Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Olivia Mangat Dhaliwal March 17, 2016

Cloning, Expression, and Characterization of a HEAT Repeat Solenoid Protein

by

Olivia M. Dhaliwal

Dr. Vincent Conticello
Advisor

Department of Chemistry

Dr. Jeremy Weaver

Advisor

Dr. Emily Weinert

Committee Member

Dr. Sandra Blakely

Committee Member

2016

Cloning, Expression, and Characterization of a HEAT Repeat Solenoid Protein

By

Olivia M. Dhaliwal

Dr. Vincent Conticello

Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Chemistry

2016

Abstract

Cloning, Expression, and Characterization of a HEAT Repeat Solenoid Protein
By Olivia M. Dhaliwal

A HEAT-like solenoid protein was chosen for this project due to published evidence of its thermodynamic stability and flexibility in sequence and structure. Two DNA sequences were designed to clone varying lengths of a repeated HEAT-like protein for expression in *E. coli*, without N- and C- cap repeated units to permit self-assembly. Expressed proteins were purified in an inclusion-body washing protocol and affinity column purification procedures. Purified proteins were assayed for thermodynamic stability, secondary structure, and homogeneity in assemblies. Proteins that had undergone a refolding protocol in a series of refolding buffers were compared to proteins that had been only dialyzed in deionized water. Refolded proteins had more intense and consistent alpha-helical CD signatures and displayed longer, more ordered tubular assemblies when imaged using TEM. Non-refolded proteins had noisy, inconsistent CD signatures and their assemblies were fragmented and disordered.

Cloning, Expression, and Characterization of a HEAT Repeat Solenoid Protein

By

Olivia M. Dhaliwal

Dr. Vincent Conticello

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Chemistry

2016

Acknowledgements

I arrived at Emory as a junior transfer student with absolutely no research experience or appreciable knowledge of biochemistry. Despite this, Dr. Vincent Conticello took a risk on me and gave me a chance I was worried I would never get. I cannot thank him enough for giving me the amazing opportunity to work in his research lab. Vince, you have inspired me to continue to pursue research and also to be willing to trust my own future students.

I owe many thanks to two people in my lab. Rebecca Anne Bartlett, thank you for patiently training me for the past two years. Thank you for painstakingly showing me so many times how to do things correctly, and for your advice and reassurance that I could do this project. Spencer Chauncey Hughes, thank you for taking me under your wing this semester. I am so grateful for the energy and time you have put into helping me get this done. I could not have done this without your support and knowledge.

This thesis is dedicated to my late father, Dr. Ranjit Singh Dhaliwal. Dad, I miss you every single day. I wish you could be here to see how far I've come. Thank you for always supporting me, loving me, and providing for me so that I could have the privilege of attending Simon's Rock and Emory.

Table of Contents

Figures

Introduction

Tandem repeat proteins (TRPs) are proteins whose repeated units vary in number. Some TRPs can even have different sequences within each repeated unit.[1,2] They exist in both prokaryotic and eukaryotic organisms, and their repeated units are generally between 3 and 40 amino acids in length. The HEAT repeat, the TRP of interest in this project, is an alpha-helical solenoid TRP whose repeat unit is composed of 30-40 amino acids. It is abundant in thermophilic organisms, so many of the HEAT proteins are thermostable to high temperatures. It has piqued the interest of biochemists because of its sequence flexibility,[3] which increases its potential for customization to develop *de novo* proteins using the solenoid's elongated binding surfaces that are aptly suited for interactions with binding partners ranging from other proteins and biological molecules to metals.[2,3] Their versatility gives them vast potential in clinical applications. For example, they could be used to create antibodies or to deactivate damaging proteins by irreversibly binding them. Additionally, the HEAT protein's assembled structure has a large curvature, which is of interest to our research group because we aim to create a closed loop with the curved assemblies. The tubular assembly, once created, can be functionalized to bind things inside it.

Other groups have successfully designed solenoid proteins that target damaging biological pathways[4] by binding proteins to deactivate them. A very successful example of therapeutic solenoid TRPs are DARPins (**D**esigned **A**nkyrin **R**epeat **P**roteins). Currently, these proteins are undergoing clinical trials to treat retinal disease, macular degeneration, and diabetic macular edema[4].

Previous studies of the HEAT1 sequence and several mutants show that they exhibit thermostable properties and assemble into alpha-helical, elongated structures. This project

explores how the covalently-linked concatemers of the HEAT1 sequence compare in stability and assembly to the other concatemers and to the synthesized monomer peptide. It is important to explore the scope of these characteristics because their flexibility and stability renders them appealing for biomaterials chemistry. If covalently linked repeats are a way to create larger and more stable assemblies, then it is important to optimize methods to produce these proteins with greater yields than peptide synthesis. Expression using prokaryotic organisms is an attractive alternative because of the high yield and the fidelity of translated protein.[5]

There are several aims for this project. The first is to design prokaryotic DNA sequences for cloning that permit the addition of repeats of the HEAT1 consensus sequence with a 6x repeated Histidine tag for purification. This sequence should be adaptable (i.e., able to accommodate cloning fewer or greater numbers of repeats than anticipated). The second aim is to create covalently-linked elongated repeats of the helix-turn-helix protein (in this case, a dimer, tetramer, and hexamer), express the proteins using *Escherichia Coli* and isolate properly folded proteins for characterization. The third and final aim of this project is to characterize the proteins using transmission electron microscopy and circular dichroism to determine if and how the numbers of repeats influence their assemblies and stabilities.

The HEAT1 amino acid sequence was designed from a published consensus sequence[3] that was also synthesized and characterized as a peptide in our laboratory. Two new DNA sequences were designed containing this consensus sequence in order to clone the concatemers: one Adapter, into which the repeat units could be added, and one Insert that contained the repeat unit sequence. The dimer, tetramer, and hexamer were successfully cloned and expressed in the inclusion bodies of *E. Coli* BL21 cells. They were solubilized with denaturants and isolated

using affinity columns and inclusion body washing protocols, and characterized using CD and TEM.

The literature has indicated that concatemers of TRPs with greater numbers of repeats yield more thermostable assemblies. It is expected that the protein with the greatest number of repeats will be thermostable and reliably form homogenous assemblies. It is also expected that the concatemers will stack consecutively on top of one another to form elongated assemblies that resemble solenoids. Because they are composed of a helix-turn-helix repeated unit,[3] the produced concatemers should be alpha-helical in their secondary structures.

Literature Review

Published work on TRPs is important to consider as context for this project because it highlights the versatility and stability of these proteins. The idea that the repeated units are "building blocks" is a theme that will be explored often throughout this review. The HEAT project builds upon existing TRP research, expanding knowledge in many important fields of biochemistry, from sequence-to-structure determination to optimizing production of inclusion body proteins. This literature review will discuss past and current research regarding tandem repeat proteins, particularly focusing on the solenoid family of repeat proteins.

Solenoid proteins are a class of proteins composed of many repeated units that usually form assemblies that stack the repeated units to form extended structures. These assemblies are stabilized through hydrophobic interactions between the repeated units, sometimes referred to as the hydrophobic core.[6] Solenoid proteins fall under the category of tandem repeat proteins (TRPs), which are found in many different environments *in vivo* because of their great diversity and adaptability in sequence, size, and function.[1] Additionally, the folding patterns of solenoids

can frequently be reliably predicted because the repeated units lack long-term interactions.[6] The highly conserved residues in the repeated units function to maintain the repeated unit structure of the proteins, but the residues that are not conserved often have high sequence variation,[7] permitting adjustments of polarity, charge, or hydrophobicity. These adjustments can be designed to alter the solvent-accessible portion of the solenoid to recognize a specific sequence in its binding partner protein or small molecule.[6]

The synthetic versions of repeat domains can be designed to be more thermodynamically stable than their naturally occurring domains.[8] This is another example of how solenoid proteins can be fine-tuned to function in various environments, making them attractive materials to use for designing highly specific protein-protein interactions.

Though TRP proteins encompass many different structural motifs,[6] solenoid proteins have been given a lot of attention in recent studies because of their potential for therapeutic, clinical, and biological applications. Current research goals for solenoid proteins include designing peptide-binding scaffolds, understanding the effects of mutagenesis on the morphologies of the proteins, and targeting binding partners by functionalizing the binding surfaces of the solenoids.

Designed solenoid proteins that can target and bind biologically significant partner proteins are important because they could interrupt damaging biological pathways by binding and disabling an essential protein in the pathway. Additionally, they could be used to bind and disable foreign proteins. HEAT, ankyrin, and armadillo (ARM) proteins are elongated and flexible in sequence and structure, and these characteristics render them ideal for surface functionalization with binding potential. The sequence similarities in their repeated units are illustrated in Figure 1.

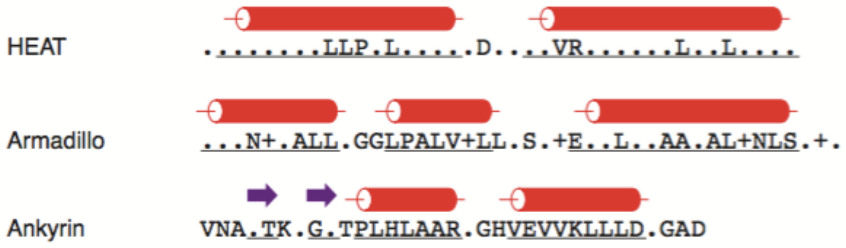


Figure 1. Consensus sequences of repeat units present in three common solenoid proteins (+, positively charged amino acid). Positions of helices (red cylinders) and β-strands (magenta arrows) are shown above the underlined portions of the sequences. Consensus sequences have been taken from the literature and might have therefore been derived using different criteria for different protein motifs. Reprinted with permission from Reference 7. © 2000 Elsevier Ltd.

