these data suggest that Oct-4-2 forms a mixture of semistable complexes in continual equilibrium with each other, which is the behavior predicted for a fusion protein that is sterically precluded from forming a stable complex because of a linker that is too short.

In contrast, analysis of the SEC fractions of Ni-purified Oct-4-4 paints a much rosier picture (Figure 5.9). While fraction 13, taken from 11.25-11.75 ml, eluted at 11.8 ml, and fraction 14, taken from 11.75-12.25 ml, eluted at 12.3 ml, fractions 15, 16, and 17, taken from 12.25-12.75 ml, 12.75-13.25 ml, and 13.25-13.75 ml respectively, all eluted at 12.5 ml. Additionally, native PAGE shows only one visible band in fractions 16 and 17. When these fractions were concentrated, the native PAGE of the resulting concentrate shows one very sharp band with two minor bands above it (Figure 5.6 - Lane 5), and this species is stable at 4 °C

for months. Further characterization is necessary to determine if this isolated single species of Oct-4-4 with what appears to be the correct size for an octahedron, is actually the 8-trimer octahedron.

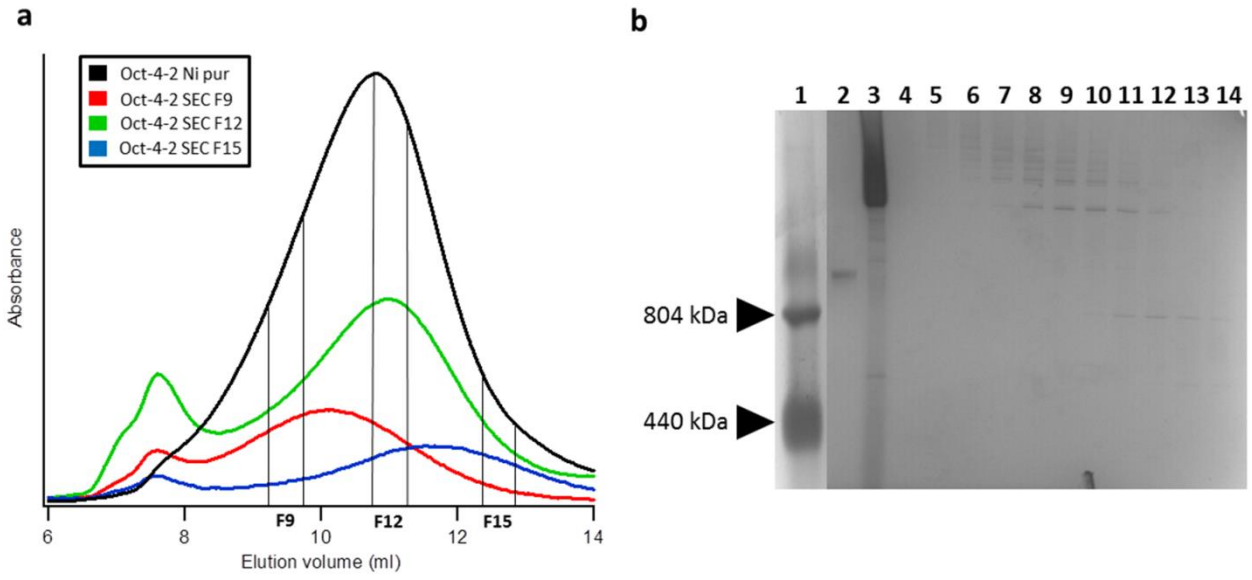


Figure 5.7. Analysis of fractions of the SEC purification of Oct-4-2. a) Re-chromatography of SEC fractions from Oct-4-2. Three 0.5 ml fractions from Ni-purified Oct-4-2 (black) were re-chromatographed on Superose 6 column. All three fractions had elution volumes close to where they were collected, but all were shifted towards the elution volume of the Ni-purified Oct-4-2, indicating that re-equilibriation had occurred. b) Native PAGE of SEC purification fragments from Oct-4-2. Lane 1: standard proteins GroEL (top, 804 kDa) and ferritin (bottom, 440 kDa). Lane 2: SEC-purified Oct-4-4. Lane 3: Ni-purified Oct-4-2. Lane 4: SEC fraction 6 of Oct-4-2. Lanes 5-14: SEC fractions 8-17 of Oct-4-2.

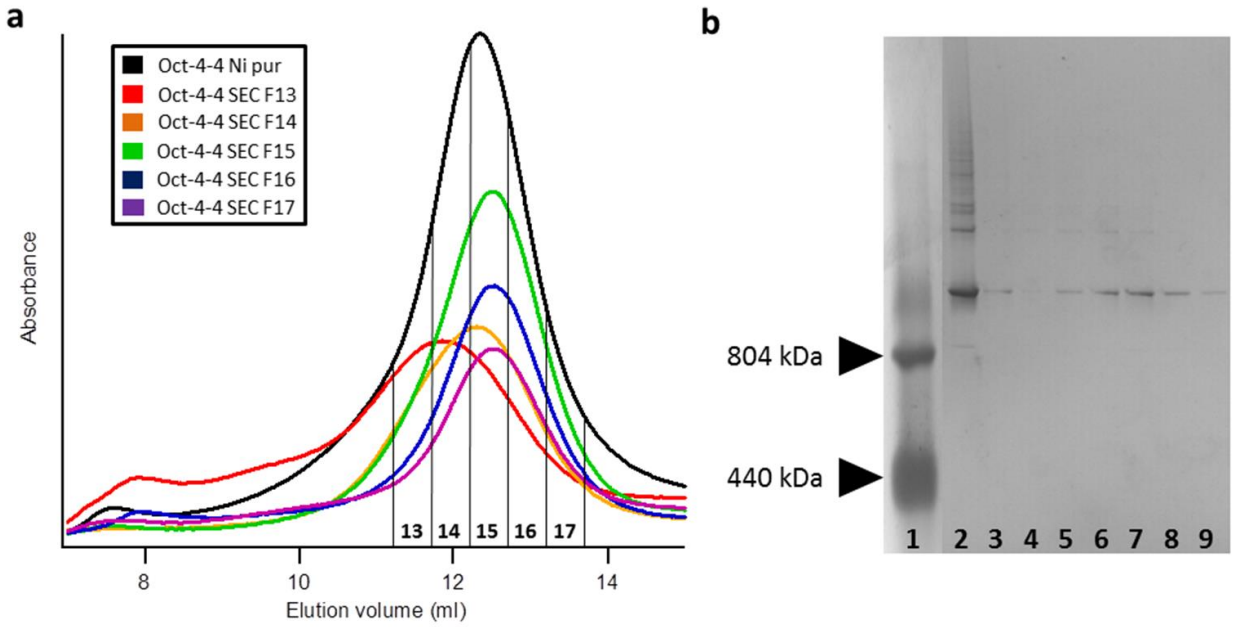


Figure 5.8. Analysis of fractions of the SEC purification of Oct-4-4. a) SEC fractions from Ni-purified Oct-4-4 (black) were re-run on a Superose 6 column. Fractions 15, 16, and 17 all had the same elution volume, indicating that the major species in Ni-purified Oct-4-4 is also the smallest. b) Native PAGE of SEC purification fractions from Oct-4-4. Lane 1: protein standards GroEL (top, 804 kDa) and ferritin (bottom, 440 kDa). Lane 2: Ni-purified Oct-4-4. Lanes 3-9: SEC fractions 11-17 of Oct-4-4.

5.4 - Analytical Ultracentrifugation of Oct-4 Constructs

The major species isolated from Oct-4-4 was further analyzed by sedimentation velocity analytical ultracentrifugation (Figure 5.10). Using the program sedfit, we could identify two distinct species, the first having a sedimentation coefficient of 11.9 and comprising 74% of the sedimenting species, the second with a sedimentation coefficient of 15.2 and comprising 18% of the sedimenting species. Using a fitted frictional ratio of 1.37, this species at 11.9 S corresponds to a molecular weight of 645 kDa, or roughly six trimers. However, it should be noted that the r.m.s.d. of the sedimentation analysis using this fitted frictional ratio (0.00270) is only marginally better than the r.m.s.d. of the sedimentation analysis where $f/f_0 = 1.70$ (0.00273), which correlates to a molecular weight of 890 kDa, much closer to the expected

molecular weight of the octahedron. The latter frictional ratio would also be consistent with the hypothesis proposed in the previous chapter, that hollow protein cages should increase in frictional ratio as their size increases.

The Oct-4-2 construct was also analyzed by SV-AUC, with results that matched well with native PAGE, and with a good r.m.s.d. of 0.0026. Analysis by sedfit identifies (at least) five distinct peaks with sedimentation coefficients of 16.7, 21.1, 25.8, 29.4, and 32.5 S. The peaks at 21.1, 25.8, and 29.4 S are well-defined and correspond to relative concentrations of 22%, 32%, and 20%, respectively. Fitting the frictional ratio via the Marquis-Levenburg algorithm resulted in an impossible frictional ratio of 0.69, so for a rough molecular weight calculation we used a frictional ratio of 1. This yields molecular weights of 943, 1260, and 1560 kDa for each of the major species, which roughly corresponds to complexes of 10-16 trimers if the assumed frictional ratio is accurate. Given the degree of uncertainty that still remains from the sedfit analysis of the Oct-4 complexes, we turned to Ultrascan for a more detailed analysis of both constructs.

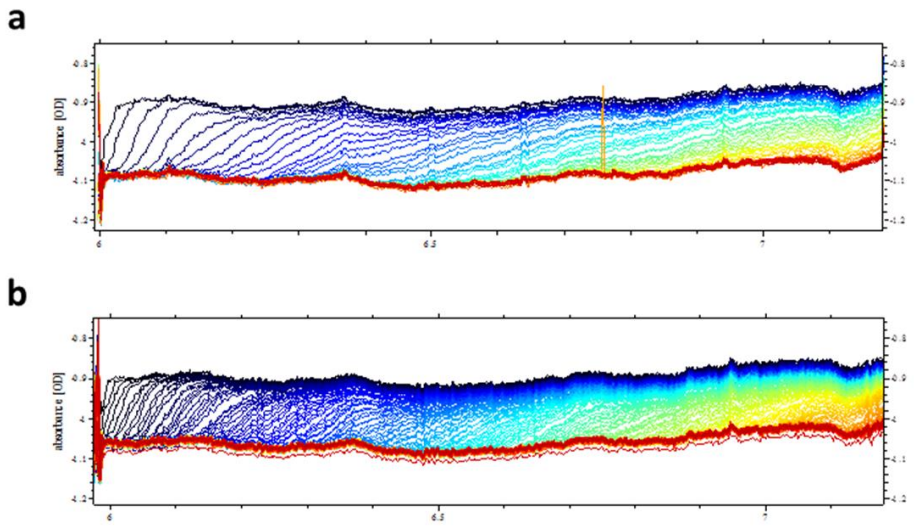


Figure 5.9. Raw SV-AUC data for SEC-purified Oct-4-2 (a) and Oct-4-4 (b). Violet traces represent the first scan, red traces represent the final scan.

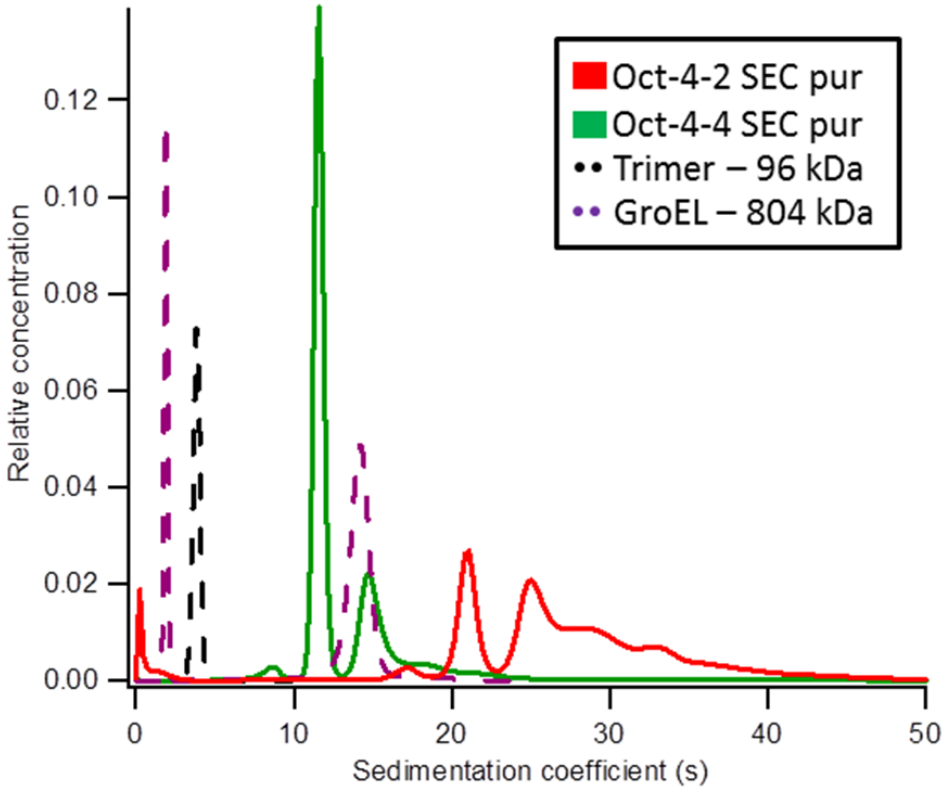


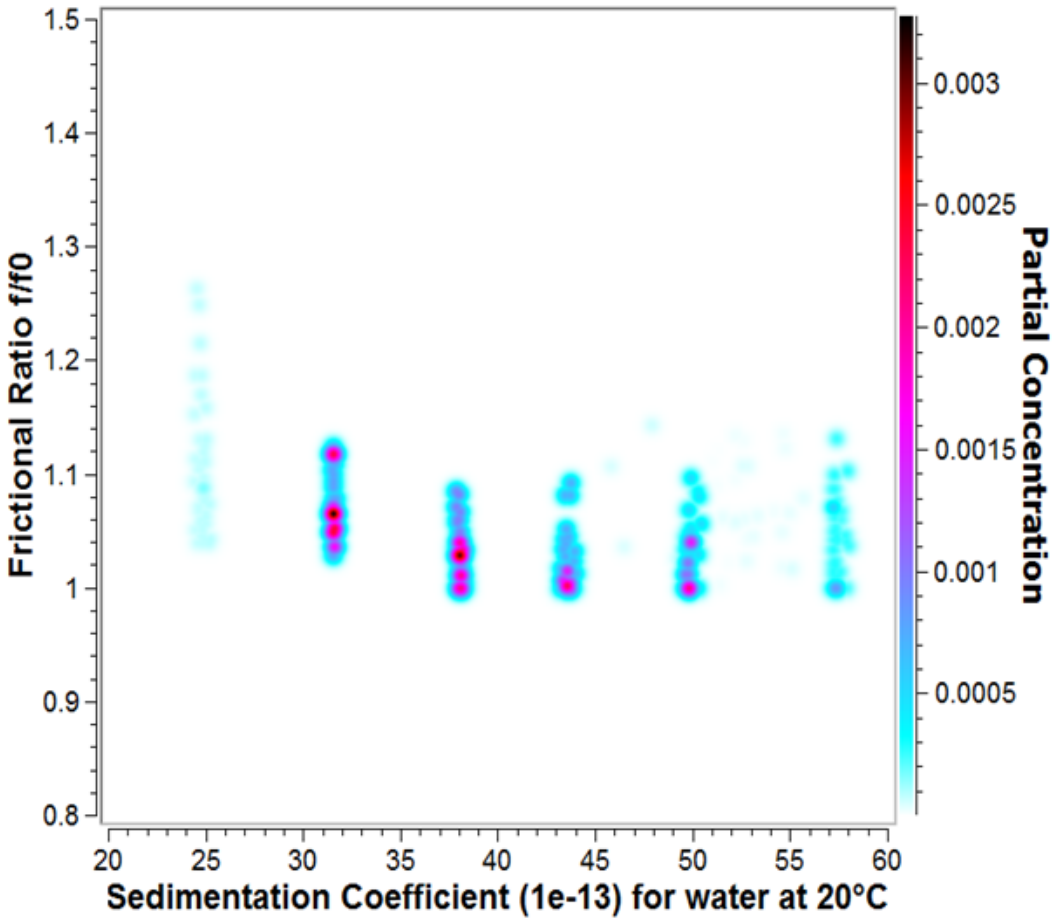
Figure 5.10. Sedimentation traces of Oct-4 constructs analyzed with sedfit. Oct-4-2 (red) has at least four distinct peaks with sedimentation coefficients larger than GroEL (purple dashed). Oct-4-4 (green) has one major peak and one minor peak with sedimentation coefficients near GroEL. Neither fusion protein has a species that sediments near the unmodified trimeric esterase (black dashed).

5.5 - 2-Dimensional Sedimentation Analysis

Both the Oct-4-2 and Oct-4-4 data were analyzed by the program Ultrascan, which yielded detailed 2-dimensional sedimentation plots for each construct. Six distinct species in Oct-4-2 could be identified, all with frictional ratios approaching 1 (Figure 5.11). These species had $s_{20,w}$ coefficients of 24.8, 31.5, 38.0, 43.5, 49.8, and 57.3 S, which calculated to molecular weights of 650, 905, 1120, 1357, 1681, and 2113 kDa respectively, and relative concentrations of 2.2%, 21.5%, 27.4%, 20.3%, 13.9%, and 7.4%, respectively. These roughly correspond to the molecular masses of complexes with six (642 kDa), eight (856 kDa), ten (1070 kDa), twelve (1284 kDa), sixteen (1712 kDa), and twenty (2140 kDa) trimers, though it is impossible to verify the true oligomerization states of these complexes. Regardless of the degree of accuracy that can be obtained from 2DSA, this confirms that Oct-4-2 forms a mix of multiple soluble complexes that range from 6-20 trimers in size, and that none of these complexes are well-formed with a hollow interior. This is another important piece of information, because as discussed in section 1.4, a C_3+C_4 symmetry pair with a restricted dihedral angle still has the potential to form a hollow 16-mer with icosahedral symmetry. Being able to distinguish between these larger, but well-formed complexes and large, misfolded complexes is an important part of future characterization of designed fusion protein constructs.

