NORTHWESTERN UNIVERSITY

Design and Production of Protein-Based Polymers for Application as "Drag-Tags" in Free-Solution DNA Sequencing

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemical Engineering

By

Jennifer Sue Lin

EVANSTON, ILLINOIS

DECEMBER 2008

ABSTRACT

Design and Production of Protein-Based Polymers for Application as "Drag-Tags" in Free-Solution DNA Sequencing

Jennifer S. Lin

End-Labeled Free-Solution Electrophoresis, or ELFSE, is an alternative strategy for DNA sequencing that proposes to eliminate the need for a viscous sieving matrix for size-based DNA separation. In this bioconjugate method, a perturbing entity or "drag-tag" is attached to differently sized DNA fragments produced by the Sanger reaction. This drag-tag alters the overall charge-to-friction ratio so that DNA separation by size can be accomplished by free-solution electrophoresis. Rapid separations with long read lengths are theoretically possible using ELFSE. Application of ELFSE to integrated "lab-on-a chip" microfluidic devices currently under development is facilitated by the absence of a viscous polymer solution.

For successful DNA sequencing, this perturbing entity needs to be large, water-soluble, preferably uncharged, monodisperse, and also have a point for unique attachment to DNA. A non-natural repetitive polypeptide, or "protein polymer", has the potential to meet these numerous requirements for optimal ELFSE performance. A small (127 amino acids) drag-tag was previously used to demonstrate that ELFSE sequencing is possible using a protein polymer as the drag-tag. Obtaining a sufficiently large and monodisperse protein polymer drag-tag has been the focus of this research and achieving this goal proved to be more challenging than originally expected.

Charged and uncharged protein polymers of varying lengths were evaluated for their potential as drag-tags. Sequences incorporating additional hydrophobic residues were also investigated as well as an alternative purification strategy using self-cleaving affinity tags. Proteins expressed with an *N*-terminal affinity tag were compared to those expressed with a *C*-terminal affinity tag to determine the best method for obtaining monodisperse protein polymers. The biophysical properties of these drag-tags were measured via spectroscopy. Extensive investigation into aspects of protein polymer design, production, and purification was required to finally produce a sufficiently monodisperse drag-tag of double the previous size, bringing us closer towards the goal of achieving long-read ELFSE sequencing.

Acknowledgments

I would like to thank my advisor, Dr. Annelise Barron, for her support and guidance over the years as well as her endless enthusiasm for the ELFSE Project. I would also like to thank our collaborator, Dr. Gary Slater and his research group at the University of Ottawa who have been advancing the theoretical half of ELFSE. I would also like to acknowledge the assistance of my thesis committee members, both past and present, Dr. William Miller, Dr. Lonnie Shea, Dr. Andreas Matouschek, Dr. Guillermo Ameer, and Dr. Sandip Ghosal.

I would also like to thank the Barron research group for their support over the years and the opportunity to work alongside such a diverse group of people. The ability to easily tap into the different fields of expertise available in the group has no doubt aided in advancing this research project. In particular, I would like to thank past and present members of the ELFSE team, Dr. Wyatt Vreeland, Dr. Jong-In Won, Dr. Felicia Bogdan, Dr. Robert Meagher, Dr. Russell Haynes, , Jennifer Coyne, Xiaoxiao Wang, Jordan Bertram, and Loiusa Carr who have all helped contribute to the ELFSE project over the years and continue to push the research forward. In fact, many of the results would not have been possible without the assistance of Dr. Robert Meagher and Jennifer Coyne who analyzed the many drag-tag-DNA conjugates by free-solution capillary electrophoresis. Additionally, Nicolynn Davis introduced several different and valuable molecular biology techniques when she joined our protein polymer group from the Koltover lab. I am especially grateful to Dr. James Broering, Dr. Matthew Kerby, and Jennifer Coyne for their assistance in reviewing this document.

Finally, I would like to thank my family for their love and support over the many years. I would not have made it through all these years of schooling if not for their continued encouragement.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction to End-Labeled Free-Solution Electrophoresis (ELFSE) of DNA

1.1 Importance of DNA sequencing technology

On April 14th, 2003, 50 years after the deduction of the DNA double helical structure, the International Human Genome Sequencing Consortium announced that a finished sequence (99%) of the human genome had been completed with 99.99% accuracy [1-4]. The project, which began in 1990, cost a total of $3 billion [5, 6]. In 2004, the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH), launched a race to develop new technologies that would reduce the cost of sequencing a single human genome from the current price of $10-20 million to $100,000 in the near future and ultimately to $1000 [7]. A reduction in cost to $100K is key to the success of the National Cancer Institute's (NCI) The Cancer Genome Atlas (TGCA) Project, announced in 2005, which seeks to gather comprehensive genetic data on all major human cancers by sequencing the DNA of multiple patient tumor samples [8]. 15,000 tumor samples are expected to be fully analyzed in this effort within a budget of $1.5 billion, which amounts to $100K per sample [9]. A further cost reduction to $1000 per genome would allow individual genomes to be sequenced as part of general medical care. Additionally, in October 2006, the Archon X Prize for Genomics was announced, promising $10 million to the first group who can sequence 100 human genomes in under 10 days [10, 11].

Although a composite human genome has already been sequenced there is still a need to sequence large, complex genomes [9]. The ultimate goal is personalized medicine, where one's medical care is tailored to an individual's genes and genomic sequence analysis is part of a regular physical exam. Personalized medicine would lead to better medication choices, safer dosages, improvements in drug development, and reduced healthcare costs [3]. Over 2 million people suffer serious adverse drug reactions each year in the U.S. leading to as many as 137,000 deaths [12]. Many of these adverse drug reactions are due to variations in genes which code for enzymes. Approximately half of all drugs are metabolized by the cytochrome P450 family of enzymes [13]. Over 30 different forms of these enzymes exist, each coded by a different gene. Variations in these genes can lead to decreased or increased metabolism of certain drugs. The Food and Drug Administration (FDA) approved in 2005 the AmpliChip cytochrome P450 test which detects variations in two key cytochrome P450 genes, CYP2D6 and CYP2C19 [14, 15]. This test is one of a growing number of personalized medicine drugs, treatments, and diagnostics currently being used in the medical field.

For the whole genome shotgun method, genomic DNA is fragmented randomly and cloned into *E. coli* to produce a random library, for example, of 2, 10, or 50 kbp inserts [2]. These clones are sequenced at both ends producing paired-end reads and the results are assembled by computer to form the complete genome by aligning matching sequences [16]. The assembly process is facilitated by knowledge of the distance between paired-end reads based on the known library size [17]. To fully characterize the genomic variations that contribute to an individual's genetic identity will require more than single nucleotide polymorphism (SNP) genotyping or

directed resequencing. Human genome resequencing projects should include both repetitive and protein-encoding regions to provide a comprehensive assessment [9]. Approximately 600-700 bases would be needed by existing algorithms to properly assemble large repeat-rich genomes [18].

1.2 Background

1.2.1 Size-based DNA separation: DNA sequencing by Sanger reaction

DNA is a polymer composed of deoxyribose sugars linked by phosphodiester bonds through their 3' and 5' carbons. Each sugar group is attached, at the 1' position, to one of four nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The negatively charged phosphate groups of each base dominate the overall charge of the DNA such that the negative charge of DNA scales linearly with chain length or the number of bases. Under the influence of an electric field, in an aqueous solution with pH > 3.0, DNA migrates towards the anode. Positively charged counterions present in the aqueous buffer surround the negatively charged DNA, and move in the opposite direction during electrophoresis. This movement of cations in the opposing direction shields hydrodynamic interactions between DNA segments [19]. The result is that each DNA monomer contributes equally to the overall hydrodynamic drag during electrophoresis (frictional forces are local) so that the net molecular friction scales linearly with the chain length and number of bases.

Therefore, in free-solution electrophoresis, with charge (q [Coulombs]) and friction (f [kg/sec]) both increasing linearly with DNA length, DNA separation is not possible [20]. The following equation approximates the electrophoretic mobility of DNA (μ [cm^2/V-sec]):

$$\mu \sim \frac{q}{f} = \mu_0 \qquad \text{where } \mu_0 \text{ is a constant.} \qquad (1\text{-}1)$$

Hence, all DNA regardless of size will migrate at the same velocity in an electric field in free solution.

A polymer matrix or gel provides an "obstacle course" for the DNA to migrate through during electrophoresis allowing separation on the basis of length. Larger DNA is slowed down more by frictional interactions with the polymer matrix than smaller DNA, so the smaller DNA molecules elutes first, followed by larger DNA. There are limits to the range of applicability of this separation technique with respect to DNA size, because of electric field-induced band broadening and molecular orientation effects (biased reptation) that cause large DNA to align with the electric field [21-24]. An average capillary sequencing run yields approximately 650-700 bases in 1-2 hours [9]. 1300 bases in 2 hours appears to be the upper limit for a highly optimized system [25].

The DNA sequencing technique that has been widely used in genomic centers for many years is called the Sanger dideoxy chain termination reaction [26]. In this method, a template DNA is copied by extension of an annealed primer. A DNA polymerase incorporates deoxyribonucleotide triphosphates (dNTPs) into the growing chain until a dideoxyribonucleotide triphosphate (ddNTP), present at low concentration, is randomly added in place of the corresponding dNTP approximately ~1% of the time [16]. Because ddNTPs lack a hydroxyl group on the 3' carbon (required for DNA polymer extension), the DNA chain can no longer grow after a ddNTP is incorporated. Additionally, a fluorescent label is attached to the ddNTP terminator, with an emission spectrum that corresponds to a particular base (A, T, C, or G). This

distribution of DNA fragments differing by one base is separated according to size by electrophoresis through a polymer network. A laser-induced fluorescence (LIF) detector and base-calling software translate the raw data into a sequence of bases.

1.2.1.1 Capillary array electrophoresis

Automated capillary array electrophoresis (CAE) has replaced the previous, time-consuming method of using ultra-thin slab gels as the preferred method for high throughput DNA sequencing [27]. Multiple samples are run in parallel by CAE and the polymer matrix required for size-based DNA separation are loaded and replaced between sequencing runs. In addition, the cylindrical capillary geometry allows for efficient dissipation of Joule heat, which results from the passage of current required to move the charged DNA molecules [22]. Typically, applied voltages for DNA sequencing by CAE are 150-175 V/cm, while a typical current is ~ 10 μA. Dilute solutions of uncrosslinked, linear polyacrylamide (LPA) or polydimethylacrylamide (pDMA) are used as the separation matrix. Expanding capillary arrays to include more capillaries (present instruments have as many as 384 in parallel) in order to increase throughput creates additional challenges for controlling matrix and sample injection into the array as well as for detection. The use of a lower electric field or longer capillaries can increase the read length beyond the 650-700 bases which is presently typical but at the expense of much longer running times (*i.e.,* >> 1-2 hours) [9].

A new sequencing technology that increases both the speed and read length of DNA sequencing is desirable. Even with additional optimization of current capillary instruments, either by reduction in the amount of sample and polymer consumed or with further increases in

automation, genome sequencing centers are unlikely to see a dramatic increase in throughput or cost savings.

1.2.1.2 Microchip-based electrophoresis

Microfabricated chips have the potential to become the preferred platform for the next generation of DNA sequencing technology. These glass or plastic microfluidic devices offer the potential for much greater throughput than capillary arrays due to the miniaturization of the separation apparatus. Very narrow sample injection zones are possible using simple cross or "double-T" injectors [28-30], potentially allowing for rapid, long DNA sequencing reads over a much shorter separation distance (*i.e.,* 7.5-15 cm as opposed to 50-60 cm for CAE).

The first four color Sanger sequencing on a microfluidic chip was achieved in 1995 by Woolley and Mathies when 200 bases were sequenced in 10 minutes in a 3.5 cm long channel using crosslinked polyacrylamide gel [31]. More recently, Fredlake, *et al*., has demonstrated sequencing of 600 bases in 6.5 minutes on a microchip [32]. In comparison, an average capillary sequencing run yields approximately 650-700 bases in 1-2 hours [9].

As in a capillary, the high surface area to volume ratio of chip microchannels provides improved heat transfer, minimizing the effects of electrical current-induced Joule heating during electrophoresis. Even though microchannels can be made easily by standard photolithography techniques with dimensions as small as 2 μm, for DNA sequencing applications they are typically 75-100 μm wide and 25-40 μm in depth, which is similar to the dimensions of conventional sequencing capillaries (50-75 μm ID). Larger channels can be loaded with a greater volume of sequencing sample, thereby allowing easier detection of fluorescently labeled

DNA but at the expense of resolution. In addition, the loading of viscous polymer solutions would be challenging with narrower channels. The typical viscosity of a high-performance DNA sequencing matrix is around 100,000 cP (LPA) although matrices based on pDMA have recently been used with viscosities ranging from 1,000 to 10,000 cP [32]. By comparison, the viscosity of water at 20°C is 1 cP.

It is necessary to load fresh sequencing matrix between each sample, to avoid sample carryover and breakdown of the sieving matrix. A quick and easy method of loading the matrix is critically important for automated DNA sequencing. Plastic chips have a limit of ~ 50 psi of applied pressure [33], whereas glass microchips have a limit of ~ 200 psi, above which the thermally bonded glass chips will fail [9]. On the other hand, a capillary array electrophoresis instrument, can routinely load the polymeric separation matrix at 1000 psi.

"Lab-on-a-chip" DNA sequencing devices, which integrate sample preparation, cleanup, and detection on a single platform are being actively researched. These devices promise to reduce time and costs that are associated with off-line sample preparation. Examples of this technology include the Landers group whose work focuses on on-chip sample cleanup, PCR amplification, and product detection for forensic analysis and pathogen detection [9].

1.2.2 Alternative sequencing techniques

Several groups are pursuing work on non-Sanger and non-electrophoretic methods of DNA sequencing. 454 Life Sciences sells an instrument that utilizes massively parallel, short reads generated by "pyrosequencing," an enzymatic sequencing-by-synthesis (SBS) method, to obtain 25 million bases of raw sequencing data in 4 hours [34, 35]. So far individual read

lengths are on average ~250 bases using the current GS-FLX instrument [36]. These numbers are still short of the required 600-700 bases needed by existing algorithms to properly assemble large repeat-rich genomes [18], requiring the development of new tools to manage and analyze the large amounts of short read sequencing data produced by the next generation sequencing platforms [37]. Furthermore, raw sequences generated by the 454 instrument tend to be far less accurate in comparison to electrophoresis-based methods particularly for repetitive regions [38]. The short read lengths (without paired-end reads) had limited this technology to the analysis of bacterial or viral genomes (<2 Mbp) which contain little repetitive DNA or small DNA such as expressed sequence tags (ESTs) [9, 32].

However, recently 454 has published a diploid genome of James D. Watson, sequenced with an average 7.4X coverage in two months for less than $1 million although more detailed cost information is not provided [39]. In contrast, the cost to sequence Craig Venter's diploid genome using standard Sanger sequencing technology [40] reportedly was $100 million [39]. The HGP reference genome was used by 454 as a guide to help reassemble the short snippets of DNA obtained with their technology. To date, no large, complex genome has been sequenced using 454's technology without a preexisting reference genome. That accomplishment would be the next challenge for this technology [41].

Several other sequencing methods are also under development, but so far they are still limited to short reads of ~ 25-40 bases [9, 36, 42, 43]. Besides 454, other recently released, commercial sequencing platforms are Applied Biosystems's (ABI) SOLiD and Illumina's Genome Analyzer [43-45]. Recently, Helicos BioSciences has demonstrated single-molecule

DNA sequencing of a viral genome, showing promise for their non-PCR based method [46]. Pacific Biosciences is actively investigating single-molecule real-time sequencing by following the progress of a DNA polymerase as it incorporates complementary nucleotides labeled with a fluorescent tag.

Another alternative technology that has been more challenging to develop is the use of nanopores to sequence polynucleotides as they move through narrow channels under an applied electric field [47, 48]. The measured current exhibits a significant decrease upon passage of the polynucleotide through the narrow pore. The degree of current reduction and other parameters depend on the composition of the polynucleotide [9].

1.2.3 Improvement over current sequencing technologies

For *de novo* sequencing of large repeat-rich genomes, microchannel electrophoresis and in particular, an integrated device, still shows promise in becoming the preferred method of analysis. Estimates show that although a significant yearly cost reduction is not seen when moving from CAE instruments with 96 capillary arrays to a 96-lane glass chips for microchannel electrophoresis, many more sequencing reads are produced per instrument. A 1-2 hour sequencing run by CAE could be performed in 7-10 minutes on a chip-based sequencer. Therefore, on a per read basis, the cost reduction is ~ 90% [9]. Further cost reductions may be achieved if plastic chips are used. Various technical issues have so far prevented microfabricated systems (both glass and plastic) from replacing CAE as the method of choice although, since both methods are based on Sanger sequencing, little change will need to be implemented to established genome sequencing strategies [9].

1.3 End-Labeled Free-Solution Electrophoresis

A novel method for microchannel DNA sequencing that would eliminate the need for a viscous polymer matrix is End-Labeled Free-Solution Electrophoresis, or ELFSE. Elimination of the viscous polymer solution would save time, reduce costs and avoid challenges associated with loading and replacing the polymer matrix in microchannel electrophoresis and especially microfabricated devices. An aqueous buffer could simply be loaded into a microchip. Free-solution electrophoretic separation of DNA can also be easily integrated into "lab-on-a-chip" devices, previously mentioned in Section 1.2.1.2.