Three studies have explored the binding potentials of HEAT, ARM, and ankyrin proteins by exploiting the high variability permitted in the repeated units. Firstly, Minard *et al*[3] report the production of a new family of artificial proteins based on HEAT-like repeats that are thermostable, have defined structures, and are able to be mutated to create specific binding surfaces. The second study, by Plückthun *et al,*[9] shows how mutated armadillo repeat building blocks can be stacked on top of each other to create an elongated scaffold with a custom binding surface. Finally, Hollenbeck *et al*[2] report the creation of a similar multivalent recognition scaffold using ankyrin repeat proteins.

All of these studies probe the biophysical characteristics of these solenoid proteins, revealing patterns in structure and function that could permit the eventual production of even more complicated custom-designed proteins.

In addition to elucidating the biophysical properties of the designed mutants, the study done by Minard *et al*[3] enhances a broader understanding of the effects between mutagenesis and

changes in structure and function. The simplicity and high variability of the repeated HEAT-like motif permits single mutations and the study of resulting structural effects. The library of mutated clones, named αRep (Alpha-helicoidal Repeat proteins), contained zero to seven repeated units of a sequence based on the HEAT repeat.

The positions they chose to mutate were chosen by identifying the highly variable positions in natural sequences related to an idealized consensus sequence previously established in the literature (from a HEAT repeat protein). The variable residues are shown as spheres in Figure 2, and they form the solvent-exposed surface of the second helix.

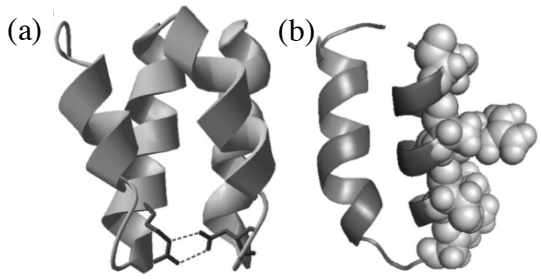


Figure 2. (a) Conserved hydrogen bonds between two repeats in the α-rep mutants. (b) Variable residues in the αRep mutants shown as spheres. Reprinted with permission from Reference 2. © 2010 Elsevier Ltd.

They found that most of their in-frame sequences created soluble proteins, which is significant because it permits large-scale expression to generate sizable amounts of product protein without cumbersome purification steps.[5] The CD spectra of the soluble mutants showed characteristic alpha-helical signatures that increased in intensity with greater numbers of repeats, which indicated that increased numbers of repeat units lead to more well-defined secondary structures. Added numbers of repeats have also been found to increase the stability of the folded

protein in other studies.[8] Moreover, most of these mutants were stable up to 95°C, and their thermal unfolding was reversible.[3] Their study successfully demonstrated the versatility and stability of solenoid proteins, emphasizing that future applications can utilize the massive surface area of these elongated proteins to design recognition sites for binding partners.

It is important to recognize that these findings from Minard *et al*[3] are relevant not only for solenoid proteins, but for all proteins in general. Furthering the understanding of structural prediction methods is important because prediction methods build the current understanding of the relationships between sequence, folding and function.

A study by Plückthun *et al*[9] took the idea of a mutant library one step further: they developed a stable, monomeric consensus sequence using armadillo repeats instead of HEAT. While armadillo and HEAT are similar in that they are both alpha-solenoid proteins,[3,9] armadillo proteins have a repeat unit that is composed of three alpha-helices, as opposed to the two helices present in one HEAT repeat. The armadillo repeat forms a three-dimensional structure that resembles a double helix; this shape enables it to wrap around a peptide.[9]

The mutants were generated with the idea of building blocks in mind. The group's idea can be described like this: if several functionalized blocks are available, then they can be assembled together in a stack to bind a large surface, with the side chains from each block in any custom order required for the binding partner. Figure 3 shows an overview of this idea.

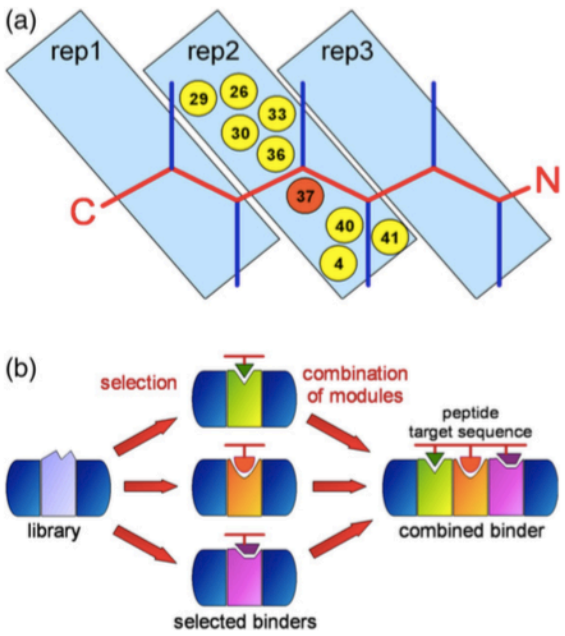


Figure 3. Binding of target peptides. (a) Schematic drawing of an armadillo repeat protein binding a peptide. N and C indicate termini of the peptide being bound. The orange position indicates a conserved Asn residue of the armadillo repeat responsible for binding the target peptide. Yellow residues are those involved in recognition of the target peptide. (b) Building block strategy overview; these potentially allow for the selection of single repeats, combined in a specific stacked manner to bind peptides. Reprinted with permission from Reference 9. © 2007 Elsevier Ltd.

During the construction of these "building block" functionalized repeated units, they further improved their mutants by making additional substitutions to stabilize the hydrophobic core.[9] By identifying the key residues that were involved in the stabilization of the core and ensuring that they were fully conserved, they were able to develop effective mutants with which to build a library.

Their work presents a way to create armadillo repeat libraries that have functionalized, solvent-exposed residues, without compromising the stability of the overall folded protein. These findings can be expanded to other solenoid proteins, as the hydrophobic core also plays an important role in the HEAT and ankyrin repeat proteins.[2,3]

The previous two studies described ways to generate a peptide-binding scaffold using HEAT and armadillo repeat proteins,[3,9] but they only aspired to bind one specific molecule. A study by Hollenbeck *et al*[2] reported successful creation of scaffolds for multivalent recognition using ankyrin proteins.

Their research revealed a method to create a well-defined, chemically homogenous scaffold upon which other proteins could be bound and displayed. To do this, they modified unconserved residues at the beta-turn portion of the ankyrin modules to accommodate reactive thiols. The mutants containing cysteine had similar thermal stabilities and CD spectra when compared with the consensus protein, suggesting that the stabilities of the recombinant proteins were not compromised by the changes. The thiol sites (at which recognition epitopes and small molecules were added to the scaffold) were chosen not only for their high variabilities, but also because the side chains extended away from the hydrophobic core (Figure 4).

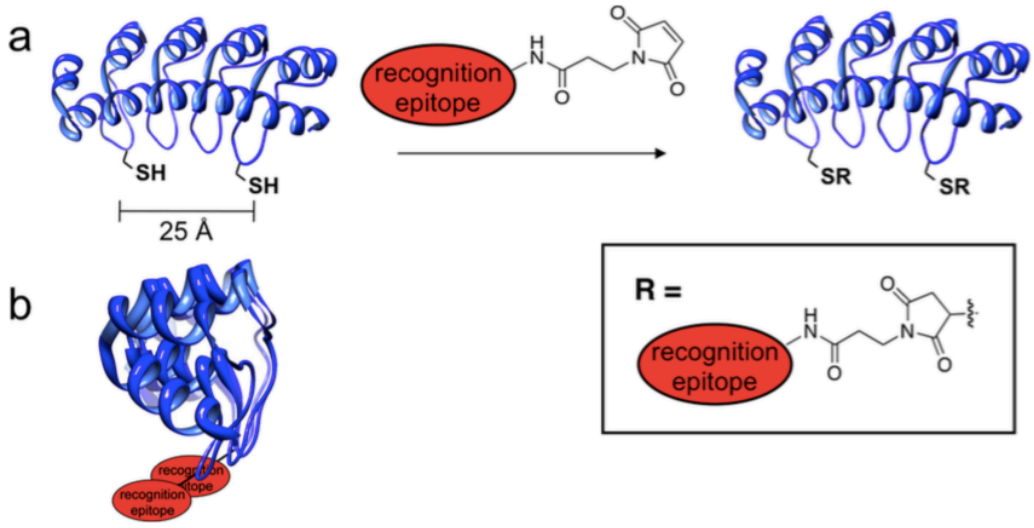


Figure 4. (a) Designed ankyrin protein with three internal repeats containing two cysteines as points of attachment. (b) Side view of the protein, demonstrating that the attached groups are oriented away from the hydrophobic core. Reprinted with permission from Reference 2. © 2012 American Chemical Society.

They found that the secondary structure of the protein was largely unaffected by the addition of the recognition epitopes, reaffirming the stabilizing influence of the hydrophobic core,[3,9] in agreement with the observations made by Plückthun *et al*.[9]

One significant difference in this design is that the cysteine resides are only reactive at a specific pH,[2] so they are able to highly specifically and covalently attach antigens or other small molecules directly to the molecular scaffold. The distance between each displayed molecule can be altered by attaching it to a cysteine on a repeat further away (Figure 4).

These three studies highlight the versatile nature of solenoid proteins. Their alpha-helical secondary structures enable them to have a high degree of flexibility in their tertiary structures. The elongated, solvent-exposed surfaces formed by solenoid proteins are highly variable in sequence, and it has been shown that this is an ideal surface to bind recognition epitopes. This is

particularly evidenced by the *in vivo* applications of the designed ankyrin repeat proteins (DARPins, by Plückthun), which are currently being tested in clinical applications.[4] There is immense potential for these proteins to be used in other biochemical and biomedical applications, not only to target proteins or small molecules of interest, but also to facilitate *in vitro* interactions by placing groups in close proximities to one another to increase rates of reactions.

Materials and Methods

All chemicals were sourced from Sigma Aldridch (St. Louis, MO, USA) or Thermo Fischer Scientific (Waltham, MA, USA). DNA was ordered through DNA2.0 (Newark, CA, USA) and sequencing was performed by GENEWIZ (South Plainfield, NJ, USA). DNA kits, enzymes, and biologicals were sourced from Millipore (Merck Millipore, Billerica, MA, USA) and/or New England Biolabs (Ipswich, MA, USA). All buffers were filter-sterilized.