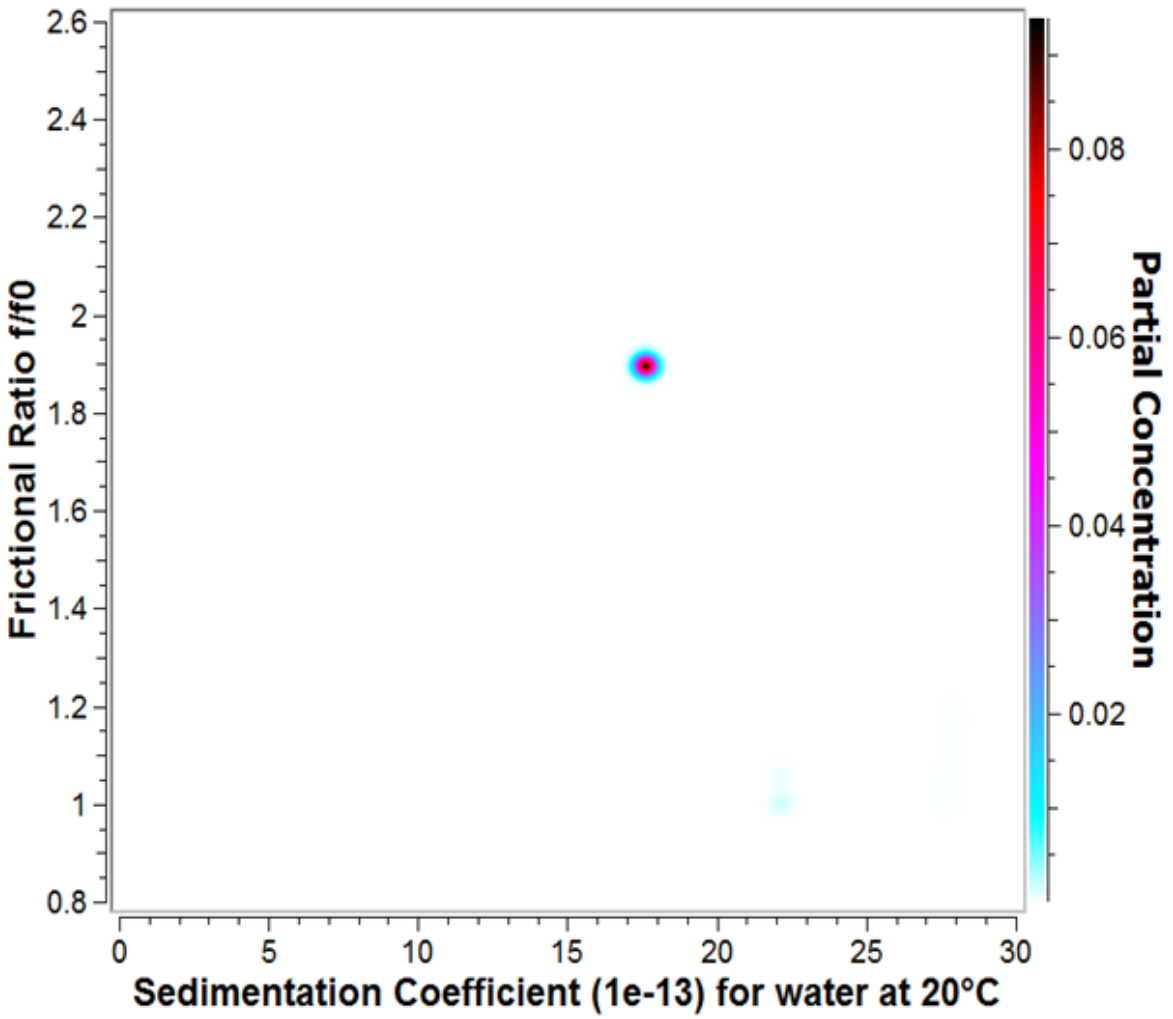
Ultrascan analysis of Oct-4-4 identified one major species and three minor species (Figure 5.12). The three minor species had larger sedimentation coefficients than the major species, having $s_{20,w}$ coefficients of 22.1, 27.7, and 37.3 S and relative concentrations of 18.6%, 4.4%, and 2.3%. All three minor species had frictional ratios close to 1, and have calculated molecular weights of 489, 726, and 1144 kDa, respectively, indicating that these are not well-

formed complexes, or perhaps are transition states. The major species, comprising 73.3% of the sample, has a sedimentation coefficient of 17.6, and a frictional ratio of 1.89. This calculates to a molecular weight of 886 kDa, very close to the expected molecular weight of the octahedron. The frictional ratio of 1.89 is much higher than all other complexes measured, and is consistent with a large hollow complex or a highly elongated complex. To investigate the meaning of this high frictional ratio, we turned to negative stain TEM to directly visualize the particles. As discussed in section 5.3, Oct-4-4 was determined to be highly spherical by TEM, thereby providing further evidence for the hypothesis that large, hollow protein complexes have high frictional ratios.



Oct-4-2	$s_{20,w}$	M.W. (kDa)	f/f_0	Conc. %
Species 1	24.80 ± 0.36	649 ± 96	1.09 ± 0.11	2.2%
Species 2	31.55 ± 0.63	905 ± 67	1.07 ± 0.05	21.5%
Species 3	38.03 ± 0.16	1,128 ± 63	1.03 ± 0.04	27.4%
Species 4	43.55 ± 0.33	1,357 ± 93	1.01 ± 0.05	20.3%
Species 5	49.82 ± 0.81	1,681 ± 174	1.02 ± 0.06	13.9%
Species 6	57.34 ± 1.14	2,113 ± 199	1.03 ± 0.07	7.4%

Figure 5.11. 2D-sedimentation analysis of Oct-4-2 by Ultrascan. Four major and two minor species were isolated, the characteristics of which are tabulated above. All species had frictional ratios near 1, indicating none of these species is well-formed and has a hollow interior. These species correlate to complexes of 6-20 trimers.



Oct-4-4	$s_{20,w}$	M.W. (kDa)	f/f_0	Conc. %
Species 1	17.59 ± 0	885 ± 14	1.89 ± 0.02	73.3%
Species 2	22.11 ± 0.07	489 ± 26	1.01 ± 0.04	18.5%
Species 3	27.74 ± 0.30	728 ± 114	1.05 ± 0.10	4.5%
Species 4	37.33 ± 0.20	1145 ± 194	1.06 ± 0.09	2.3%

Figure 5.12. 2D-sedimentation analysis of Oct-4-4 by Ultrascan. One major and three minor species were isolated, the characteristics of which are tabulated above. The three minor species have frictional ratios consistent with collapsed species, while the major species is a well-formed, hollow complex with a molecular weight slightly higher than what is expected for an 8-trimer complex.

5.6 - Transmission Electron Microscopy

Transmission electron microscopy images of SEC-purified Oct-4-2 and Oct-4-4 were obtained (Figures 5.13, 5.14). These show that both constructs exist as primarily globular complexes, with diameters of 15-25 nm for Oct-4-2 and 17-19 nm for Oct-4-4. While particles of Oct-4-2 have no distinguishable pattern of geometry, in Oct-4-4 there are several particles that have a clearly visible 4-fold symmetry axis that connects 4 trimers at their vertices. As with Oct-3-4, the resolution isn't good enough on either of these constructs to identify density that correlates to coiled-coils.

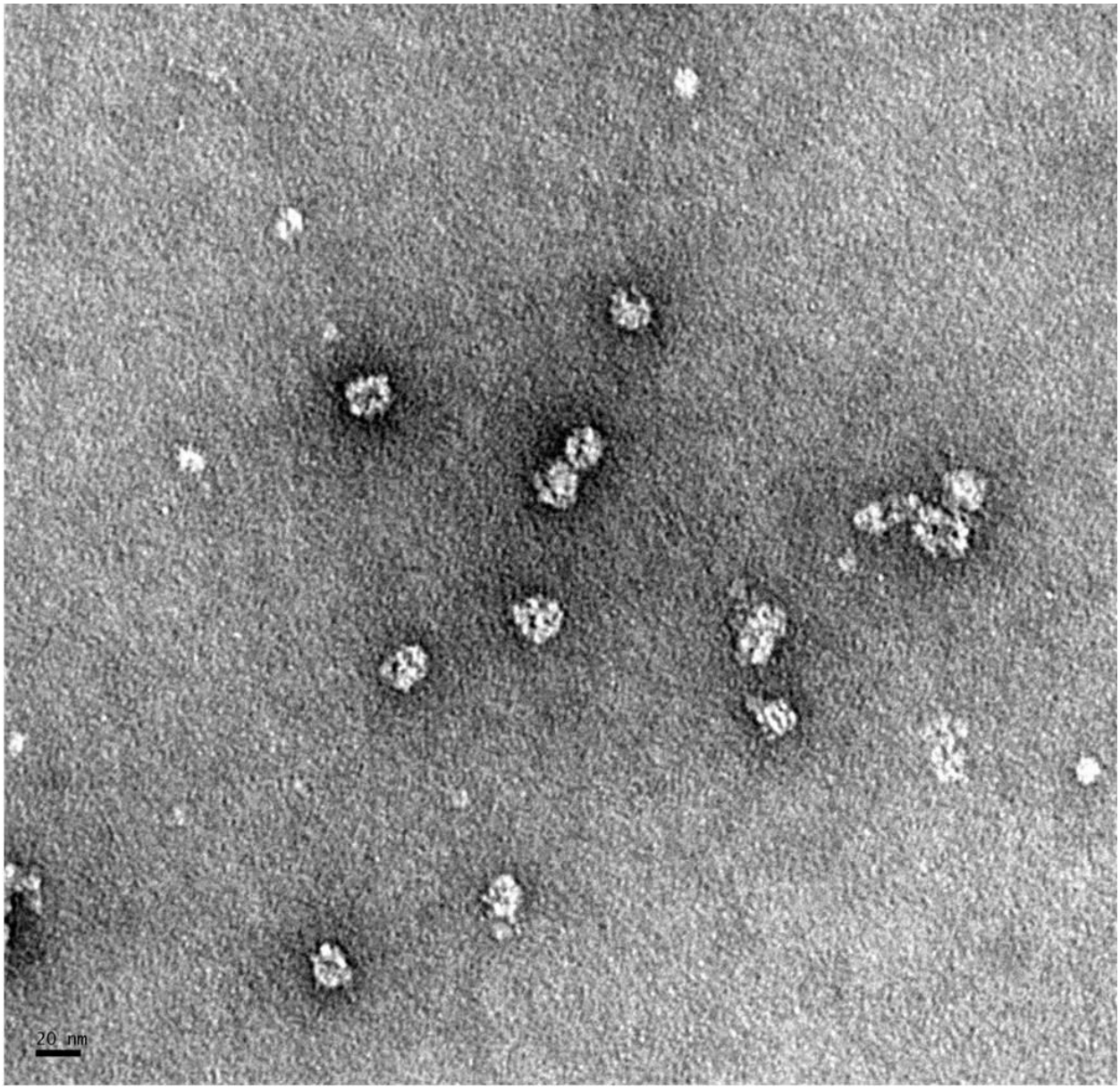


Figure 5.13. Transmission electron micrographs of SEC-purified Oct-4-2. Most particles observed are globular, though it appears some particles consist of multiple associated globular species. No particle has an observable symmetry. Particle diameter varies significantly from 15-25 nm for single globular species, larger for multiple associated species. Scale bar is 20 nm.

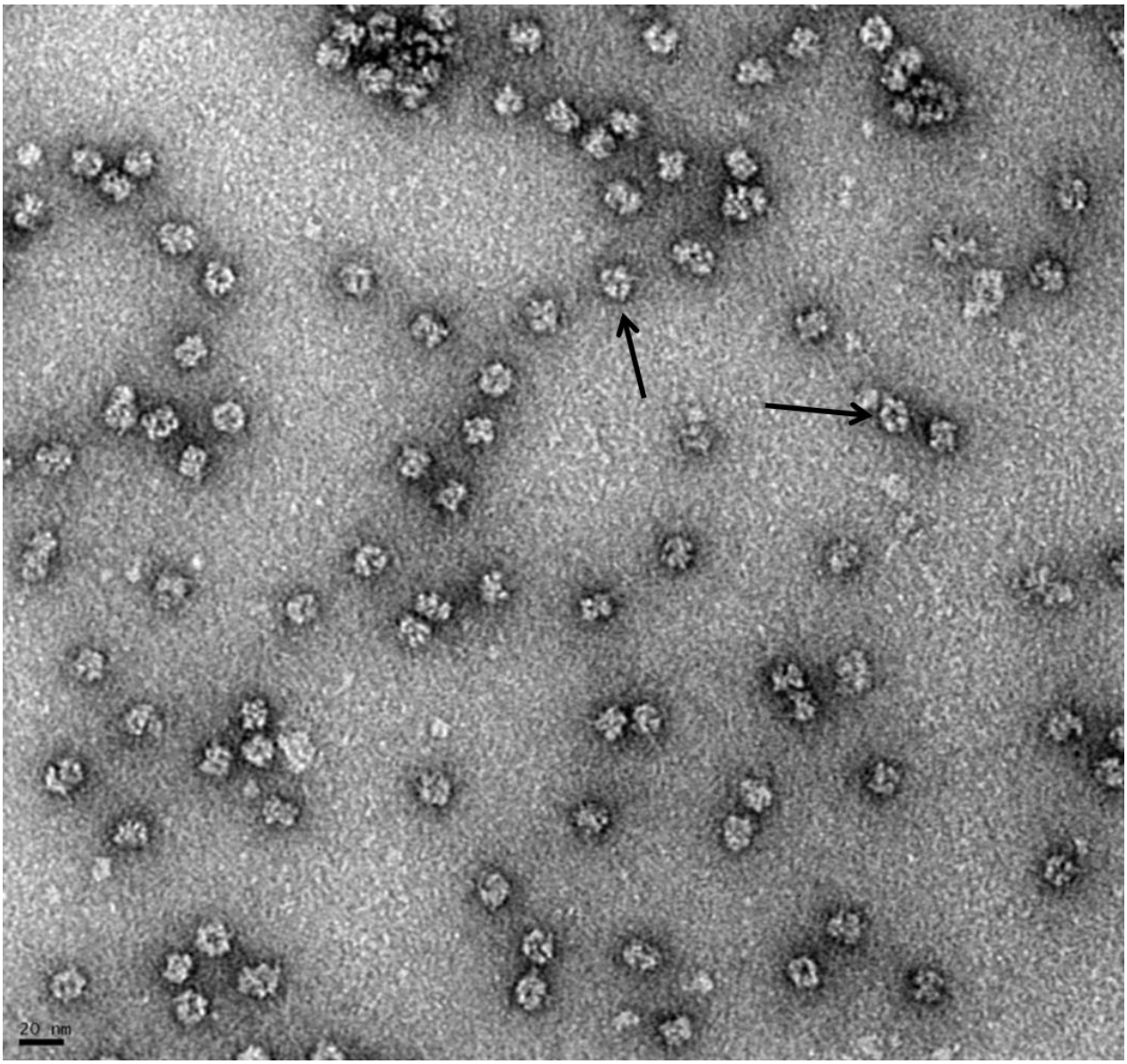


Figure 5.14. Transmission electron micrographs for SEC purified Oct-4-4. All observed particles are globular, with a narrow range of diameters from 17-19 nm. In several particles highlighted with arrows, C_4 geometry can be seen connecting trimers. Scale bar is 20 nm.

Since the TEM images of Oct-4-4 show particles that are highly regular, it was possible to perform cryo-EM studies and obtain class averages and reconstruct a 3D model of the purported octahedral complex. In collaboration with Dr. Min Su, we selected 44,856 particles imaged on cryogenically frozen carbon grids of SEC purified Oct-4-4 and subjected these to reference-free class averaging. 406 class averages were generated, of which 95 class averages

were selected for 3D electron reconstruction on the basis of resolution and cage distortion (Figure 5.15). In many of these class averages, the 4-fold symmetry axis is apparent, and unexpectedly, many of these class averages have at the center of their 4-fold axis a region of strong electron density. This additional density is presumed to correspond to the coiled-coil, but density correlating to the coiled-coil cannot be seen at any peripheral location on almost all of the class averages. There are several explanations for this – there may be too much flexibility in the linker, or that the coiled-coils point into the interior of the cage structure, and are thus only visible when viewed from the top-down. Regardless, the end result is significant – in the initial 3D electron density reconstruction and symmetrization into octahedral space, this region appeared as a dense sphere of density located in the supposedly hollow protein cage interior (Figure 5.16a). This rough model was further improved by fitting the crystal structure of the trimeric esterase to the electron density, which yielded a much improved reconstruction that has a resolution of 17 Å at the 0.5 level of Fourier shell correlation. This improved model shows the trimeric esterases oriented to align their C-termini with the C_4 axis, in close agreement with the computational model generated in section 4.1. This is a detailed enough reconstruction to confirm the identity of the reconstructed species as an octahedron arranged with distinctly identifiable C_3 and C_4 symmetry sites, so although there are still many open questions about the location and orientation of the coiled-coils, those questions are outside the scope of this dissertation and are appropriate for future directions.

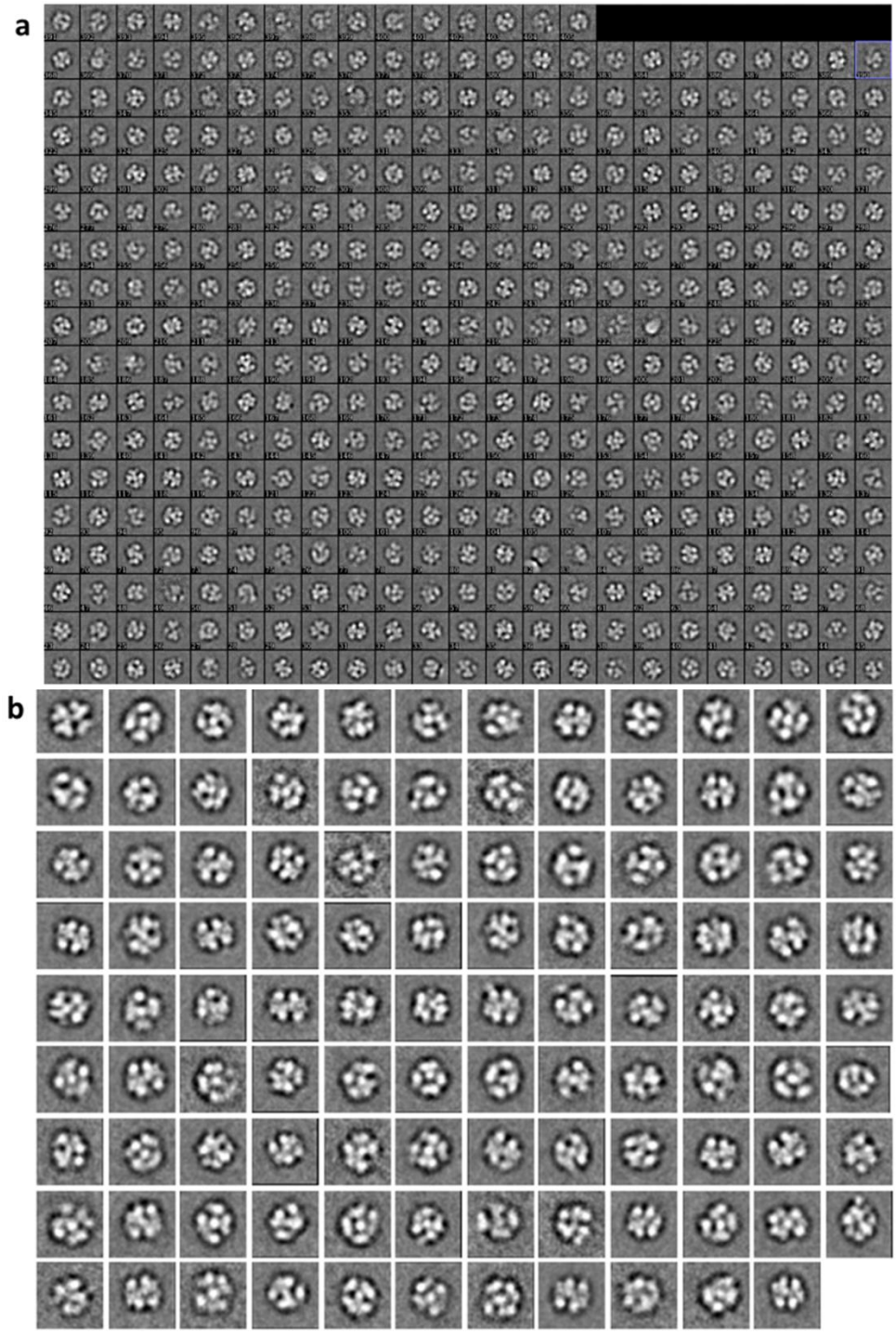


Figure 5.15. Reference-free 2-D class averages of particles of Oct-4-4 imaged by cryo-EM. a) The full set of 406 class averages generated from the 44,856 particles imaged. b) The 95 class averages used in the 3-D electron density reconstruction of Oct-4-4.