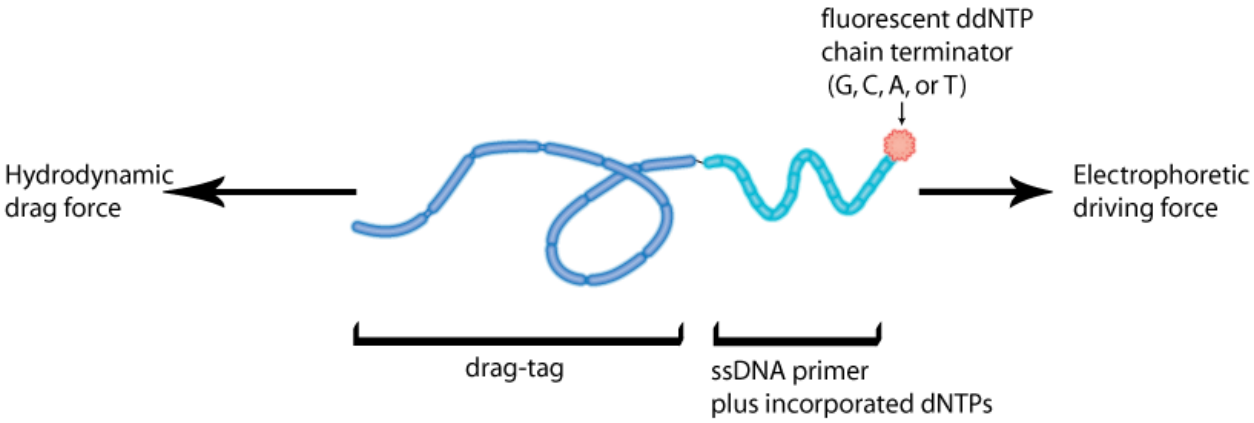


Figure 1-1: Diagram of ELFSE concept showing the drag-tag conjugated to the DNA sequencing fragment

1.3.1 Basic theory of ELFSE

In this method conceptually proposed by Noolandi in 1992 [49], a monodisperse perturbing entity (*e.g.,* a protein, virus, or microsphere), which has a different charge-to-friction ratio than DNA, is attached to DNA to break the symmetry between charge and friction that prevents DNA separation in free solution. The perturbing entity, such as a protein [50, 51], is

attached to each DNA molecule and has the primary function of adding hydrodynamic drag (Figure 1-1). The presence of this "drag-tag" alters the electrophoretic mobility of each DNA molecule in a size-dependent manner, allowing the separation of these bioconjugates to occur in free solution. The basic analytical theory behind ELFSE was first developed by Mayer, Slater and Drouin [50] and is summarized below.

If N is the number of DNA bases, and α is the total frictional drag of the drag-tag in units of the friction, ξ, of one DNA base, then the molecular friction of the hybrid molecule of DNA and drag-tag will be $\xi(N+\alpha)$. If $-\beta$ is the electrostatic charge of the drag-tag in units of the charge, ρ, on one DNA base, then the total charge of the hybrid molecule will be $\rho(N-\beta)$. Therefore the electrophoretic mobility of the hybrid molecule is (approximately):

$$\mu(N) = \frac{v}{E} \approx \frac{q}{f} = \frac{\rho(N-\beta)}{\xi(N+\alpha)} = \mu_0 \frac{(N-\beta)}{(N+\alpha)} \qquad (1\text{-}2)$$

The electrophoretic mobility of the hybrid molecule in free solution electrophoresis will therefore be size-dependent when $\alpha \neq -\beta$. For neutral tags ($\beta = 0$), the equation for mobility simplifies to:

$$\frac{\mu}{\mu_0} = \frac{1}{(1+\alpha/N)} \qquad (1\text{-}3)$$

In ELFSE, the elution order of DNA fragments with respect to size is reversed compared to matrix-based sequencing, with larger DNA fragments eluting first because their greater electromotive force is less affected by a given drag-tag than a smaller piece of DNA. A family of drag-tags of different sizes would need to be designed in order to provide excellent resolution for various DNA sizes. In ELFSE, there is no polymer network to minimize DNA diffusion,

therefore smaller microchannels would be advantageous to minimize dispersion. Higher electric fields for faster separations are also a possibility as biased reptation is not a concern with no matrix present. It is predicted that 670 bases can be achieved in 480 seconds (< 8 minutes) using reasonably optimistic numbers such as α = 400 bases and V = 40,000 Volts with a 20 cm long channel [52].

1.3.2 Ideal drag-tag properties

The ideal drag-tag would have a large value of α. With a large drag, higher resolution and performance can be achieved for the separation of large sequencing fragments (*i.e.,* > 200 bases). In addition to a high α value, several other properties are needed for the ideal drag-tag:

1. Complete monodispersity
2. Water-solubility
3. The tag should be uncharged, or nearly so
4. Unique and stable attachment to DNA
5. Minimal adsorption to or non-specific interaction with microchannel walls

DNA sequencing is done in aqueous solution so the potential drag-tag must be water-soluble. The drag-tag would preferably be uncharged as well. Positively charged drag-tags may have undesirable electrostatic interactions with the negatively charged DNA and microchannel walls. On the other hand, negatively charged drag-tags would effectively increase the mobility of the tag compared to an uncharged one, thereby decreasing the amount of separation possible. Also, it is essential that the DNA and drag-tag can be linked together in a unique and stable manner (preferably end-on) in order to ensure that each DNA is attached to only one drag-tag.

One of the most important properties for an ideal drag tag is complete monodispersity. In other words, every tag needs to be identical in charge and drag. If a polydisperse molecule is used as a drag-tag and the DNA-drag-tag conjugate is analyzed by capillary electrophoresis, the resulting peak pattern is ambiguous (*i.e.,* peaks could correspond to either different DNA sizes, different drag-tags, or both). Accurate DNA sequencing very likely would not be possible. This requirement for total monodispersity eliminates as possibilities all commonly available synthetic polymers and microparticles as useful drag-tag candidates for DNA sequencing. This has been demonstrated experimentally. Using a method closely related to ELFSE, called Free-Solution Conjugate Electrophoresis (FSCE), Vreeland *et al.* characterized the polydispersity of a synthetic (low-polydispersity) poly(ethylene glycol) (PEG) sample by attaching a monodisperse, fluorescently labeled DNA 20mer (end-on) to the PEGs [53]. Capillary electrophoresis with LIF detection showed > 110 peaks (one for each PEG size), in an overall Gaussian distribution. This result demonstrated the importance of using a monodisperse drag-tag since only slight differences in PEG size (*i.e.,* one monomer) were enough to produce distinct peaks resolvable by microchannel electrophoresis.

In contrast to PEGs, proteins can be completely monodisperse, large enough to provide sufficient drag, and are produced under tight molecular weight control by living cells. It should be noted though that in some cases post-translational modifications in eukaryotes can lead to heterogeneous protein products. However, natural proteins have several drawbacks that make them non-ideal drag-tag candidates. In aqueous solution, most natural proteins are folded into compact shapes and typically present numerous surface charges. Even if DNA sequencing were

performed at the isoelectric point of a protein where the net charge is zero, the protein would still be likely to have local interactions with the DNA or microchannel walls. Adsorption of the drag-tag to the microchannel surface could result in poor resolution and non-Gaussian peak shapes that would complicate the interpretation of DNA sequencing data. Additionally, natural proteins typically contain multiple reactive groups (*e.g.,* amino, thiol, carboxylic acid) on their surface, making unique attachment to DNA difficult. Finally, proteins possess flexible polyamide backbones, which can adopt various conformations in the denaturing conditions under which DNA sequencing is performed. This is not necessarily a disadvantage, but a property that must be understood since it will affect the net hydrodynamic friction a given tag provides.

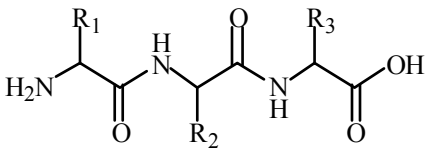
1.4 Previous, relevant work on ELFSE

1.4.1 DNA separations with streptavidin drag-tag

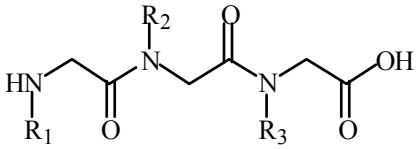
The first proof of principle for DNA separation by ELFSE was obtained with a protein drag-tag. Heller *et al.* demonstrated that DNA separation was possible using the ELFSE method by using the natural protein streptavidin as the protein drag-tag for low-resolution separation of double-stranded DNA (dsDNA) by CE [54]. Streptavidin is a tetramer of 159-residue polypeptides that binds strongly to biotin. Later, Ren *et al.* obtained 110 bases of DNA sequence in less than 18 minutes, also by using streptavidin as the drag-tag for ssDNA fragments [55]. Streptavidin was easily conjugated to biotinylated DNA after the Sanger cycle sequencing reaction was complete. Unfortunately, the protein itself is fairly heterogeneous due to different degrees of proteolysis and glycosylation. Electrophoretic tube-gel purification was performed that improved upon but could not achieve complete homogeneity of the protein.

1.4.2 DNA separations with synthetic polypeptoid drag-tags

Protein mimics called peptoids or poly-*N*-substituted glycines [56-58] have also been tested as drag-tags for ELFSE. These peptoids (Figure 1-2) are chemically synthesized on solid phase supports in a similar manner to polypeptides by the sequential addition of individual monomers to a growing chain. PEG-like poly-*N*-methoxyethylglycines (polyNMEG) peptoids have been produced in our lab and purified to monodispersity by reversed-phase high-performance liquid chromatography (RP-HPLC). Peptoids ranging from 10-60 monomers have been used to separate 20- to 21-base ssDNA in free solution [59]. Additionally, peptoids have been used for free-solution separation of single-base extension (SBE) reactions [60] as well as multiplexed SBE analysis of p53 gene mutations in both capillary and microfluidic devices [61]. Branched polypeptoids, consisting of a 30mer polyNMEG "backbone", have been used to separate ssDNA sizes of 50, 75, 100, and 150 bases by free-solution capillary electrophoresis in under 10 minutes [62]. The largest of these branched polypeptoids (70mer with 5 octamer branches) had an α value of 17.



Polypeptide



Polypeptoid

Figure 1-2: Structural comparison between polypeptide and polypeptoid trimers

1.5 Protein polymers

1.5.1 Advantages of protein polymers as drag-tags

The work of Ren *et al*. was the first demonstration of the ELFSE sequencing method [55]. Even though the sequencing results were poor compared to results obtained by capillary array electrophoresis with a polymer matrix (*i.e.,* the read length was short), a monodisperse protein drag-tag that was uncharged and unstructured, and which had a larger value of α, would theoretically provide better sequencing results. As a natural protein, streptavidin has areas of positive and negative charge on its surface which may adversely affect its ability to be an effective drag-tag. It was also necessary to maintain the folded conformation of streptavidin for biotin binding, so sequencing was performed at a relatively low temperature of 35°C, in only 3M urea instead of in the standard fully denaturing conditions of 7 M urea and 55°C for DNA. The α value for streptavidin was also relatively small (25-45 depending on buffer conditions).

While polypeptoids can be purified to monodispersity, they are still limited due to the low α values that have been achieved so far. For the moment, 60 NMEG monomers appears to be the limit for obtaining reasonable yields of highly pure peptoids since the coupling efficiency for solid-phase synthesis is less than 100% [59]. Currently work is in progress to produce even longer, linear polypeptoid molecules. Peptoids with a length of 100-125 monomers would have to be created to reach an α value comparable to streptavidin (α = 0.2-0.25 per peptoid monomer) [59, 63].

Therefore, a non-natural "protein polymer" with the requirements described above may be the best drag-tag for ELFSE and obtaining such a molecule in highly pure form has been the focus of the research presented here. The properties of an engineered protein-based drag-tag can be fully customized.

1.5.2 Current research in protein polymer engineering

"Protein polymers" are produced by genetic engineering and consist of a repeating sequence of amino acid monomers. The repeating sequence can be a mimic of a natural sequence motif or it can be highly non-natural and designed specifically for a particular structure and/or function. The properties of these polymers can be customized according to interest by using different DNA sequences that code for the final protein sequence. The only limitations are those of the genetic code (*i.e.,* 20 amino acids to choose from) although proteins have been engineered that incorporate non-canonical amino acids [64-67]. In addition, specific chain lengths of the desired sequence can be obtained. Unlike conventional, synthetic polymerization techniques, protein-based materials produced in biological systems, such as the bacterium *Escherichia coli*, allow for much better control over the properties of the final product [64].

Table 1-1: Summary of Current Protein Polymer Research

Research Group	**Example Sequences**	**Applications**
Cappello (Protein Polymer Technologies) [68-70]	Silk-elastin-like proteins (SELP) $(GVGVP)_4GEGVP(GVGVP)_3GAGAGS$ [71]	Drug delivery and biomedical applications
Chilkoti (Duke) [72-79]	Elastin-like polypeptides (ELPs = VPGXG where X is any amino acid except proline)	Drug delivery, protein separation/purif-ication, hydrogels
Conticello (Emory) [80-85]	ELPs [80, 85] α-helical coiled-coils/spider dragline silk $(AEAEAKAK)_2AG(GPGQQ)_6GS$[82]	Self-assembly, tissue engineering

Harden (Johns Hopkins) [86, 87]	Self-assembling hydrogels (Tirrell coiled-coil design described below) containing RGD sequence [87]	Tissue engineering
Heilshorn (Stanford)	No sequence data published yet.	Biomaterials for regenerative medicine
Hubbell (EPFL) [88-90]	Fibrinogen (LRGDFSSANNRDNTY) and collagen (GPQGIAGQRGV) motifs including RGD, plasmin sites, MMP cleavage sequence	Tissue regeneration and hydrogels
Kaplan (Tufts) [91-95]	Spider silk mimics SGRGGLGGQGAGAAAAAGGAGQGGY-GGLGSQGT [93]; SGPGGYGPGQQT [96]	Structure-function elucidation, assembly, tissue engineering
	Sericin-like SSTGSSSNTDSNTDSNSNSVGSSTSGGSSTYGYSS-NSRDGSV [92]	
	Chimeric silkworm silk ASA_{18}TSGVGAGYGAGAGYGVGAGYGAGVGYGA-GAGYTS [97]	
Kiick (Delaware) [98-103]	Helical alanine-rich protein polymers with multiple functional groups (*e.g.,* glycopolymers) AAAQEAAAAQAAAQAAQAAQ [100]	Control of cellular responses (*e.g.,* immune response, toxin inhibition), drug delivery
Kopeček (Utah) [104-107]	Coiled-coil self-assembling hydrogels $(AG)_3$PEG and VSSLESK domains[106]	Biomaterials and drug delivery
Tirrell (Caltech) [66, 108-112]	ELPs containing REDV sequence [112] Coiled-coil self-assembling proteins of $(AG)_3$PEG and SGDLENEVAQLEREVRSLEDEAAELEQKVSRLKNE IEDLKAE domains [111]	Tissue regeneration and hydrogels
Urry (Minnesota) [113-117]	ELPs (GVGVP)	Thermodynamics of model protein-based polymers

Table 1-1 provides a summary of current research being done in the protein polymer engineering field by various groups. As can be seen, the field is still small, with most of the focus on medical applications; however, recent research on protein polymers has expanded to the areas of cosmetic formulations and ink applications [118, 119]. Work such as ours to create monodisperse protein polymers for DNA sequencing drag-tags was begun by us first, and except for a brief and unsuccessful effort at Applied Biosystems (ABI) in 1999, we have been the only ones taking this approach. In addition, our sequences are not based on a natural protein or existing design and are unique.

1.6 Current research in ELFSE

Chapter 2 describes the general protocols used to design, construct, and produce the protein polymer drag-tags. Dr. Jong-In Won, a previous graduate student in our lab, developed several sequences that became the first generation of protein polymer drag-tags (PZ-1 through PZ-6 and BB-1). These early sequences are illustrated in Table 1-2 and their corresponding gene sequences are detailed in Appendix A. For various reasons [120-122], these sequences were not satisfactory as drag-tags for ELFSE, leading to the PZ-7 and PZ-8 designs. In the meantime, work on alternative sequences and purification strategies (Chapters 3 and 4 respectively) began. Once PZ-8 proved to be a sequence that could be successfully used for sequencing [123], PZ-8 and its variants became the focus of future work. A new challenge then arose: achieving complete monodispersity (especially with longer length protein polymers). Chapters 5 and 6 discuss studies of the early PZ-8 sequences and experiments done in the pursuit of obtaining totally monodisperse drag-tags. Chapter 7 summarizes recent work done with several PZ-8

variants while Chapter 8 reviews the biophysical studies performed to date with these unique protein polymers. Conclusions based on the work presented here and recommendations on future research directions are detailed in Chapter 9.

Table 1-2: Summary of Previous Drag-Tag Sequences (Designed by Dr. Barron and former Barron lab graduate student Jong-In Won)

Name	**Repeating sequence**
PZ-1	Gly-Ser-Gly-Gln-Gly-Glu-Ser
PZ-2	Gly-Ala-Gly-Gln-Gly-Glu-Ala
PZ-3	Gly-Val-Gly-Gln-Gly-Glu-Val
PZ-4	Gly-Leu-Gly-Gln-Gly-Glu-Leu
PZ-5	Gly-Ala-Gly-Gln-Gly-Asn-Ala
PZ-6	Gly-Ala-Gly-Gln-Gly-Ser-Ala
BB-1	Gly-Lys-Gly-Ser-Ala-Gln-Ala
PZ-7	Gly-Ala-Gly-Ser-Gly-Ser-Ala
PZ-8	Gly-Ala-Gly-Thr-Gly-Ser-Ala

Chapter Two

General Experimental Protocols for the Production, Expression, and Purification of Repetitive Protein Sequences

2.1 Introduction

The overall process to generate these protein polymers can be divided into four sequential steps: designing the sequence, cloning the gene, expressing it, and finally purifying the desired protein. Because of their size, protein polymers are typically synthesized in biological systems such as *E. coli* and then later purified using methods such as affinity chromatography. The sequence designs are based on the ideal drag-tag criteria outlined earlier in Section 1.3.2. Additional considerations include selection of favorable codons specific to the host (*i.e., E. coli*) and expected protein secondary structure. Chemical synthesis of peptides is limited to the production of short chains (< 50 amino acids). Similarly, it is also difficult to synthesize long DNA polynucleotides. For this reason, a large repetitive gene used for the expression of a given protein polymer is built up by joining (ligating) several DNA "monomers" (macromonomers as large as 100 bases). The ends of the DNA being joined must be complementary with respect to their base-pair sequence (*i.e.,* have cohesive termini or "sticky ends") in order to be properly ligated together by T4 DNA ligase enzyme.