The major aim of this project was to create a closed-loop structure using a repeated building block unit of a HEAT-like sequence. It has been shown in the literature that these building blocks can be covalently linked to create larger proteins. Our lab observed that the HEAT1 sequence self-assembles into elongated tubular structures, and we hypothesized that covalently-linked concatemers would assemble into longer and more stable tubular assemblies. The concatemers were created by cloning prokaryotic DNA sequences to express in *E. coli*. Isolation and purification protocols were optimized to improve the yield of bioactive, properly refolded protein. These protocols are described in detail below.

Sequence Design, Cloning, and Expression

The 31-amino acid sequence of the repeated helix-turn-helix unit was copied directly from a study by Minard *et al* in 2010[3] with the exception of position 30, in which the glutamic acid residue was replaced with a lysine residue. Position 30 was a highly variable position in the sequence, and the change was made to prevent the isoelectric point of the protein from being too acidic. A 6x Histidine tag was attached to the C-terminal of the sequence for affinity column purification. Figure 5 shows the designed HEAT1 Dimer sequence.

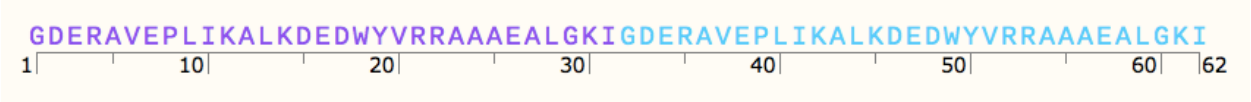


Figure 5. Single-letter amino acid code of the HEAT1 Dimer repeat unit.

Two prokaryotic DNA sequences were designed for cloning purposes. The Adapter sequence (Figure 6) was designed with EcoR I and HinD III overhangs surrounding the expressed portion of the DNA sequence. These overhangs matched our laboratory's pQE80 expression vector. Following the EcoR I restriction site and the ribosomal binding site, two Bsa I insertion sites were included. These restriction sites allowed elongation of the DNA sequence to create the concatemers. The HEAT1 dimer sequence was included in the adapter sequence immediately following the Bsa I restriction sites.

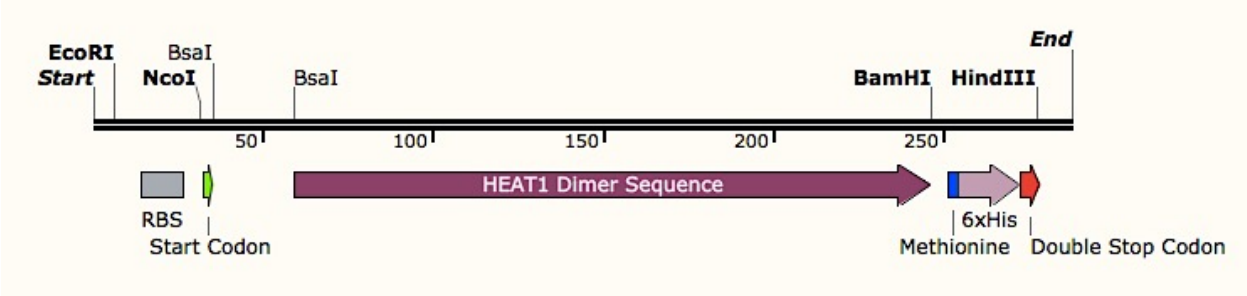


Figure 6. HEAT1 Adapter sequence with insertion sites and features.

A few features were included in the Adapter sequence to facilitate adjustments in case of the His tag interfering with self-assembly: a BamH I insertion site was placed immediately preceding a Methionine and 6x Histidine tag at the C-terminus of the transcribed DNA sequence. The His tag could be removed chemically (using the Methionine for cleavage with cyanogen bromide) or by cloning to create a DNA sequence that did not contain the His tag.

The Adapter was designed as a master template, to which more repeats of the HEAT1 Dimer sequence could be inserted. The Insert sequence was designed to produce copies of the repeated sequence to add to the Adapter. The Insert was designed with two sets of restriction sites flanking the sequence of the HEAT1 Dimer. These extra sets of cut sites were included in case the intended digestions did not work. For this project, the Bbs I and BsmB I restriction sites were the only sites used. The sticky ends from digesting at those sites matched the sticky ends of the HEAT1 Adapter sequence when digested with Bsa I. Figure 7 shows the Insert sequence with the designed restriction sites labeled.

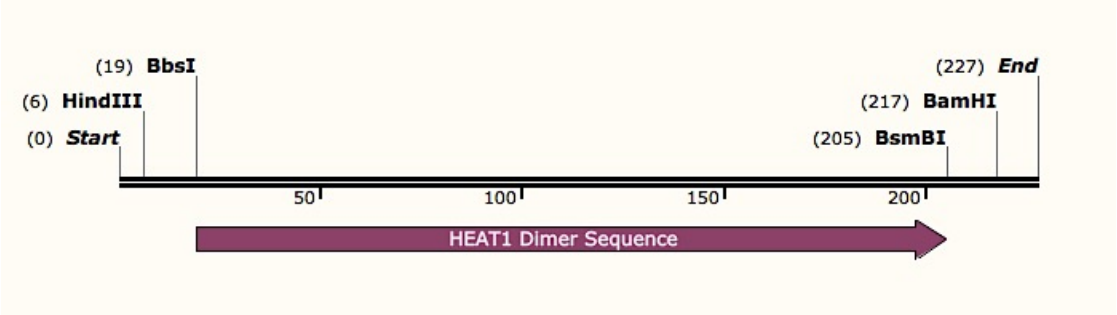


Figure 7. Insert sequence with restriction sites.

The Insert (pBB24) and Adapter (pBB25) sequences were delivered in Kanamycin-resistant pJ201 plasmids. Cloning was carried out in these plasmids transformed into Top 10 F' cells, but the plasmids were expressed in an Ampicillin-resistant pQE80 expression vector, pBB23, using BL21 expression cells.

All digestions and ligations were run on DNA gels (1% agarose with ethidium bromide) with control samples of undigested or unligated DNA and a ladder (O'GeneRuler 1 kb 250-10,000 bp, Thermo Scientific).

Three concatemers were created by cloning using the designed DNA sequences. With the exception of the dimer, the Adapter and Insert were used together to create the concatemers. To create the HEAT1 dimer, the Adapter in pJ201 plasmid (pBB25) was digested with Bsa I and ligated to remove the insertion site. The DNA fragment in Figure 9 was ligated into the pQE80 vector (previously digested with EcoR I and HinD III). These two digested fragments were then ligated to create a plasmid containing the ribosomal binding site, HEAT1 dimer sequence, and 6x-His tag. This plasmid was named pBB31 and was used to express the HEAT1 dimer.

The tetramer and hexamer were cloned by the following protocol: the Adapter in vector pJ201 was digested with *Bsa* I, while the Insert in pJ201 was sequentially digested with BsmB I and then Bbs I. To prevent the Adapter from ligating to itself, the 3' ends of the Adapter were

dephosphorylated with phosphatase. The dephosphorylated Adapter in pJ201 and the digested HEAT1 Insert DNA fragment were then ligated together. The Insert was added in a much higher concentration than the Adapter (1x Adapter:3x Insert and 1x Adapter:5x Insert). Variations of these concentrations resulted in the creation of the tetramer and hexamer. Aliquots of the ligated DNA were sent for sequencing to confirm that they were correct. The sequence was then digested with EcoR I and HinD III and ligated into the similarly digested pQE80 expression vector. The tetramer plasmid was named pBB34, and the hexamer plasmid was named pBB35. The final DNA sequences are shown in Figures 8 and 9. See Supplemental Information for complete DNA sequences.

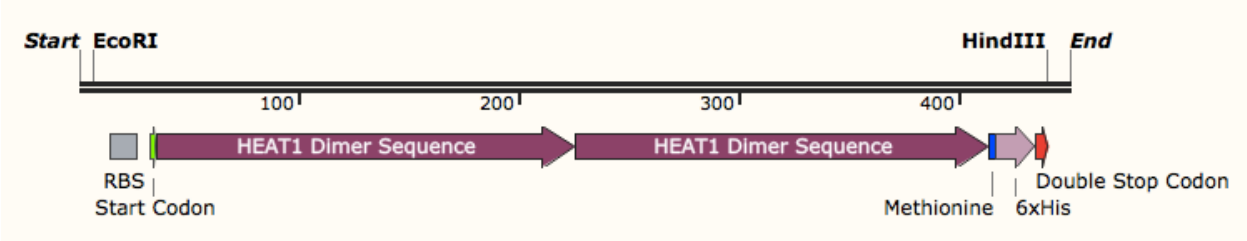


Figure 8. HEAT1 tetramer sequence with restriction sites for insertion into pBB23.

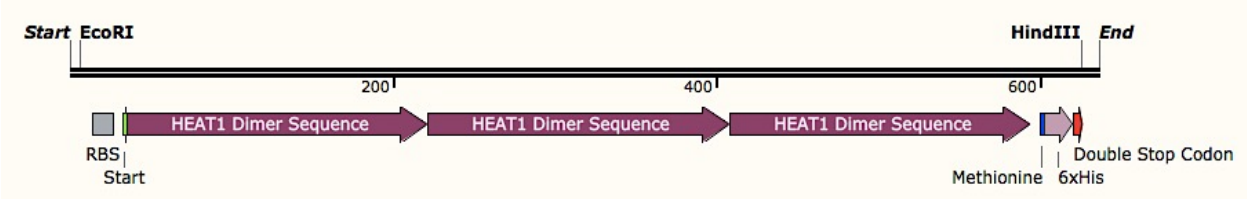


Figure 9. HEAT1 hexamer sequence with restriction sites for insertion into pBB23.