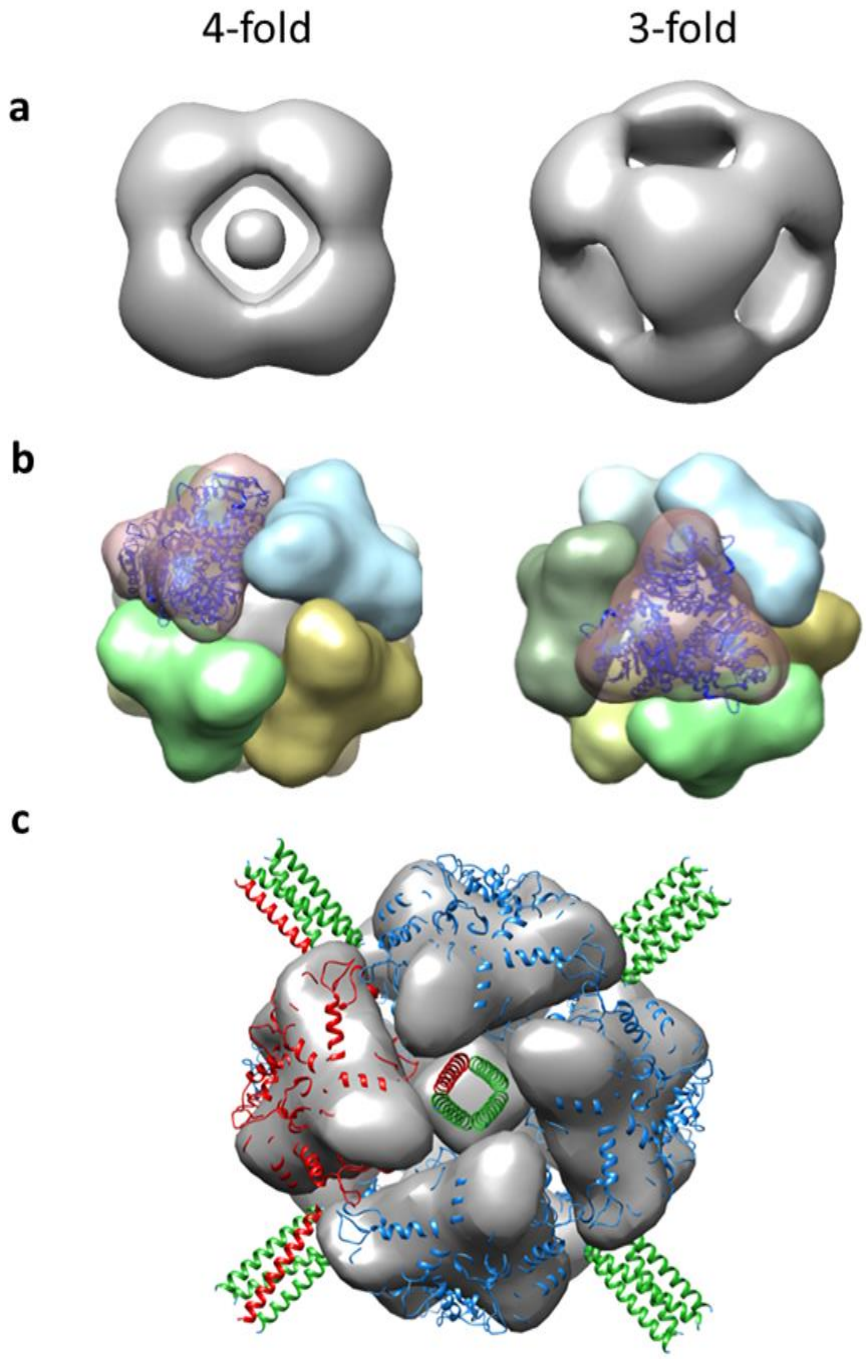


Figure 5.16. 3D electron density reconstruction of Oct-4-4. a) Initial electron density map after selected particles were symmetrized into octahedral space, serving as a starting point for further refinements. b) Electron density map after refinement with the crystal structure of the trimeric esterase, overlaid with the crystal structure of a single trimeric esterase (Image credit: Min Su and Georgios Skiniotis). c) Overlay of the Rosetta-generated model of the octahedron with the 3-dimensional electron density map of Oct-4-4.

5.7 - Ion Mobility-Mass Spectrometry of Oct-4-4

The previously-described data collected for the major species of Oct-4-4 provides very good evidence that it assembles into an octahedral cage. It is therefore an interesting target for analysis by a broad range of techniques, which would help gain an understanding of the strengths and limitations of each analytical technique as it applies to other designed fusion constructs going forward. IM-MS is a promising new technology that is designed to assay the molecular weight and surface area of designed protein constructs in their native states, and has been used previously to determine the molecular weight of designed fusion protein cages with similar masses.[1]

We attempted to analyze SEC-purified Oct-4-4 by IM-MS, but the conditions required may not be feasible for analysis. IM-MS requires that buffers and constructs be carefully prepared to remove all salt from the sample, involving multiple buffer exchanges into ammonium acetate solution. Every sodium ion attached to a native protein will mask a charge on that protein, complicating the calculation of molecular weight by comparing m/z peaks. On a large molecule such as the octahedron this is difficult, and considering the octahedron has 6 bundles of 4 coiled-coils, with 3 lysines and an arginine on each coil, it proved to be impossible to remove enough salt from Oct-4-4 to see definitive m/z peaks. Despite this, we were able to see a broad peak that was detected in the 10,000-12,000 m/z range, that consisted of multiple smaller, but poorly-defined peaks (Figure 5.17). Calculating the mass from these m/z peaks gives a mass of 855-890 kDa, an encouraging result despite the low data quality. Further attempts to outcompete the remaining salt molecules by increasing the concentration of ammonium acetate to 1 M during buffer exchange were unsuccessful in improving the mass

resolution. These results are preliminary, and a higher resolution could be obtained with a more concentrated sample of Oct-4-4 or a more careful preparation to remove salt.

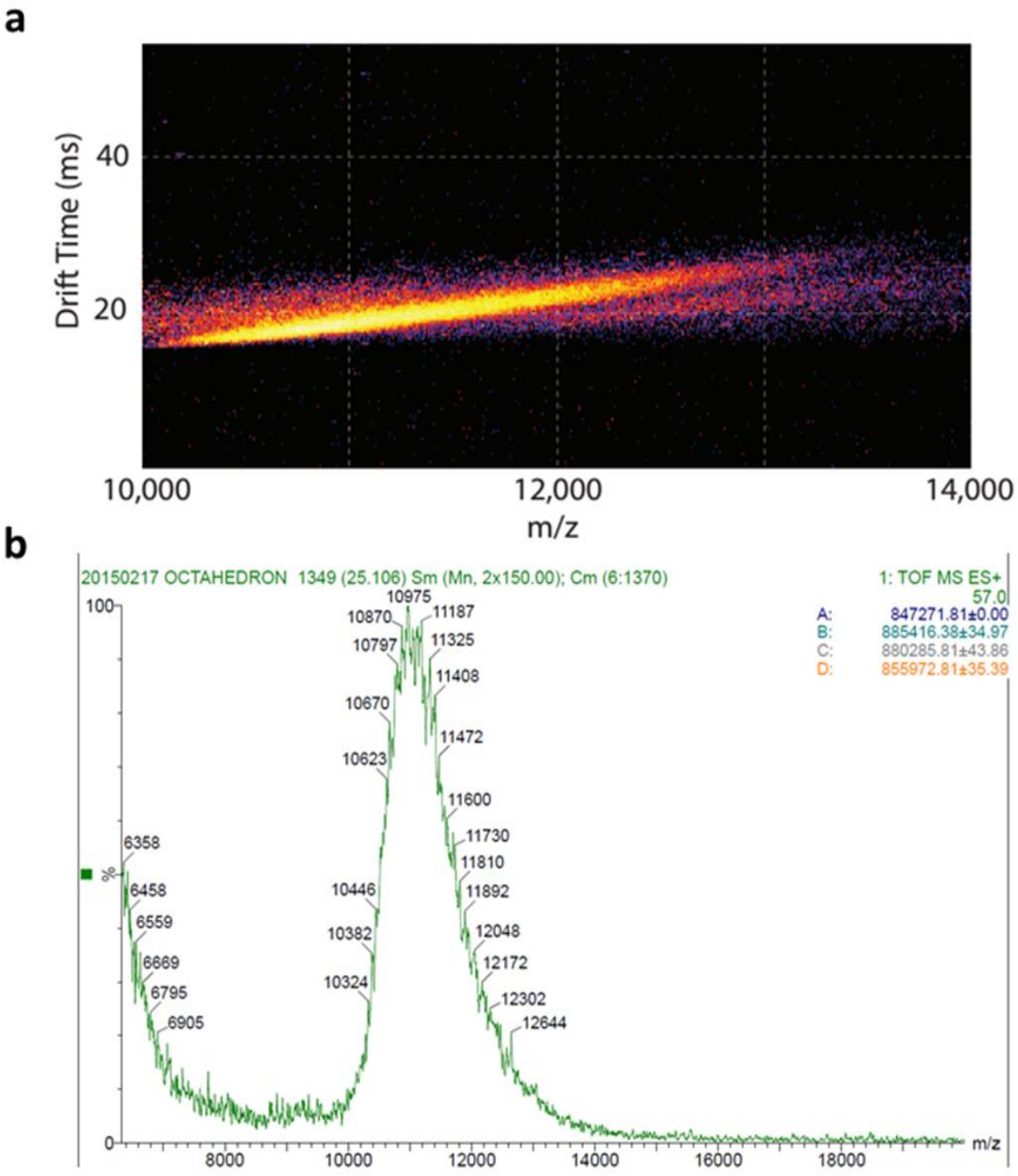


Figure 5.17. Ion-mobility mass spectrographs of Oct-4-4. a) Plot of m/z against drift time shows a distribution characteristic of a single species, but without distinctly identifiable m/z peaks. b)

Mass calculations from m/z peaks yield a molecular weight very close to the predicted molecular weight of the octahedron.

5.8 - Effect of Urea on Oligomerization State of Oct-4-4

A common theme throughout this project has been the relationship between the binding strength of the coiled-coils and the oligomerization state of the assembled species. We saw previously during the analysis of Oct-1 that attaching a weakly-associating coil to the trimer led to a broadly oligomerizing species in constant equilibrium, but it is also easy to picture a fusion protein complex where an attached coiled-coil that binds too tightly may become kinetically trapped into large and irregular species. We therefore explored the possibility of using urea to loosen the hydrophobic interactions in the coiled-coil domain that mediate assembly. Research on this topic was also limited by project constraints but data was gathered in two areas – the effects of adding urea during lysis and Ni-affinity purification, and the effects of adding urea to SEC-purified octahedron.

Addition of urea to the lysis buffer was considered early on in the analysis of Oct-4-4 as the prior buffer conditions for Ni purification (0-100 mM NaCl, 0 M urea, 40 mM imidazole) led to nonspecific binding of other proteins on the column resin, including GroEL. As noted in section 2.3, GroEL is a problematic contaminant because it co-elutes with the octahedron during size exclusion purification. Thus, cells expressing Oct-4-4 were lysed and nickel-affinity purified under four different buffer conditions. The first two had lysis buffer that contained no urea, and either 0 M or 300 mM NaCl. The second two had lysis buffer that contained 1 M urea and either 0 M or 300 mM NaCl. All four preparations were eluted with the elution buffer described in chapter 2, containing 300 mM NaCl, 50 mM HEPES, 500 mM imidazole and 5%

glycerol, but which did not contain urea. Native PAGE of concentrated, dialyzed, Ni-purified samples from each of these four preparations revealed that in both buffers containing urea, GroEL was removed, while it was present in both buffers lacking urea (Figure 5.18). Raising the urea content of the lysis/wash buffer to 2 M, however, resulted in greatly reduced yield, and dialyzing the Ni-purified Oct-4-4 into dialysis buffer that contained 1 M urea and storing at 4 °C resulted in all of the protein precipitating out over the course of several weeks.

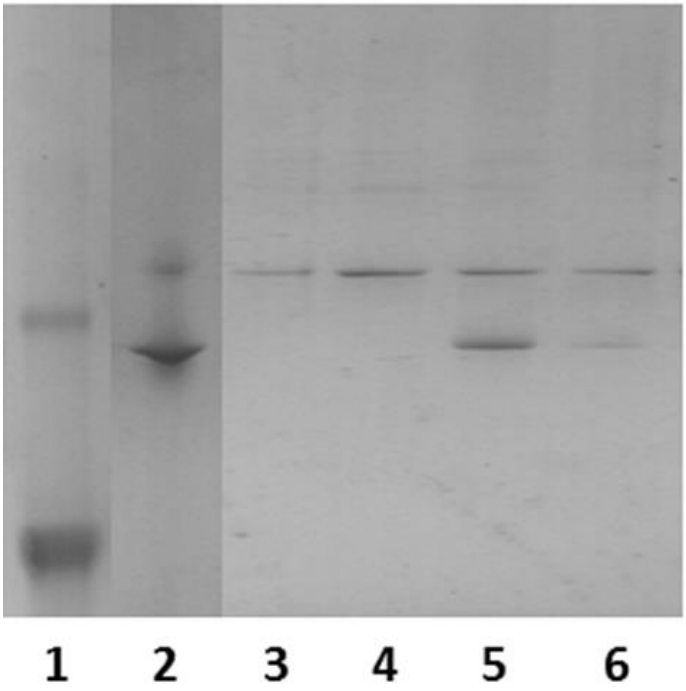


Figure 5.18. Native PAGE of four different buffer conditions for lysis and Ni-affinity purification for Oct-4-4. Lane 1: ferritin (440 kDa). Lane 2: GroEL (804 kDa). Lane 3: Concentrated Oct-4-4 after Ni purification in buffer containing 1 M urea and 300 mM NaCl. Lane 4: Concentrated Oct-4-4 after Ni purification in buffer containing 1 M urea and 0 mM NaCl. Lane 5: Concentrated Oct-4-4 after Ni purification in buffer containing 0 M urea and 300 mM NaCl. Lane 6: Concentrated Oct-4-4 after Ni purification in buffer containing 0 M urea and 0 mM NaCl. Both preparations with 1 M urea did not contain the contaminant GroEL.

The second set of experiments concerned the effects of Oct-4-4 after size exclusion chromatography purification. Size exclusion profiles were collected one day after concentrated, SEC-purified Oct-4-4 samples had 0.5, 1, 2, or 3 M urea added to them (Figure 5.19). The

resulting elution profiles show that adding 0.5 M urea causes little perturbation in the oligomerization state of Oct-4-4, but at higher concentrations of urea, the coiled-coil connections loosen enough to re-equilibrate into larger complexes, although this may be also caused by general unfolding and aggregation of the fusion proteins. This is an interesting effect, but its relevance to the design of future studies is unclear.

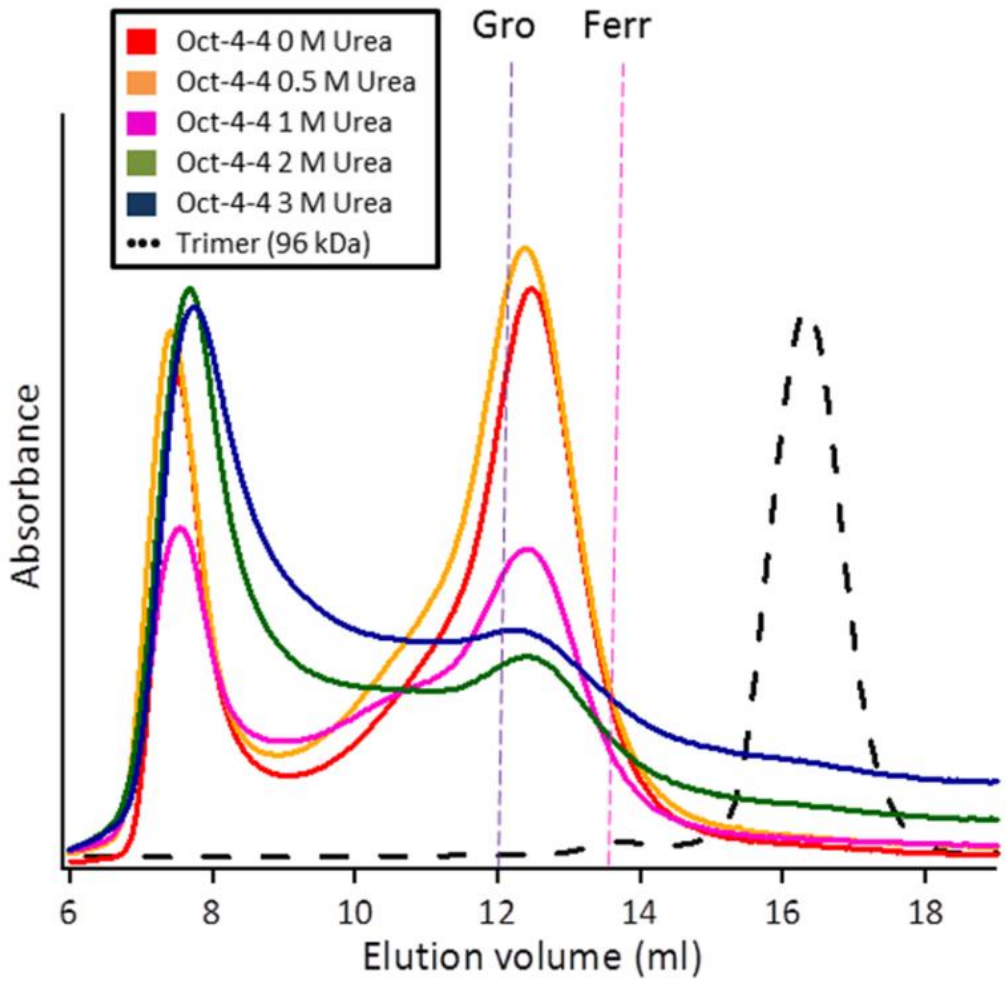


Figure 5.19. SEC elution profiles of Ni-purified Oct-4-4 after addition of different concentrations of urea. Oct-4-4 with up to 0.5 M urea shows no change in absorption, but upon increased addition of urea the peak at 12 ml corresponding to the octahedron is reduced and broadened, indicating that equilibrium between the octahedron and larger species was induced.

5.9 - Conclusions

In this chapter, we have finally put all the pieces together, attaching a tightly-binding tetrameric coiled-coil to the trimeric esterase via an optimized linker. I designed three fusion protein constructs with 2, 3, or 4 residues in the linker, named Oct-4-2, Oct-4-3, and Oct-4-4. Oct-4-2 expressed and purified as a mix of collapsed, soluble, globular aggregates consisting of 6-20 trimers. Oct-4-3 expressed mainly as inclusion bodies and could not be purified in high enough quantities for a detailed analysis, but displays some characteristics expected for an octahedral protein cage. Oct-4-4 expressed as soluble protein that purifies as multiple large species by Ni-affinity, of which the smallest species can be purified by size exclusion to 73% purity. This species was characterized by AUC as an approximately spherical, hollow protein complex with a molecular weight of slightly more than 8 trimers. By TEM the C_4 axis could be observed connecting the vertices of 4 trimeric proteins, and a cryo-EM class averaging and 3-D electron density reconstruction of Oct-4-4 resulted in a 3-dimensional model with modest resolution, but which nonetheless confirmed the presence of both trimeric esterases and coiled-coils in an octahedral configuration. These results confirm the first successful synthesis and purification of an octahedral protein cage built from a *de novo* designed fusion protein with a C_4+C_3 symmetry pair, as well as one of only a handful of soluble protein cages incorporating a flexible linker. The successful oligomerization of this construct opens up a smorgasbord of potential avenues of research as well as the possibility of functionalization for industry application.

5.10 - References

1. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* **6**, 1065-1071 (2014).

Chapter 6 Conclusions and Future Directions

In this dissertation, I have evaluated a general method for the design, optimization, and purification of protein cages with arbitrary symmetries. Like previously designed protein cages, this method consists of selecting two C_n-symmetric protein building blocks and linking them by genetic fusion. But unlike the majority of these protein cages, the two symmetric building blocks are connected by a flexible linker. The flexible linker allows us to avoid one of the major difficulties in protein cage design: significant computational power must be exhausted to design mutants that have precisely oriented dihedral angles between symmetry axes and many of these designed proteins do not express as soluble, assembled protein complexes. In contrast, employing a flexible linker only requires consideration of the length of the linker to avoid unwanted steric hindrance or self-association. One of the major disadvantages of using a flexible linker is that a flexibly-linked, designed fusion protein can assemble into a broad range of stable, closed (though potentially porous) complexes, provided that the range of dihedral angles required for that specific complex can be imparted by the flexibility of the linker. However, it was recently shown that a rigidly linked fusion protein designed to assemble into a cube, instead formed a variety of symmetries, indicating that the inherent flexibility of proteins may complicate the design of rigid symmetric protein cages.[1] The difficulty of rigid protein cage design most likely increases with the size of the protein cage – to wit, only one example currently exists of a rigid fusion protein construct that assembled into a protein cage with a higher symmetry than a tetrahedron.[2] Therefore, we set out to connect a C_3 symmetric building

block with a C_4 symmetric building block with the end goal to synthesize, purify, and characterize an octahedral protein cage complex.