The target gene is then placed in an expression plasmid containing a bacteriophage T7 promoter, coding regions for a specific antibiotic resistance, and an affinity tag sequence. The plasmid DNA is inserted or "transformed" into *E. coli* cells possessing a chromosomal copy of

the T7 RNA polymerase under *lacUV5* control. Protein expression is induced upon the addition of a lactose analogue, isopropyl-β-D-thiogalactopyranoside (IPTG) to the culture [124, 125]. Multiple copies of T7 RNA polymerase are produced which in turn transcribe the expression plasmid containing the T7 promoter sequence. These transcripts are then translated into proteins by the ribosomes. The desired protein can constitute up to 50% of the total cellular protein after a few hours of induction [125].

The cells are lysed by sonication or other means and then the target protein is isolated and purified from the cell contents by affinity chromatography. It is important to note that because we are designing the sequence *de novo*, it is difficult to predict beforehand whether a sequence will be well expressed, be soluble in water, or have favorable biophysical properties for the application of interest.

2.2 Materials and General Methods

All molecular biology protocols described here were adapted from standard protocols [124] or from instructions provided by the manufacturers. While gel electrophoresis in hand-cast gels was most commonly used in this work, an Agilent 2100 Bioanalyzer was used for some DNA and protein electrophoretic analyses. This is a commercially available system which uses a disposable microfluidic chip to analyze samples by electrophoresis through a polymer solution. Results can be presented in either an electropherogram (fluorescence vs. time) format or in a "pseudo-gel" image.

E.coli strains BLR(DE3) and NovaBlue, carbenicillin antibiotic, and the plasmid pET-19b were obtained from Novagen (Madison, WI). Plasmid pUC18, ultra-pure agarose, and

ProBond nickel-chelated resin were obtained from Invitrogen (Carlsbad, CA). Synthetic oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA) or Oligos, Etc. (Wilsonville, OR). Talon cobalt-chelated resin was obtained from Clontech Laboratories, Inc. (Mountain View, CA). *Pfu* DNA polymerase was obtained from Stratagene (La Jolla, CA). *Taq* polymerase was purchased from Promega (Madison, WI). All other enzymes including *Sap* I and *Ear* I were purchased from New England BioLabs (Ipswich, MA). Oligonucleotide and plasmid purification kits and the penta-His antibody were obtained from Qiagen (Valencia, CA). Anti-mouse IgG HRP antibody, Hybond-ECL nitrocellulose membrane, and ECL reagents were purchased from GE Healthcare (Piscataway, NJ). Membranes for protein dialysis and GelCode Blue protein stain were purchased from Pierce Biotechnology, Inc. (Rockford, IL). Protein molecular weight standards and 30% acrylamide/bis solution were obtained from BioRad Laboratories (Hercules, CA). General reagents for cloning, protein expression, and purification were obtained from Fisher Scientific (Fairlawn, NJ) or VWR (West Chester, PA).

2.3 Choosing the gene

Careful consideration is needed when choosing the target sequence as it is a lengthy and laborious process to finally obtain the purified protein polymer from a concatemer of the 100-bp oligonucleotide.

2.3.1 Amino acid composition

As mentioned previously, the ideal drag-tag characteristics are monodispersity, water solubility, lack of charged residues, a site for unique covalent attachment to DNA, and minimal adsorption to or non-specific interaction with glass microchannel walls. Based on these

requirements, phenylalanine, tyrosine, and tryptophan were eliminated from consideration as possible amino acids due to their strongly hydrophobic, aromatic groups [126]. Valine, leucine, and isoleucine are quite hydrophobic as well. However, valine and leucine were included in PZ-3 and PZ-4 (refer to Table 1-2) in order to study the effects of increasing hydrophobicity in drag-tag designs. Cysteine is not desirable in a drag-tag because its thiol side chain is highly reactive and may oxidize to form disulfide bonds. The charged amino acids (lysine, arginine, histidine, aspartic acid, and glutamic acid) were also used minimally or not at all. In PZ-1 through PZ-4, a glutamic acid was included to facilitate the water solubility of the protein polymer. This left eight of the natural 20 amino acids that can be used in the drag-tag design. Each protein "monomer" unit was more or less arbitrarily chosen to consist of 3 repeats of a 7-amino acid sequence, such as those listed previously in Table 1-2. This region is encoded by a DNA molecule that can be synthetically produced and obtained commercially. Sequences were designed to be as diverse and non-repetitive as possible with the use of such a limited monomer sequence.

2.3.2 Protein secondary structure prediction

For a drag-tag, an unstructured, random-coil configuration is preferred to α-helix or β-sheet structures, as this exposes more of the protein to the solvent [127]. Additionally, β-sheet structures have a tendency to aggregate, an undesirable property for a drag-tag. A program called GOR IV [128] (http://pbil.ibcp.fr/) was used to predict protein secondary structure for the designed sequences. Amino acid sequences are entered into the GOR IV program which then returns a prediction of protein structure in terms of percentage of α-helix, β-sheet, or random-

coil regions. This algorithm uses a probabilistic model derived from the Protein Data Bank (PDB) to determine the expected secondary structure of each residue based upon its type and nearest neighbors [129]. Even though the program is not especially accurate with a reported average accuracy of 64.4%, it provides some guidance when designing a new sequence, as it is likely to identify sequences with a very high tendency to form helices or β-sheets.

2.3.3 Codon selection

After the amino acid sequence for a designed protein polymer has been selected, the corresponding gene sequence must be designed as well and this is a very important part of the project. In the genetic code, three DNA bases (a codon) code for a particular amino acid or stop codon. The genetic code is degenerate, meaning that several different codons can correspond to the inclusion of the same amino acid. Depending on species, the cell will favor some codons over others in protein production.

When designing the DNA sequence to encode the desired protein polymer, it is important to use codons preferred by the particular species, in this case *E. coli*, while also maintaining variety (*i.e.,* a different codon is used for the same amino acid when two or more of the same amino acid is present in the sequence) [130]. Using more of the preferred codons while also maintaining codon variety would, in principle, aid in protein expression [131]. Since we are expressing repetitive protein sequences, this is particularly important, as the expression of a simple, repeating sequence can place extra strain on the protein production machinery of the cell.

2.3.4 DNA melting temperature and other considerations

Single-stranded DNA has a tendency to form secondary structures such as loops and hairpins if there is enough complementarity between two regions of the sequence. Such secondary structures increase the melting temperature (T_m) of double-stranded DNA regions. A large number of GC pairings can also increase the T_m of a given DNA sequence. A high T_m will complicate PCR reactions by making it difficult to separate annealed strands. The T_m's for each sequence were designed to be below 70°C. In addition, DNA sequences were verified to exclude recognition sites for restriction enzymes to be used later in the cloning process.

2.4 Creating the full-length gene

After the sequence has been designed, the "monomer" polynucleotide (~ 100 bp of ssDNA) is purchased from an outside company which chemically synthesizes and purifies the polynucleotide. This DNA then forms the basis for the assembly of the final gene of the desired length, so that it is ready to be used in the *E. coli* cells for protein expression.

Several cloning methods have been developed to generate the large repetitive gene needed for a protein polymer. Capello and McGrath first demonstrated in 1990-92 the use of head-to-tail enzymatic linkage (concatemerization) in creating a repetitive gene using nonpalindromic restriction enzymes to generate the necessary cohesive ends for ligation by T4 DNA ligase [68, 108, 109]. A disadvantage of this method is the strict sequence requirement for the 5' and 3' termini of gene sequences which are limited to only non-palindromic enzyme recognition sites. In 1999, McMillan *et al.* reported an improved technique called "seamless cloning" which

eliminates the prior dependence on nonpalindromic enzymes by using *Eam1104* I, an enzyme that cleaves downstream of its recognition site [80].

However, it was still difficult in some cases to obtain large concatemer genes (*e.g.,* > 400 bases) because of their limited yield after DNA macromonomer ligation. Most of the DNA multimers obtained would be smaller than desired. Other cloning strategies have been developed recently (2002, 2005-06) using an iterative and recursive methodology of gene construction [74, 132] or assembly by "modules of degenerate codons" [86].

The method we have adopted for our purposes was developed in our lab and is called "controlled cloning." It was developed by Dr. Jong-In Won in our group in the process of constructing the early drag-tag sequences and is a variation on the "seamless cloning" strategy [133]. Controlled cloning allows the directed generation of large DNA concatemers (an advantage over seamless cloning) while still not being restricted to any specific sequence requirement. This method also can be used to join together two or more different sequences to create multidomain repetitive polypeptides (*e.g.,* block copolymers). Figure 2-1 illustrates the general cloning strategy for producing the long, repetitive gene that encodes the protein polymer. A later figure, Figure 2-3, will illustrate "controlled cloning."

2.4.1 Generating a concatemer gene "ladder" from the synthetic DNA "monomer"

First, PCR with the primers 5'-GCT AGC CAT ATG CTC TTC AGG-3'and 5'-ACT AGT GGA TCC CTC TTC AAC-3' [133] is used to amplify the 100-bp ssDNA "monomer," which is designed to code for a 21-amino acid gene sequence. The sequence is flanked by *Ear* I and *Sap* I restriction enzyme recognition sites, two key enzymes in our cloning strategy. Note

that the controlled cloning method was initially developed using the restriction enzyme, *Eam1104* I, which is an isoschizomer of *Ear* I (*i.e.*, identical recognition and cleavage sites). DNA monomer sequences were amplified via PCR using high-fidelity *Pfu* polymerase. After an initial 5 minute denaturing step at 95°C, 30 cycles of amplification consisting of 1 minute denaturation at 95°C, 1 minute annealing at 55°C, and 2 minute extension at 72°C were used, followed by a final extension step of 72°C for 10 minutes.

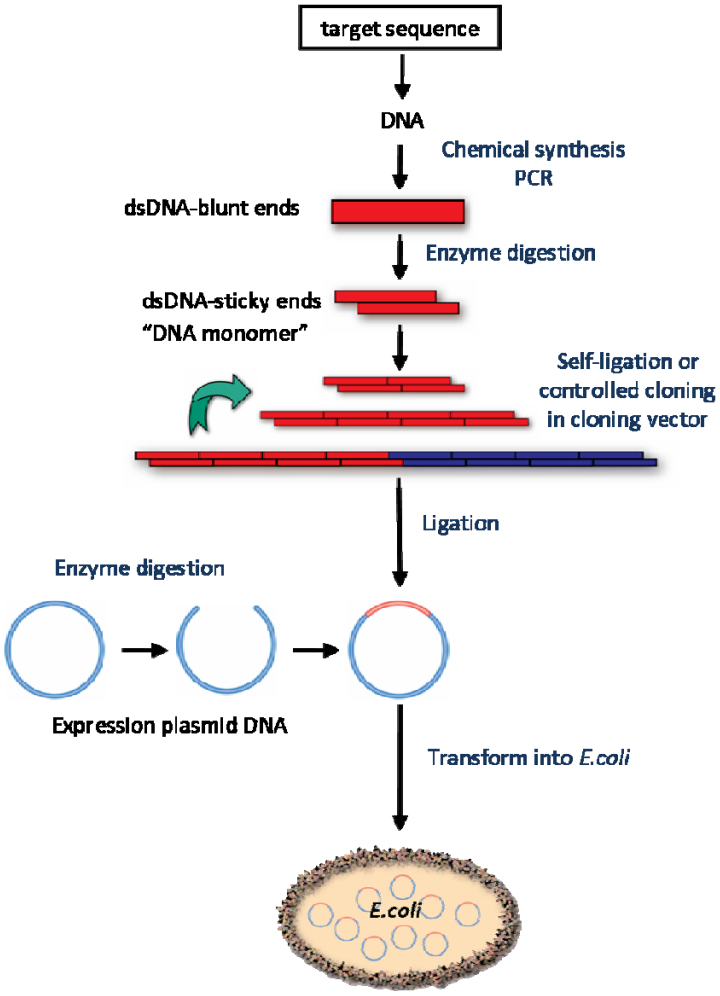


Figure 2-1: General protein polymer cloning strategy

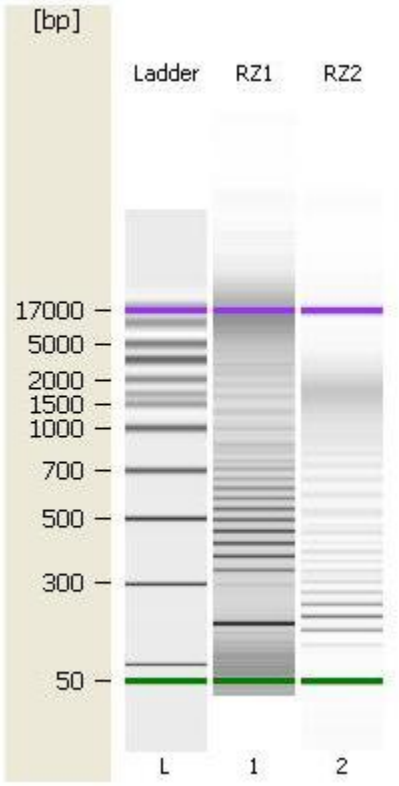


Figure 2-2: Concatemer ladders for RZ-1 and RZ-2 sequences as seen in "pseudo-gel" format using the Agilent 2100 Bioanalyzer. The colored 50 bp and 17,000 bp bands are size markers used for alignment of each sample run.

After amplification, the ends of the PCR product are digested with the restriction enzyme *Ear* I (four hours or overnight at 37°C) to produce cohesive DNA termini. The desired, fully digested 63-bp DNA fragment is then isolated by agarose slab gel electrophoresis and then purified using a Qiagen QIAEX II gel extraction kit. Due to the low activity of the *Ear* I enzyme, it is often necessary to repeat the digestion more than once to obtain > 85% fully digested DNA. T4 DNA ligase is used to concatemerize the DNA monomer at 16°C (16 hours), producing a ladder of differently sized DNA ligation products. Two concatemer ladders are represented in Figure 2-2 for the RZ-1 and RZ-2 sequences, which will be discussed in more detail later in Chapter 3.

2.4.2 Transformation of synthetic DNA ladder once ligated to cloning vector

This concatemer ladder is then ligated into a modified pUC18 cloning vector [133], to be used for gene construction and amplification, with easy subsequent enzymatic excision, and transformed into *E. coli* NovaBlue cells by heatshock, allowing uptake of the plasmid DNA into the cells. Transformed cells are then plated onto LB (Luria-Bertani) agar plates containing the antibiotic carbenicillin at a concentration of 50 μg/mL and tetracycline at a concentration of 12.5 μg/mL. The NovaBlue strain is resistant to tetracycline but not carbenicillin. Only NovaBlue cells that successfully

acquired the pUC18 plasmid containing the carbenicillin resistance gene will be able to grow on the agar plate.

2.4.3 Colony screening and miniprep DNA isolation

Transformed colonies were screened for the presence of desired concatemers using colony PCR. Each colony should be "cloned," (*i.e.*, have one identical DNA insert). PCR with *Taq* polymerase used primers 5'-TTA ATG AAT CGG CCA ACG CGC-3' and 5'-GGA TAA CCG TAT TAC CGC CTT-3,' which flank the insert region in the pUC18 vector, using the same thermal cycling protocol outlined in Section 2.4.1. *E. coli* cells are directly mixed into the mixture of aqueous reagents for the PCR reaction by touching a toothpick or pipet tip to the colony and then dipping it into the reaction mixture. The subsequent high temperature cycling is sufficient to break apart the bacteria and release their plasmid DNA, which then acts as the template DNA for amplification. The size of the amplified concatemers is then identified by agarose slab gel electrophoresis or by the Agilent Bioanalyzer. Typically, the largest concatemers obtained after checking numerous colonies are either 5mers or 6mers of the 63-bp DNA fragment. Many colonies contain monomers of the sequence. The occurrence of these monomers can be reduced by partially separating the concatemer ladder by gel electrophoresis prior to plasmid insertion, a method that was developed in our lab by Jong-In Won.

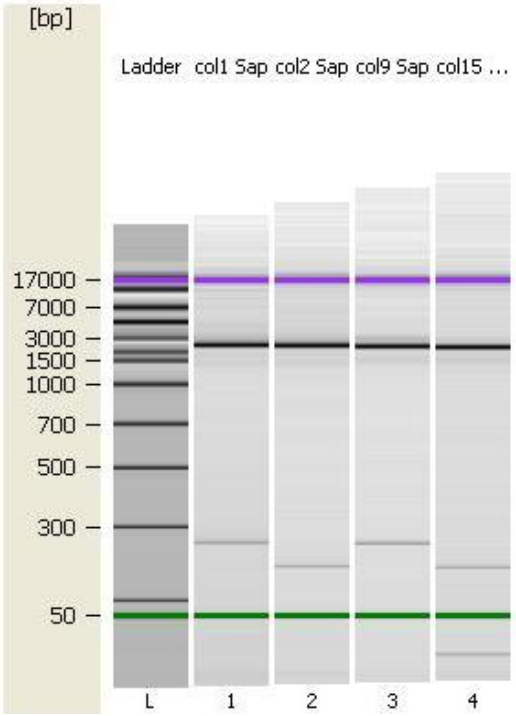


Figure 2-3: Agilent Bioanalyzer "pseudo-gel" of four pUC18 plasmid DNA digestions using the *Sap* I enzyme. Colonies #1 and #9 are tetramers of RZ-1 (252 bp) whereas #2 and #15 are trimers (189 bp).

Plasmids from colonies containing the largest concatemers we were interested in were obtained from overnight cultures using a Qiagen QIAprep spin miniprep kit, which uses alkaline lysis to break apart the cells and a silica gel membrane to recover the plasmid DNA. Insert sizes can be verified by digesting the plasmid DNA at flanking restriction enzyme recognition sites and running gel electrophoresis (Figure 2-3). In the case of the cloning vector, pUC18, *Sap* I is used for the digestion. For the expression vector pET-19b, *Nde* I and *BamH* I are used.