The proteins were expressed in *E. Coli* BL21 cells in the following manner: from a single colony on an LB/Amp plate, 15 mL cultures were grown overnight at 37°C. The cultures were

then added to 700 mL expression flasks of LB/Amp (17.5 g/700μL) and grown to an OD_{600} of 0.6-0.8. They were then induced with 0.5 mM IPTG and over-expressed overnight at 30°C. Cells were harvested by centrifugation at 1,000 rpm for 2 hours at 4°C.

It is important to note that expressions are usually spun down at 4,000 rpm for 20 minutes at 4°C; however, the centrifuge used for this protocol could not reach such speeds. This change probably adversely affected the overall yield.

Lysis, inclusion body washing, and solubilization

Cells were resuspended in Lysis Buffer A (50 mM Tris-HCl, 0.1 mM EDTA, 5% glycerol, 0.1 mM NaCl, pH 8.0) and subjected to three freeze-thaw cycles at -80°C and 37°C, respectively. The following were then added to the cell solution: 4 mL lysozyme (50 mg/mL), 10 μL Benzonase nuclease, 200 μL 1 M $MgCl_2$, and 200 μL Protease Inhibitor Cocktail. The solution was shaken overnight at 30°C.

Prior to separating the soluble and insoluble portions of the cell lysate, 2 mL Triton-X 100 was added to the solution. The solution was shaken at 4°C for one hour, and then sonicated for 30 minutes (9 seconds on, 9 seconds off, 4°C probe tip temperature, 21% amplitude). The solution was then spun down at 12,000 rpm for 20 minutes at 4°C. The supernatant and pellet were called “IB Wash 1 Super” and “IB Wash 1 Pellet”.

The IB Wash 1 Pellet was then resuspended in 120mL Lysis Buffer B (50 mM Tris-HCl, 0.1 M NaCl, 5% glycerol, 1% Triton-X 100, pH 8.0) and shaken at 4°C for two hours. This solution was sonicated for 30 minutes (9 seconds on, 9 seconds off, 4°C probe tip temperature, 21% amplitude), and spun down at 12,000 rpm for 20 minutes at 4°C. The supernatant and pellet were called “IB Wash 2 Super” and “IB Wash 2 Pellet”.

The IB Wash 2 Pellet was resuspended in 80mL Lysis Buffer B, shaken overnight at 4°C, and sonicated under the same conditions as the previous step. The sonicated solution was spun down at 12,000 rpm for 20 minutes at 4°C, and the supernatant and pellet were called “IB Wash 3 Super” and “IB Wash 3 Pellet”.

The IB Wash 3 Pellet was resuspended in 80mL Lysis Buffer C (50 mM Tris-HCl, 0.1 M NaCl, 5% glycerol, pH 8.0) and shaken for at least 30 minutes at room temperature. The solution was then spun down at 12,000 rpm for 20 minutes at 4°C. These were labeled “IB Wash 4 Super” and “IB Wash 4 Pellet”. This was the final step of inclusion body washing.

Aliquots of pellet and supernatant were taken at each of the four washing steps and assayed via SDS-PAGE. The gels showed that the pelleted inclusion bodies became more homogeneously composed of the protein of interest.

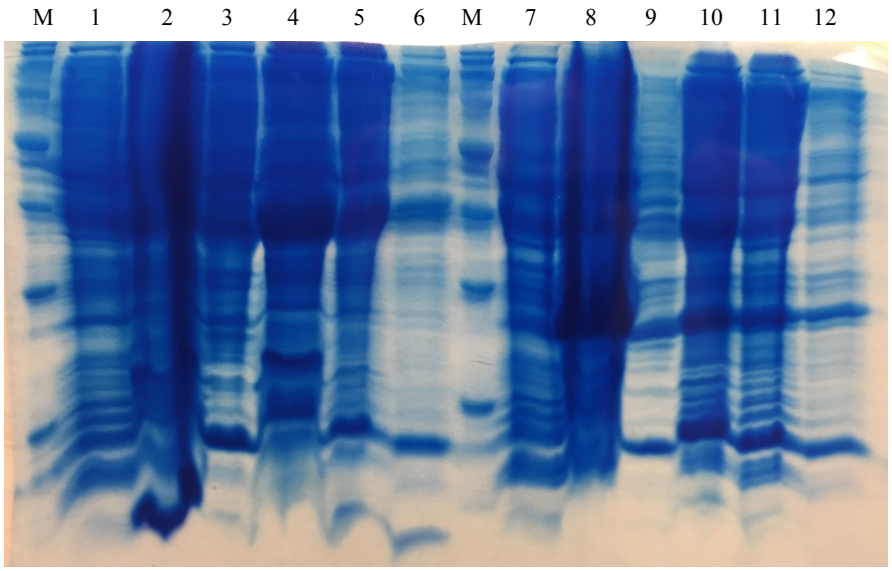


Figure 10. 16% SDS-PAGE: HEAT1 Dimer and HEAT1 Hexamer Inclusion Body Washing Supernatant Aliquots and Expression Aliquots. Key: **(M)** Perfect Protein Marker Ladder, 10-225 kDa; **(1)** HEAT1 Dimer, 0hr expression aliquot; **(2)** HEAT1 Dimer, 20hr expression aliquot; **(3-**

6) HEAT1 Dimer, IB Wash 1, 2, 3, 4 Supernatants; **(7)** HEAT1 Hexamer, 0hr expression aliquot; **(8)** HEAT1 Hexamer, 20hr expression aliquot; **(9-12)** HEAT1 Hexamer, IB Wash 1, 2, 3, 4, Supernatants.

The IB Wash 4 Pellet was then assayed for solubility in 2 M Urea Solubilization Buffer (50 mM NaH_2PO_4, 100 mM NaCl, 5% glycerol, 2 M urea, pH 8.0) and 4 M Urea Solubilization Buffer (same as previous buffer but with 4 M urea). The dimer, tetramer, and hexamer IB Wash 4 Pellets were soluble in the 2 M Urea Solubilization Buffer, as evidenced by SDS-PAGE (See Supplemental Information). They were also soluble in the 4 M Urea Solubilization Buffer, but it is preferable to solubilize in a lower concentration of denaturant, because the denaturant needs to be dialyzed out before characterizing the protein of interest. Therefore, solubilization in the 2M Urea Solubilization Buffer was the better choice. The solubilized proteins were isolated using affinity column purification.

Affinity column purification

HisPur™ Cobalt Resin (crosslinked 6% beaded agarose support, $\geq$10 mg/mL binding capacity; Thermo Scientific, Waltham, MA, USA) was used to purify His tagged proteins. Each column contained 2-3 mL of resin, and four columns were used to purify each 2.1 L expression. Columns were prepared by equilibrating with 30 mL dH_2O followed by 10-15 mL 2 M Urea Wash Buffer (50 mM NaH_2PO_4, 100 mM NaCl, 20 mM Imidazole, 2 M urea, pH 8.0). After equilibration, the protein was loaded onto the columns. The columns were washed again with 7-10 mL 2 M Urea Wash Buffer, and then the protein was eluted with 10-12 mL 2 M Urea Elution Buffer. To prepare the columns for storage, the beads were rinsed with 40 mL 20 mM MES

buffer (20 mM 2-(N-morpholine)-ethanesulfonic acid, 0.1 M NaCl, pH 5.0) followed by 30 mL dH_2O. Columns were stored in 1:1 volume of resin:ethanol (20% in water) at 4°C. Resin beads were re-used for up to three purifications and then discarded. Aliquots of purification samples were assayed via SDS-PAGE to confirm presence of the protein in the eluent. Figure 11 shows the aliquots of the collected eluents.

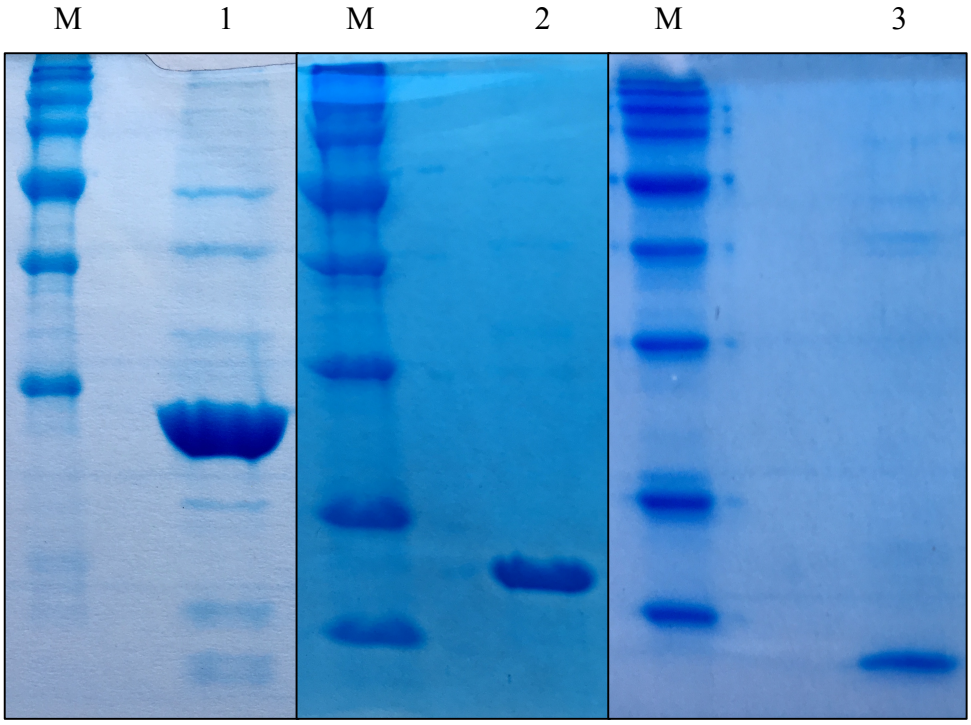


Figure 11. 16% SDS-PAGE: eluents from affinity column purifications. Key: **(M)** Perfect Protein Marker Ladder, 10-225 kDa; **(1)** HEAT1 Hexamer, 2 M Urea eluent; **(2)** HEAT1 Tetramer, 2 M Urea eluent; **(3)** HEAT1 Dimer, 2 M Urea eluent.