The initial protein construct, dubbed Oct-1, consisted of a trimeric esterase building block connected to a tetrameric coiled-coil by a 12 residue glycine-rich linker. Oct-1 could be purified and was found to consist of multiple complexes of various sizes, the majority being smaller than the octahedron, and a small percentage of Oct-1 existed as an unbound trimer. Additionally, these complexes were shown to rapidly re-equilibrate to a broad range of complexes following isolation of specific fractions from a size exclusion column. This behavior was assumed to be caused by the tetrameric coiled-coil, which had a low T_m and was judged to be weakly-associating. We replaced this weakly-associating tetrameric coiled-coil with a strongly-associating coiled-coil that has been known to form either a dimer or a trimer, depending on environmental conditions and sequence identity. This construct, dubbed Oct-2, could not be purified in large amounts, but TEM images showed small, randomly-associated complexes of 2-5 trimers.

We then redesigned the flexible linker using Rosetta in collaboration with Dr. Neil King, modeling the two protein building blocks replicated to their respective coordinates in octahedral symmetry. Each set of building blocks was then independently rotated and translated until a configuration was found with the lowest inter-terminus distance between the C-terminus of the C_3 building block and the N-terminus of the C_4 building block. This distance was 9.1 Å, or roughly three residues long. The Oct-2 construct was thus redesigned and three new fusion proteins were expressed with 3, 4, and 5 residues in the linker, termed Oct-3-3, Oct-3-4, and Oct-3-5. All three of these fusion proteins could be purified and each characterized as a

different mixture of complexes between approximately 5 and 16 trimers in size. The complexes formed by these three fusion protein constructs, unlike the previous generations of Oct proteins, were stable in solution and appeared roughly globular by TEM with identifiable symmetry elements.

After discovering the error in the sequence of the coiled-coil in the second and third generations of Oct proteins, we inserted the correct coil that was previously verified to oligomerize into a tightly-bound tetramer. Three fusion proteins were designed and expressed with 2, 3, and 4 residues in the linker, termed Oct-4-2, Oct-4-3, and Oct-4-4. Of these three, Oct-4-3 expressed almost entirely as insoluble protein, and Oct-4-2 formed a mixture of collapsed, globular proteins between approximately 6 and 20 trimers in size. Oct-4-4, however, could be purified into a single species that was shown by 2DSA to be hollow and well-formed, with a molecular weight very close to the predicted molecular weight of an octahedron. Further analysis of Oct-4-4 by TEM showed particles with a narrow size distribution of 17-19 nm, in the range of the size of the complex modeled by Rosetta, and the C_4 symmetry site could be visualized in several of the particles. A cryo-EM electron reconstruction of Oct-4-4 resulted in a visualization of the octahedral complex with moderate resolution, in which the trimeric building blocks are oriented with their termini aligned with the C_4 symmetry axis.

While the individual characterization techniques used in this dissertation do not provide the resolution that can be obtained from crystallography of rigid protein complexes, I have shown that a fusion protein consisting of two symmetry elements connected by a flexible linker can definitively assemble into a single, characterizable protein cage with defined symmetric features. This octahedral fusion protein complex is the first ever *de novo* designed symmetric

cage protein to utilize a C_4 symmetry element and the second *de novo* designed octahedral protein cage to be successfully purified and characterized. This marks the beginning of a new chapter into the design of novel protein cages, since the general method used to create the octahedron described here is in principle applicable to any combination of symmetry elements provided the linker length has been optimized.

Future directions for this project are numerous. The scope of this dissertation has only covered a small range of the possible constructs and conditions that could be used to optimize the assembly of a protein cage. Immediately, one could explore more fully the relationship between linker length and protein cage formation, as all we have done is found the minimum distance required for an octahedron to form. It is an open question as to how many residues can be added to the linker before two coils on a single trimer can self-associate to generate misformed protein complexes that are smaller than the octahedron. This research has interesting implications for designing porous materials, as the pore size of he assembled complex will increase linearly with the linker length.

Another vein of research involves varying the length and strength of the tetrameric coiled-coil. It is well-known that the number of heptads in a coiled-coil is correlated to the association behavior of that coiled-coil, with more heptads inducing a more strongly-bound coil. For medicinal purposes that must be conducted at body temperatures, a coiled-coil that is resistant to heat may be necessary to prevent the therapeutic protein cage from dissociating, and so it is a worthwhile pursuit to investigate a variety of homotetrameric coiled-coils for their binding strength, and to correlate that data with the association behavior of protein cages with those coils attached. It should be noted to this end that the Oct series of proteins as purified

are all unstable at room temperature, falling out of solution as aggregates within a few days, so there is a lot of intellectual space that could be devoted to improving the thermal stability of these designed protein cages.

While all of the soluble Oct proteins have been found to be enzymatically active in a simple kinetic assay involving the de-esterification of para-nitrophenol acetate, the exact effect of cage formation on the enzymatic activity needs to be explored further. Currently there is only one other study that has investigated the effect of cage formation on the kinetic activity of the building block enzyme,[3] and exploring this topic should have implications regarding industrial applications of enzymes, particularly if it were shown that protein cages have properties that made them superior for a specific application than the building block enzymes.

Furthermore, significantly more work has to be done to test the generalizability of this method, as its major benefit is that it can be theoretically applied to any combination of two symmetric proteins, provided the linker length has been sufficiently optimized. Most immediately, we can substitute the tetrameric coiled-coil in Oct-4 for a well-characterized dimeric, trimeric, or a pentameric coiled-coil, to induce formation of prismatic, tetrahedral, or icosahedral protein cages, respectively. Of particular interest is protein cages with icosahedral geometries, as currently there are no examples of a *de novo* designed icosahedral cage, and the large cavity size in the interior of an icosahedron could prove useful for industrial and medical purposes. Additionally, we could look at substituting a large, symmetric protein building block in the place of the coiled-coil, to attempt to create a protein cage with a much greater size than currently purified. We can also look into replacing the esterase trimer with another building

block, as there are two symmetry pairs (C_4+C_2 and C_5+C_2) that are inaccessible by using the trimeric esterase.

Further work could also focus on the design of environmentally-responsive coiled-coils, for example, there are several parallel trimeric coiled-coils that have been designed to dissociate or associate in a pH-dependent manner or have assembly-directing metal binding sites.[4-7] This would allow for the possibility of controllably inducing cage formation, which would have use in various sensing applications as well as for *in vivo* applications that are designed to dissociate upon crossing the cell membrane.

Finally, any and all of these protein cages could be functionalized for any of the purposes described in section 1.2. Controllably altering the linker length or protein building blocks will allow us control of the size of the interior cavity, which has direct implications in the field of nanoparticle formation and other syntheses that proceed on the interior of the protein cage. Designed protein cages can be functionalized to carry targeting or therapeutic molecules, and there exists a strong possibility that these protein cages, by nature of their repeat units, can be used to induce immune responses to either the protein building blocks or to attached epitopes. Though at this moment it is unclear which application would gain the most from using *de novo* designed protein cages over natural protein cages, it is clear that there are a broad range of possibilities that are now able to be explored.

References:

1. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* **6**, 1065-1071 (2014).
2. King, N. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **6085**, 1171-1174 (2012).
3. Patterson, D.P., Desai, A.M., Holl, M.M.B. & Marsh, E.N.G. Evaluation of a symmetry-based strategy for assembling protein complexes. *RSC Adv.* **1**, 1004-1012 (2011).

4. Kiyokawa, T. et al. Binding of Cu(II) or Zn(II) in a de novo designed triple-stranded α-helical coiled-coil toward a prototype for a metalloenzyme. *The Journal of Peptide Research* **63**, 347-353 (2004).
5. Suzuki, K., Yamada, T. & Tanaka, T. Role of the buried glutamate in the alpha-helical coiled coil domain of the macrophage scavenger receptor. *Biochemistry* **38**, 1751-6 (1999).
6. Zimenkov, Y. et al. Rational design of a reversible pH-responsive switch for peptide self-assembly. *J. Am. Chem. Soc.* **128**, 6770-6771 (2006).
7. Suzuki, K., Hiroaki, H., Kohda, D., Nakamura, H. & Tanaka, T. Metal Ion Induced Self-Assembly of a Designed Peptide into a Triple-Stranded α-Helical Bundle: A Novel Metal Binding Site in the Hydrophobic Core. *J. Am. Chem. Soc.* **120**, 13008-13015 (1998).

Appendices

Appendix A

Design of a coiled-coil control system using GFP

A.1 – Design of the GFP fusion construct

As it became increasingly clear that the coil inserted into the Oct-2 and Oct-3 constructs was not oligomerizing as intended, it was deemed necessary to design a control system wherein we could probe the oligomerization state of coiled-coils prior to attaching them to the trimeric esterase. To this end, we designed a fusion protein construct which substituted out the trimeric esterase for a monomeric protein, allowing us to characterize the mass of the resulting protein complex and determine the oligomerization state of the attached coiled-coil. This system is similar to the Oct protein system in that the coiled-coils are formed by *in vivo* expression, whereas the coiled-coils that have been crystallographically verified were synthesized by stitching together amino acids. This is an important difference, as a particular coiled-coil sequence may be toxic to the expression system or be a target for proteolysis. As we can see with Oct-4-3, some designed fusion proteins may be expressed almost exclusively as inclusion bodies despite having optimized components. Therefore, this control system has the additional purpose of screening coiled-coils for their compatibility with the expression vectors used for designed fusion protein cage constructs.

For this study, we used green fluorescent protein (GFP) as a monomer, for two reasons. First, this protein is well-studied and is known to be a monomer at concentrations up to 50 μM. Second, GFP is highly fluorescent, and as such it is trivial to visually confirm that the GFP-coil

fusion protein has expressed, or has bound to a column during purification. We inserted sequences encoding coiled-coils to GFP in two different locations – either at the N-terminus of GFP or at the C-terminus, in both cases with a linker consisting of 6 glycine residues. Full peptide sequences for the constructs are reported in appendix B. Along with monomeric GFP, seven fusion protein constructs were designed with the coiled-coil at the N-terminus, and four fusion protein constructs were designed with the coiled-coil at the C-terminus. These are denoted by the nomenclature "GFP **–** residues at the *a* and *d* positions on the coiled-coil – number of heptad repeats – location of coil", so a GFP fusion construct with a C-terminal coiled-coil identical to the one inserted into Oct-4 is named "GFP-LI-4-C", as the coil in Oct-4 has 4 heptad repeats, and leucine and isoleucine residues at the *a* and *d* positions in its heptad repeat.

The vector that encodes for monomeric GFP is pMCSG18, which is designed for C-terminal addition of GFP to a target nucleotide sequence. Therefore, it has a plethora of restriction sites in the open reading frame on the N-terminal side of the GFP construct, whereas on the C-terminal side of the GFP construct, there are several restriction sites immediately after the stop codon, but none immediately before the stop codon, where we would be inserting the coiled-coil. The closest restriction site that is still within the ORF is 100 nucleotides away from the stop codon. When this project was first designed, we were assembling fusion protein constructs by annealing two single stranded DNA pieces together, generating a dsDNA fragment with complementary overlaps to the double digested vector. Adding a coiled-coil to the C-terminus of pMCSG18 would have required a ssDNA length that was too large (>200 nt) to order from standard suppliers. Thus, we opted to instead insert the coiled-coil at the N-

terminus, situated between a 6xHis tag and a thrombin cleavage site. We designed seven N-terminal coiled-coil constructs: One with the F,F coil motif from Oct-1 (GFP-FF-7-N), one with the L,L coil motif from Oct-2 and Oct-3 with 4 heptads (GFP-LL-4-N), two with the L,I coil motif from Oct-4 with 3 and 4 heptads (GFP-LI-3-N and GFP-LI-4-N), two with a crystallographically verified trimeric I,I coil motif with 4 and 5 heptads (GFP-II-4-N and GFP-II-5-N), and one seven heptad coil identical to the one in Oct-1 but with all the interior residues replaced with tryptophans (GFP-WW-7-N).

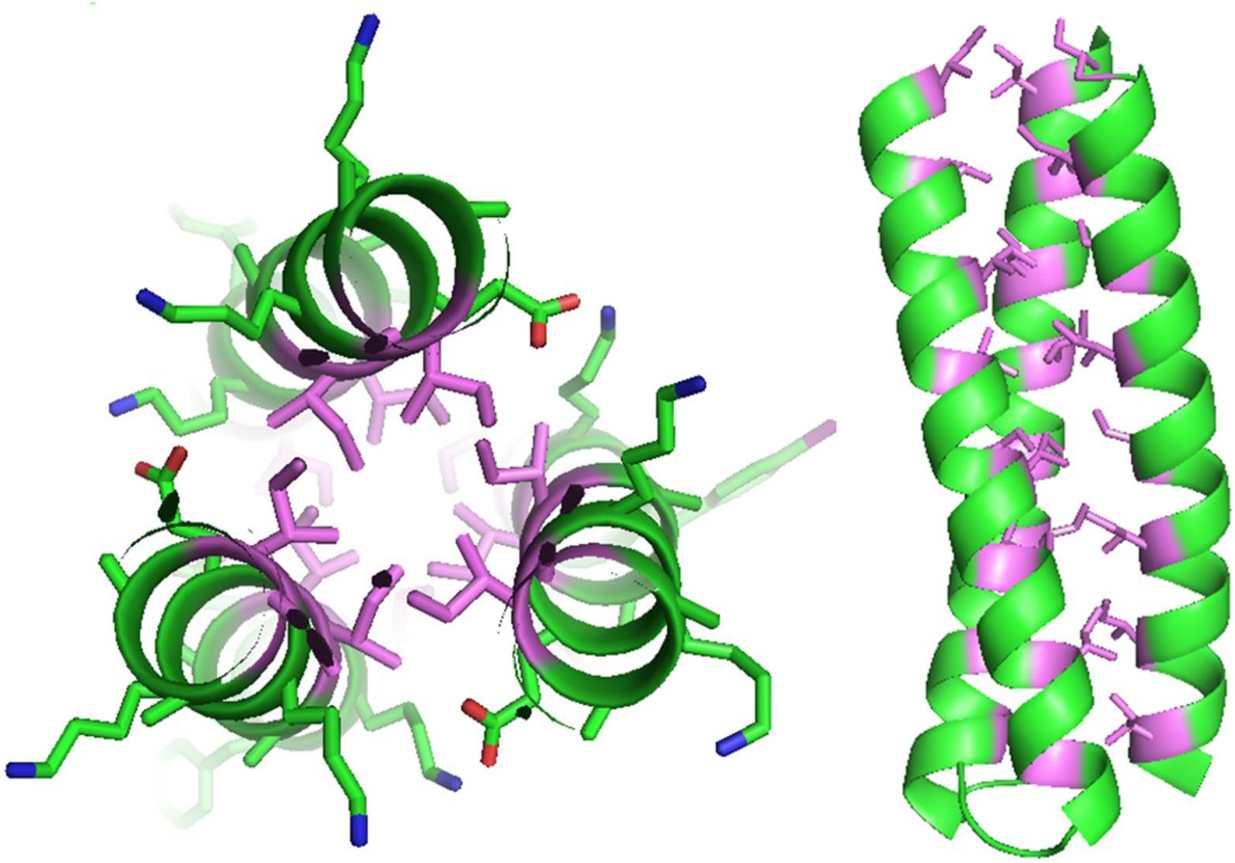


Figure A.1. Crystal structure of the parallel trimeric coiled-coil motif to be inserted into GFP. In this coil (PDB ID 4DZL), the *a* and *d* positions in the heptad repeat consist of isoleucine residues (purple).

Figure A.2. Schematic of design of N-terminal GFP constructs. The His-tag is located on the N-terminus. The crystallographically-determined oligomerization state is in parenthesis. Bolded residues represent oligomerization-determining *a* and *d* positions.

A.2 – Characterization of the N-terminal GFP constructs

All seven fusion protein constructs were transformed into *E.coli* BL21(DE3) cells, and all but the GFP-WW-7-N protein construct expressed as soluble protein, as measured by the green color of the cells after overnight expression. Yields for soluble protein varied from 30-150 mg/L cell media, a vastly increased yield compared to Oct protein constructs. GFP constructs were purified by Ni-affinity and size exclusion chromatography by the same methods as described in chapter 2, except using a Superdex 200 column for size exclusion instead of a Superose 6, in lieu of the difference in molecular weight between GFP and Oct constructs. After purification, GFP constructs were stable at 4 °C for several months.

The six soluble N-terminal GFP constructs could all be purified to remove nearly all contaminants detectable by SDS-PAGE (Figure A3). When these purified complexes are injected onto the size exclusion column, their behavior is curious and unexpected: GFP-LL-4-N, GFP-FF-7-

N, GFP-LI-4-N, and GFP-II-5-N all have similar elution volumes, with no discernable pattern present: GFP-FF-7-N and GFP-LI-4-N are both based off tetrameric motifs, and have elution volumes of 11.7 and 12.1 mL, respectively. GFP-LL-4-N and GFP-II-5-N are both based off trimeric motifs, and have elution volumes of 12.4 and 12.8 mL, respectively. However, GFP-LI-3-N and GFP-II-4-N elute at 14.3 and 15.4 mL, closer to the elution volume of the unmodified GFP of 15.4 mL. While the 3 heptad coiled-coil in GFP-LI-3-N may be expected to not oligomerize, the 4 heptad coiled-coil in GFP-II-4-N is based off a coiled-coil that is well-known to oligomerize into a trimer. That the GFP-II-4-N fusion construct elutes as the same size of the unmodified monomeric GFP is cause for some concern, as this system is supposed to faithfully replicate the oligomerization state of these coiled-coils.

Native PAGE proved to be unwieldy and yielded the same level of accuracy as size exclusion chromatography. Due to time and resource constraints, only a few N-terminal constructs were analyzed by native PAGE, which showed that GFP-LL-4-N and GFP-LI-4-N had similar R_fs, and both fusion constructs ran similarly to the unmodified trimeric esterase used as the building block for the Oct series of proteins (96 kDa). The monomeric GFP (32 kDa), on the other hand, had a similar R_f to BSA (66 kDa). Although native PAGE shows single sharp bands for all three GFP constructs, these results show a total disconnect between protein size and electrophoretic behavior, and as such conditions for native PAGE were not optimized further.

Analytical ultracentrifugation yielded significantly higher resolution data than size exclusion chromatography or native PAGE. After a sedfit analysis, patterns in the oligomerization states for the six GFP constructs could be observed. GFP-LI-3-N and GFP-II-4-N had s-values of 2.0 and 1.7 respectively, close to the s-value of unmodified monomeric GFP of

1.7. This indicates that the coils on these two constructs do not oligomerize at the concentrations (0.2-1.0 mg/ml) tested. Using a fitted frictional ratio of 1.28, the monomeric GFP is calculated to have a molecular weight of 32 kDa, close to its expected molecular weight of 32.4 kDa. GFP-LL-4-N has two peaks, one at 2.0 and one at 2.8 s, indicating that this construct may be weakly bound and exists as both a monomer and an oligomer. Using a fitted frictional ratio of 1.24, this species at 2.8 s correlates to a molecular weight of 67.4 kDa, which is close to the hypothetical molecular weight of a GFP dimer (70 kDa). GFP-LI-4-N and GFP-II-5-N have similar s-values of 3.1 and 3.2, which calculate to molecular weights of 85.2 and 87.7 kDa using the fitted frictional ratio of 1.27. This is also a highly unusual result, as it suggests that the LI coil motif, long known to oligomerize as a tetramer, actually oligomerizes as a trimer when attached to GFP. Finally, GFP-FF-7-N has an s-value of 4.2, which calculates to a molecular mass of 131 kDa using a fitted frictional ratio of 1.31. This is in line with what is expected from a tetrameric coiled-coil.

We further investigated the curious oligomerization states of GFP-coil constructs by IM-MS, as this technique should give much more accurate masses that do not rely on assumptions about frictional ratios. Four constructs were re-purified by SEC into ammonium acetate buffer and analyzed by IM-MS as described in Chapter 2. Two of these four constructs – GFP-mono and GFP-II-4-N had masses of 28 and 32 kDa, consistent with monomers. The third construct, GFP-LL-4-N, had a mass of 61-66 kDa, consistent with a dimer. The fourth construct, GFP-LI-4-N, had a measured mass of 101-106 kDa, consistent with a trimer.