Recovered plasmid DNA molecules are sequenced by SeqWright, Inc. (Houston, TX) using the above mentioned primers that flank the insert region in the pUC18 vector. For very long genes (> 1500 bp) it is not possible to sequence the entire repetitive insert due to limitations on read length, even with both forward and reverse analyses being performed. As this is a highly repetitive gene, there are no unique internal regions to act as additional sites for sequencing primer annealing.

2.4.4 Doubling of select concatemer genes by controlled cloning

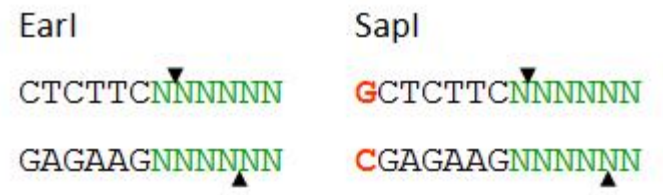


Figure 2-4: *Ear* I and *Sap* I restriction enzyme recognition sites

Controlled cloning [133] is used when a longer gene than is typically obtained by simple concatemerization is desired. Most often, this method has been used for doubling of the same gene, but it can be used to combine genes of different lengths and/or sequences. The method utilizes two different restriction enzymes, *Ear* I and *Sap* I. Figure 2-4 shows the recognition and cleavage sites, the latter indicated by black arrows, for these two enzymes. *Sap* I has the same restriction site as *Ear* I, but has an additional one-base recognition requirement (shown in red). Hence, *Ear* I can recognize and digest all *Sap* I sites, but not vice versa. Both enzymes cleave downstream of their recognition sites.

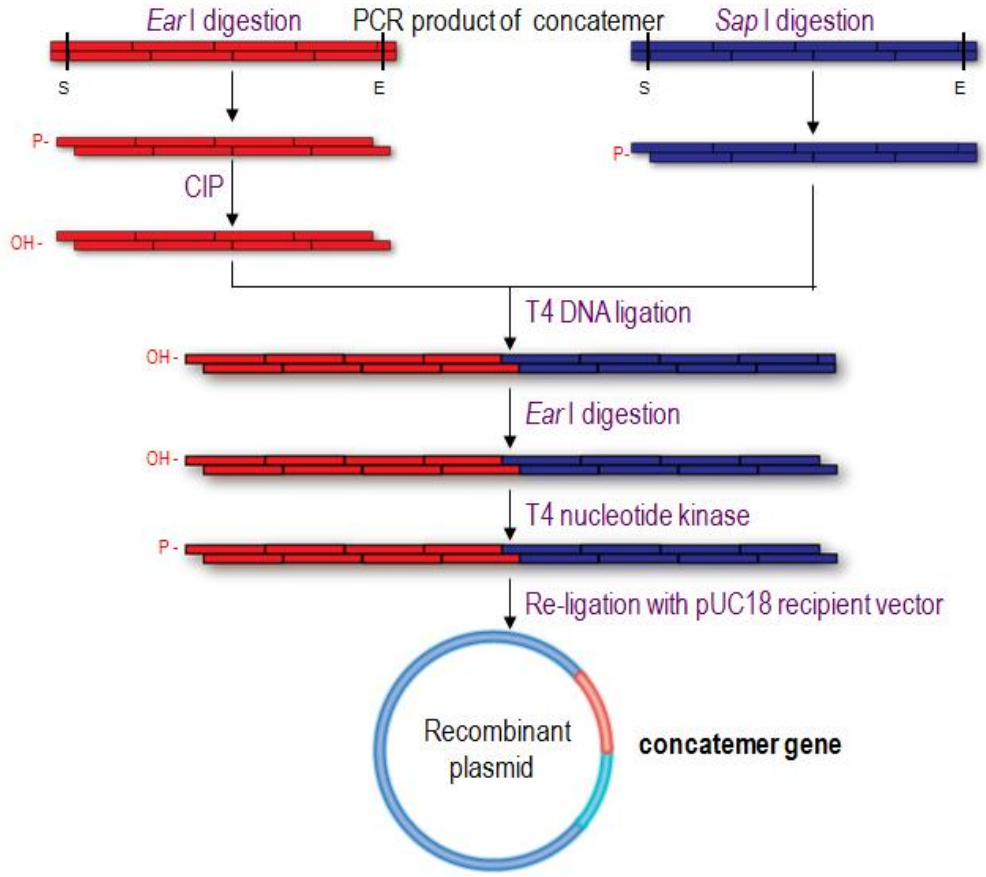


Figure 2-5: Controlled cloning strategy

The controlled cloning strategy is illustrated in Figure 2-5. PCR is used to amplify a concatemer gene already inserted into the pUC18 vector that contains two *Sap* I sites, each on opposite ends of the inserted gene. The forward primer, 5'-TTA ATG AAT CGG CCA ACG CGC-3', is perfectly complementary to the vector, whereas the reverse primer, 5'-TGA GCG AGG AAC TCT TCA GGT-3', contains one mismatched base which changes one *Sap* I site to an *Ear* I site. The amplified product is then split into two tubes. In one tube, the product is digested with *Sap* I, while in the other tube the sample is digested with *Ear* I. Incubation with

Ear I yields cohesive ends on both termini, whereas incubation with *Sap* I yields one cohesive end and one blunt (non-digested) end. Reaction of the *Ear* I product with calf intestinal phosphatase (CIP) then dephosphorylates the DNA fragment at the 5' end, preventing unwanted intramolecular recircularization. Since the *Sap* I product contains one blunt end, it will not cyclize.

Ligation of these two gene products can therefore be carried out in a controlled and predictable manner. For example, if the original two products were both 6mers (= 6 x 63 bp or 378 bp), the final product would be a 12mer (756 bp). Before ligating the gene into the pUC18 vector, T4 polynucleotide kinase (PNK) is used to rephosphorylate the gene. The cloning vector can then be transformed and grown in *E. coli* to generate large amounts of the inserted concatemer gene for further manipulation (*e.g.,* sequencing, insertion into expression vector, or concatemerization with another gene). Figure 2-6 is an example of controlled cloning where RZ2-5 is doubled to create the RZ2-10 gene.

2.5 Expression of the Protein Polymers

Once the gene is confirmed to have the correct sequence, it is then cloned into the modified expression plasmid (pET-19b) [133] for protein expression. *Sap* I digestion excises the gene from the pUC18 vector and the insert DNA is isolated and purified using preparative gel electrophoresis and the Qiagen QIAquick Gel Extraction Kit. The desired insert is then ligated into the expression vector. Note that typically *Ear* I is not used for the digestion as the cloning vector contains multiple *Ear* I recognition sites in its sequence.

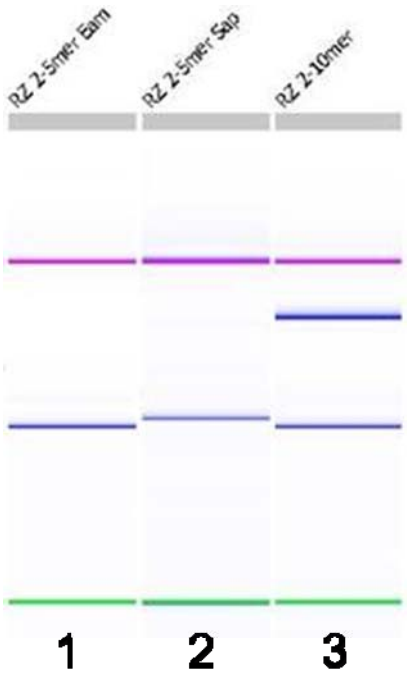


Figure 2-6: Agilent Bioanalyzer "pseudo-gel" image of RZ2-10 created by controlled cloning. Lane 1: PCR-amplified pentamer gene (317 bp) after *Eam1104* I (*Ear* I) digestion; lane 2: PCR-amplified pentamer gene after *Sap* I digestion (334 bp); lane 3: ligation products. Note that due to differences in DNA concentration, a portion of the *Eam1104* I-digested DNA fragment remains unligated.

Once the desired gene has been transferred to the expression vector, the vector is transformed into NovaBlue cells. Like the modified pUC18 vector we use, we started with a pET-19b that contains a carbenicillin resistance gene, and modified it using dangled primers to accept genes produced by controlled cloning [133]. Almost all colonies will contain the desired gene of correct length. In rare cases, two ligated copies of the original gene will be inserted into the same vector in series, resulting in a doubled gene. The plasmid DNA is sequenced again for verification and then transformed into BLR(DE3) *E. coli* cells, using the manufacturer's protocol, a strain that contains the T7 RNA polymerase gene needed for high levels of protein

production. BLR(DE3) is a tetracycline resistant *recA*- derivative of BL21 that may help stabilize target plasmids containing repetitive sequences [125], and is commonly used by other protein polymer research groups.

Note that Dr. Jong-In Won later improved the method used previously to generate the modified pET-19b vector [133]. The vector is digested with *Ear* I following PCR using dangled primers to obtain the proper cohesive ends. Therefore it is necessary to alter preexisting *Ear* I sites in the vector to prevent unwanted cleavages. The original approach was to perform the dangled primer PCR in the presence of 5-methyldeoxycytosine to protect other *Ear* I sites contained within the plasmid from enzymatic digestion. However, this method was not completely reliable as not all sites would incorporate the 5-methyldeoxycytosine. In a subsequently revised protocol, one base was changed in the single preexisting *Sap* I recognition site of pET-19b. As there were no more *Sap* I sites left in the vector except those deliberately incorporated using the dangled primers, the proper cohesive ends could be obtained by *Sap* I digestion. We have not observed any change in protein expression levels between using the pET-19b vector and the newer MpET-19b vector.

2.5.1 Growth and induction protocols

A small starter culture of LB media is inoculated with BLR(DE3) cells containing the expression plasmid and allowed to grow overnight at 37°C. A portion of this starter culture is then used to inoculate a larger-volume culture grown during the day at approximately a 1:40 volume : volume ratio. The exact volumes used depends on the size of the test expression or large-scale expression to be performed, discussed in further detail in Sections 2.5.5 and 2.5.6,

respectively. LB or Terrific Broth (TB) is typically used as the media for the day culture. Cells are grown under both carbenicillin and tetracycline selection. Once the cells reach the mid-log growth phase (usually 2.5-3 hours) as evidenced by an optical density at 600 nm ranging from OD_{600} = 0.5-0.8, protein expression is induced by adding IPTG. This initiates production of T7 RNA polymerase and subsequently of the target protein. After 3-4 hours of expression the cells are harvested by centrifugation at 6500 rpm, and the used growth media is discarded.

Typically an IPTG concentration of 1 mM in the culture is used for expression although we have induced expression at alternative IPTG concentrations (0.1, 0.5, 5 mM). We have also explored expression using no inducer. In this situation, the "leaky" nature of the T7 promoter (*i.e.,* the low basal level of protein expression that occurs with no inducer) is exploited to provide protein expression over a long period of time (24-48 hours) at room temperature. The results that are obtained on protein expression without an inducer will be discussed in detail in Sections 6.4.3.4 and 6.4.3.9.2.

2.5.2 Cell lysis

The harvested cells are then resuspended in denaturing lysis buffer containing 8 M urea or 6 M guanidine at pH 7.0. Cells are lysed to release their internal contents through a combination of freeze/thaw cycles, followed by sonication, which uses high frequency sound waves to shear cells. Detergent formulations (commercially available) can also be used to break the lipid barrier which surrounds cell membranes by solubilizing proteins and disrupting lipid:lipid, protein:protein and protein:lipid interactions. Our most common method was

freeze/thaw cycles combined with sonication to lyse the cells. Afterwards, centrifugation at 10,000 rpm separates the cell debris from the liquid (clarified) lysate.

2.5.3 SDS-PAGE protein analysis

Discontinuous SDS (sodium dodecyl sulfate) polyacrylamide gel electrophoresis (SDS-PAGE) was frequently used to analyze cell lysates or other protein mixtures. SDS unravels proteins and binds to them at a ratio of 1.4 g SDS/g of protein [134]. Binding of the negatively charged SDS detergent masks any intrinsic charge of the protein. Proteins treated with SDS have similar shapes (linear) and charge-to-mass ratios as a result. This allows protein separation by electrophoresis to occur strictly according to chain length (molecular mass), in theory.

Our proteins all contain a 8X or 10X polyhistidine tag for affinity purification. We have observed that the presence of a His tag causes proteins to migrate at a higher apparent molecular weight on the gel when compared to their actual molecular weight; so clearly, there is some sequence dependent mobility. We have also observed abnormal protein migration on a gel for our PZ8 (GAGTGSA) sequences, which contain no charged residues. Coomassie blue staining is used to visualize the proteins on the PAGE gel. For completely uncharged proteins, such as PZ8 or variants containing only a few arginines (~ 2 - 8 total), protein visualization is difficult if the positively charged histidine tag is removed. A negative zinc stain has also been tested but with no difference in detection sensitivity.

2.5.4 Dot blot antibody-based detection

The use of a so-called dot blot is another way to detect the presence of protein in a sample if it contains a specific antigen for which an antibody is available. The protocol

described below was obtained from the Koltover lab (through Nicolynn Davis). Protein samples are dotted onto a nitrocellulose membrane. After incubation with a milk solution to block nonspecific binding sites, the membrane is incubated with the primary antibody at a 1:10,000 dilution. This antibody binds to proteins containing a penta-His (HHHHH) sequence. A secondary, anti-IgG antibody becomes bound to the primary anti-His antibody in the next incubation step (1:8000 dilution) due to their similar mouse IgG isotypes. Horseradish peroxidase (HRP), which is an enzyme that is conjugated to the secondary antibody, then allows for chemiluminescent detection of the histidine-tagged protein. The membrane is briefly soaked in the ECL (enhanced chemiluminescence) reagents from an ECL kit (GE Healthcare, Piscataway, NJ). This elicits a peroxidase-catalyzed oxidation of luminol and subsequently enhanced chemiluminescence where the HRP antibody is present on the membrane. The resulting light is detected by exposure to Kodak BioMax XAR film in a dark room for seconds or minutes depending on signal intensity.

2.5.5 Small-scale test expression of protein polymers

When a new protein sequence is being expressed for the first time, expression levels from a few colonies are checked to determine which colony is best to use for large-scale expressions. Although all colonies nominally contain the same expression vector, mutations may exist elsewhere in the plasmid or cell that can be beneficial or detrimental to protein expression levels. The protocol described below was obtained from the Koltover lab (through Nicolynn Davis).

Typically, 10 mL of TB media are inoculated from an overnight culture and grown until it reaches the optimal optical density range. For best comparison results, cultures are induced at

similar optical densities. 5 mL of the culture is induced with IPTG while the other half serves as the control. The OD_{600} of both the control and induced cultures are taken after induction is complete. 3 mL of each sample is centrifuged and the media is discarded. The cells are resuspended in denaturing lysis buffer according to the following formula: $OD_{600}/0.05*3$ = µL of buffer. Afterwards, the cell lysates are compared by SDS-PAGE and if necessary, a dot blot, to assess relative protein expression levels. Figure 2-7A is a Coomassie blue-stained SDS-PAGE gel of purified PZm8-6 (to be discussed in Chapter 5) protein from a large-scale expression and control and IPTG-induced cell lysates from test expressions of longer length proteins. Figure 2-7B is a dot blot of the same samples in Figure 2-7A.

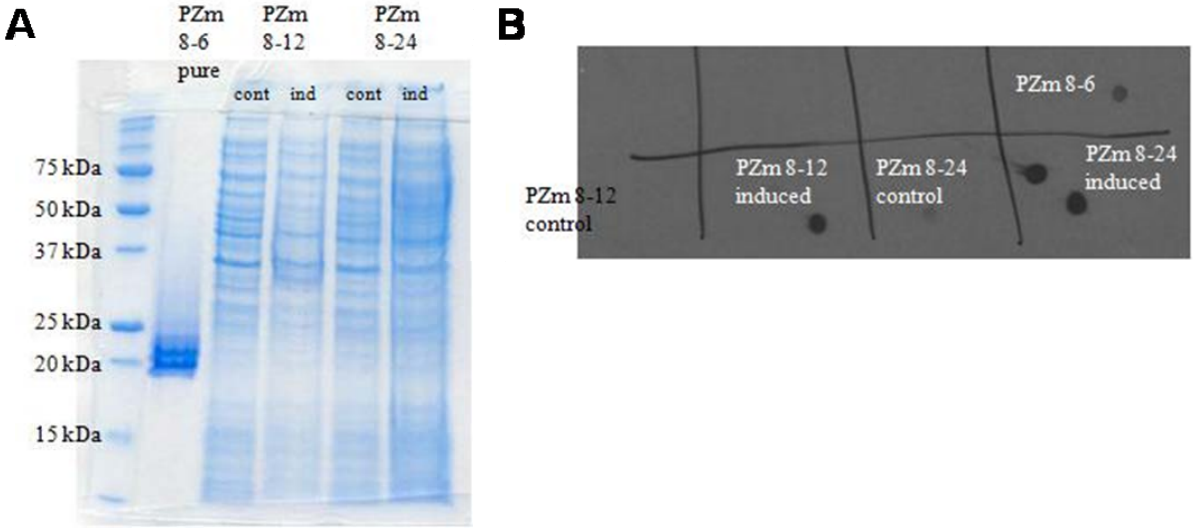


Figure 2-7: Purified PZm8-6 protein and test expressions of PZm8-12 and PZm8-24 A) SDS-PAGE results compared to B) dot blot results of the same protein samples. Typically in SDS-PAGE gels the expressed protein band is indistinguishable from native proteins in the cell lysate as can be seen by the nearly identical control and induced samples.

2.5.6 Large-scale cultures

Large-scale expressions are performed in 2 L to 8 L batches in 2800 mL Fernbach shaker flasks, using the growth and induction conditions and cell lysis techniques described earlier in Section 2.5.1 and 2.5.2, respectively.