Refolding and assembly

The solubilized, purified proteins were assembled by dialyzing the 2 M Urea Eluent in a series of Refolding buffers with decreasing concentrations of urea. Proteins were dialyzed for 12 hours minimum per buffer, with stirring at 4°C.

The refolding protocol was as follows: aliquots of the 2 M urea eluent from affinity column purification were placed in appropriate MWCO SnakeSkin™ Dialysis Tubing (Thermo Scientific, Waltham, MA, USA) and immersed in Refolding Buffer 1 (1 M urea, 50 mM NaH_2PO_4, 100 mM NaCl, 5% glycerol, pH 8.0). Eluent was dialyzed for 12 hours minimum and then the buffer was changed to Refolding Buffer 2 (same as Refolding Buffer 1 but with 0.5 M urea). After $\geq$12 hours, the buffer was changed to Refolding Buffer 3 (same as Refolding Buffer 1 but with 0 M urea). After this dialysis step, proteins were dialyzed in dH_2O and lyophilized. Proteins labeled "Refolded" means that they were refolded using the series of Refolding buffers prior to dialysis in water and lyophilization. Proteins labeled "Not Refolded" or "Non-Refolded" were not dialyzed in the Refolding buffers, but were directly dialyzed in water and then lyophilized.

Proteins were assembled in 10 mM buffers at pH 5, 6, 7, and 8.

Characterization

The isolated proteins were characterized using circular dichroism (CD) using a JASCO J-810 Spectropolarimeter (Oklahoma City, OK, USA) to screen for presence of the expected alpha-helical secondary structure. Spectra were taken at 20°C with a 2 nm bandwidth, 2 second response, 100 nm/min scanning speed, standard sensitivity, and an accumulation of 3.

Transmission Electron Microscopy (TEM) was carried out using a CEH02.2EM Hitachi HT-7700 Electron Microscope Unit (80 kVp, Copper grids, 7.6 μAmp beam current, 5kX magnification; Chiyoda City, Tokyo, Japan) and was used to image assemblies of the concatemers. Samples were made using carbon film 200 mesh copper grids (CF200-CU, Electron Microscopy Sciences, Hatfield, PA) and stained with both Nano-Van® (methylamine

vanadate stain; Nanoprobes, Yaphank, NY, USA) and Nano-W® (methylamine tungstate stain; Nanoprobes, Yaphank, NY, USA) in a 1:1 ratio.

Results and Discussion

While working on this project, I became focused on the effects of the solubilization and isolation protocols on the assemblies of the proteins. The proteins I worked with were expressed in the inclusion bodies, which means that during expression, they became insoluble aggregates that lacked functional activity.[10] Solubility is an extremely important factor when expressing and purifying protein, because a lack of solubility is usually the first indication that a protein is improperly folded.[11]

There are two major components to recovering native, functional protein from inclusion bodies: solubilization and refolding. Usually, solubilization is done in a similar fashion to the IB Solubilization protocol I used; detergents are used to gently wash the inclusion bodies several times by vortexing and centrifugation. Using this protocol, my recombinant proteins were soluble in significantly lower concentrations of urea. Prior to the IB protocol, I used a general solubilization protocol directly after lysis and the lysed pellet was only soluble in 6-8 M urea. The IB solubilization protocol permitted solubilization in a concentration of only 2 M urea.

The second step to recover protein is to refold it. To refold protein, the denaturant used for solubilization must be entirely removed. Because I was able to solubilize in a much lower concentration of urea using the IB Washing Protocol, I was able to remove the denaturant much more easily. Dialysis in the pH-controlled Refolding buffers was used to slowly dial down the concentration of urea until it was completely removed. Starting at a concentration of only 2 M

urea, I was able to significantly shorten my dialysis time, which is advantageous because extended periods of dialysis risks leaking which often results in the loss of a lot of product.

I developed a refolding protocol that involved four refolding buffers, each with a 0.5 M stepwise decrease in urea. I found that the samples that had been refolded in the series of Refolding buffers prior to dialysis in water formed more ordered and homogenous assemblies than the samples that not been refolded. The Refolding buffers stabilized the assemblies of the proteins because the denaturant was removed slowly, in a controlled environment. Direct dialysis in water removed the denaturant quickly, which can cause proteins to misfold. This happens because hydrophobic residues become improperly exposed on the surface of the protein when the folding process is rushed,[10] creating a sticky area that increases the chances for other proteins to misfold, because it attracts other hydrophobic residues on surrounding proteins.[11]

This discussion section will compare and contrast the secondary structures and assemblies of proteins that were dialyzed in the Refolding buffers and those that were not refolded. The most alpha-helical and intense CD signatures were observed in the refolded proteins. Samples that had been directly dialyzed in water had noisy CD spectra that was not consistent pre- and post-annealing, suggesting that the proteins were not in a thermodynamically stable conformation after dialysis in water. It is reasonable to suggest that the Refolding buffers encouraged proper folding, because the pre- and post-annealing spectra of the refolded aliquots were consistently alpha-helical with similar maxima and minima, suggesting that they were in a thermodynamically stable folded state immediately after refolding.

The EM images also support this hypothesis. The refolded aliquots consistently had larger and more homogenous assemblies that contained fewer fragments. Aliquots directly

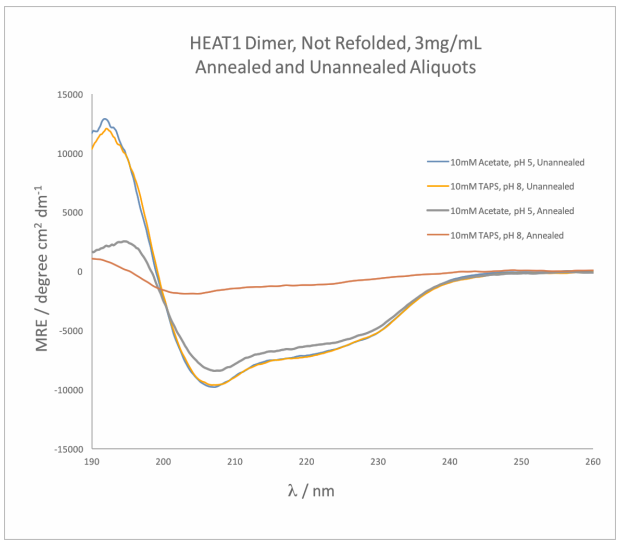
dialyzed in water were fragmented, lacked homogeneity across the grid, and the larger assemblies present lacked organization and order.

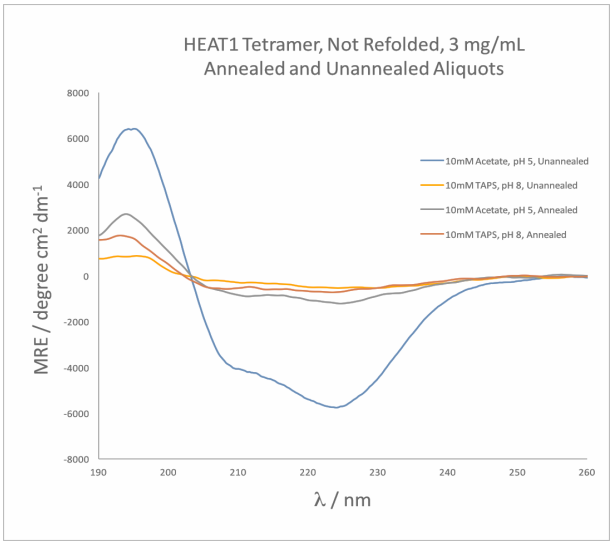
Non-Refolded Proteins

Aliquots of the IB Solubilized and affinity column purified concatemers were directly dialyzed in deionized water. Protein precipitate was visible during dialysis for each concatemer. After dialysis, the soluble portions were assayed for presence of secondary structure, but there was no appreciable CD signature for any of the samples, suggesting that the protein was in the precipitate. To isolate this precipitate, the contents of the dialysis tubing enclosures were lyophilized. The lyophilized samples were dissolved in water and assembled.

Because these proteins were not refolded, they precipitated more dramatically during dialysis when the denaturant was suddenly removed. Additionally, the non-refolded concatemers were less soluble than the refolded concatemers even after lyophilization, further supporting the idea that the Refolding buffers support proper folding. All three of the concatemers are predicted to have good water solubility, so in their properly folded forms they should be water-soluble.

The CD spectra of each protein is shown below (Figures 12-14) at pH 5 (10 mM Acetate) and pH 8 (10 mM TAPS), annealed and unannealed. Close to their isoelectric points (all concatemers' PI's were between pH 5.5-6.5), the proteins tended to precipitate out of solution and display distorted CD spectra. It follows that the most alpha-helical signatures were observed in environments with pHs distant from the isoelectric points of the proteins.





Figures 12 and 13. CD spectra of HEAT1 dimer, not refolded, annealed and unannealed aliquots at pH 5 and 8; CD spectra of HEAT1 tetramer, not refolded, annealed and unannealed aliquots at pH 5 and 8.