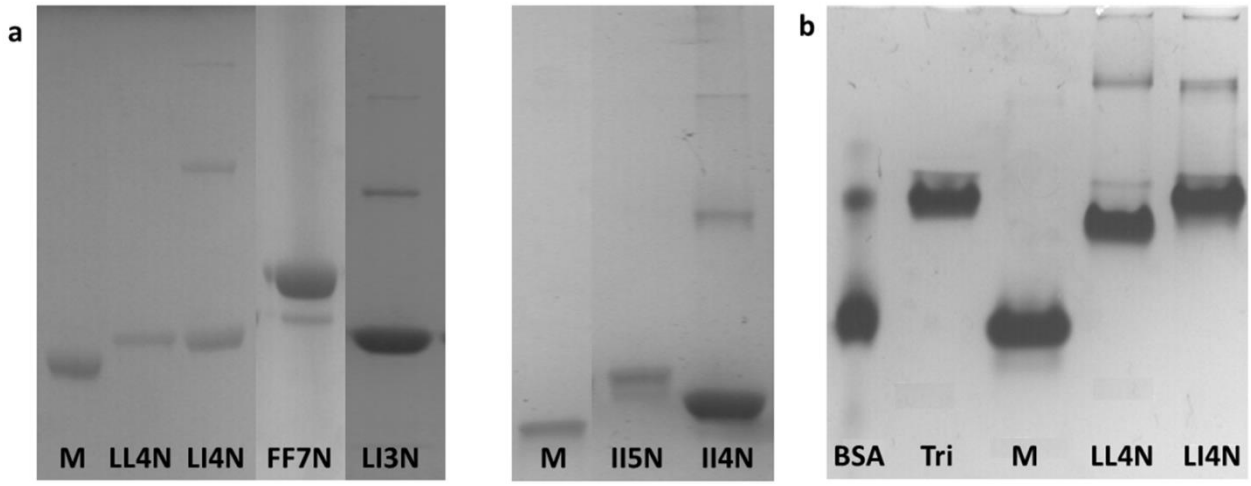


Figure A.3 Purification of N-terminal GFP constructs. a) SDS-PAGE shows that all six soluble N-terminal GFP constructs could be purified to high purity. B) Native PAGE shows that the three N-terminal constructs tested migrate as mainly a single band, but the migration patterns have little relation with the expected molecular weight.

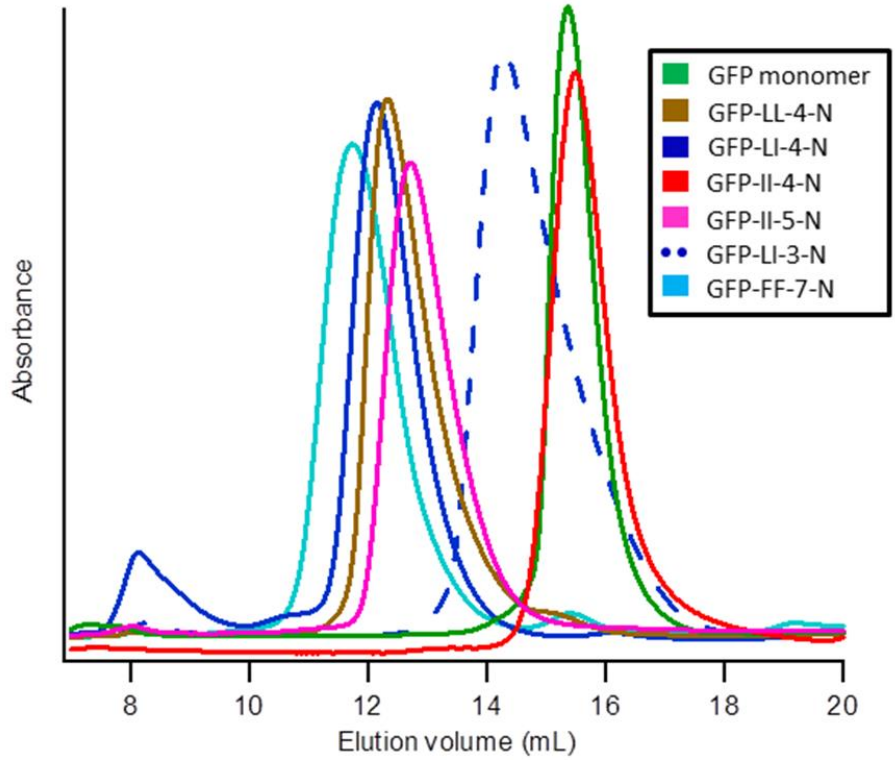


Figure A.4. Size exclusion chromatography profiles of N-terminal GFP constructs. Elution volumes of the GFP constructs can be roughly divided into two clusters – the first comprised of GFP-FF-7-N (light blue), GFP-LI-4-N (dark blue), GFP-LL-4-N (brown), and GFP-II-5-N (magenta),

and the second comprised of monomeric GFP (green), GFP-LI-3-N (blue dashed), and GFP-II-4-N (red).

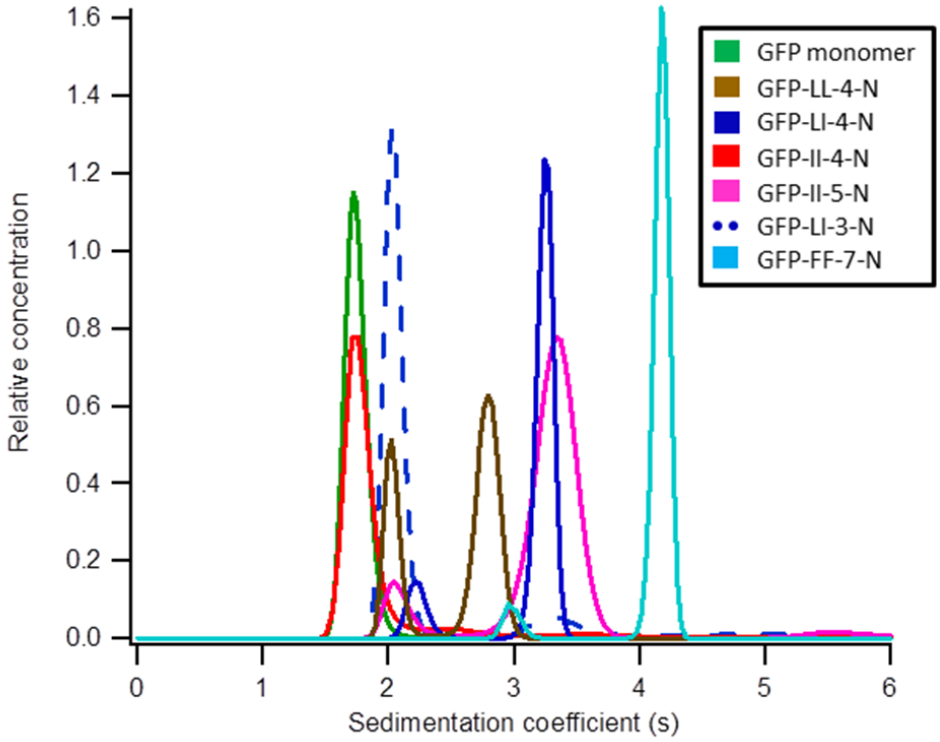


Figure A.5. Analytical ultracentrifugation of N-terminal GFP constructs. Constructs have sedimentation coefficients corresponding to one of four oligomerization states. GFP-LI-3-N (blue dashed) and GFP-II-4-N (red) both have s-values similar to the monomeric GFP (green), GFP-LL-4-N (brown) sediments as a dimer, GFP-LI-4-N (blue) and GFP-II-5-N (magenta) both sediment as trimers, and GFP-FF-7-N (light blue) sediments as a tetramer.

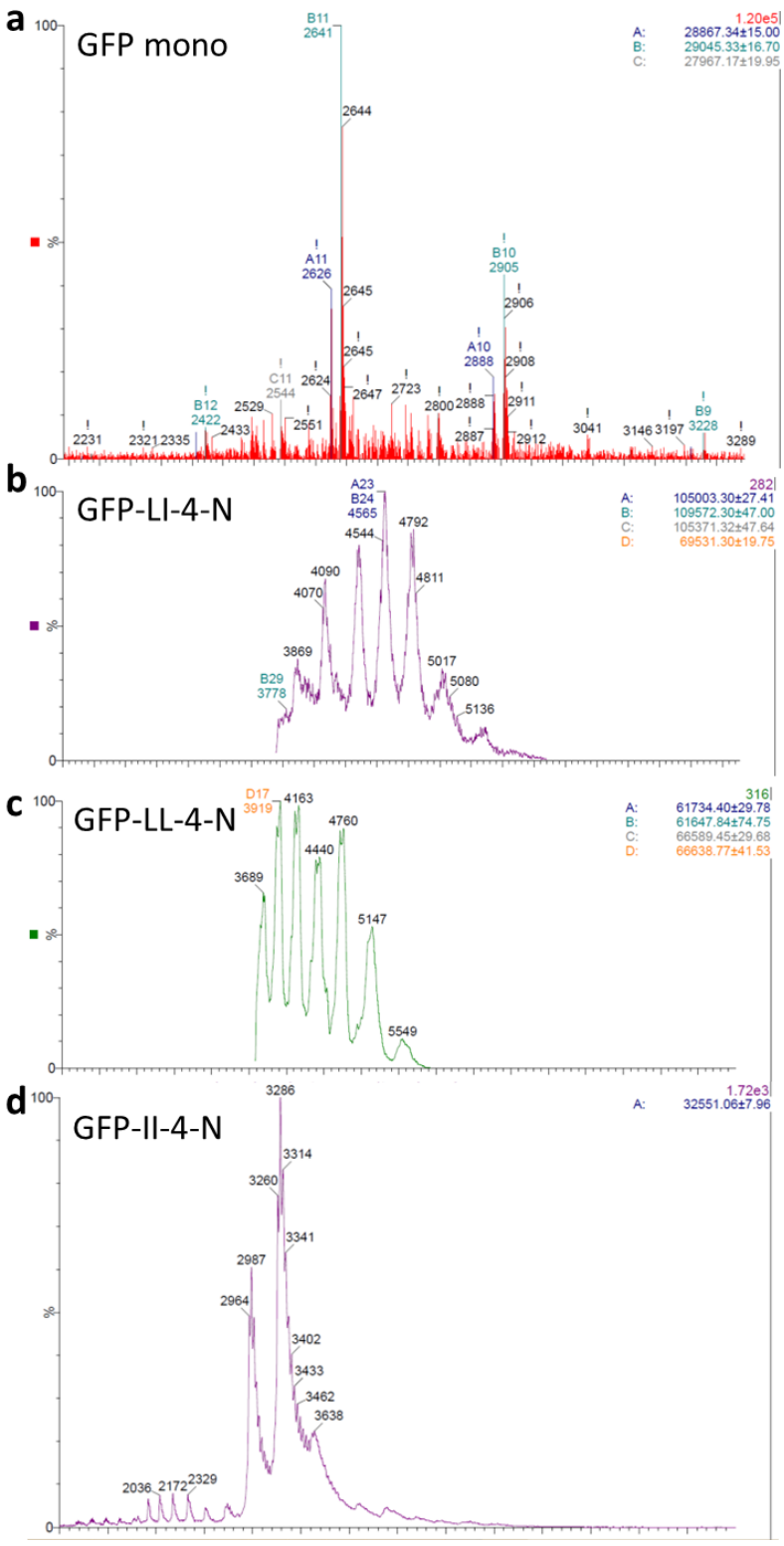


Figure A.6. Ion mobility-mass spectrometry of N-terminal GFP constructs. Correlative with AUC data, GFP-II-4-N (d) has a native mass similar to the monomeric GFP (a), while GFP-LL-4-N (c)

has a native mass of around two subunits and GFP-LI-4-N (b) has a native mass of around three subunits.

A.3 – Design and Characterization of C-terminal GFP Fusion Protein Constructs

After getting unexpected oligomerization states from multiple constructs, we decided to re-attach the coiled-coil onto the C-terminus of GFP to determine if these unexpected results were caused by the coil being too close to the His-tag, or were independent of coil location. Using Gibson assembly, we were able to design larger dsDNA sequences that retained the 120 nucleotides that were excised from the C-terminal double digestion of pMCSG18, in addition to the glycine linker and coiled-coil insert. Four constructs were designed, one with the L,L motif and 4 heptads (GFP-LL-4-C), one with the L,I motif and 4 heptads (GFP-LI-4-C), and two with the I,I motif and either 4 or 5 heptads (GFP-II-4-C and GFP-II-5-C). DNA encoding these constructs was successfully ligated and transformed into BL21(DE3) *E.coli* cells, and purified as above.

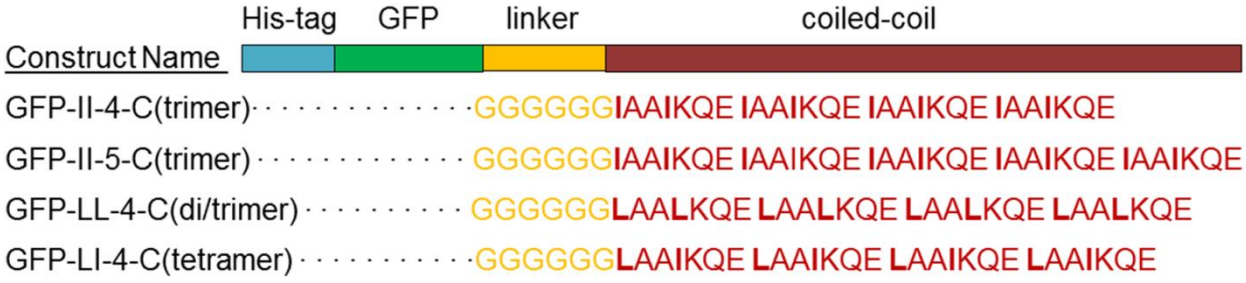


Figure A.7. Schematic of design of C-terminal GFP constructs. The His-tag is located on the N-terminus. The crystallographically-determined oligomerization state of the attached coils is in parenthesis.Bolded residues represent oligomerization-determining *a* and *d* positions.

All four of these constructs could be purified to remove all contaminants detectable by SDS-PAGE. Analysis of SEC-purified GFP constructs by size exclusion showed oligomerization in all four constructs. The monomeric GFP had an elution volume of 16.5 mL, while GFP-LL-4-C had an elution volume of 13.9 mL, and the other three constructs had elution volumes of 13.3 mL.

These data are backed up by native PAGE, which shows that GFP-LL-4-C migrates between monomeric GFP and the three other C-terminal GFP constructs, although just like the native PAGE of N-terminal GFP constructs, the oligomerizing constructs migrate very close to each other, with no relation to the migration patterns of standard proteins. Sedfit analysis of analytical ultracentrifugation data confirms these oligomerization states. GFP-LL-4-C has an s-value of 2.8, identical to the s-value of GFP-LL-4-N, while the other three C-terminal constructs all have major peaks with identical s-values of 3.6. These almost certainly correlate to a dimer and three trimers, which was confirmed by IM-MS analysis. IM-MS analysis of the four C-terminal constructs showed that GFP-LI-4-C, GFP-II-4-C, and GFP-II-5-C all had molecular masses between 97 and 99 kDa, consistent with formation of a trimer, while GFP-LL-4-C had multiple m/z peaks, the major peak correlated to the dimeric species, and smaller signals correlating to the monomeric and the trimeric species. This is most likely an artifact from the process of native mass spectrometry, as minor peaks correlating to larger or smaller species can be found in all spectra.

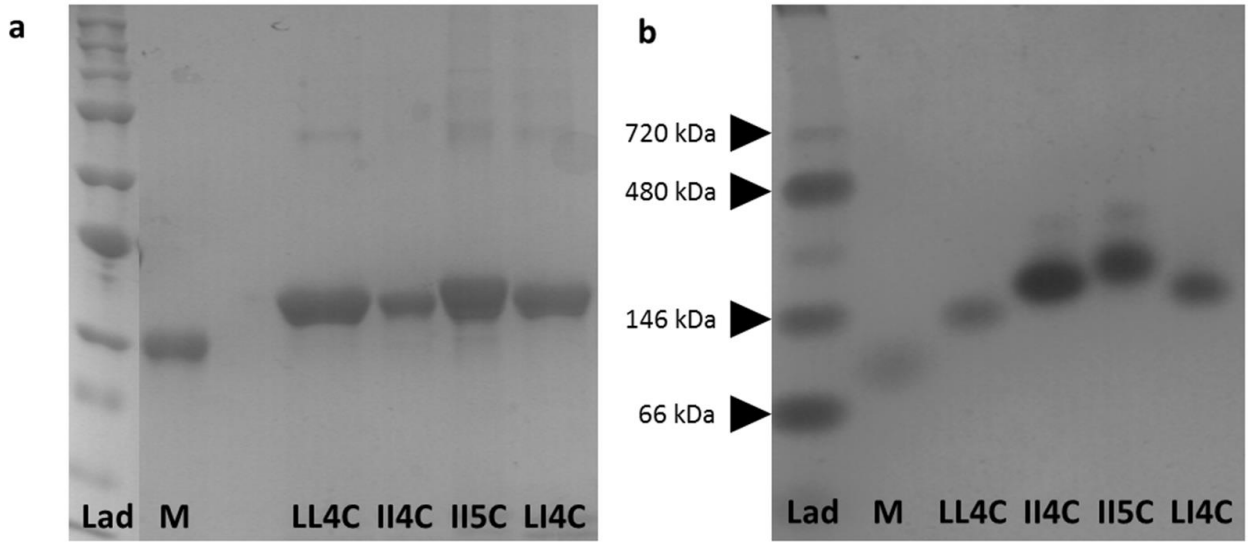


Figure A.8. Purification of C-terminal GFP fusion constructs. a) SDS-PAGE shows that all four C-terminal GFP constructs could be purified to high purity. B) Native PAGE shows that all four C-terminal constructs migrate as mainly a single band, but the migration patterns have little relation with the expected molecular weight.

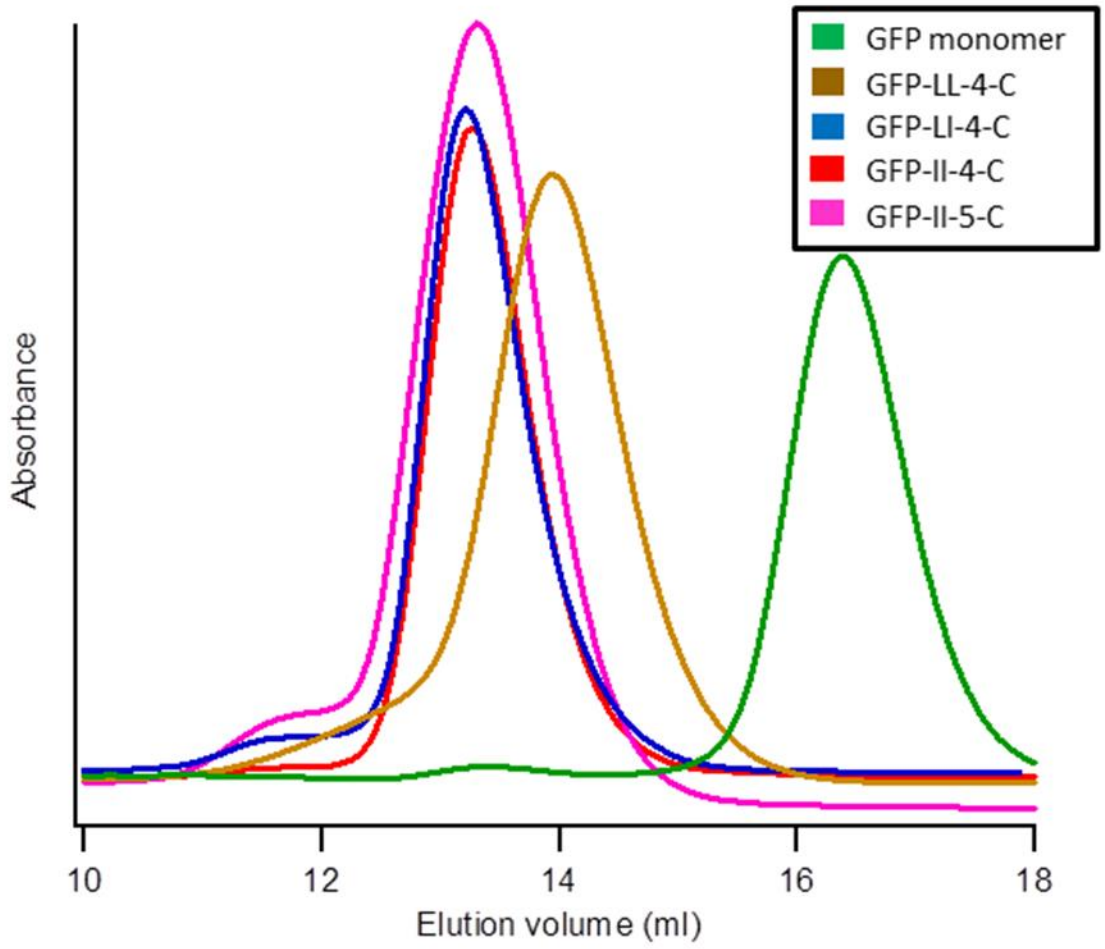


Figure A.9. Size exclusion chromatography of C-terminal GFP constructs. All four constructs have sharp peaks and significantly smaller elution volumes than the monomer, with GFP-LL-4-C (brown) eluting at 13.9 mL, and the other three constructs eluting at 13.3 mL.