2.6 Purification of non-natural repetitive polypeptides

As with any synthesized molecule, purification is an important and often the most time-consuming step in the production process. The protein polymer must be isolated and purified from a complex mixture of native *E. coli* proteins, lipids, DNA, and other cellular components. The method we have chosen to use is affinity chromatography, a commonly used method of protein purification [135]. The target protein is expressed with an affinity tag attached. This tag binds to a particular ligand contained on the surface of a resin. The fusion protein binds to the column while, ideally, all other unwanted cellular components flow through.

2.6.1 Immobilized metal affinity chromatography (IMAC)

Specifically, in immobilized metal affinity chromatography (IMAC) protein polymers are expressed with a His tag consisting of 8-10 consecutive histidines. The imidazole side chains on the histidines have an affinity for divalent metal ions such as nickel, zinc, and cobalt [136, 137]. Under physiological pH, histidine binds by sharing electron density of the imidazole nitrogen with the electron-deficient orbitals of the transition metal ion [138]. Three histidines can bind transition metals under certain conditions but six histidines can reliably bind transition metals even in the presence of strong denaturants. The small size of the tag and the ability to do purification under denaturing conditions are two advantages of using the His tag.

The clarified cell lysate obtained from a large-scale expression is bound to nickel-chelating or cobalt-chelating resin (to which the His tag binds) for a minimum of one hour on a rotator. The entire mixture is then loaded onto a column fitted with a filter, which allows the passage of buffer but not resin. In the case of the nickel resin (ProBond), the column is subsequently washed with 3 column volumes each of buffers of gradually decreasing pH (pH 7.8, pH 6.0, and pH 5.3) to remove non-specifically bound native proteins from the column. A 6 column volume buffer wash at pH 4.0 then elutes the desired protein from the column. The imidazole nitrogen becomes protonated at this pH and the positively charged ammonium ion is repelled by the positively charged metal ion. For the cobalt resin (Talon), buffer pH is maintained at 7.0. Non-specifically bound proteins are removed by buffer washes (10 column volumes each) containing zero to low amounts of imidazole (5-20 mM). The protein is then eluted in 3 column volumes of buffer containing 150 mM imidazole. Imidazole is a binding competitor to histidine as it is structurally identical to the histidine side chains. These were the buffer and protocols of choice outlined in the respective manufacturer's protocols. However, imidazole elution can be used with nickel resin and likewise pH elution can be used with the cobalt resin. Some adjustment in concentration or pH may be needed to account for the different metal ions.

Column purification fractions are analyzed by SDS-PAGE to determine the purity of the final eluted protein. If necessary, column purification is repeated to obtain a purer product. Higher purity protein can sometimes be obtained using both pH- and imidazole-based purifications sequentially. The protein solution is then dialyzed against water for 3 days to remove salts and small contaminants then lyophilized to a dry powder. Dialysis membrane

molecular weight cutoffs (MWCO) range from 3500 to 14,000 Da depending on the expected protein polymer size. A typical yield of a protein polymer expressed with an *N*-terminal His tag ranges from 25 to 35 mg/L of culture. Lower yields have been obtained using a *C*-terminal His tag (Chapter 6).

2.6.2 Removal of the affinity tag by cyanogen bromide cleavage

The *N*-terminal His tag in the pET-19b vector contains several undesirable, charged amino acids, such as lysine and aspartic acid, that need to be removed before the protein polymer can be used as a drag-tag. The His tag can be removed by chemical cleavage at the *N*-terminal methionine residue, using cyanogen bromide in 70% formic acid for 24-48 hours [139]. Our repetitive sequences otherwise contain no methionine residues, so this method will cleave at one specific point. Proteins are dissolved in the reaction mixture at a final concentration of ~ 1 mg/mL. Cyanogen bromide is added at approximately 5 mg/mg protein. After nitrogen purging, the entire mixture is covered with aluminum foil and gently mixed for several hours. A rotary evaporator is then used to remove volatiles and dry the solution under vacuum. The product is resuspended in water and the cleaved His tag can be removed either by dialysis or spin column ultrafiltration (2.92 kDa size). A second column chromatography purification also can be performed to separate successfully cleaved protein from protein still attached to a His tag. Typically >70% of the protein is successfully cleaved within 24 hours. Longer reaction times lead to a higher percent of cleaved protein, but consequently increases the potential for more side reactions to occur.

2.6.3 Molecular mass (MALDI-TOF)

The mass of the purified protein can be measured using matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) mass spectrometry on a Voyager-DE PRO instrument (Analytical Services Laboratories, Northwestern University and the Protein and Nucleic Acid Facility, Stanford University). The protein is dissolved in a 50% water/50% acetonitrile with a 0.1% trifluoroacetic acid solution of sinapinic acid and allowed to dry on the surface of the metal target plate. A laser hits the sample which is under vacuum. The matrix absorbs most of the energy, preventing unwanted fragmentation of the protein sample. The protein becomes ionized and is accelerated in an electric field. The time-of-flight analyzer separates the ions according to mass-to-charge ratios (m/z). In this method, predominantly singly charged protein ions are produced [140].

2.6.4 Purity (RP-HPLC)

Reversed-phase High-Performance Liquid Chromatography (RP-HPLC) is a well-established technique for the isolation and analysis of peptides and proteins [140]. In gradient elution, proteins are essentially retained according to their hydrophobic character with more hydrophobic molecules being retained longer. The stationary bed is nonpolar (hydrophobic) while the mobile phase is a polar liquid.

Small amounts of protein are dissolved in water at 1 mg/mL and then analyzed for purity by reversed-phase HPLC () on C4 and C18 columns (Vydac, 5mm, 300 Å, 2.1 x 250 mm) with a 0-95% acetonitrile gradient in water for 50 minutes at 58°C. Peaks are detected by UV

absorbance at 220 nm (absorption of the amide bond). Preparative RP-HPLC can be used to further purify large amounts of a protein after affinity chromatography.

2.6.5 Amino acid analysis

Select proteins were sent to the Yale W.M. Keck Facility for amino acid compositional analysis. The proteins are hydrolyzed by HCl into their individual amino acids constituents and the product is analyzed on a Beckman Model 7300 ion-exchange instrument. During acid hydrolysis asparagine will be converted to aspartic acid and glutamine to glutamic acid. Serine and threonine are also partially destroyed during hydrolysis [141]. Molar ratios/percentages of each amino acid can be estimated from the resulting chromatogram.

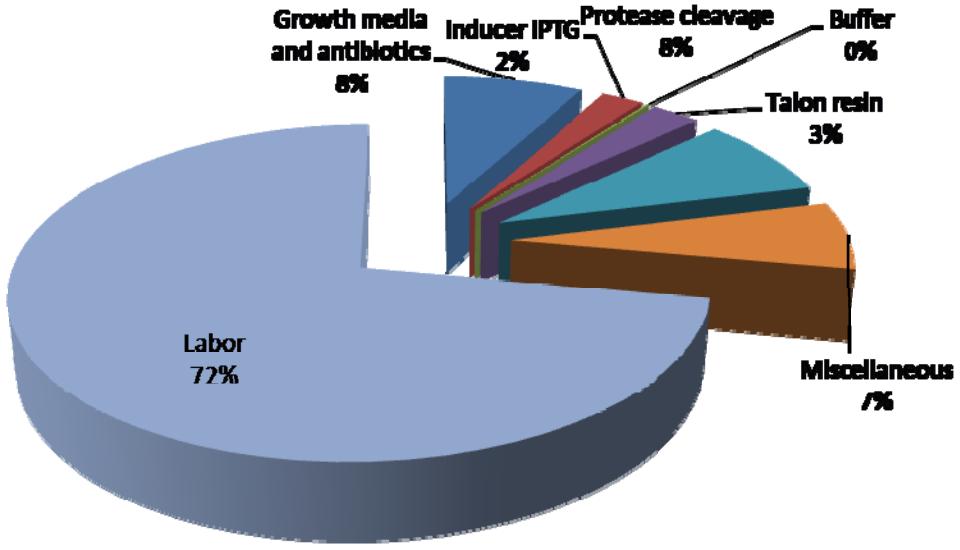
2.7 Protein polymer production: cost analysis

The process of obtaining the purified protein polymer from the initial ssDNA sequence is clearly lengthy, but once the desired gene is in an expression plasmid, the protein is relatively easy to produce. The material and labor cost of producing a batch of protein polymers using the protocols outlined in this chapter is discussed in this section. Estimates are based on items used in our lab and purchased with list pricing from Fisher Scientific (Fairlawn, NJ) or VWR (West Chester, PA). The median hourly earnings of a biological technician ($17.17) in May 2006, as reported by the U.S. Department of Labor, was used to determine labor costs [142]. Equipment such as an autoclave, sonicator, centrifuge, UV-Vis spectrophotometer, and incubators are all assumed to be available. Reusable items such as glassware and centrifuge bottles were also excluded from the cost analysis. Spreadsheet values used for the cost calculations are presented in Appendix B. Figure 2-8 illustrates the various costs involved in producing a 4 L batch of

protein polymer with a low yield of 7.5 mg or 1.9 mg/L (obtained for a 253-amino acid protein polymer using only a *C*-terminal His tag). It is also assumed that the protein is being cleaved using a protease instead of cyanogen bromide, followed by additional sample cleanup to remove the protease and uncleaved protein. Miscellaneous costs include the cost of pipet tips, gloves, dialysis membrane, and sterile filtering of buffers.

Figure 2-8 shows that the amount of man-hours involved and consequently, the associated labor costs, dominate the overall cost of protein production in this example, accounting for 72% of the total cost. The raw materials cost is $19.01 per mg out of a total of $68.23. With a 25 mg/L yield (Figure 2-9), the cost per mg protein drops to $7.40. Despite increases in materials cost to process the larger amount of protein (*i.e.,* resin, buffer, and protease), the overall cost per mg is reduced to $12.12. Note that the price of performing protease cleavage does become more significant with the higher protein yield.

A

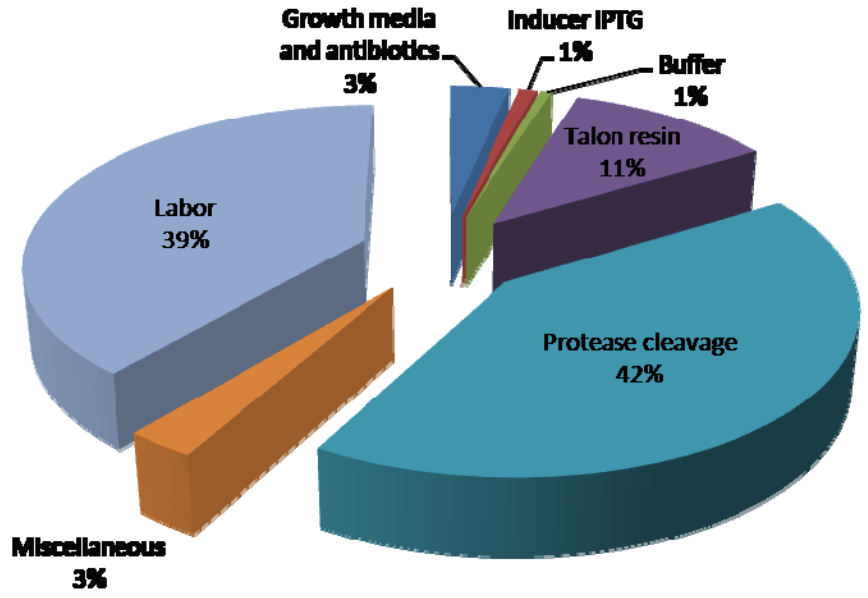


B

	per 4L batch	per mg
Growth media and antibiotics	$ 40.13	$ 5.35
Inducer IPTG	$ 11.92	$ 1.59
Buffer	$ 0.77	$ 0.10
Talon resin	$ 13.94	$ 1.86
Protease cleavage	$ 41.41	$ 5.52
Miscellaneous	$ 34.44	$ 4.59
Labor	$ 369.16	$ 49.22
Total	$ 511.75	$ 68.23

Figure 2-8: Cost analysis for 7.5 mg/ 4 L culture A) pie chart showing % contribution B) cost table of actual dollar amounts

A



B

	per 4L batch	per mg
Growth media and antibiotics	$ 40.13	$ 0.40
Inducer IPTG	$ 11.92	$ 0.12
Buffer	$ 7.67	$ 0.08
Talon resin	$ 139.40	$ 1.39
Protease cleavage	$ 506.19	$ 5.06
Miscellaneous	$ 34.44	$ 0.34
Labor	$ 472.18	$ 4.72
Total	$ 1,211.92	$ 12.12

Figure 2-9: Cost analysis for 100 mg/ 4 L culture A) pie chart showing % contribution B) cost table of actual dollar amounts

Fortunately, for our purposes, gram quantities of protein polymer are not needed. Only ~ 0.05 mg of (maleimide-activated) drag-tag is used in a single DNA conjugation reaction, as discussed in more detail in Section 5.2.1 [63, 123]. Of this conjugated drag-tag, 4.2 pmol (or 83 ng for a 264-amino acid protein with a molecular weight of 19.72 kDa) are used in a single

Sanger sequencing reaction. This amounts to a per sequencing reaction cost of 0.57¢ and 0.10¢ for the low and high yields being analyzed, respectively.

Automation of processes such as affinity chromatography, which is certainly feasible, would reduce the amount of man-hours required. Costs may be further reduced if protease cleavage and use of an inducer were unnecessary. These strategies will be discussed in later chapters.

2.8 Future Recommendations

Many steps are involved in producing these protein polymer drag-tags. The cloning, expression, and purification of proteins is a vast field and we have only touched upon and utilized a number of the techniques that exist for making proteins, and applied them to our protein polymer system.

2.8.1 Design of amino acid sequences for protein polymers

As in the course of our research, more and more different protein sequences are generated by these methods, our knowledge of what works and does not work in a drag-tag sequence is improving. As will be discussed in later chapters, our original criteria of completely avoiding all positively charged amino acids has proven to be unnecessary. We were concerned about ionic interactions with DNA and also glass surfaces. The inclusion of a limited number of arginine residues not only improves protein polymer water solubility but also increases the hydrodynamic drag α of the protein relative to uncharged sequences. These benefits were achieved with none of the expected detrimental interactions with the negatively charged DNA or microchannel walls. Research is currently being undertaken by another graduate student, Xiaoxiao Wang, to

determine the maximum number/percentage of arginine monomers that can be incorporated into a drag-tag sequence before we begin to observe unfavorable interactions. These designs consist of 1 Arg in 18 up to 1 Arg in 8 amino acids. The production of block copolymer drag-tag designs would allow the study of protein polymers that have different regions of charge density, length, and sequence.

2.8.2 Cloning

As the protein polymer sequences we express become longer and longer, we have observed a corresponding decrease in the amount of the desired PCR product, as well as an increase in the number of side products generated during amplification. Consequently, larger volume PCR reactionsmust be done to obtain the same amount of starting DNA. Using insert DNA that is directly excised from the pUC18 by *Ear* I or *Sap* I digestion easily generates large amounts of the DNA needed for half of the controlled cloning protocol, thus avoiding one of the PCR steps. Additionally, this DNA can be self-ligated prior to insertion into cloning or expression vectors, improving the chances of obtaining longer sequences. These modified methods were developed by Xiaoxiao Wang.

2.8.3 Expression and Purification

We have chosen to use a His tag for affinity purification of our proteins because of its widespread use, simple sequence, small size, and ability to bind under both native and denaturing conditions. However, many other affinity tags [137, 143] and purification strategies [141] exist for protein purification. For example, affinity tags based on glutathione S-transferase (GST), maltose binding protein (MBP), calmodulin, and *Strep*-tag II (WSHPQFEK), which utilizes

streptavidin-biotin binding, are also available. MBP and GST can also be used to improve overall recombinant protein solubility. A T7 tag (MASMTGGQQMG), S-tag (KETAAAKFERQHMDS), or FLAG (DYKDDDDK) tag can be used for antibody-based affinity chromatography. Additionally, variants on the His tag such as MAT (HNHRHKH) and HAT (KDHLIHNVHKEEHAHAHNK) have been developed. These variants are still compatible with nickel and cobalt resins and are purported by the manufacturers to be more "natural" by reducing the number of consecutive, positively charged histidine residues and consequently improving solubility of the recombinant protein. Besides affinity chromatography and RP-HPLC, ion-exchange and size-exclusion chromatography (SEC) can be used to further purify a protein polymer product, if necessary, by utilizing a different separation mechanism to remove remaining impurities.

2.8.4 Cost of the process

New techniques now are being applied in our lab that could reduce the cost of the traditional strategy of protein expression followed by affinity chromatography and protease cleavage. Inteins, which are self-cleaving proteins, have been adapted for affinity tags that are self-cleaving, requiring no expensive protease step. Two such systems will be discussed in further detail in Chapter 4.

Inverse transition cycling (ITC) is a technique that relies on the phase transition behavior of elastin (ELP) tags [144]. These polypeptides are sensitive to salt and temperature, allowing purification by mild temperature shifts and addition of salt. The ELP tag will self-assemble into an insoluble precipitate and the entire fusion protein can be separated from the soluble fraction

and collected by centrifugation or filtration. A new strategy combines ELP tags with a self-cleaving intein, a purification technique that uses no chromatography resin or protease [145-148]. However, theses processes can be very time-consuming, while being lower in cost.

Another novel purification system developed by the Wood lab has the cells synthesize the "resin" simultaneously with the target protein [149]. A self-cleaving phasin tag is expressed in fusion with the desired protein. Another plasmid in the cell codes for 3 enzymes involved in the synthesis of polyhydroxybuterates (PHBs) from metabolic acetyl CoA. These biodegradable polymers form small intracellular granules when expressed and have an affinity for phasin. The granules along with bound protein are easily recovered and the protein is released using the self-cleaving intein.