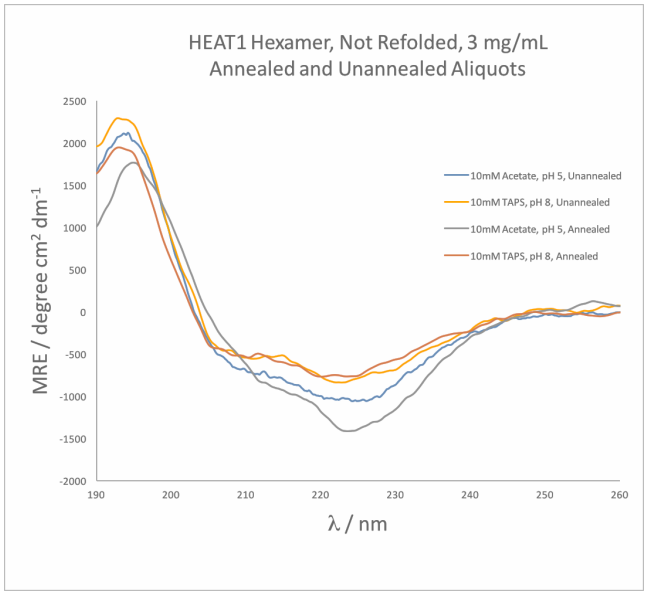
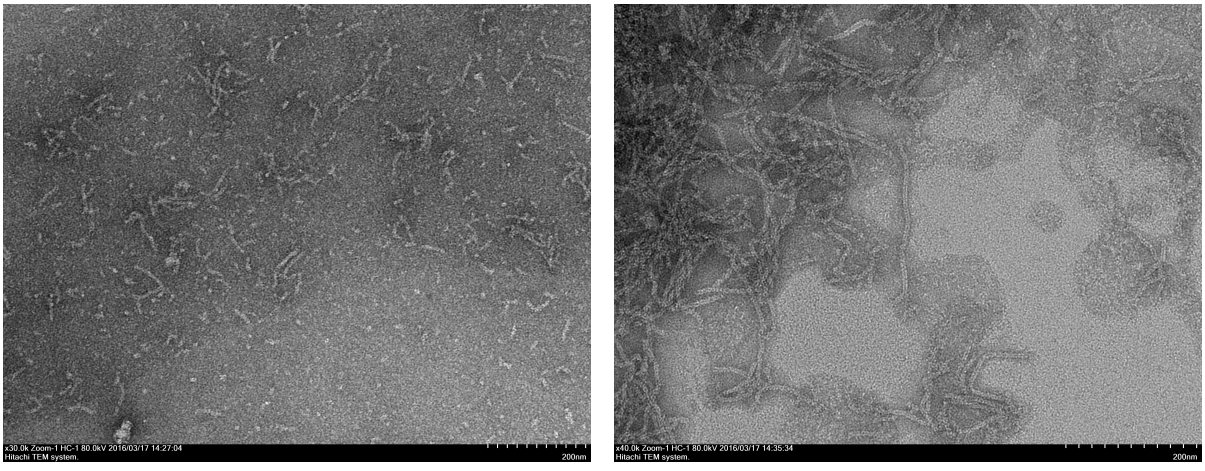


Figure 14. CD spectra of HEAT1 hexamer, not refolded, annealed and unannealed aliquots at pH 5 and 8.

All CD signatures of the non-refolded proteins are at least weakly alpha-helical. The hexamer shows consistent alpha-helicity in the annealed and unannealed samples, but this is not true of the dimer and tetramer whose signatures are dramatically different pre- and post-annealing. These differences suggest that dialysis in water caused the proteins to fold into a thermodynamically unstable state. Additionally, the signatures were noisy and ill-defined, particularly for the hexamer.

Images of the tetramer assemblies are displayed in Figures 15-17. Tubular assemblies were formed most successfully at pH 5 for all proteins. Figures 15 and 16 contrast the assemblies formed pre- and post-annealing. Pre-annealing, the tetramer is fragmented and disordered; post-annealing, there are far fewer fragments present, and elongated tubes are clearly visible. Lateral association is present in the assembled tubes, but this is not surprising because lateral association was found in all assemblies of the HEAT1 sequence with both the covalently-linked concatemers and the synthesized peptide monomer.



Figures 15 and 16. Left to right: HEAT1 tetramer, 10 mM Acetate, pH 5, Unannealed; HEAT1 tetramer, 10 mM Acetate, pH 5, Annealed.

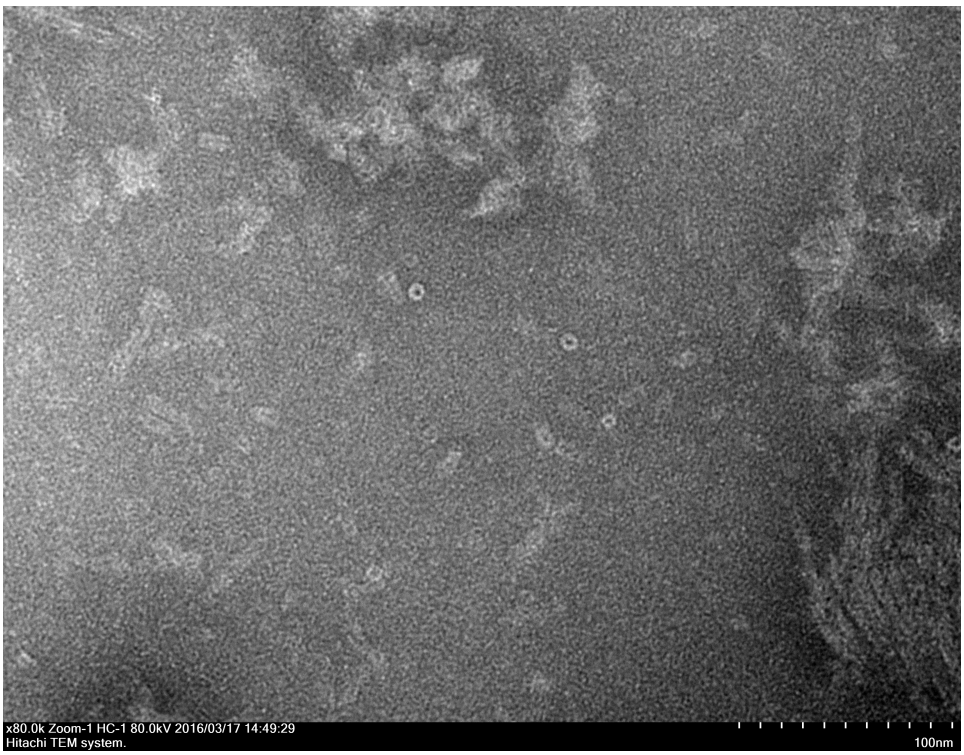


Figure 17. HEAT1 tetramer, 10 mM Acetate, pH 5, Annealed.

Figure 17 is of particular interest because of the clearly defined and abundant rings. These rings illustrate a cross-sectional view of the assembly of the tubes.

Images of the non-refolded dimer and hexamer can be seen in the Supplemental Information. Their assemblies were fragmented and lacked order and length pre-annealing. Post-annealing, they were more homogenous and less fragmented, but their assemblies were far more impressive when they were refolded.

Refolded Proteins

Due to the time constraints of this project and its deadline, the tetramer was not examined under refolding conditions.

The first step to characterize the refolded dimer and hexamer was to look at their CD signatures. Figures 18 and 19 show the CD spectra. Note the consistency in intensities of the signatures pre- and post-annealing, and the well-defined minima and maxima that are indicative of alpha-helicity.

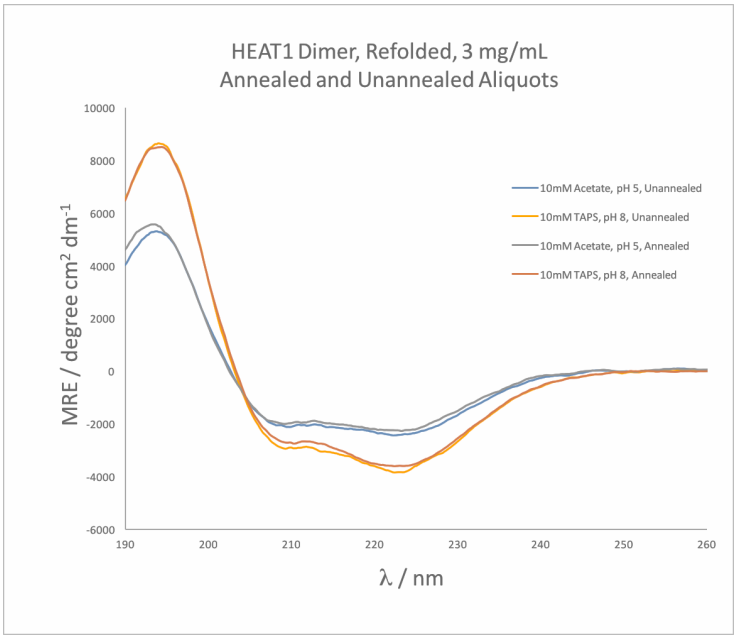


Figure 18. HEAT1 dimer, refolded, annealed and unannealed aliquots at pH 5 and 8.