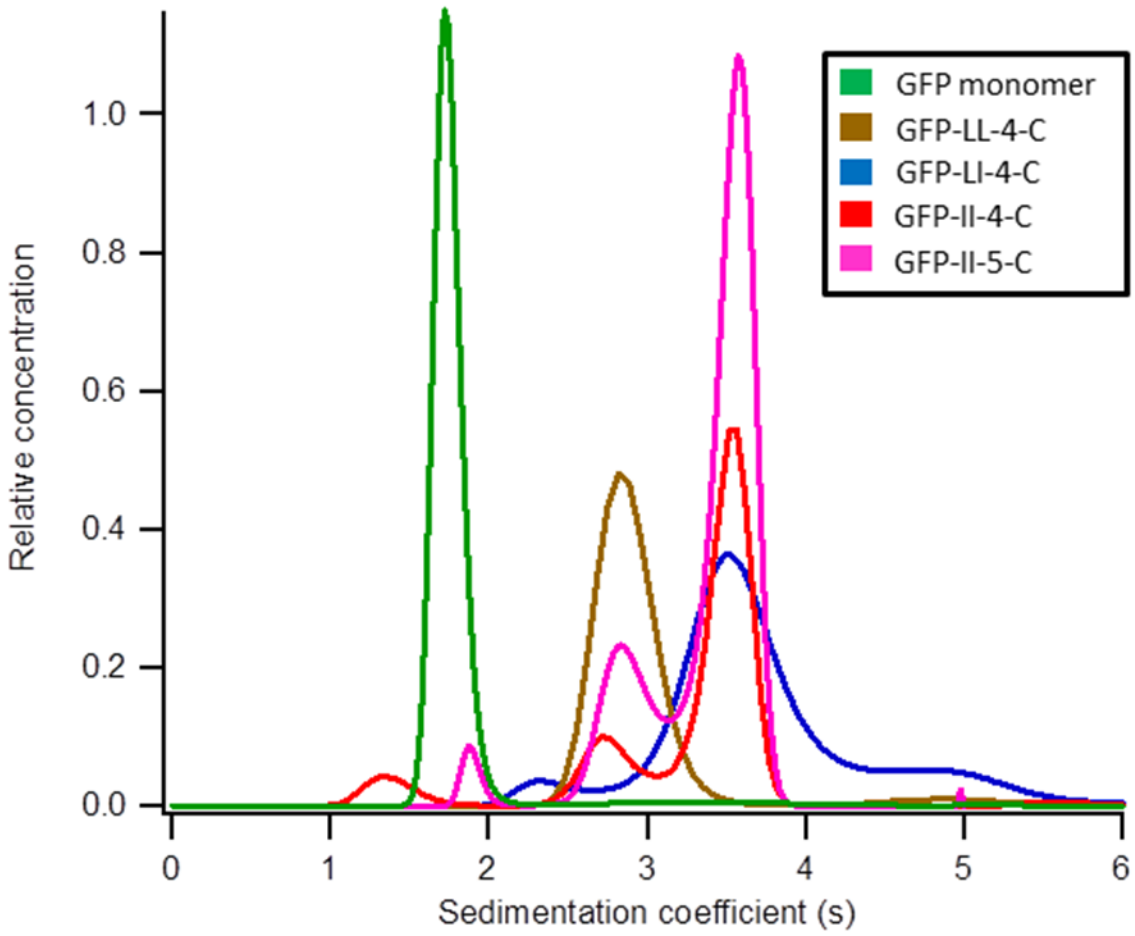


Figure A.10. Analytical ultracentrifugation of C-terminal GFP constructs. Correlative with results from size exclusion chromatography, GFP-LL-4-C sediments as a dimer, while the other three constructs sediment mainly as trimers.

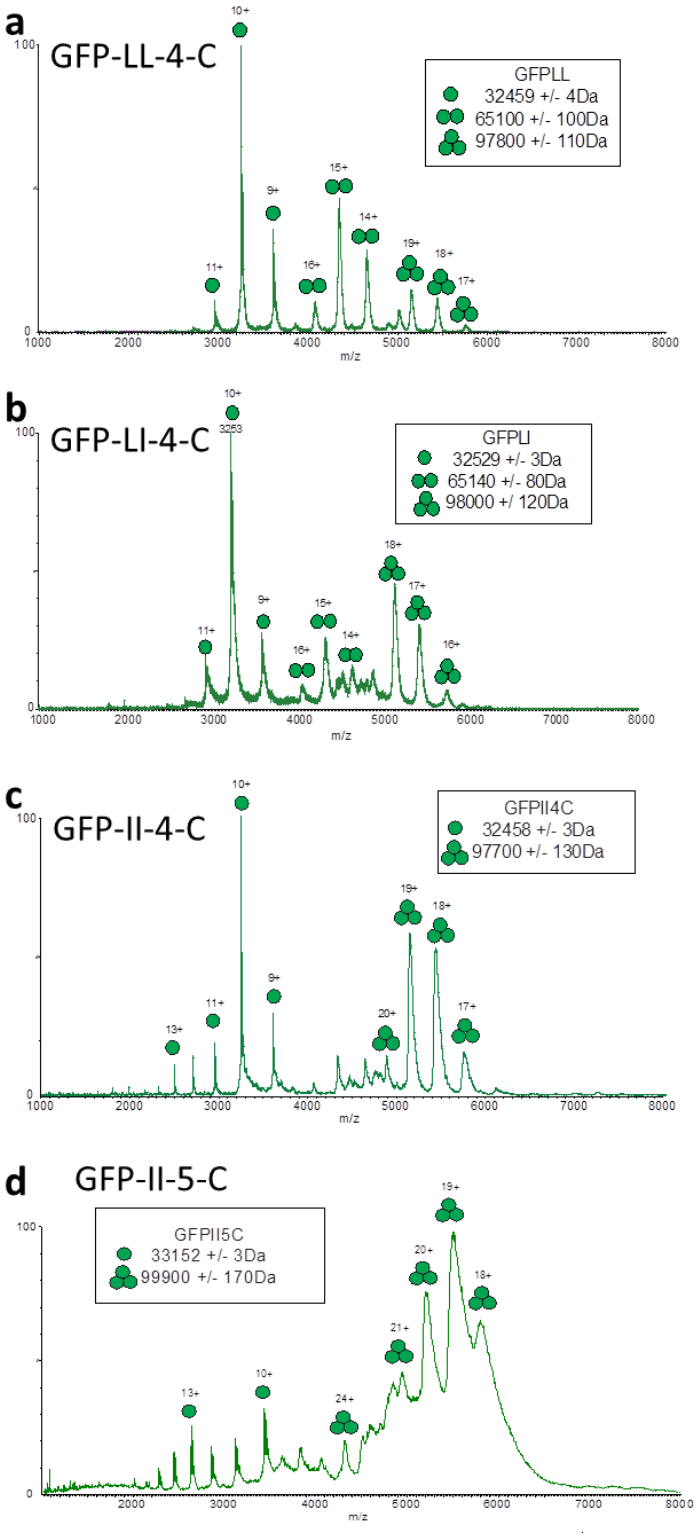


Figure A.11. Ion mobility-mass spectrometry of C-terminal GFP constructs. Though other species are present the major peak in GFP-LL-4-C (a) corresponds to a dimer, while the other three GFP constructs (b-d) have major peaks corresponding to trimers.

A.4 - Conclusions

The goal of creating a GFP-coil fusion construct was to develop a system that would test and confirm the oligomerization of coiled-coils that have been previously characterized exclusively *in vitro*. It was hypothesized that the attachment of these coils onto a protein or the biological expression of this fusion construct might effect a change in the coils' oligomerization states. In this chapter I designed and characterized eleven different GFP fusion constructs, with varying degrees of success. Many of the GFP constructs tested had different oligomerization states than were expected for the attached coiled-coil, in particular GFP-LI-4, which was trimeric when attached to both the C- and N-terminus. This contradicts multiple crystal structures of coiled-coils with leucine and isoleucine residues at the *a* and *d* positions, as well as cryo-EM data from the Oct-4-4 construct that shows an octahedron formation with a C_4 axis at the location of the coiled-coil. Additionally, it was shown that this system is sensitive to small perturbations – while GFP-II-4-N could be shown to not oligomerize at all, adding an extra heptad (GFP-II-5-N) or relocating the coil to the C terminus (GFP-II-4-C) induced trimerization of the coiled-coil.

The major success of this experiment was to demonstrate a robust method for characterizing the oligomerization states of fusion constructs, which will undoubtedly be useful for characterization of more exotic coiled-coils, for example, a coil whose association can be controlled by environmental conditions such as metal availability or pH. With this system, we could ascertain the extent of association at various environmental conditions, and thus determine if a coil is appropriate to attach to the trimeric esterase to yield a protein cage with the desired properties.

Appendix B

DNA and protein sequences of protein building blocks and fusion constructs

In this appendix, I will detail the DNA and protein sequences of each protein construct that was analyzed in the previous thesis and appendix. Restriction sites in the open reading frames of the DNA are noted where applicable. All Oct proteins were ligated into the pet28b vector and all GFP proteins were ligated into the pMCSG18 vector.

Oct-1 DNA sequence:

```
1       ---------+---------+---------+---------+---------+---------+    60
        atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
61      ---------+---------+---------+---------+---------+---------+   120
        atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
121     ---------+---------+---------+---------+---------+---------+   180
        gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
181     ---------+---------+---------+---------+---------+---------+   240
        cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
241     ---------+---------+---------+---------+---------+---------+   300
        ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
301     ---------+---------+---------+---------+---------+---------+   360
        gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
361     ---------+---------+---------+---------+---------+---------+   420
        ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
421     ---------+---------+---------+---------+---------+---------+   480
        atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
481     ---------+---------+---------+---------+---------+---------+   540
        tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
541     ---------+---------+---------+---------+---------+---------+   600
        gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
601     ---------+---------+---------+---------+---------+---------+   660
        atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
661     ---------+---------+---------+---------+---------+---------+   720
        gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
721     ---------+---------+---------+---------+---------+---------+   780
        gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
781     ---------+---------+---------+---------+---------+---------+   840
        aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
841     ---------+---------+---------+---------+---------+---------+   900
        acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggtggc
                                                          KpnI
901     ---------+---------+---------+---------+---------+---------+   960
        ggtggcggtggcggtggcggtactagttccaacgcaaaatttgaccagttctcatcggat
                           SpeI
961     ---------+---------+---------+---------+---------+---------+  1020
        tttcaaaccttcaatgcgaaatttgaccagttcagtaacgatatgaatgcctttcgttcc
1021    ---------+---------+---------+---------+---------+---------+  1080
        gattttcaggcatttaaagacgattttgctcgtttcaaccaacgctttgataatttcgcg
```

```
 1081    ---------+---------+---------+---------+---------+---------+    1140
         accaaatatcgctaatttaaatagggatcc
                                BamHI
```

Protein sequence for Oct-1:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGG
       310        320        330        340        350        360
GGGGGGGTSS NAKFDQFSSD FQTFNAKFDQ FSNDMNAFRS DFQAFKDDFA RFNQRFDNFA
```

Calculated molecular weight: 39,842 kDa

Oct-2 DNA sequence:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
 61      ---------+---------+---------+---------+---------+---------+    120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
 121     ---------+---------+---------+---------+---------+---------+    180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
 181     ---------+---------+---------+---------+---------+---------+    240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
 241     ---------+---------+---------+---------+---------+---------+    300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
 301     ---------+---------+---------+---------+---------+---------+    360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
 361     ---------+---------+---------+---------+---------+---------+    420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
 421     ---------+---------+---------+---------+---------+---------+    480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
 481     ---------+---------+---------+---------+---------+---------+    540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
 541     ---------+---------+---------+---------+---------+---------+    600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
 601     ---------+---------+---------+---------+---------+---------+    660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
 661     ---------+---------+---------+---------+---------+---------+    720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
 721     ---------+---------+---------+---------+---------+---------+    780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
 781     ---------+---------+---------+---------+---------+---------+    840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
 841     ---------+---------+---------+---------+---------+---------+    900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggtggc
                                                           KpnI
 901     ---------+---------+---------+---------+---------+---------+    960
         ggtggcggtggcggtggcggtactaggctggcggccctgaagcaggaactggcagctctg
 961     ---------+---------+---------+---------+---------+---------+    1020
         Cggtccgaactggcagcactgaagcacgagctggcagctctgaagcaagatggc
```

Protein sequence for Oct-2:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGG
       310        320        330
GGGGGGGTRL AALKQELAAL RSELAALKHE LAALKQDG
```

Calculated molecular weight of Oct-2: 36,310 Da

DNA sequence for trimeric esterase:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
 61      ---------+---------+---------+---------+---------+---------+   120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
121      ---------+---------+---------+---------+---------+---------+   180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
181      ---------+---------+---------+---------+---------+---------+   240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
241      ---------+---------+---------+---------+---------+---------+   300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
301      ---------+---------+---------+---------+---------+---------+   360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
361      ---------+---------+---------+---------+---------+---------+   420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
421      ---------+---------+---------+---------+---------+---------+   480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
481      ---------+---------+---------+---------+---------+---------+   540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
541      ---------+---------+---------+---------+---------+---------+   600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
601      ---------+---------+---------+---------+---------+---------+   660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
661      ---------+---------+---------+---------+---------+---------+   720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
721      ---------+---------+---------+---------+---------+---------+   780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
781      ---------+---------+---------+---------+---------+---------+   840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
841      ---------+---------+---------+---------+---------+---------+   900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtacc
```

Protein sequence for trimeric esterase:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGT
```

Calculated molecular weight of trimeric esterase: 32485 Da

DNA sequence for Oct-3-3:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
 61      ---------+---------+---------+---------+---------+---------+    120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
 121     ---------+---------+---------+---------+---------+---------+    180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgc
 181     ---------+---------+---------+---------+---------+---------+    240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
 241     ---------+---------+---------+---------+---------+---------+    300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
 301     ---------+---------+---------+---------+---------+---------+    360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
 361     ---------+---------+---------+---------+---------+---------+    420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
 421     ---------+---------+---------+---------+---------+---------+    480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
 481     ---------+---------+---------+---------+---------+---------+    540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
 541     ---------+---------+---------+---------+---------+---------+    600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
 601     ---------+---------+---------+---------+---------+---------+    660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
 661     ---------+---------+---------+---------+---------+---------+    720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
 721     ---------+---------+---------+---------+---------+---------+    780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
 781     ---------+---------+---------+---------+---------+---------+    840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
 841     ---------+---------+---------+---------+---------+---------+    900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggcctg
                                                          KpnI
 901     ---------+---------+---------+---------+---------+---------+    960
         gcagcactgcggtccgaactggccgcactgaagcaggaactggcggccctgaaacaagaa
 961     ---------+---------+---------+---------+---------+---------+    1020
         Ctggcagctctgaagcaagatggataag
```

Protein sequence for Oct-3-3:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGL
       310        320
AALRSELAAL KQELAALKQE LAALKQDG
```

Calculated molecular weight for Oct-3-3: 35588 Da.

DNA sequence for Oct-3-4:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
 61      ---------+---------+---------+---------+---------+---------+    120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
 121     ---------+---------+---------+---------+---------+---------+    180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
 181     ---------+---------+---------+---------+---------+---------+    240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
 241     ---------+---------+---------+---------+---------+---------+    300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
 301     ---------+---------+---------+---------+---------+---------+    360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
 361     ---------+---------+---------+---------+---------+---------+    420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
 421     ---------+---------+---------+---------+---------+---------+    480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
 481     ---------+---------+---------+---------+---------+---------+    540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
 541     ---------+---------+---------+---------+---------+---------+    600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
 601     ---------+---------+---------+---------+---------+---------+    660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
 661     ---------+---------+---------+---------+---------+---------+    720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
 721     ---------+---------+---------+---------+---------+---------+    780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
 781     ---------+---------+---------+---------+---------+---------+    840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
 841     ---------+---------+---------+---------+---------+---------+    900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggcggt
                                                          KpnI
 901     ---------+---------+---------+---------+---------+---------+    960
         ctggcagcactgcggtccgaactggccgcactgaagcaggaactggcggccctgaaacaa
 961     ---------+---------+---------+---------+---------+---------+    1020
         Gaactggcagctctgaagcaagatggataag
```

Protein sequence for Oct-3-4:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGG
       310        320
LAALRSELAA LKQELAALKQ ELAALKQDG
```

Calculated molecular weight for Oct-3-4: 35645 Da.

DNA sequence for Oct-3-5:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
 61      ---------+---------+---------+---------+---------+---------+   120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
121      ---------+---------+---------+---------+---------+---------+   180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
181      ---------+---------+---------+---------+---------+---------+   240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
241      ---------+---------+---------+---------+---------+---------+   300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
301      ---------+---------+---------+---------+---------+---------+   360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
361      ---------+---------+---------+---------+---------+---------+   420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
421      ---------+---------+---------+---------+---------+---------+   480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
481      ---------+---------+---------+---------+---------+---------+   540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
541      ---------+---------+---------+---------+---------+---------+   600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
601      ---------+---------+---------+---------+---------+---------+   660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
661      ---------+---------+---------+---------+---------+---------+   720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
721      ---------+---------+---------+---------+---------+---------+   780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
781      ---------+---------+---------+---------+---------+---------+   840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
841      ---------+---------+---------+---------+---------+---------+   900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggcggt
                                                          KpnI
901      ---------+---------+---------+---------+---------+---------+   960
         ggcctggcagcactgcggtccgaactggccgcactgaagcaggaactggcggccctgaaa
961      ---------+---------+---------+---------+---------+---------+  1020
         caagaactggcagctctgaagcaagatggataag
```

Protein sequence for Oct-3-5:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGG
       310        320        330
GLAALRSELA ALKQELAALK QELAALKQDG
```

Calculated molecular weight for Oct-3-5: 35702 Da.

DNA sequence for Oct-4-2:

```
 1      ---------+---------+---------+---------+---------+---------+    60
        atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                              NdeI
 61     ---------+---------+---------+---------+---------+---------+   120
        atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
121     ---------+---------+---------+---------+---------+---------+   180
        gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
181     ---------+---------+---------+---------+---------+---------+   240
        cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
241     ---------+---------+---------+---------+---------+---------+   300
        ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
301     ---------+---------+---------+---------+---------+---------+   360
        gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
361     ---------+---------+---------+---------+---------+---------+   420
        ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
421     ---------+---------+---------+---------+---------+---------+   480
        atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
481     ---------+---------+---------+---------+---------+---------+   540
        tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
541     ---------+---------+---------+---------+---------+---------+   600
        gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
601     ---------+---------+---------+---------+---------+---------+   660
        atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
661     ---------+---------+---------+---------+---------+---------+   720
        gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
721     ---------+---------+---------+---------+---------+---------+   780
        gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
781     ---------+---------+---------+---------+---------+---------+   840
        aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
841     ---------+---------+---------+---------+---------+---------+   900
        acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccctggca
                                                        KpnI
901     ---------+---------+---------+---------+---------+---------+   960
        gcaatcaagtccgaactggccgcaatcaagcaggaactggcggccatcaaacaagaactg
961     ---------+---------+---------+---------+---------+---------+  1020
        gcagctatcaagcaagatgga
```

Protein sequence for Oct-4-2:

```
         10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
         70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
        130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
        190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
        250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTLA
        310        320
AIKSELAAIK QELAAIKQEL AAIKQDG
```

Calculated molecular weight for Oct-4-2: 35502 Da.

DNA sequence for Oct-4-3:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                              NdeI
 61      ---------+---------+---------+---------+---------+---------+   120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
 121     ---------+---------+---------+---------+---------+---------+   180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
 181     ---------+---------+---------+---------+---------+---------+   240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
 241     ---------+---------+---------+---------+---------+---------+   300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
 301     ---------+---------+---------+---------+---------+---------+   360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
 361     ---------+---------+---------+---------+---------+---------+   420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
 421     ---------+---------+---------+---------+---------+---------+   480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
 481     ---------+---------+---------+---------+---------+---------+   540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
 541     ---------+---------+---------+---------+---------+---------+   600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
 601     ---------+---------+---------+---------+---------+---------+   660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
 661     ---------+---------+---------+---------+---------+---------+   720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
 721     ---------+---------+---------+---------+---------+---------+   780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
 781     ---------+---------+---------+---------+---------+---------+   840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
 841     ---------+---------+---------+---------+---------+---------+   900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggcctg
                                                        KpnI
 901     ---------+---------+---------+---------+---------+---------+   960
         gcagcaatcaagtccgaactggccgcaatcaagcaggaactggcggccatcaaacaagaa
 961     ---------+---------+---------+---------+---------+---------+  1020
         ctggcagctatcaagcaagatgga
```

Protein sequence for Oct-4-3:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGL
       310        320
AAIKSELAAI KQELAAIKQE LAAIKQDG
```

Calculated molecular weight for Oct-4-3: 35560 Da.

DNA sequence for Oct-4-4:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgggcagcagccatcatcatcatcatcacagcagcggcctggtgccgcgcggcagccat
                                                                NdeI
 61      ---------+---------+---------+---------+---------+---------+   120
         atgagttatgtcaccacgaaagatggcgtgcagatcttttataaagactggggtccgcgt
 121     ---------+---------+---------+---------+---------+---------+   180
         gatgcgccggtgatccatttccatcacggttggccgctgtccgcagatgactgggatgca
 181     ---------+---------+---------+---------+---------+---------+   240
         cagctgctgtttttcctggcgcacggttatcgtgtggttgcacatgaccgtcgcggtcac
 241     ---------+---------+---------+---------+---------+---------+   300
         ggtcgtagctctcaagtctgggatggccatgacatggatcactacgcggatgacgttgcg
 301     ---------+---------+---------+---------+---------+---------+   360
         gccgtcgtggcacatctgggcattcagggtgctgtgcatgttggtcactctaccggcggt
 361     ---------+---------+---------+---------+---------+---------+   420
         ggcgaagttgtccgttatatggcccgccacccggaagataaagttgcgaaagcagtcctg
 421     ---------+---------+---------+---------+---------+---------+   480
         atcgcagctgtgccgccgctgatggttcaaacgccgggtaacccgggtggcctgccgaaa
 481     ---------+---------+---------+---------+---------+---------+   540
         tcagtgtttgacggtttccaggcgcaagttgcctcgaatcgtgcacagttttaccgcgat
 541     ---------+---------+---------+---------+---------+---------+   600
         gtgccggctggcccgttctatggttacaaccgtccgggcgttgaagcaagcgaaggcatt
 601     ---------+---------+---------+---------+---------+---------+   660
         atcggtaattggtggcgccagggcatgattggtagcgcaaaagctcattatgatggtatc
 661     ---------+---------+---------+---------+---------+---------+   720
         gtggctttttctcaaaccgacttcacggaagatctgaaaggcattcagcaaccggtcctg
 721     ---------+---------+---------+---------+---------+---------+   780
         gtgatgcatggtgatgacgatcagatcgttccgtacgaaaacagcggcgtcctgtctgcg
 781     ---------+---------+---------+---------+---------+---------+   840
         aaactgctgccgaatggtgccctgaaaacctataaaggctacccgcatggtatgccgacc
 841     ---------+---------+---------+---------+---------+---------+   900
         acgcacgccgacgttattaacgcagatctgctggcttttatccgcagtggtaccggcggt
                                                          KpnI
 901     ---------+---------+---------+---------+---------+---------+   960
         ctggcagcaatcaagtccgaactggccgcaatcaagcaggaactggcggccatcaaacaa
 961     ---------+---------+---------+---------+---------+---------+  1020
         gaactggcagctatcaagcaagatgga
```

Protein sequence for Oct-4-4:

```
        10         20         30         40         50         60
MGSSHHHHHH SSGLVPRGSH MSYVTTKDGV QIFYKDWGPR DAPVIHFHHG WPLSADDWDA
        70         80         90        100        110        120
QLLFFLAHGY RVVAHDRRGH GRSSQVWDGH DMDHYADDVA AVVAHLGIQG AVHVGHSTGG
       130        140        150        160        170        180
GEVVRYMARH PEDKVAKAVL IAAVPPLMVQ TPGNPGGLPK SVFDGFQAQV ASNRAQFYRD
       190        200        210        220        230        240
VPAGPFYGYN RPGVEASEGI IGNWWRQGMI GSAKAHYDGI VAFSQTDFTE DLKGIQQPVL
       250        260        270        280        290        300
VMHGDDDQIV PYENSGVLSA KLLPNGALKT YKGYPHGMPT THADVINADL LAFIRSGTGG
       310        320
LAAIKSELAA IKQELAAIKQ ELAAIKQDG
```

Calculated molecular weight for Oct-4-4: 35617 Da.

DNA sequence for GFP-mono:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcatcattcttctggtgtagatctgggtaccgagaacctgtacttc
                                        BglII  KpnI
 61      ---------+---------+---------+---------+---------+---------+    120
         caatccaatattggaagtggattactggctagcaaaggagaagaactcttcactggagtt
 121     ---------+---------+---------+---------+---------+---------+    180
         gtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtgga
 181     ---------+---------+---------+---------+---------+---------+    240
         gagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggc
 241     ---------+---------+---------+---------+---------+---------+    300
         aaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttt
 301     ---------+---------+---------+---------+---------+---------+    360
         tcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggt
 361     ---------+---------+---------+---------+---------+---------+    420
         tatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaa
 421     ---------+---------+---------+---------+---------+---------+    480
         gtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaag
 481     ---------+---------+---------+---------+---------+---------+    540
         gaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgtatac
 541     ---------+---------+---------+---------+---------+---------+    600
         atcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacatt
 601     ---------+---------+---------+---------+---------+---------+    660
         gaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggc
 661     ---------+---------+---------+---------+---------+---------+    720
         cctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatccc
                                                          BstBI
 721     ---------+---------+---------+---------+---------+---------+    780
         aacgaaaagagagaccacatggtccttcttgagtttgtaacagctgctgggattacacat
 781     ---------+---------+---------+---------+---------+---------+    840
         ggcatggatgaactgtacaactga
```

Protein sequence for GFP-mono:

```
        10         20         30         40         50         60
MHHHHHHSSG VDLGTENLYF QSNIGSGLLA SKGEELFTGV VPILVELDGD VNGHKFSVSG
        70         80         90        100        110        120
EGEGDATYGK LTLKFICTTG KLPVPWPTLV TTLCYGVQCF SRYPDHMKRH DFFKSAMPEG
       130        140        150        160        170        180
YVQERTIFFK DDGNYKTRAE VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY
       190        200        210        220        230        240
IMADKQKNGI KVNFKTRHNI EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP
       250        260
NEKRDHMVLL EFVTAAGITH GMDELYN
```

Calculated molecular weight for GFP-mono: 30017 Da.

DNA sequence for GFP-II-4-N:

```
1       ---------+---------+---------+---------+---------+---------+    60
        atgcaccatcatcatcatcattcttctggtgtagatcatattgccgcaatcaagcaggaa
61      ---------+---------+---------+---------+---------+---------+   120
        attgccgcaatcaagcaggaaattgccgcaatcaagcaggaaattgccgcaatcaagcag
121     ---------+---------+---------+---------+---------+---------+   180
        gaaggtggcggtggcggtggtaccgagaacctgtacttccaatccaatattggaagtgga
                          KpnI
181     ---------+---------+---------+---------+---------+---------+   240
        ttactggctagcaaaggagaagaactcttcactggagttgtcccaattcttgttgaatta
241     ---------+---------+---------+---------+---------+---------+   300
        gatggtgatgttaacggccacaagttctctgtcagtggagagggtgaaggtgatgcaaca
301     ---------+---------+---------+---------+---------+---------+   360
        tacggaaaacttaccctgaagttcatctgcactactggcaaactgcctgttccatggcca
361     ---------+---------+---------+---------+---------+---------+   420
        acactagtcactactctgtgctatggtgttcaatgcttttcaagatacccggatcatatg
421     ---------+---------+---------+---------+---------+---------+   480
        aaacggcatgacttttcaagagtgccatgcccgaaggttatgtacaggaaaggaccatc
481     ---------+---------+---------+---------+---------+---------+   540
        ttcttcaaagatgacggcaactacaagacacgtgctgaagtcaagtttgaaggtgatacc
541     ---------+---------+---------+---------+---------+---------+   600
        cttgttaatagaatcgagttaaaaggtattgacttcaaggaagatggcaacattctggga
601     ---------+---------+---------+---------+---------+---------+   660
        cacaaattggaatacaactataactcacacaatgtatacatcatggcagacaaacaaaag
661     ---------+---------+---------+---------+---------+---------+   720
        aatggaatcaaagtgaacttcaagacccgccacaacattgaagatggaagcgttcaacta
721     ---------+---------+---------+---------+---------+---------+   780
        gcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaac
781     ---------+---------+---------+---------+---------+---------+   840
        cattacctgtccacacaatctgccctttcgaaagatcccaacgaaaagagagaccacatg
                                  BstBI
841     ---------+---------+---------+---------+---------+---------+   900
        gtccttcttgagtttgtaacagctgctgggattacacatggcatggatgaactgtacaac
901     ---------+---------+---------+---------+---------+---------+   960
        tga
```

Protein sequence for GFP-II-4-N:

```
        10         20         30         40         50         60
MHHHHHHSSG VDHIAAIKQE IAAIKQEIAA IKQEIAAIKQ EGGGGGGTEN LYFQSNIGSG
        70         80         90        100        110        120
LLASKGEELF TGVVPILVEL DGDVNGHKFS VSGEGEGDAT YGKLTLKFIC TTGKLPVPWP
       130        140        150        160        170        180
TLVTTLCYGV QCFSRYPDHM KRHDFFKSAM PEGYVQERTI FFKDDGNYKT RAEVKFEGDT
       190        200        210        220        230        240
LVNRIELKGI DFKEDGNILG HKLEYNYNSH NVYIMADKQK NGIKVNFKTR HNIEDGSVQL
       250        260        270        280        290        300
ADHYQQNTPI GDGPVLLPDN HYLSTQSALS KDPNEKRDHM VLLEFVTAAG ITHGMDELYN
```

Calculated molecular weight for GFP-II-4-N: 33342 Da.

DNA sequence for GFP-II-5-N:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcatcattcttctggtgtagatcatattgccgcaatcaagcaggaa
 61      ---------+---------+---------+---------+---------+---------+   120
         attgccgcaatcaagcaggaaattgccgcaatcaagcaggaaattgccgcaatcaagcag
121      ---------+---------+---------+---------+---------+---------+   180
         gaaattgccgcaatcaagcaggaaggtggcggtggcggtggtaccgagaacctgtacttc
                                                 KpnI
181      ---------+---------+---------+---------+---------+---------+   240
         caatccaatattggaagtggattactggctagcaaaggagaagaactcttcactggagtt
241      ---------+---------+---------+---------+---------+---------+   300
         gtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtgga
301      ---------+---------+---------+---------+---------+---------+   360
         gagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggc
361      ---------+---------+---------+---------+---------+---------+   420
         aaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttt
421      ---------+---------+---------+---------+---------+---------+   480
         tcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggt
481      ---------+---------+---------+---------+---------+---------+   540
         tatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaa
541      ---------+---------+---------+---------+---------+---------+   600
         gtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaag
601      ---------+---------+---------+---------+---------+---------+   660
         gaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgtatac
661      ---------+---------+---------+---------+---------+---------+   720
         atcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacatt
721      ---------+---------+---------+---------+---------+---------+   780
         gaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggc
781      ---------+---------+---------+---------+---------+---------+   840
         cctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatccc
                                                       BstBI
841      ---------+---------+---------+---------+---------+---------+   900
         aacgaaaagagagaccacatggtccttcttgagtttgtaacagctgctgggattacacat
901      ---------+---------+---------+---------+---------+---------+   960
         ggcatggatgaactgtacaactga
```

Protein sequence for GFP-II-5-N:

```
        10         20         30         40         50         60
MHHHHHHSSG VDHIAAIKQE IAAIKQEIAA IKQEIAAIKQ EIAAIKQEGG GGGGTENLYF
        70         80         90        100        110        120
QSNIGSGLLA SKGEELFTGV VPILVELDGD VNGHKFSVSG EGEGDATYGK LTLKFICTTG
       130        140        150        160        170        180
KLPVPWPTLV TTLCYGVQCF SRYPDHMKRH DFFKSAMPEG YVQERTIFFK DDGNYKTRAE
       190        200        210        220        230        240
VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY IMADKQKNGI KVNFKTRHNI
       250        260        270        280        290        300
EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP NEKRDHMVLL EFVTAAGITH

GMDELYN
```

Calculated molecular weight for GFP-II-5-N: 34096 Da.

DNA sequence for GFP-LI-4-N:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcatcattcttctggtgtagatcatctggccgcaatcaagcaggaa
 61      ---------+---------+---------+---------+---------+---------+   120
         ctggccgcaatcaagcaggaactggccgcaatcaagcaggaactggccgcaatcaagcag
121      ---------+---------+---------+---------+---------+---------+   180
         gaaggtggcggtggcggtggtaccgagaacctgtacttccaatccaatattggaagtgga
                           KpnI
181      ---------+---------+---------+---------+---------+---------+   240
         ttactggctagcaaaggagaagaactcttcactggagttgtcccaattcttgttgaatta
241      ---------+---------+---------+---------+---------+---------+   300
         gatggtgatgttaacggccacaagttctctgtcagtggagagggtgaaggtgatgcaaca
301      ---------+---------+---------+---------+---------+---------+   360
         tacggaaaacttaccctgaagttcatctgcactactggcaaactgcctgttccatggcca
361      ---------+---------+---------+---------+---------+---------+   420
         acactagtcactactctgtgctatggtgttcaatgcttttcaagatacccggatcatatg
421      ---------+---------+---------+---------+---------+---------+   480
         aaacggcatgacttttcaagagtgccatgcccgaaggttatgtacaggaaaggaccatc
481      ---------+---------+---------+---------+---------+---------+   540
         ttcttcaaagatgacggcaactacaagacacgtgctgaagtcaagtttgaaggtgatacc
541      ---------+---------+---------+---------+---------+---------+   600
         cttgttaatagaatcgagttaaaaggtattgacttcaaggaagatggcaacattctggga
601      ---------+---------+---------+---------+---------+---------+   660
         cacaaattggaatacaactataactcacacaatgtatacatcatggcagacaaacaaaag
661      ---------+---------+---------+---------+---------+---------+   720
         aatggaatcaaagtgaacttcaagacccgccacaacattgaagatggaagcgttcaacta
721      ---------+---------+---------+---------+---------+---------+   780
         gcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaac
781      ---------+---------+---------+---------+---------+---------+   840
         cattacctgtccacacaatctgccctttcgaaagatcccaacgaaaagagagaccacatg
                                  BstBI
841      ---------+---------+---------+---------+---------+---------+   900
         gtccttcttgagtttgtaacagctgctgggattacacatggcatggatgaactgtacaac
901      ---------+---------+---------+---------+---------+---------+   960
         tga
```

Protein sequence for GFP-LI-4-N:

```
        10         20         30         40         50         60
MHHHHHHSSG VDHLAAIKQE LAAIKQELAA IKQELAAIKQ EGGGGGGTEN LYFQSNIGSG
        70         80         90        100        110        120
LLASKGEELF TGVVPILVEL DGDVNGHKFS VSGEGEGDAT YGKLTLKFIC TTGKLPVPWP
       130        140        150        160        170        180
TLVTTLCYGV QCFSRYPDHM KRHDFFKSAM PEGYVQERTI FFKDDGNYKT RAEVKFEGDT
       190        200        210        220        230        240
LVNRIELKGI DFKEDGNILG HKLEYNYNSH NVYIMADKQK NGIKVNFKTR HNIEDGSVQL
       250        260        270        280        290        300
ADHYQQNTPI GDGPVLLPDN HYLSTQSALS KDPNEKRDHM VLLEFVTAAG ITHGMDELYN
```

Calculated molecular weight for GFP-LI-4-N: 33342 Da.

DNA sequence for GFP-LI-3-N:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcatcattcttctggtgtagatcatctggccgcaatcaagcaggaa
 61      ---------+---------+---------+---------+---------+---------+    120
         ctggccgcaatcaagcaggaactggccgcaatcaagcaggaaggtggcggtggcggtggt
                                                                KpnI
 121     ---------+---------+---------+---------+---------+---------+    180
         accgagaacctgtacttccaatccaatattggaagtggattactggctagcaaaggagaa
 181     ---------+---------+---------+---------+---------+---------+    240
         gaactcttcactggagttgtcccaattcttgttgaattagatggtgatgttaacggccac
 241     ---------+---------+---------+---------+---------+---------+    300
         aagttctctgtcagtggagagggtgaaggtgatgcaacatacggaaaacttaccctgaag
 301     ---------+---------+---------+---------+---------+---------+    360
         ttcatctgcactactggcaaactgcctgttccatggccaacactagtcactactctgtgc
 361     ---------+---------+---------+---------+---------+---------+    420
         tatggtgttcaatgcttttcaagatacccggatcatatgaaacggcatgacttttcaag
 421     ---------+---------+---------+---------+---------+---------+    480
         agtgccatgcccgaaggttatgtacaggaaaggaccatcttcttcaaagatgacggcaac
 481     ---------+---------+---------+---------+---------+---------+    540
         tacaagacacgtgctgaagtcaagtttgaaggtgatacccttgttaatagaatcgagtta
 541     ---------+---------+---------+---------+---------+---------+    600
         aaaggtattgacttcaaggaagatggcaacattctgggacacaaattggaatacaactat
 601     ---------+---------+---------+---------+---------+---------+    660
         aactcacacaatgtatacatcatggcagacaaacaaaagaatggaatcaaagtgaacttc
 661     ---------+---------+---------+---------+---------+---------+    720
         aagacccgccacaacattgaagatggaagcgttcaactagcagaccattatcaacaaaat
 721     ---------+---------+---------+---------+---------+---------+    780
         actccaattggcgatggccctgtccttttaccagacaaccattacctgtccacacaatct
 781     ---------+---------+---------+---------+---------+---------+    840
         gccctttcgaaagatcccaacgaaaagagagaccacatggtccttcttgagtttgtaaca
              BstBI
 841     ---------+---------+---------+---------+---------+---------+    900
         gctgctgggattacacatggcatggatgaactgtacaactga
```

Protein sequence for GFP-LI-3-N:

```
        10         20         30         40         50         60
MHHHHHHSSG VDHLAAIKQE LAAIKQELAA IKQEGGGGGG TENLYFQSNI GSGLLASKGE

        70         80         90        100        110        120
ELFTGVVPIL VELDGDVNGH KFSVSGEGEG DATYGKLTLK FICTTGKLPV PWPTLVTTLC

       130        140        150        160        170        180
YGVQCFSRYP DHMKRHDFFK SAMPEGYVQE RTIFFKDDGN YKTRAEVKFE GDTLVNRIEL

       190        200        210        220        230        240
KGIDFKEDGN ILGHKLEYNY NSHNVYIMAD KQKNGIKVNF KTRHNIEDGS VQLADHYQQN

       250        260        270        280        290
TPIGDGPVLL PDNHYLSTQS ALSKDPNEKR DHMVLLEFVT AAGITHGMDE LYN
```

Calculated molecular weight for GFP-LI-3-N: 32588 Da.