The above mentioned methods all show considerable reduction (up to 11 fold) in the material cost of protein expression and purification when compared to the traditional methods we have used so far [145, 150, 151]. However, it has been shown that the sequence of the protein being fused to an ELP tag can have a dramatic affect on its phase-transition behavior [152]. This would necessitate optimization of the ELP tag (sequence and length) for each protein polymer being expressed, which could be a time-consuming process. Simplicity and economics are the most touted advantages of these systems. Nevertheless, our protein polymers must be completely pure in order to be used as drag-tags, regardless of the cost. It would be worth exploring whether these lower-cost techniques can achieve similar or improved levels of purity compared to our current methods, if at some time the cost of the drag-tags becomes an important factor in keeping the cost of sequencing low.

Chapter Three

New Drag-Tag Sequence Designs: RZ-1 and RZ-2

3.1 Introduction

One of the most significant challenges in designing a non-natural repetitive polypeptide as a drag-tag is that there is an enormous sequence space to choose from using the 20 naturally occurring amino acids. Assumptions or arbitrary choices have to be made based on the anticipated desirable properties of the drag-tag. Here we present the results we obtained for two designs radically different from the PZ-1 through PZ-6 and BB-1 series listed in Table 3-1. At the time of this work, we were still seeking a suitable drag-tag sequence that could be successfully used for DNA sequencing.

3.2 Previous drag-tag designs

All the designs that had been previously produced by Jong-In Won were based on a seven-amino acid repeating sequence. Three repeats made up a "monomer" of 21 amino acids. Based on the ideal drag-tag criteria outlined in Section 1.3.2, the amino acids used were limited to glycine, alanine, serine, glutamine, glutamic acid, lysine, valine, leucine, and asparagine. Despite its negative charge, glutamic acid was included in some sequences because of concerns that the protein polymer might not be water-soluble if it were completely uncharged. Accordingly, only two hydrophobic residues were used in these designs.

Table 3-1: Previously Studied Drag-Tag Designs

Name	**Repeating sequence**	**Comments**
PZ-1	Gly-Ser-Gly-Gln-Gly-Glu-Ser	Good expressibility and solubility; too hydrophilic for RP-HPLC purification; contain glutamine and glutamic acid; similar mobility to DNA makes these poor drag-tags
PZ-2	Gly-Ala-Gly-Gln-Gly-Glu-Ala	
PZ-3	Gly-Val-Gly-Gln-Gly-Glu-Val	not well expressed by *E. coli*
PZ-4	Gly-Leu-Gly-Gln-Gly-Glu-Leu	
PZ-5	Gly-Ala-Gly-Gln-Gly-Asn-Ala	impure; both asparagine and glutamine included; soluble only with addition of 1M urea
PZ-6	Gly-Ala-Gly-Gln-Gly-Ser-Ala	contains glutamine (potentially unstable)
BB-1	Gly-Lys-Gly-Ser-Ala-Gln-Ala	positive/negative charge interaction and presence of lysine requires derivatization first; glutamine also present

PZ-1 and PZ-2 both expressed well but were too hydrophilic. As a result, they did not resolve well by RP-HPLC, and were difficult to obtain in pure form. PZ-2 also had a similar electrophoretic mobility to DNA, which is undesirable in a drag-tag for DNA separation. The similar mobility resulted from the inclusion of one negatively charged glutamic acid residue in every seven amino acids. DNA happens to have a similar 1 in 7 overall charge due to Manning

condensation of buffer cations on the DNA strands [122, 153]. We were unable to obtain expressed protein samples of PZ-3 and PZ-4, apparently due to poor expression of the two sequences in the *E. coli* host cells. The PZ-5 protein required ~ 1 M urea to remain soluble in the presence of buffer salts. In addition, we suspected glutamine and asparagine were readily deamidating to form glutamic acid and aspartic acid, respectively, under certain conditions such as low pH [120, 154, 155]. During protein purification of the His-tagged protein, the buffer pH is lowered to 4.0 to elute the bound protein off the nickel resin. Furthermore, during cyanogen bromide cleavage to remove the His tag, the protein is dissolved in 70% formic acid with a pH below 1. Even slightly basic pH conditions can promote glutamine deamidation as well [155]. We believed that this uncontrolled deamidation resulted in a mixture of drag-tags with various charges that was observed for the PZ-6 sequence during extended cyanogen bromide reaction times [120]. Although several designs were attempted none had yet proven suitable for DNA sequencing [121]; this was winter 2002.

3.3 Selection of new protein sequences to pursue

Since the previous designs had complications that made them unsuitable for use as a drag-tag in ELFSE, we examined new drag-tag sequence designs with the hope that they might result in more suitable drag-tags. The guidelines described in Chapter 2 were used to design the RZ sequences.

3.3.1 Choice of amino acids

To avoid any possible concerns regarding deamidation, glutamine, asparagine, glutamic acid, and aspartic acid were eliminated from the list of amino acid residues. In order to counter

the excessive hydrophilicity of the earlier PZ sequences which prevented good HPLC purification, the new sequences, called RZ, included glycine, serine, and alanine as before, as well as additional hydrophobic residues such as isoleucine, phenylalanine, tyrosine, and threonine, in addition to valine and leucine. Cysteine was again avoided because of the possible formation of disulfide bonds. Since a seven-amino acid repeat may be stressful on the cell machinery (low diversity), in addition to limiting the variety of amino acids that can be used, we also decided on a longer repeating sequence of 21 amino acids to replace the three repeats of seven previously used. The selected hydrophobic residues were used sparingly in the new sequence (*i.e.,* one each of isoleucine, leucine, threonine, phenylalanine, valine, and tyrosine). Eight glycines, 3 alanines, and 4 serines were also included. No charged amino acids were used.

3.3.2 Gene design

As stated previously in Chapter 2, for a drag-tag, an unstructured, random-coil configuration is preferred to α-helix or β-sheet structures, as this exposes more of the protein to the outside environment [127]. The GOR IV [128] program was used to predict protein secondary structure. Two hundred random sequences using different permutations of the 21 selected amino acid residues were generated using Matlab™. These sequences were entered into the GOR IV program, which then returned a prediction of protein structure in terms of percentage of α-helix, β-sheet, or random-coil regions.

Table 3-2: Fifteen sequences with the highest predicted % random-coil secondary structure as determined by GOR IV. Sequences with increasing random-coil tendency as length increased are shown in yellow.

	Random-coil Tendency %			
Sequence	**1x**	**3x**	**6x**	**12x**
SAGGLIVGGSTYGGFAGGSSA	61.9	71.43	73.81	75
TSGAAGGSAGGGIVSFGLSYG	61.9	71.43	73.81	75
LSGIGSSSAGFTGAYGGGAVG	61.9	71.43	73.81	75
GSFGIGASSLGVGSYGAGATG	71.43	71.43	71.43	71.43
LAGSGIYGGSASSAGGGFGTV	66.67	69.84	70.63	71.03
GYGSGSSATGAGGILGAVSFG	66.67	66.67	66.67	66.67
SYTGAGIGSGASAGSVLGGGF	61.9	65.08	65.87	66.27
VLGGSGASTAGSAFSGYGGIG	61.9	65.08	65.87	66.27
YVSSTSGASGGGAILGGGAGF	66.67	63.49	62.7	62.3
SIYGTGAGASFLGSSGAGVGG	61.9	61.9	61.9	61.9
AFSTYIGGGASGAGSGGGLVS	66.67	60.32	58.73	57.94
GLAASGGVGGSSTAGYGIGFS	71.43	58.73	55.56	53.97
ISTYAGGSSFLVAGGGGGGAS	66.67	57.14	54.76	53.57
VTAGGSGGGSGYGAISFASGL	66.67	57.14	54.76	53.57

Only two sequences yielded random-coil predictions higher than 70%. Sequences for which greater than 60% random coil was predicted for the "monomers" were tested for random-coil tendencies as trimers, hexamers, and dodecamers. The top 15 results are presented in Table 3-2. Sequences that showed a decreasing trend in random-coil prediction as chain length increased were discarded. The remaining sequences were further reworked by repositioning residues to avoid placement of identical amino acids in adjacent positions and to also spread out the hydrophobic residues as much as possible. To be compatible with currently available plasmids and primers in our lab, all sequences were changed to position a glycine at the amino terminus. Each reworked sequence was then rerun through GOR IV. Most of the time,

rearrangement decreased the random-coil prediction of the sequence. In a few cases the sequence rearrangement increased the random-coil tendencies. The two sequences with the highest random-coil tendencies were chosen. The selected sequences are GSFGIGSASLGVGSYGAGATG and GYSTAGIGSGASAGSGVGLFG or RZ-1 and RZ-2, respectively (Figure 3-1A). Even though as a monomer both had random-coil predictions of only 61.9%, these proteins yielded gradually increasing random-coil tendencies as they became larger multimers (Figure 3-1B). As 24mers, both had random-coil predictions of over 75%. By comparison, as a 24mer, PZ-6 has a predicted random-coil tendency of 99.6%.

Designing the actual gene sequence that encodes the desired protein was the next stage. For amino acids appearing only once in the sequence, we utilized the codon with the highest frequency of usage in *E. coli*. For amino acids appearing more than once in the repeat, the number of times a codon was used was based on the relative frequency of use in *E. coli*. For example, there are eight glycines in the RZ sequences and four separate codons that encode glycine [156]. Based on this information, the first two codons, GGC and GGU, would be used 3 times each whereas, the other two, GGG and GGA, would be used only once each. To be

consistent and compatible with previous work, GGT was used to encode the first glycine.

A

RZ-1 (GSFGIGSASLGVGSYGAGATG):
5'-GCTAGCCATATGCTCTTCAGGTAGCTTTGGCATTGGGAGCGCGTCTCTGGGTGTGG
GCAGTTATGGCGCCGGTGCAACTGGAGGTTGAAGAGGGATCCACTAGT-3'

RZ-2 (GYSTAGIGSGASAGSGVGLFG):
5'-GCTAGCCATATGCTCTTCAGGTTATAGCACGGCGGGCATTGGATCTGGTGCAAGTGC
CGGCAGCGGGGTGGGTCTGTTTGGCGGTTGAAGAGGGATCCACTAGT-3'

B

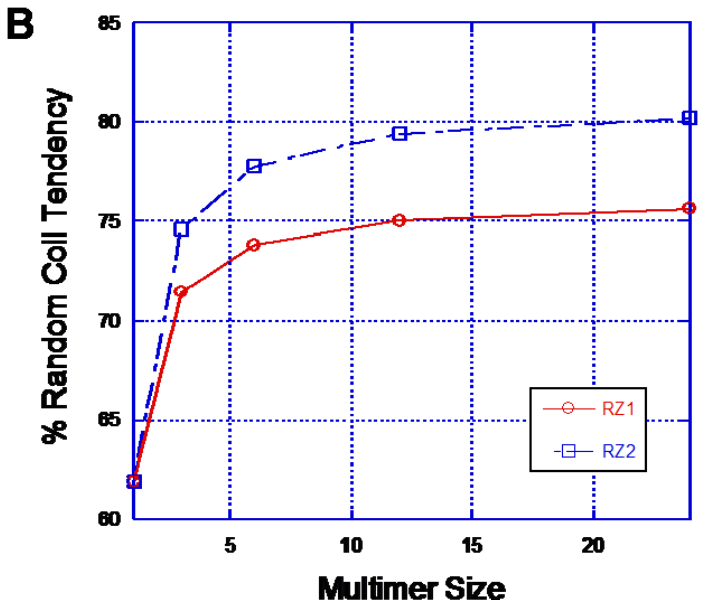


Figure 3-1: A) Amino acid and gene sequence for RZ-1 and RZ-2 designs B) GOR IV predicted % random-coil tendency for RZ-1 and RZ-2 as multimers of the 21-amino acid repeating sequence

The dsDNA melting temperatures for both sequences were designed to be below 70°C. To achieve this, the codon used for threonine had to be changed to one that was less favorable. Both DNA sequences were verified to exclude the recognition sites of enzymes used in the cloning process within the repeating sequence (*i.e., Ear* I, *Sap* I, *Nde* I and *BamH* I).

3.4 Cloning of the new RZ designs

Approximately 20 extra bases were added to each side of the 63-bp gene sequence. These bases encode flanking restriction enzyme sites (*Sap* I/*Ear* I on one end and only *Ear* I on the other) as well as *Nde* I and *BamH* I sites for direct insertion into the expression vector, if desired. Identical extra sequences used previously [133], were implemented for the RZ designs, allowing use of the same primers for PCR amplification and controlled cloning as before.

The 104-bp DNA sequences encoding RZ-1 and RZ-2 were purchased from Oligos, Etc. (Wilsonville, OR). Cloning of the two RZ genes was performed as discussed earlier in Section 2.4. Self-ligation of the gene generated a ladder of differently size concatemers as shown in Figure 3-2. The largest concatemer size obtained after checking numerous (36) colonies by colony PCR screening and plasmid digestion in both cases was a pentamer.

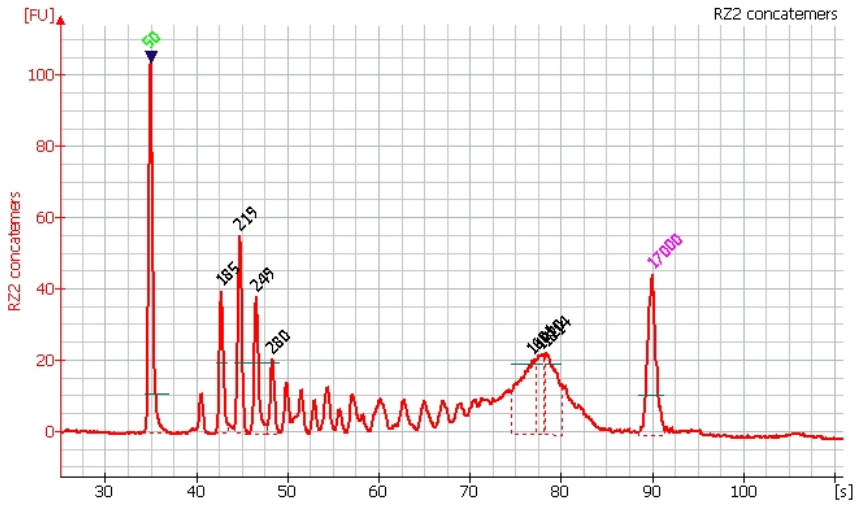


Figure 3-2: Electropherogram of the RZ-2 concatemer ladder (Agilent 2100 Bioanalyzer using the DNA 12000 kit). 50 bp and 17,000 bp peaks correspond to the lower and upper markers used to align each run against the ladder sample for size determination. Figure 2-2 (lane 2) is a gel version of the same concatemer ladder presented here.

3.5 RZ pentamer protein expressions

3.5.1 Small-scale test expression of RZ 1-5 and RZ 2-5

Expression of both proteins on a small scale (5 mL) was attempted; however, SDS-PAGE with Coomassie staining did not identify any bands that were unique to the induced samples compared to the controls. Occasionally these protein polymers do not stain well (even with the presence of the His tag) so a larger (50 mL) scale expression for both proteins was performed. Again, no apparent band could be found in the elutions after column purification. No protein product was found after dialysis and lyophilization of the elutions. MALDI-TOF analysis could not detect the presence of any expressed RZ-1 or RZ-2 protein.

3.5.2 Large-scale RZ 2-5 protein expression

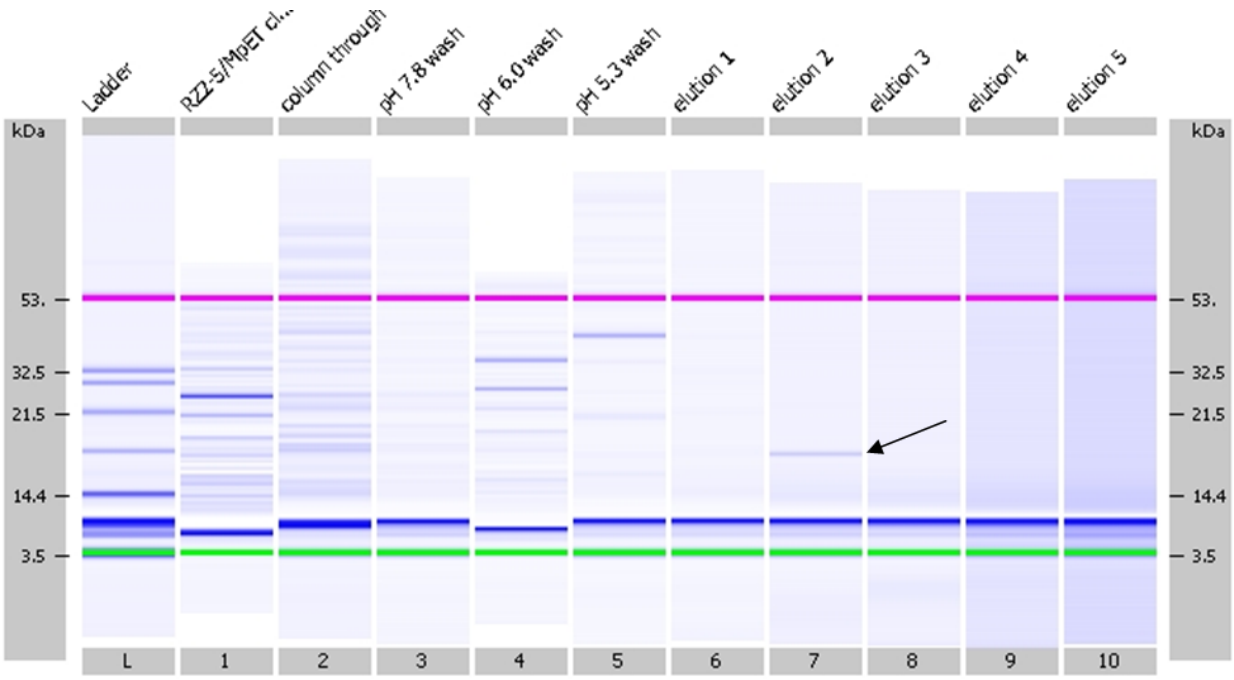


Figure 3-3: Affinity chromatography fractions of RZ 2-5 500 mL expression (Agilent 2100 Bioanalyzer pseudo-gel image using Protein 50 kit). Product appears in 2nd elution. Green and pink bands are the lower and upper markers used to align runs with the ladder sample (far left). The intense blue band present in all samples at 6 kDa is an artifact of the system and is not indicative of actual protein at that size.