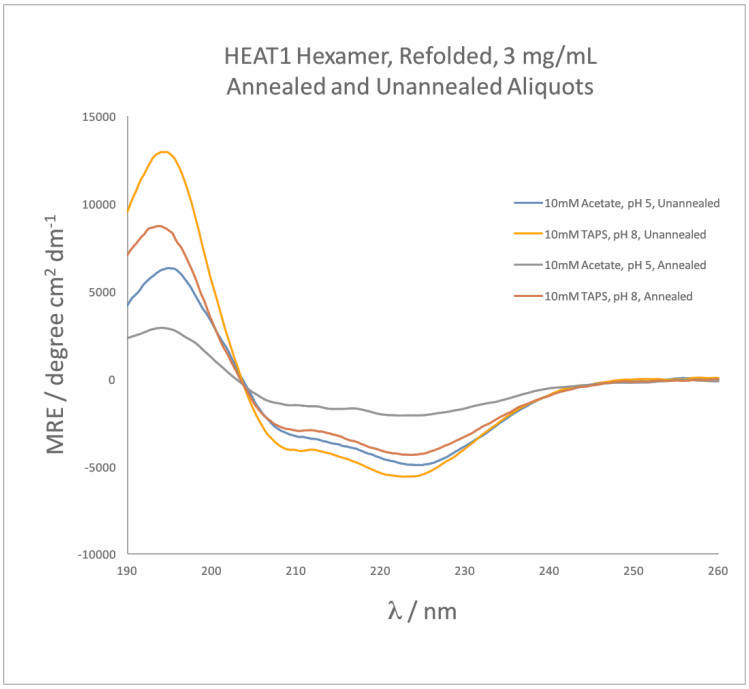
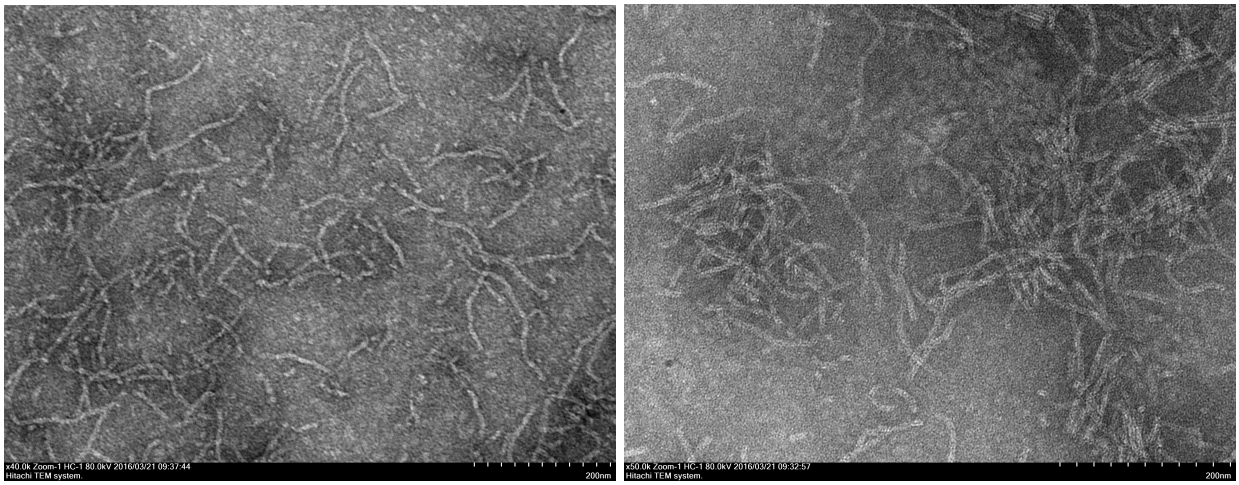
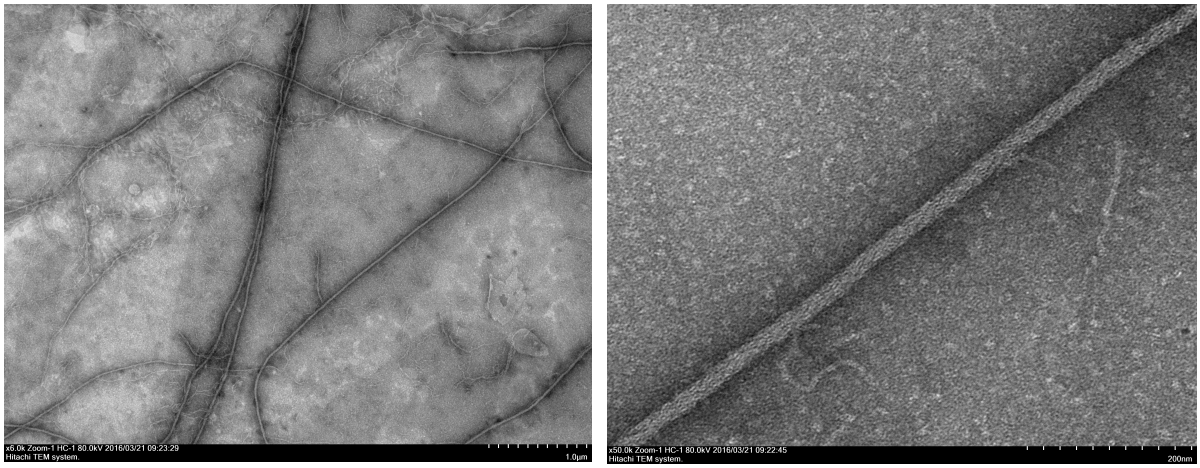


Figure 19. HEAT1 hexamer, refolded, annealed and unannealed aliquots at pH 5 and 8.

Although it appears that the dimer is more consistent pre- and post-annealing in its CD signatures, the hexamer has more intense minima and maxima that suggest it is more alpha-helical in its secondary structure than the dimer. Both the dimer and hexamer have the most intense maxima and minima at pH 8. The samples were assembled and imaged using TEM, and their assemblies pre- and post-annealing are in Figures 20-23.



Figures 20 and 21. Left to right: HEAT1 dimer, refolded, unannealed, 10 mM Acetate, pH 5; HEAT1 dimer, refolded, annealed, 10 mM Acetate, pH 5.



Figures 22 and 23. Left to right: HEAT1 hexamer, refolded, unannealed, 10 mM Acetate, pH 5; HEAT1 hexamer, refolded, annealed, 10 mM Acetate, pH 5.

The unannealed dimer has tubular structures at pH 5. This suggests that the refolding protocol does have a stabilizing effect on the formation of tubes because assemblies of this size and order are not observed in the non-refolded unannealed aliquots. After annealing, there are

fewer fragments present around the tubes formed, which implies that the structures formed post-annealing are more ordered (Figure 21).

The unannealed hexamer forms very long tubular structures at pH 5 (Figure 22); tubes longer than one micron in length are observed consistently throughout the grid. However, lateral assembly seems to be necessary to form these structures. No tube is found by itself. Post-annealing, these tubes seem to interact with one another, forming bundled assemblies that almost appear to be braided (Figure 23). Lateral association is increased post-annealing, suggesting that the bundled structures are more thermodynamically stable.

Additional images of the dimer and hexamer assemblies at pH 8 are shown in the Supplemental Information.

Conclusion and Future Aims

This thesis project focused on a consensus sequence in the literature that had been shown in our lab to produce homogeneous, thermostable tubular assemblies from a simple HEAT-like peptide sequence. This discovery is significant for protein materials biochemistry because it shows that an unsophisticated sequence can form complex and stable structures. This raises a fundamental research goal in protein materials chemistry: being able to predict the structure of a three-dimensional protein from its sequence. Cataloging the effects of point mutations on the assemblies and secondary structures of peptides and proteins is essential to better understanding what makes proteins assemble into their highly specific native structures, and why protocols like the refolding protocol I designed have such vivid effects on the characteristics of proteins.

The HEAT1 concatemer project is an extension of a project focused on understanding the effects of single-residue mutations on assembly formation. The covalently-linked concatemers

were created because they are more stable building blocks for assembly, enabling them to create larger and longer assemblies than the synthesized monomer peptide. By utilizing a prokaryotic host to express the concatemers quickly and in large quantities and developing a method to isolate and refold the proteins of interest, we have created more stable assemblies that are better-suited for large-scale applications, such as binding drugs or therapeutics within the center of the HEAT1 tubes.

Future directions for this project are to investigate the effects of different refolding environments: with and without glycerol present, and in the presence of other chemicals that enhance the formation of alpha-helical proteins, such as L-arginine. A more effective purification protocol, such as HPLC, would be helpful to prove that the protein of interest has indeed been isolated, so that its presence can be definitively proved with MALDI. Additionally, our laboratory's project focusing on the mutations of the HEAT1 consensus sequence has revealed sequences that better form the desired tubular assemblies without lateral association. Cloning and expression of these more promising mutant sequences is another direction this project could take.

Finally, this project could lead to the development of similarly designed projects focused on investigating assemblies formed from solenoid TRPs. They present an attractive template for protein design and engineering because they are flexible in sequence and adaptable to various environments. The HEAT1 motif is one of many solenoid proteins with the potential to form complex assemblies from a simple repeated sequence. Investing time and resources into researching these proteins could provide the context needed to unearth necessary clues about how and why specific combinations of amino acids lead to different structures and functions of proteins. If researchers can more reliably understand how protein folding and assembly occurs, it will be easier to develop therapeutic proteins to cure or delay the onset of debilitating diseases.

Supplemental Information

Heat 1 Adapter Nucleotide Sequence (287nt)
GCTTAGAATTCATTAAAGAGGAGAAATTAACCATGGGTGCGAGACCGTCGACGGTC
TCCGGTGATGAACGCGCAGTGGAACCACTGATTAAAGCATTGAAAGATGAGGACTG
GTATGTACGTCGTGCGGCCGCGGAGGCGCTGGGTAAGATTGGCGATGAGCGTGCCG
TAGAACCGCTTATTAAAGCGCTTAAAGATGAAGATTGGTATGTACGTCGCGCAGCCG
CGGAAGCGCTGGGTAAGATTGGATCCATGCATCACCACCATCACCATTAATAAGCTT
GGTCC

Heat Insert Nucleotide Sequence (227nt)
GCATCAAGCTTGAAGACTTGGTGATGAACGCGCAGTGGAACCACTGATTAAAGCAT
TGAAAGATGAGGACTGGTATGTACGTCGTGCGGCCGCGGAGGCGCTGGGTAAGATT
GGCGATGAGCGTGCCGTAGAACCGCTTATTAAAGCGCTTAAAGATGAAGATTGGTA
TGTACGTCGCGCAGCCGCGGAAGCGCTGGGTAAGATTGGTGTGAGACGGGATCCAC
TAG

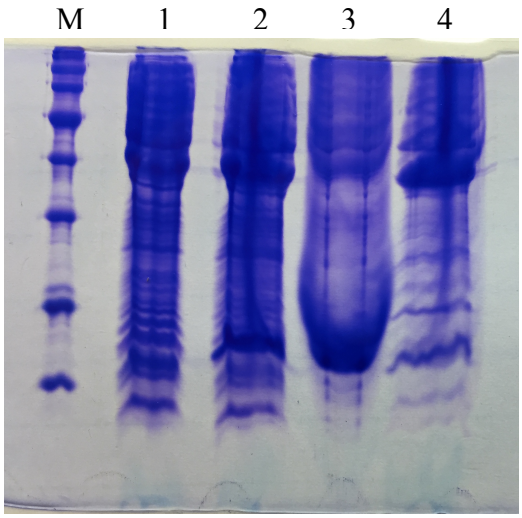


Figure S1. 16% SDS-PAGE: HEAT1 tetramer expression and solubilization aliquots. Key: (M) Perfect Protein Marker Ladder, 10-225 kDa; (1) pBB 34, 0 hr expression aliquot; (2) pBB 34, 20 hr expression aliquot; (3) H1.T, IB wash 4 pellet in 2 M urea, supernatant; (4) H1.T, IB wash 4 pellet in 2 M urea, pellet.