DNA sequence for GFP-LL-4-N:

```
1       ---------+---------+---------+---------+---------+---------+    60
        atgcaccatcatcatcatcattcttctggtgtagatcatctggccgcactgaagcaggaa
61      ---------+---------+---------+---------+---------+---------+   120
        ctggccgcactgaagcaggaactggccgcactgaagcaggaactggccgcactgaagcag
121     ---------+---------+---------+---------+---------+---------+   180
        gaaggtggcggtggcggtggtaccgagaacctgtacttccaatccaatattggaagtgga
                           KpnI
181     ---------+---------+---------+---------+---------+---------+   240
        ttactggctagcaaaggagaagaactcttcactggagttgtcccaattcttgttgaatta
241     ---------+---------+---------+---------+---------+---------+   300
        gatggtgatgttaacggccacaagttctctgtcagtggagagggtgaaggtgatgcaaca
301     ---------+---------+---------+---------+---------+---------+   360
        tacggaaaacttaccctgaagttcatctgcactactggcaaactgcctgttccatggcca
361     ---------+---------+---------+---------+---------+---------+   420
        acactagtcactactctgtgctatggtgttcaatgcttttcaagatacccggatcatatg
421     ---------+---------+---------+---------+---------+---------+   480
        aaacggcatgacttttcaagagtgccatgcccgaaggttatgtacaggaaaggaccatc
481     ---------+---------+---------+---------+---------+---------+   540
        ttcttcaaagatgacggcaactacaagacacgtgctgaagtcaagtttgaaggtgatacc
541     ---------+---------+---------+---------+---------+---------+   600
        cttgttaatagaatcgagttaaaaggtattgacttcaaggaagatggcaacattctggga
601     ---------+---------+---------+---------+---------+---------+   660
        cacaaattggaatacaactataactcacacaatgtatacatcatggcagacaaacaaaag
661     ---------+---------+---------+---------+---------+---------+   720
        aatggaatcaaagtgaacttcaagacccgccacaacattgaagatggaagcgttcaacta
721     ---------+---------+---------+---------+---------+---------+   780
        gcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaac
781     ---------+---------+---------+---------+---------+---------+   840
        cattacctgtccacacaatctgccctttcgaaagatcccaacgaaaagagagaccacatg
                                  BstBI
841     ---------+---------+---------+---------+---------+---------+   900
        gtccttcttgagtttgtaacagctgctgggattacacatggcatggatgaactgtacaac
901     ---------+---------+---------+---------+---------+---------+   960
        tga
```

Protein sequence for GFP-LL-4-N:

```
        10         20         30         40         50         60
MHHHHHHSSG VDHLAALKQE LAALKQELAA LKQELAALKQ EGGGGGGTEN LYFQSNIGSG
        70         80         90        100        110        120
LLASKGEELF TGVVPILVEL DGDVNGHKFS VSGEGEGDAT YGKLTLKFIC TTGKLPVPWP
       130        140        150        160        170        180
TLVTTLCYGV QCFSRYPDHM KRHDFFKSAM PEGYVQERTI FFKDDGNYKT RAEVKFEGDT
       190        200        210        220        230        240
LVNRIELKGI DFKEDGNILG HKLEYNYNSH NVYIMADKQK NGIKVNFKTR HNIEDGSVQL
       250        260        270        280        290        300
ADHYQQNTPI GDGPVLLPDN HYLSTQSALS KDPNEKRDHM VLLEFVTAAG ITHGMDELYN
```

Calculated molecular weight for GFP-LL-4-N: 33342 Da.

DNA sequence for GFP-WW-7-N:

```
1       ---------+---------+---------+---------+---------+---------+    60
        atgcaccatcatcatcatcattcttctggtgtagatcattcttccaacgcgaaatgggat
61      ---------+---------+---------+---------+---------+---------+   120
        cagtggtcttccgattggcagacctggaacgcgaaatgggatcagtggagcaacgattgg
121     ---------+---------+---------+---------+---------+---------+   180
        aacgcgtggcgttctgattggcaggcgtggaaagatgattgggcgcgttggaaccagcgt
181     ---------+---------+---------+---------+---------+---------+   240
        tgggataactgggcgaccggtggcggtggcggtggtaccgagaacctgtacttccaatcc
                                          KpnI
241     ---------+---------+---------+---------+---------+---------+   300
        aatattggaagtggattactggctagcaaaggagaagaactcttcactggagttgtccca
301     ---------+---------+---------+---------+---------+---------+   360
        attcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtggagagggt
361     ---------+---------+---------+---------+---------+---------+   420
        gaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggcaaactg
421     ---------+---------+---------+---------+---------+---------+   480
        cctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttttcaaga
481     ---------+---------+---------+---------+---------+---------+   540
        tacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggttatgta
541     ---------+---------+---------+---------+---------+---------+   600
        caggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaagtcaag
601     ---------+---------+---------+---------+---------+---------+   660
        tttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaaggaagat
661     ---------+---------+---------+---------+---------+---------+   720
        ggcaacattctgggacacaaattggaatacaactataactcacacaatgtatacatcatg
721     ---------+---------+---------+---------+---------+---------+   780
        gcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacattgaagat
781     ---------+---------+---------+---------+---------+---------+   840
        ggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggccctgtc
841     ---------+---------+---------+---------+---------+---------+   900
        cttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatcccaacgaa
                                                 BstBI
901     ---------+---------+---------+---------+---------+---------+   960
        aagagagaccacatggtccttcttgagtttgtaacagctgctgggattacacatggcatg
961     ---------+---------+---------+---------+---------+---------+  1020
        gatgaactgtacaactga
```

Protein sequence for GFP-WW-7-N:

```
         10         20         30         40         50         60
MHHHHHHSSG VDHSSNAKWD QWSSDWQTWN AKWDQWSNDW NAWRSDWQAW KDDWARWNQR
         70         80         90        100        110        120
WDNWATGGGG GGTENLYFQS NIGSGLLASK GEELFTGVVP ILVELDGDVN GHKFSVSGEG
        130        140        150        160        170        180
EGDATYGKLT LKFICTTGKL PVPWPTLVTT LCYGVQCFSR YPDHMKRHDF FKSAMPEGYV
        190        200        210        220        230        240
QERTIFFKDD GNYKTRAEVK FEGDTLVNRI ELKGIDFKED GNILGHKLEY NYNSHNVYIM
        250        260        270        280        290        300
ADKQKNGIKV NFKTRHNIED GSVQLADHYQ QNTPIGDGPV LLPDNHYLST QSALSKDPNE
        310        320
KRDHMVLLEF VTAAGITHGM DELYN
```

Calculated molecular weight for GFP-WW-7-N: 37184 Da.

DNA sequence for GFP-FF-7-N:

```
1       ---------+---------+---------+---------+---------+---------+    60
        atgcaccatcatcatcatcattcttctggtgtagatcatagcagcaacgcgaaatttgat
61      ---------+---------+---------+---------+---------+---------+   120
        cagtttagcagcgattttcagacctttaacgcgaaatttgatcagtttagcaacgatatg
121     ---------+---------+---------+---------+---------+---------+   180
        aacgcgtttcgcagcgattttcaggcgtttaaagatgattttgcgcgctttaaccagcgc
181     ---------+---------+---------+---------+---------+---------+   240
        tttgataactttgcgaccaaatatcgcggtggcggtggcggtggtaccgagaacctgtac
                                                  KpnI
241     ---------+---------+---------+---------+---------+---------+   300
        ttccaatccaatattggaagtggattactggctagcaaaggagaagaactcttcactgga
301     ---------+---------+---------+---------+---------+---------+   360
        gttgtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagt
361     ---------+---------+---------+---------+---------+---------+   420
        ggagagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactact
421     ---------+---------+---------+---------+---------+---------+   480
        ggcaaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgc
481     ---------+---------+---------+---------+---------+---------+   540
        ttttcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaa
541     ---------+---------+---------+---------+---------+---------+   600
        ggttatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgct
601     ---------+---------+---------+---------+---------+---------+   660
        gaagtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttc
661     ---------+---------+---------+---------+---------+---------+   720
        aaggaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgta
721     ---------+---------+---------+---------+---------+---------+   780
        tacatcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaac
781     ---------+---------+---------+---------+---------+---------+   840
        attgaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgat
841     ---------+---------+---------+---------+---------+---------+   900
        ggccctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagat
                                                        BstBI
901     ---------+---------+---------+---------+---------+---------+   960
        cccaacgaaaagagagaccacatggtccttcttgagtttgtaacagctgctgggattaca
961     ---------+---------+---------+---------+---------+---------+  1020
        catggcatggatgaactgtacaactga
```

Peptide sequence for GFP-FF-7-N:

```
        10         20         30         40         50         60
MHHHHHHSSG VDHSSNAKFD QFSSDFQTFN AKFDQFSNDM NAFRSDFQAF KDDFARFNQR
        70         80         90        100        110        120
FDNFATKYRG GGGGGTENLY FQSNIGSGLL ASKGEELFTG VVPILVELDG DVNGHKFSVS
       130        140        150        160        170        180
GEGEGDATYG KLTLKFICTT GKLPVPWPTL VTTLCYGVQC FSRYPDHMKR HDFFKSAMPE
       190        200        210        220        230        240
GYVQERTIFF KDDGNYKTRA EVKFEGDTLV NRIELKGIDF KEDGNILGHK LEYNYNSHNV
       250        260        270        280        290        300
YIMADKQKNG IKVNFKTRHN IEDGSVQLAD HYQQNTPIGD GPVLLPDNHY LSTQSALSKD
       310        320
PNEKRDHMVL LEFVTAAGIT HGMDELYN
```

Calculated molecular weight for GFP-FF-7-N: 37069 Da.

DNA sequence for GFP-II-4-C:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcattcttctggtgtagatctgggtaccgagaacctgtacttc
                                       BglII  KpnI
 61      ---------+---------+---------+---------+---------+---------+   120
         caatccaatattggaagtggattactggctagcaaaggagaagaactcttcactggagtt
 121     ---------+---------+---------+---------+---------+---------+   180
         gtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtgga
 181     ---------+---------+---------+---------+---------+---------+   240
         gagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggc
 241     ---------+---------+---------+---------+---------+---------+   300
         aaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttt
 301     ---------+---------+---------+---------+---------+---------+   360
         tcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggt
 361     ---------+---------+---------+---------+---------+---------+   420
         tatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaa
 421     ---------+---------+---------+---------+---------+---------+   480
         gtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaag
 481     ---------+---------+---------+---------+---------+---------+   540
         gaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgtatac
 541     ---------+---------+---------+---------+---------+---------+   600
         atcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacatt
 601     ---------+---------+---------+---------+---------+---------+   660
         gaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggc
 661     ---------+---------+---------+---------+---------+---------+   720
         cctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatccc
                                                        BstBI
 721     ---------+---------+---------+---------+---------+---------+   780
         aacgaaaagagagaccacatggtccttcttgagtttgtaacagccgcggggattacaggt
 781     ---------+---------+---------+---------+---------+---------+   840
         ggtggcggaggtggcgagatcgcggcgatcaaacaggagatcgcagcgatcaaacaggaa
 841     ---------+---------+---------+---------+---------+---------+   900
         attgccgcaattaaacaggaaattgctgcaattaaacaa
```

Peptide sequence for GFP-II-4-C:

```
        10         20         30         40         50         60
MHHHHHHSSG VDLGTENLYF QSNIGSGLLA SKGEELFTGV VPILVELDGD VNGHKFSVSG
        70         80         90        100        110        120
EGEGDATYGK LTLKFICTTG KLPVPWPTLV TTLCYGVQCF SRYPDHMKRH DFFKSAMPEG
       130        140        150        160        170        180
YVQERTIFFK DDGNYKTRAE VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY
       190        200        210        220        230        240
IMADKQKNGI KVNFKTRHNI EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP
       250        260        270        280        290
```

Calculated molecular weight for GFP-II-4-C: 32415 Da.

DNA sequence for GFP-II-5-C:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcattcttctggtgtagatctgggtaccgagaacctgtacttc
                                       BglII  KpnI
 61      ---------+---------+---------+---------+---------+---------+   120
         caatccaatattggaagtggattactggctagcaaaggagaagaactcttcactggagtt
 121     ---------+---------+---------+---------+---------+---------+   180
         gtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtgga
 181     ---------+---------+---------+---------+---------+---------+   240
         gagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggc
 241     ---------+---------+---------+---------+---------+---------+   300
         aaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttt
 301     ---------+---------+---------+---------+---------+---------+   360
         tcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggt
 361     ---------+---------+---------+---------+---------+---------+   420
         tatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaa
 421     ---------+---------+---------+---------+---------+---------+   480
         gtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaag
 481     ---------+---------+---------+---------+---------+---------+   540
         gaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgtatac
 541     ---------+---------+---------+---------+---------+---------+   600
         atcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacatt
 601     ---------+---------+---------+---------+---------+---------+   660
         gaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggc
 661     ---------+---------+---------+---------+---------+---------+   720
         cctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatccc
                                                      BstBI
 721     ---------+---------+---------+---------+---------+---------+   780
         aacgaaaagagagaccacatggtccttcttgagtttgtaacagccgcggggattacaggt
 781     ---------+---------+---------+---------+---------+---------+   840
         ggtggcggaggtggcgagatcgcggcgatcaaacaggagattgcagccattaaacaagaa
 841     ---------+---------+---------+---------+---------+---------+   900
         atcgcagcgatcaaacaggaaattgccgcaattaaacaggaaattgctgcaattaaacaa
```

Peptide sequence for GFP-II-5-C:

```
        10         20         30         40         50         60
MHHHHHHSSG VDLGTENLYF QSNIGSGLLA SKGEELFTGV VPILVELDGD VNGHKFSVSG
        70         80         90        100        110        120
EGEGDATYGK LTLKFICTTG KLPVPWPTLV TTLCYGVQCF SRYPDHMKRH DFFKSAMPEG
       130        140        150        160        170        180
YVQERTIFFK DDGNYKTRAE VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY
       190        200        210        220        230        240
IMADKQKNGI KVNFKTRHNI EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP
       250        260        270        280        290        300
NEKRDHMVLL EFVTAAGITG GGGGGEIAAI KQEIAAIKQE IAAIKQEIAA IKQEIAAIKQ
```

Calculated molecular weight for GFP-II-5-C: 33169 Da.

DNA sequence for GFP-LI-4-C:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcattcttctggtgtagatctgggtaccgagaacctgtacttc
                                       BglII  KpnI
 61      ---------+---------+---------+---------+---------+---------+    120
         caatccaatattggaagtggattactggctagcaaaggagaagaactcttcactggagtt
 121     ---------+---------+---------+---------+---------+---------+    180
         gtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtgga
 181     ---------+---------+---------+---------+---------+---------+    240
         gagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggc
 241     ---------+---------+---------+---------+---------+---------+    300
         aaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttt
 301     ---------+---------+---------+---------+---------+---------+    360
         tcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggt
 361     ---------+---------+---------+---------+---------+---------+    420
         tatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaa
 421     ---------+---------+---------+---------+---------+---------+    480
         gtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaag
 481     ---------+---------+---------+---------+---------+---------+    540
         gaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgtatac
 541     ---------+---------+---------+---------+---------+---------+    600
         atcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacatt
 601     ---------+---------+---------+---------+---------+---------+    660
         gaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggc
 661     ---------+---------+---------+---------+---------+---------+    720
         cctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatccc
                                                        BstBI
 721     ---------+---------+---------+---------+---------+---------+    780
         aacgaaaagagagaccacatggtccttcttgagtttgtaacagccgcggggattacaggt
 781     ---------+---------+---------+---------+---------+---------+    840
         ggtggcggaggtggcgagctggcggcgatcaaacaggagctggcagcgatcaaacaggaa
 841     ---------+---------+---------+---------+---------+---------+    900
         ctggccgcaattaaacaggaactggctgcaattaaacaa
```

Peptide sequence for GFP-LI-4-C:

```
        10         20         30         40         50         60
MHHHHHHSSG VDLGTENLYF QSNIGSGLLA SKGEELFTGV VPILVELDGD VNGHKFSVSG
        70         80         90        100        110        120
EGEGDATYGK LTLKFICTTG KLPVPWPTLV TTLCYGVQCF SRYPDHMKRH DFFKSAMPEG
       130        140        150        160        170        180
YVQERTIFFK DDGNYKTRAE VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY
       190        200        210        220        230        240
IMADKQKNGI KVNFKTRHNI EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP
       250        260        270        280        290
NEKRDHMVLL EFVTAAGITG GGGGGELAAI KQELAAIKQE LAAIKQELAA IKQ
```

Calculated molecular weight for GFP-LI-4-C: 32415 Da.

DNA sequence for GFP-LL-4-C:

```
 1       ---------+---------+---------+---------+---------+---------+    60
         atgcaccatcatcatcattcttctggtgtagatctgggtaccgagaacctgtacttc
                                       BglII  KpnI
 61      ---------+---------+---------+---------+---------+---------+   120
         caatccaatattggaagtggattactggctagcaaaggagaagaactcttcactggagtt
 121     ---------+---------+---------+---------+---------+---------+   180
         gtcccaattcttgttgaattagatggtgatgttaacggccacaagttctctgtcagtgga
 181     ---------+---------+---------+---------+---------+---------+   240
         gagggtgaaggtgatgcaacatacggaaaacttaccctgaagttcatctgcactactggc
 241     ---------+---------+---------+---------+---------+---------+   300
         aaactgcctgttccatggccaacactagtcactactctgtgctatggtgttcaatgcttt
 301     ---------+---------+---------+---------+---------+---------+   360
         tcaagatacccggatcatatgaaacggcatgacttttcaagagtgccatgcccgaaggt
 361     ---------+---------+---------+---------+---------+---------+   420
         tatgtacaggaaaggaccatcttcttcaaagatgacggcaactacaagacacgtgctgaa
 421     ---------+---------+---------+---------+---------+---------+   480
         gtcaagtttgaaggtgataccttgttaatagaatcgagttaaaaggtattgacttcaag
 481     ---------+---------+---------+---------+---------+---------+   540
         gaagatggcaacattctgggacacaaattggaatacaactataactcacacaatgtatac
 541     ---------+---------+---------+---------+---------+---------+   600
         atcatggcagacaaacaaaagaatggaatcaaagtgaacttcaagacccgccacaacatt
 601     ---------+---------+---------+---------+---------+---------+   660
         gaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcgatggc
 661     ---------+---------+---------+---------+---------+---------+   720
         cctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatccc
                                                        BstBI
 721     ---------+---------+---------+---------+---------+---------+   780
         aacgaaaagagagaccacatggtccttcttgagtttgtaacagccgcggggattacaggt
 781     ---------+---------+---------+---------+---------+---------+   840
         ggtggcggaggtggcgagctggcggcgctgaaacaggagctggcagcgctgaaacaggaa
 841     ---------+---------+---------+---------+---------+---------+   900
         Ctggccgcactgaaacaggaactggctgcactgaaacaa
```

Peptide sequence for GFP-LL-4-C:

```
        10         20         30         40         50         60
MHHHHHHSSG VDLGTENLYF QSNIGSGLLA SKGEELFTGV VPILVELDGD VNGHKFSVSG
        70         80         90        100        110        120
EGEGDATYGK LTLKFICTTG KLPVPWPTLV TTLCYGVQCF SRYPDHMKRH DFFKSAMPEG
       130        140        150        160        170        180
YVQERTIFFK DDGNYKTRAE VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY
       190        200        210        220        230        240
IMADKQKNGI KVNFKTRHNI EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP
       250        260        270        280        290
NEKRDHMVLL EFVTAAGITG GGGGGELAAL KQELAALKQE LAALKQELAA LKQ
```

Calculated molecular weight for GFP-LL-4-C: 32415 Da.