The entire cloning process was repeated from the beginning, starting with the initial amplification of the synthesized oligonucleotides to ensure each step was executed correctly. A 500 mL culture of RZ2-5 in MpET-19b was expressed and purified. The column purification results are presented in Figure 3-3. A 17 kDa band is present in the second elution fraction (13 kDa is the expected fusion protein size). After dialysis and lyophilization, no protein was recovered from any of the elution fractions.

3.6 Longer length RZ sequences

RZ 1-5 was doubled using controlled cloning into a 10mer and expressed, in case a longer length of RZ would be more favorable for expression.

3.6.1 RZ 1-10 test expression dot blot

Figure 3-4 is a dot blot of test expressions performed on three different RZ 1-10mer colonies. It was observed that all three colonies grew more slowly than past RZ and PZ expressions. After three hours of induction with 1 mM IPTG, only a slight difference in optical density was measured between control and induced samples, hinting at little to no protein expression occurring. Cell growth typically slows down upon induction as they devote most of their resources to producing the target protein instead of multiplying. The dot blot results of the cell lysates were surprising. Not only did all three samples show evidence of induced His-tagged protein (bottom row), but also there was strong protein presence in the 3rd control (top row). Perhaps protein was being expressed prematurely, slowing down cell growth before IPTG was added to the culture.

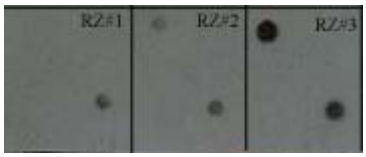


Figure 3-4: Dot blot of RZ1-10 test expressions

3.6.2 RZ 1-10 large-scale expression and purification

RZ 1-10 was expressed in 2 L of LB media using 1 mM IPTG for three hours. After affinity chromatography, the fractions were analyzed by SDS-PAGE. No protein was visible in either the second wash or any of the elution fractions. The elution fractions were combined, dialyzed, and lyophilized. White precipitate was found after the dialysis but although the entire

sample was lyophilized, only small specks of material were recovered. MALDI-TOF did not detect any protein in the RZ1-10 sample.

3.7 Conclusions and Recommendations

We were unable to determine the exact reason for the apparent failure of the RZ protein expressions. Proteins of similar length to RZ 2-5 and RZ 1-10 have successfully been expressed before (*i.e.,* not issue of length but rather sequence). At this point, work by Dr. Jong-In Won on his PZ-8 sequence (6mer) successfully produced a protein polymer that was used as a drag-tag in DNA sequencing [123]. We believe the RZ sequences were too drastic a change from previous sequences and instead based future sequence designs on PZ-8.

Protein expression of the RZ sequences may still be possible using a different expression system than the *N*-terminal His tag. One such system was tested with the RZ 2-5 sequence and the results are discussed in Chapter 4. Another option is to attempt expression in the GST/His tag double vector design discussed in Chapter 6. The large GST protein may help promote protein expression and improved solubility. Altering the growth and induction conditions to lower IPTG concentrations and incubation temperatures may also be beneficial. A lower temperature of 24°C (not 30°C or 37°C) was necessary for expression of a protein polymer consisting of hydrophobic and hydrophilic blocks [157].

Chapter Four

Intein-Based Protein Purification: A New Method for Obtaining High-Purity Proteins

4.1 Introduction

Chapter 2 discussed the standard cloning, expression, and purification scheme used to produce the protein polymer drag-tags. Each protein is inserted into an expression plasmid and purified by immobilized metal affinity chromatography (IMAC) followed by cleavage of the histidine affinity tag by cyanogen bromide. We suspected that the harsh, acidic conditions of the cyanogen bromide reaction could lead to unwanted side reactions or modifications to the protein polymer [120, 135]. While site-specific proteases such as enterokinase are commonly used for small-scale cleavages and require only mild cleavage conditions, such a technique would be expensive when processing large amounts of protein. This was evident in Section 2.7 where the proportion of production costs spent on protease cleavage increased from 8% to 42% of the total production cost when the yield of a 4 L expression increased from 7.5 mg to 100 mg. Additionally, after reaction with the protease, an extra step must be performed to separate the enzyme from the desired protein product. This cleanup is typically done using special resin to capture the protease from the reaction mixture. To obtain a monodisperse, pure protein polymer, it would be desirable to use an alternate purification method, in particular one that could easily be used both on a small scale and on a large scale, with mild conditions for affinity tag removal

and no need for a second purification step. In this chapter, we discuss cloning and expression work done using two different self-cleaving "intein" affinity tags.

4.2 Background on Inteins

In 1990, three groups independently discovered that a *Saccharomyces cerevisiae* ATPase gene (*Sce* VMA) contained an insertion sequence unrelated to other ATPases [158]. They theorized it was removed via protein splicing (a novel concept at the time). Subsequent experiments proved that this protein splicing was post-translational [158]. In 1994, a uniform nomenclature for protein splicing was established [159]. The intein (internal protein), analogous to RNA introns, is the intervening protein sequence that is removed by protein splicing. A precursor protein, the primary translation product, containing an N-extein (external protein), the intein, and C-extein, undergoes cleavage. The intein is excised, while the two exteins join (ligate) together, forming a new protein. The majority of inteins have a nucleophilic amino acid at their N-terminus (*e.g.,* Cys, Ser) and asparagine at their C-terminus and are linked to a C-extein with an *N*-terminal nucleophilic amino acid (*e.g.,* Cys, Ser, Thr) [158]. Protein splicing occurs through a four-step mechanism of sequenctial acyl rearrangements.

Researchers have harnessed this unique property of inteins to design affinity tags that undergo self-cleavage upon addition of either a reducing agent such as dithiothreitol (DTT), β-mercaptoethanol, or cysteine or through a pH shift [160-163]. Self-cleavage of the tag avoids the use of an enzyme for cleavage and a second purification step. Intein cleavage is very specific compared to a protease, the conditions are relatively mild compared to cyanogen bromide cleavage, and the process can be easily scaled-up.

Intein-based bioseparations have become widely used on the laboratory scale for protein purification and may eventually be used in large-scale protein production such as pharmaceuticals or commodity enzymes, if it can be made economically competitive [158]. The simple induced cleavage reaction may be as advantageous in large-scale bioseparations as it has been on the laboratory scale. A novel pilot-scale system called vortex flow adsorption has been developed using intein purification [164]. In addition, a simulation comparing the IMPACT system from New England BioLabs (discussed below) to a conventional process using affinity chromatography for the large-scale production of a protein in *E. coli* [158, 165] determined that improving the binding capacity and reusability of the chitin resin and reduction or elimination of the chemical inducer (by switching to pH or temperature induction) would be necessary to be economically competitive. Cleavage efficiency also heavily influenced the final cost.

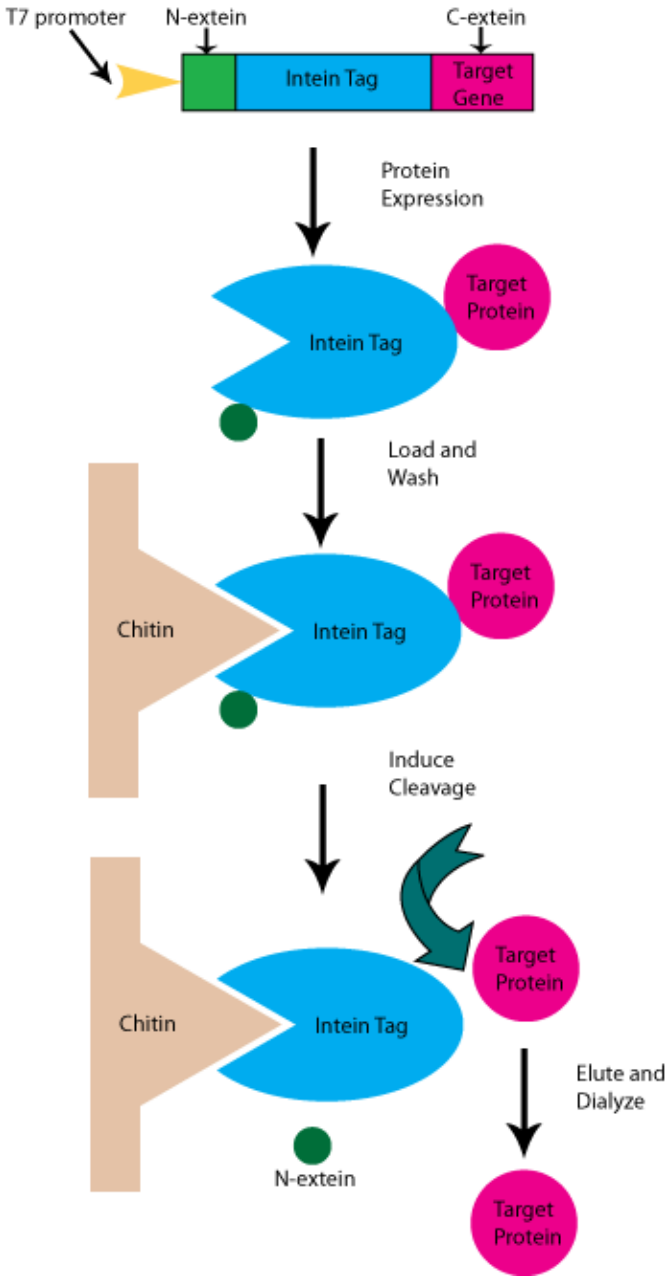


Figure 4-1: Diagram of New England BioLabs IMPACT system (*N*-terminal affinity tag) adapted from manual [166]

4.3 The IMPACT system (Intein Mediated Purification with a Chitin Tag)

New England BioLabs (NEB) has developed and commercialized intein affinity tags for protein purification [160, 167]. In the Intein Mediated Purification with an Affinity Chitin Tag (IMPACT) system (refer to Figure 4-1), a modified *Saccharomyces cerevisiae* intein (56 kDa) is attached to a chitin- binding domain taken from *Bacillus circulans*. A protein expressed with this tag binds to a column containing chitin beads. After several washes to clear the column of nonspecifically bound proteins, a thiol-containing compound such as DTT is added to induce cleavage for 16-40 hours at 4°C, 16°C, or 23°C depending on the properties of the target protein. In principle, the cleaved intein tag, along with the uncleaved precursor protein, remains bound to the column, while the cleaved protein can be eluted from the resin. For the *N*-terminal fusion vector, cleavage occurs before and after the intein tag. The small N-extein sequence (1.6 kDa) that is cleaved simultaneously with the target protein (C-extein) can be separated from the co-eluted target protein by dialysis. Protein yields have ranged from 0.8 to 20 mg/L culture using this system [166].

4.3.1 Inserting the gene into the intein expression vector

Designed drag-tag sequences are inserted into the pET-19b expression vector (Chapter 2) where the insert is flanked by *Nde* I and *BamH* I restriction enzyme sites. Drag-tag sequences were inserted into the *N*-terminal intein vector, pTYB12, by converting the *BamH* I site to *EcoR* I by PCR amplification with the appropriate primers (two bases changed).

4.3.1.1 PCR primers

The forward primer is 21 bases long and starts before the *Nde* I site (5'-CACAGCAGC GGCCATATCGAC-3') whereas the reverse primer is 29 bases (5'-TTCGGGCTTTGTTAG CAGCCGAATTCTTA-3') long. The *EcoR* I recognition site is underlined. Because of the mismatches, additional bases were added to ensure proper annealing. With both primers, the sequences had to be specific in order for them to properly anneal to the vector. The reverse primer had the additional requirement of mismatched bases to alter the enzyme recognition site.

4.3.1.2 PCR amplification

PZ 6-16 (GAGQGSA; 337 residues) was chosen as the gene to amplify since the protein expresses well in pET-19b (PZ8 had not been created yet). Initial amplification using *Taq* polymerase and the standard PCR protocol from Section 2.4.1 was unsuccessful. A control reaction using only primers and no DNA template indicated that primer amplification was occurring. A different enzyme, SureStart™ *Taq* from Stratagene (La Jolla, CA), resulted in successful PCR amplification. This enzyme is modified so that it will not activate until exposed to high temperatures for several minutes (*i.e.,* denaturing step). This prevents room-temperature amplification of a PCR mixture before it has been placed in the thermocycler. PerfectMatch™ (Stratagene) was also added to reduce side products caused by mispairings. The PCR protocol used consisted of 95°C for 12 minutes (to activate the enzyme) followed by 30 cycles of 1 minute at 95°C, 45 seconds at 63°C, and 1.5 minutes at 72°C followed by a final extension at 72°C for 10 minutes. The amplified product was digested with *Nde* I and *EcoR* I to generate "sticky ends." The DNA oligomer was then ligated into the pTYB12 recipient vector.

4.3.2 Expression and purification of PZ6-16 and MBP with the intein tag

Expression under standard conditions of 37°C and 1 mM IPTG induction for 3 hours (Section 2.5.1) resulted in insoluble protein. Note that in this system denaturants cannot be used in the buffer as that would prevent the fusion protein from binding to the chitin resin. Expression of the control plasmid, pMYB5, containing the maltose binding protein similarly resulted in insoluble protein. However, expression using one of the manufacturer's recommended conditions of 15°C and 0.5 mM IPTG overnight successfully yielded soluble protein for both PZ6-16 and the maltose binding protein.

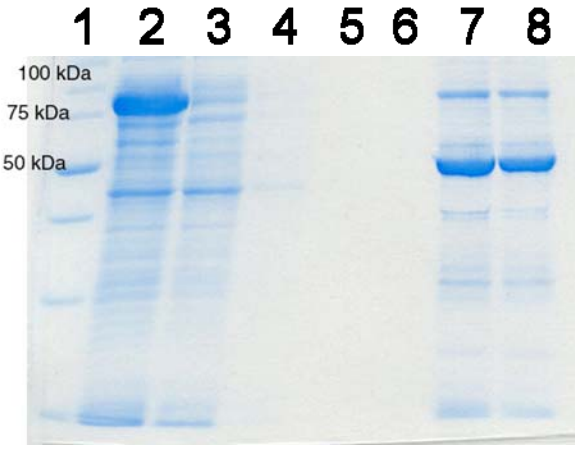


Figure 4-2: SDS-PAGE of PZ6-16 protein purification steps; lane 1: protein standards; lane 2: clarified lysate; lane 3: column flow through; lanes 4-5: buffer washes; lane 6: quick flush with 50 mM DTT; lanes 7-8: resin after 20 hours of cleavage

4.3.2.1 On-column tag cleavage of PZ6-16

A 500-mL batch of PZ 6-16 was expressed and purified on chitin resin using the lower induction temperature. Cleavage was performed according to the manufacturer's protocol and the results were analyzed on a 10% SDS-PAGE gel (Figure 4-2). After the resin was incubated for 20 hours at room temperature with DTT, it appeared that about 80% of the fusion protein was cleaved. After 40 hours, approximately 90% of the protein was cleaved on the resin. However, all that was obtained from the elutions was a tiny amount of brown powder that was too little to measure on a

balance, and was much less than what would be expected for a 500-mL expression of PZ6-16 (at least a few mg). MALDI-TOF analysis did not detect any protein in the sample.

4.3.2.2 Recovery of cleaved PZ6-16 protein

Since the protein was not in the elutions, it may have remained on the column, eluted out during the quick flush with DTT, or been lost during dialysis. MALDI-TOF analysis of the quick DTT flush and column elutions with detergent did not detect any protein; however, some cleaved protein appeared to be present in the sample taken from the buffer the column was stored in for several days while the elutions were being dialyzed. MALDI-TOF analysis identified a peak corresponding to 25.8 kDa. If this is indeed PZ 6-16, then the larger than expected mass size may be indicative of glutamine deamidation [154]. Elutions using sodium phosphate buffer followed by HEPES buffer containing 2 M urea were performed on the chitin column. If PZ 6-16 was still present on the column, urea would denature the protein so that it is soluble in solution and elute from the resin. Urea may also denature the bound intein tag and precursor protein which could then co-elute with the cleaved protein. After dialysis of the sodium phosphate/urea elutions using 7000 MWCO membrane followed by lyophilization. Seven milligrams of a white substance was obtained from the sodium phosphate elution while 15 mg of a dense, pale yellow substance was obtained from the urea elution. However, MALDI-TOF analysis did not detect the presence of proteins in either sample. Even though the PZ6-16 fusion protein bound to the column and the intein tag successfully cleaved, we were unable to recover the target PZ6-16 protein.

4.3.2.3 *In vivo cleavage of MBP protein*

In contrast to the PZ 6-16 results, intein purification of maltose binding protein did yield a cleaved protein sample. Protein expression was performed in BLR(DE3) with an overnight induction temperature of 15°C. The expressed protein band in the lysate lane (97 kDa) was less than expected for an overexpressed protein. Another band at approximately 43 kDa is in fact more intense than the band for the expressed full-length fusion protein. This 43 kDa band is also the same molecular weight of the cleaved protein present in the elutions. *In vivo* cleavage of the intein tag is a possibility, but lower induction temperatures are supposed to decrease the occurrence of *in vivo* cleavage. After dialysis followed by lyophilization, a small amount of yellow-brown material was obtained. This substance was tested by MALDI-TOF and was determined to have a molecular mass of 43.1 kDa. The correct size for MBP is 42 kDa.

In summary, the New England Biolabs (NEB) intein affinity tag required low temperature induction to be properly expressed in a soluble form. Both PZ6-16 and the control protein MBP were expressed in this system. However, we were not able to recover the cleaved PZ6-16 protein. Additionally, the majority of the MBP protein cleaved *in vivo* despite using the lowest recommended induction temperature and therefore most of the protein was not recovered.