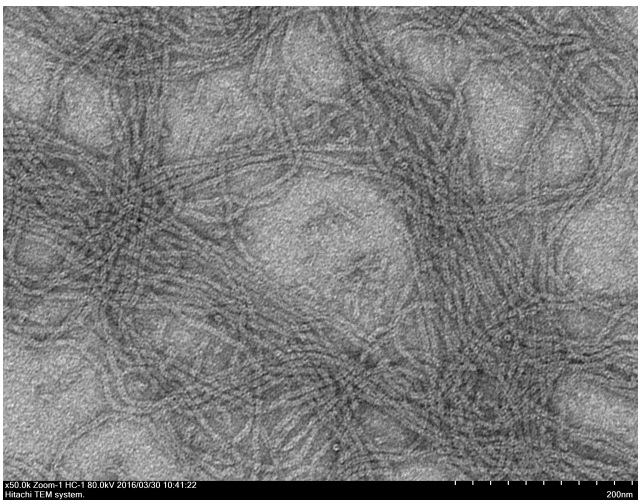
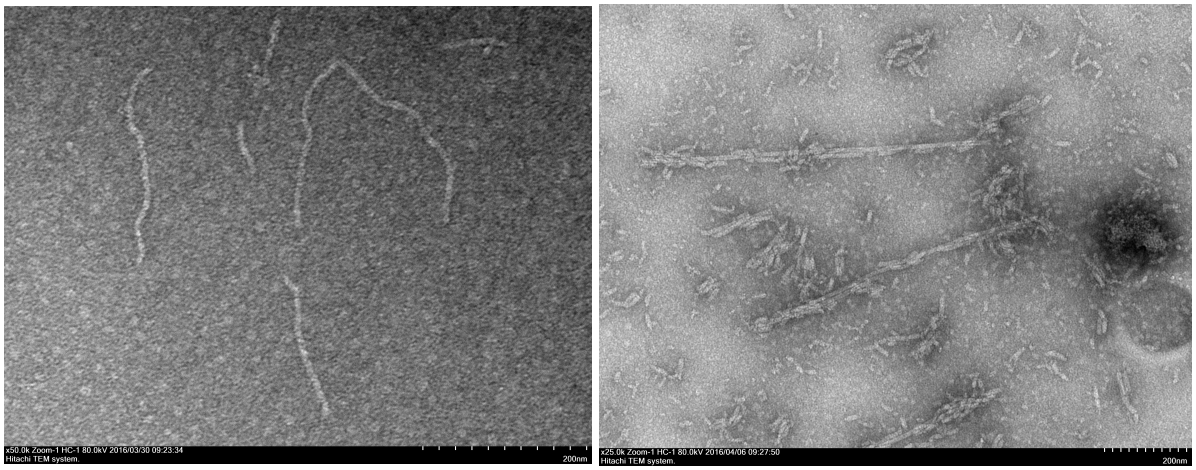


Figure S2. HEAT1 dimer, not refolded, annealed, 10 mM TAPS, pH 8.



Figures S3 and S4. HEAT1 dimer, not refolded, not annealed (left) and annealed (right), 10 mM Acetate, pH 5.

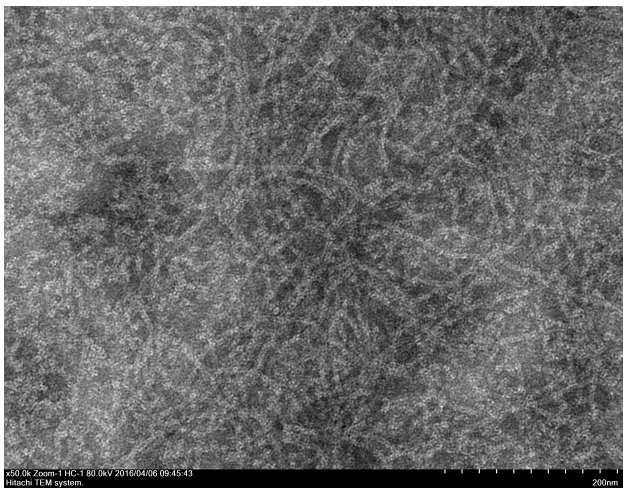


Figure S5. HEAT1 hexamer, not refolded, annealed, 10 mM TAPS, pH 8.

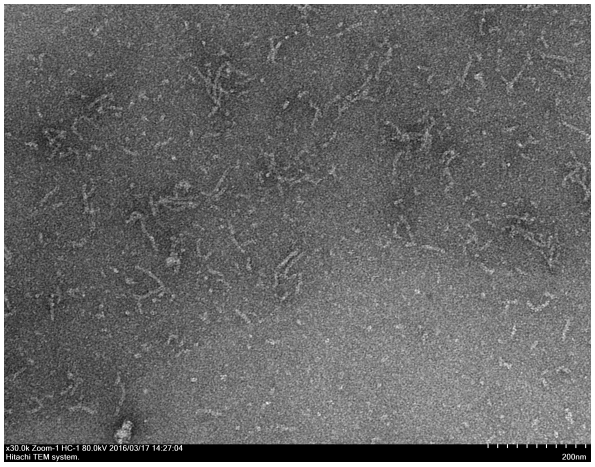


Figure S6. HEAT1 tetramer, not refolded, not annealed, 10 mM Acetate pH 5.

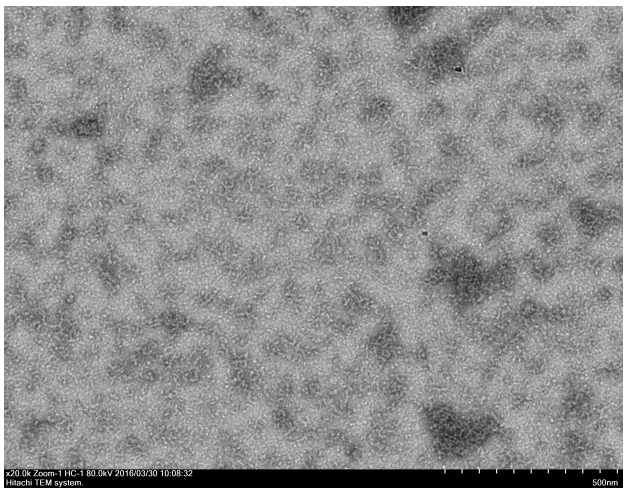
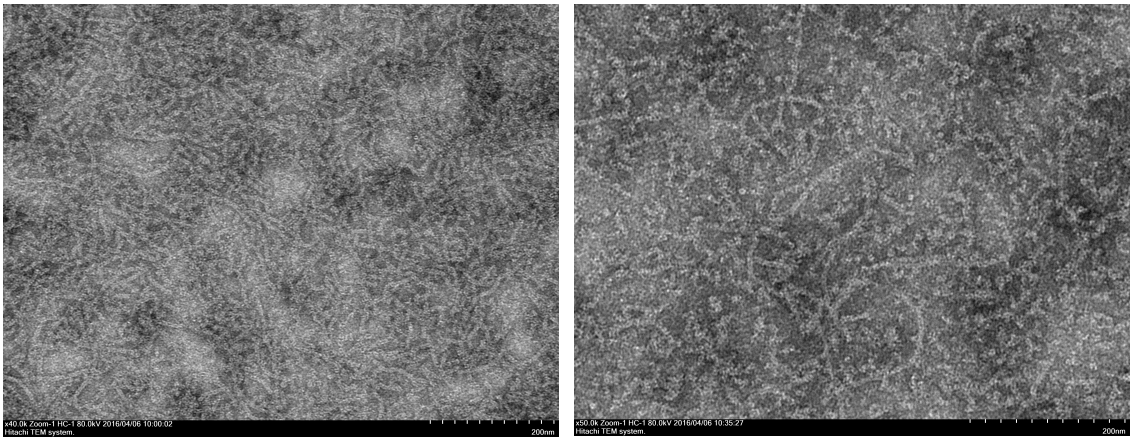


Figure S7. HEAT1 dimer, refolded, not annealed, 10 mM TAPS, pH 8.



Figures S8 and S9. HEAT1 hexamer, refolded, not annealed (left) and annealed (right), 10 mM TAPS, pH 8.

References

(1) Kobe, B. Turn up the HEAT. *Structure*. 1999. 7: R91-R97.

(2) Hollenbeck, J. J., et al. Designed Ankyrin Repeat Proteins as Scaffolds for Multivalent Recognition. *Biomacromolecules.* 2012. 13: 1996-2002.

(3) Minard, P. *et al*. Design, Production, and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats. *Journal of Molecular Biology.* 2010. 404: 307-327.

(4) Plückthun, A. Designed Ankyrin Repeat Proteins (DARPins): Binding proteins for research, diagnostics, and therapy. *Annu. Rev. Pharmacol. Toxicol.* 2015. 55: 489-511.

(5) Rosano, G.L. and Ceccarelli, E.A. Recombinant Protein Expression in *Escherichia Coli*: advances and challenges. *Frontiers in Microbiology*. 2014. 5(172): 1-17.

(6) Kajava, A. V. Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology.* 2012. 179: 279-288.

(7) Kobe, B. and Kajava, A. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *TIBS*. 2000. 25: 509-515.

(8) Tripp, K.W. and Barrick, D. Enhancing the Stability and Folding Rate of a Repeat Protein through the Addition of Consensus Repeats. *Journal of Molecular Biology.* 2007. 365: 1187-1200.

(9) Plückthun, A. *et al*. Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core. *Journal of Molecular Biology.* 2008. 376: 1282-1304.

(10) Singh, S.M. and Panda, A.K. Solubilization and Refolding of Bacterial Inclusion Body Proteins. *Journal of Bioscience and Bioengineering*. 2005. 99(4): 303-310.

(11) Tusmoto, K., Kumagai, I., et al. Practical considerations in refolding proteins from inclusion bodies. *Protein Expression and Purification*. 2003. 28: 1-8.