4.4 Intein-based purification using the pMΔI†T-CM plasmid

Due to the continuing difficulties encountered with the NEB intein purification system, we investigated a similar purification system but one based on a smaller mini-intein (18 kDa) [163, 168]. The larger intein used in the IMPACT system can cause diminished solubility and purification efficiency [168] . The mini-intein has been inserted into the NEB pMAL-C2 vector

directly after the maltose binding domain, creating the pMΔI†T-CM plasmid. Immediately following the intein is the thymidylate synthase (TS) gene which was inserted into the multiple cloning site and can be replaced with the gene of choice. Intein tag cleavage occurs with a reduction in pH (8.5 to 6.0). This engineered plasmid was provided by Marlene Belfort and Vicky Derbyshire from the Wadsworth Center in New York.

4.4.1 Inserting the gene into the expression vector by PCR amplification

In order to insert our protein sequence in frame to the intein, the gene must be inserted using *BsrG* I at the 5' end and either *Xba* I, *Sal* I, *Pst* I, or *Hind* III at the 3' end. We chose to encode *Hind* III into the 3' primer with *Xba* I as an alternative sequence. The following two primers were used to generate the RZ2-5 insert from pET-19b (Chapter 3) with the modified *BsrG* I and *Hind* III ends: 5'-GTT GTT GTA CAC AAC ATG GGT TAT AGC AC-3'and 5'-ATC TAG AAG CTT CAG CCG GAT CCT TAA CC-3'. Enzyme recognition sites are underlined. Testing of RZ2-5 was being undertaken in both the pET-19b vector and the intein vector concurrently. The gene was amplified using an initial 5 minute denaturing step at 95°C, 30 cycles of amplification consisting of 1 minute denaturation at 95°C, 45 second annealing at 57°C, and 3 minute extension at 72°C were used, followed by a final extension step of 72°C for 10 minutes.

4.4.2 Small-scale test expression of RZ2-5

Test expressions at 37°C were analyzed by SDS-PAGE which showed an intense, induced protein band at approximately 55-60 kDa, of which 13% is the RZ2-5 protein by mass (remainder is intein tag plus maltose binding domain). The broad band makes an accurate

determination of protein size difficult. The expected size of the maltose binding protein (MBP) plus the intein tag is 60 kDa. The expected size of the entire fusion protein including RZ2-5 (9 kDa) is 69 kDa. At first, *in vivo* cleavage of the intein tag was suspected. Generally use of lower induction temperatures minimizes this effect which was not done for this initial expression test.

4.4.3 Induction condition test expressions

4.4.3.1 RZ2-5 gene

Several different induction conditions were tested (37°C for 3 hours, 15°C overnight, 20°C for 2, 4, and 6 hours) to determine optimal growth conditions. Results were analyzed on the Agilent 2100 Bioanalyzer using a Protein 200 kit (Figure 4-3A). Two hour induction at 20°C appeared to have the highest protein expression level (lane 10) but the expressed protein size (~ 60 kDa) was still lower than expected (69 kDa).

4.4.3.2 Thymidylate synthase (TS) gene

The thymidylate synthase (TS) control protein was expressed with a 20°C induction temperature for 2, 4.5, and 6 hours. Figure 4-3B is an Agilent 2100 Bioanalyzer pseudo-gel image comparing the TS test expression to the previous RZ2-5 test expressions. The size of the expressed TS fusion protein was 90 kDa, smaller than the expected mass of 97 kDa. If *in vivo* cleavage of the TS protein was occurring, there should be additional bands at 60 kDa (MBP-tag) and 37 kDa (TS) and not 90 kDa. These results indicated that the reduced size of the fusion protein was due to physical properties of the intein tag which affect its migration on gels and not premature cleavage of the tag. This observation was later confirmed after directly contacting the

creators of the pMΔI†T-CM plasmid.

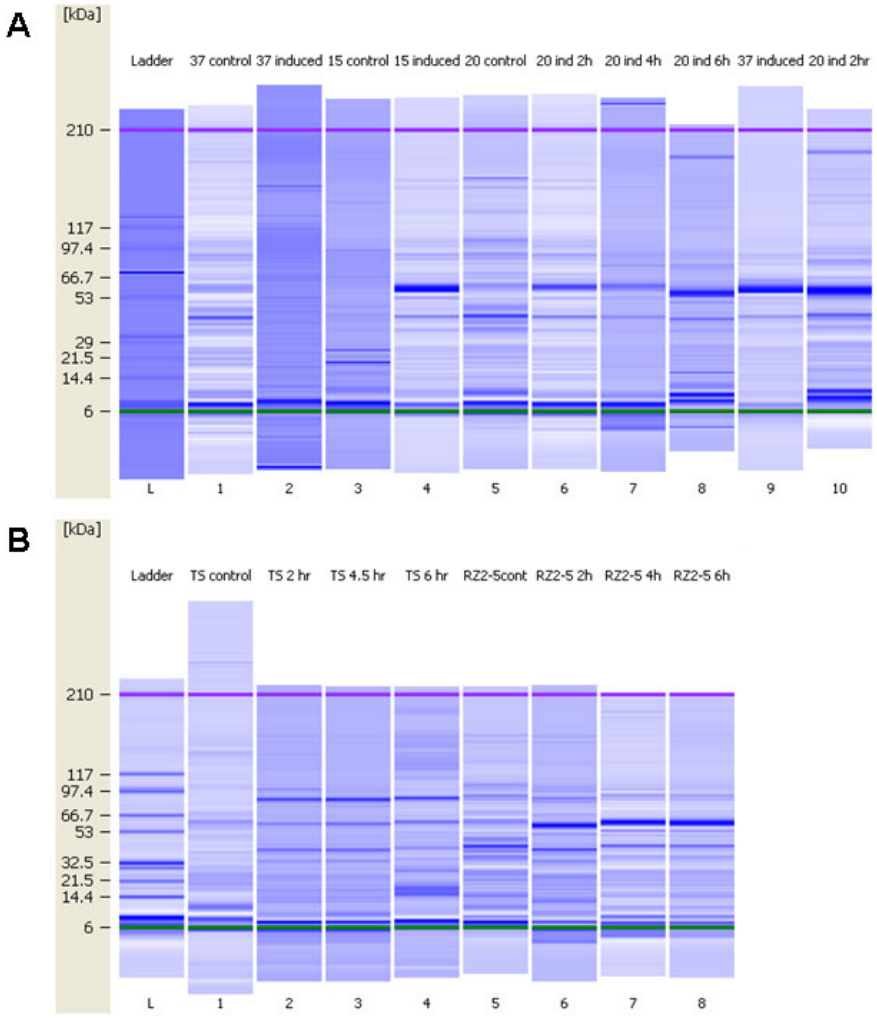


Figure 4-3: Test expressions on the Agilent 2100 Bioanalyzer. Green and pink bands are markers used to align sample runs with ladder lane for size determination. A) RZ2-5 at various times and temperatures. Lanes 9 and 10 are replicates B) TS and RZ2-5 test expressions at 20°C

4.4.4 Large-scale protein expressions

4.4.4.1 RZ2-5 gene expression and on-column cleavage

The recommended buffers and protocols for the pMΔI†T-CM vector were used for the on-column cleavage of a 500 mL RZ2-5 expression. The clarified cell lysate was bound to an amylose column for an hour followed by multiple washes of pH 8.5 buffer. Cleavage was initiated with a pH 6.0 buffer wash. After ~ 48 hours of incubation at 4°C, the protein was eluted off the column with additional pH 6.0 buffer. A final pH 8.5 wash was done with 10 mM maltose to release the bound tag from the resin. The fractions were tested on the Agilent Bioanalyzer system. An analysis of the elutions using the Protein 50 c hip (5-50 kDa size range) did not reveal any 9 kDa protein. The samples were retested with a larger size range kit as shown in Figure 4-4. The fusion protein did not bind to the amylose resin. Instead it was detected in the column flow through and subsequent washes. Maltose was mixed with the lysate sample to verify that maltose binding to the fusion protein did not affect mobility of the fusion protein (lane 3).

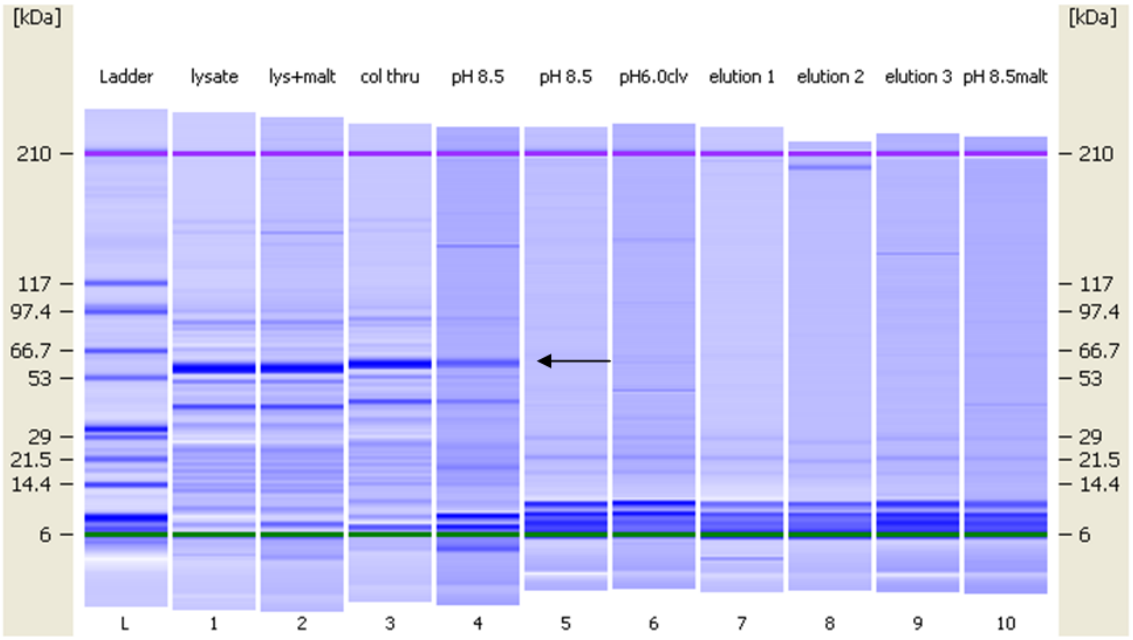


Figure 4-4: Pseudo-gel image of RZ2-5 purification and on-column cleavage as determined on Agilent 2100 Bioanalyzer. Arrow indicates product band.

The binding between maltose binding protein and amylose resin is not always guaranteed. Amylase, an enzyme encoded in the *E. coli* chromosome, can interfere with column binding and degrade the column according to NEB protocols on the pMAL plasmid. For this reason, LB media used for expression had been supplemented with 1% glucose to suppress amylase production and supposedly prevent this problem.

Another possibility is that the strain we typically use for general protein expression, BLR(DE3) was somehow incompatible with this system. The only strain specifically mentioned by name in the protocol accompanying the plasmid was ER2566, a strain of *E.coli* that was included with the NEB IMPACT kit. Test expressions of RZ2-5 and TS using the ER2566 cell strain were identical to the previous BLR(DE3) test expressions when visualized by SDS-PAGE.

Nevertheless, there may be subtle differences between the two cell strains that make ER2566 more suitable for intein expressions.

4.4.4.2 TS gene

4.4.4.2.1 TS expression and cleavage using ER2566 cell strain

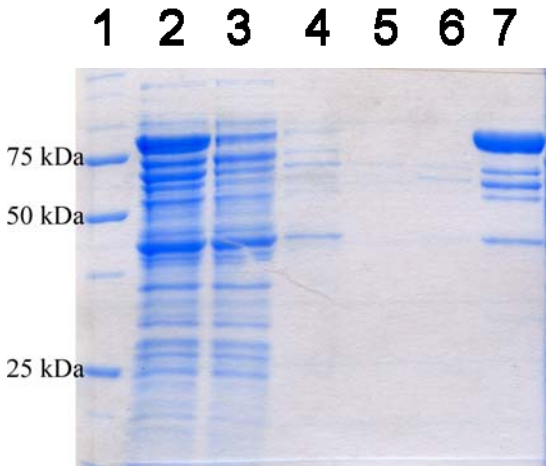


Figure 4-5: SDS-PAGE of thymidylate synthase (TS) purification steps; lane 1: protein standards; lane 2: clarified lysate; lane 3: column flow through; lanes 4-5: buffer washes; lane 6: elution; lane 7: maltose wash

To verify that this intein expression system could work properly in our hands, the original vector containing the TS protein was expressed in 1L of ER2566 cells using LB media supplemented with 1% glucose. The induction conditions were 20°C for 3 hours. The binding, washing, cleavage, and elution steps were all done at 4°C. After washing with 12 column volumes of pH 8.5 buffer and 2 column volumes of pH 6.0 buffer, the column was allowed to sit for ~ 36 hours at 4°C. After protein elution, the column was washed with pH 8.5 buffer containing 10 mM of maltose to release the bound MBP-intein tag from the resin.

The clarified lysate, column flow through, washes, and elutions were all analyzed on a 10% SDS-PAGE gel (Figure 4-5). The gel showed the correct size of the expressed fusion protein which bound completely to the column in contrast to the earlier RZ2-5 results. The presence of a thin band at ~ 40 kDa indicates some *in vivo* cleavage of the tag occurred during

expression. Unfortunately, no pH-induced cleavage was observed. There was no TS protein in the elution sample and the protein that eluted off the column in the maltose wash was the full length, uncleaved fusion protein. A few faint bands in the elution lane indicate nonspecific binding of impurities that were not completely washed off the column prior to cleavage. Cleavage conditions of 24 hours at room temperature and 96 hours at 4°C were also attempted without success.

4.4.4.2.2 TS expression with no glucose media supplement

We conferred with the research group at the Wadsworth Center in New York, who sent us the intein vector, but they did not encounter problems similar to that described above. Another 500 mL expression of the TS protein was performed using ER2566 cells in rich media (2% tryptone, 1% yeast extract, 1% NaCl w/v) with no glucose supplement. This media formulation has been used before for expression of the TS-intein vector [168]. No particular media formulation was emphasized in the original vector instructions beyond the suggestion to follow the pMAL protocols provided by NEB.

Only a single band was obtained in the elution fractions as visualized by SDS-PAGE; however, the size of this band likely corresponds to the cleaved binding domain (60 kDa). In addition, it did not appear that all of the expressed fusion protein bound to the resin. The final 10 mM maltose wash to elute any protein bound to the resin yielded three bands: the uncleaved precursor protein, the cleaved binding domain, and a very faint band that could correspond to the cleaved TS protein.

4.5 Conclusions and Recommendations

At this point, work on the intein systems was suspended due to the many problems that were encountered (*i.e.,* protein not binding and/or cleaving and issues with solubility and *in vivo* cleavage). While there was no difficulty in expressing the fusion protein, recovering quantifiable amounts of the cleaved target protein proved to be challenging. Perhaps with further work these intein systems can eventually be used successfully for protein polymer purification.

In the meantime, research on inteins for protein purification has continued to progress, leading to newer strategies such as those discussed briefly in Section 2.8.4 and elsewhere [143]. Self-cleaving elastin tags allow for purification with no chromatography resin or protease involvement [145-148]. Polyhydroxybuterate (PHB) granules are produced intracellularly and bind to phasin-tagged fusion proteins. These granules are then recovered and the target protein is released by a self-cleaving intein [149].

More recent work on protein polymer drag-tags has shown benefits from using a *C*-terminal affinity tag instead of the *N*-terminal affinity tag (Chapter 6). Only the *N*-terminal tag was tested for both of the intein systems discussed above. The NEB IMPACT kit also contains vectors for a *C*-terminal affinity tag. Its suitability for expressing protein polymer drag-tags will be evaluated by another graduate student, Xiaoxiao Wang. More testing should be done with the ER2566 strain at low induction temperatures and low IPTG concentrations (< 1 mM). The chitin binding domain and/or the maltose binding domain used in these two intein purification systems for resin binding could be replaced by a different binding domain such as the more familiar histidine tag [169].

Chapter Five

Results of Electrophoretic Analysis of Protein-DNA Bioconjugates and the Refinement of Drag-tag Purification Techniques

5.1 Introduction

After designing, cloning, expressing and purifying the protein polymers they must finally be evaluated to determine their suitability as drag-tags for ELFSE. Protein polymers are chemically conjugated to fluorescently labeled ssDNA primers and analyzed in free solution by capillary electrophoresis to assess their hydrodynamic drag as well as monodispersity. This conjugation step must produce a unique and stable link between a single DNA molecule and one drag-tag to ensure there are no ambiguous results. The development of this conjugation strategy and the subsequent electrophoretic analyses using ELFSE were performed by Dr. Robert Meagher and now by another graduate student in our lab, Jennifer Coyne.

Obtaining a protein polymer that is completely monodisperse is a difficult task and proved more challenging as we began creating and evaluating longer-length protein polymers. While not completely monodisperse, a 127-amino acid protein polymer was obtained that was pure enough to be used for DNA sequencing in free solution by ELFSE [123]. This result proved that protein polymers could be used as drag-tags while also motivating us to seek drag-tags with greater hydrodynamic drag in order to improve the read length (~ 180 bases for the 127-amino acid drag-tag with an α of 25).

5.2 Experimental protocols for ELFSE analysis

5.2.1 Bioconjugation of DNA oligomers to protein polymers

Of the naturally occurring amino acids, primary amines, thiols, and carboxylic acids are potential reactive groups that are likely to be present in a polypeptide . However, carboxylic acid coupling was avoided due to its relatively low reactivity in aqueous solutions [170]. Additionally, the water-soluble drag-tags and DNA would be sparingly soluble in organic solvents that are favorable for carboxylic acid coupling. Not only have cysteines (thiol) been avoided in our protein polymer sequences, but also all protein polymers already include a primary amine at the N-terminus. Therefore the thiol functionality has been introduced to the DNA oligonucleotide, an easily obtainable modification that can be requested upon ordering the DNA primer from a commercial vendor. The general approach that was taken is to activate the amino terminus of the drag-tag with a thiol-reactive group, maleimide, which can then react specifically and quantitatively with thiolated DNA [170].

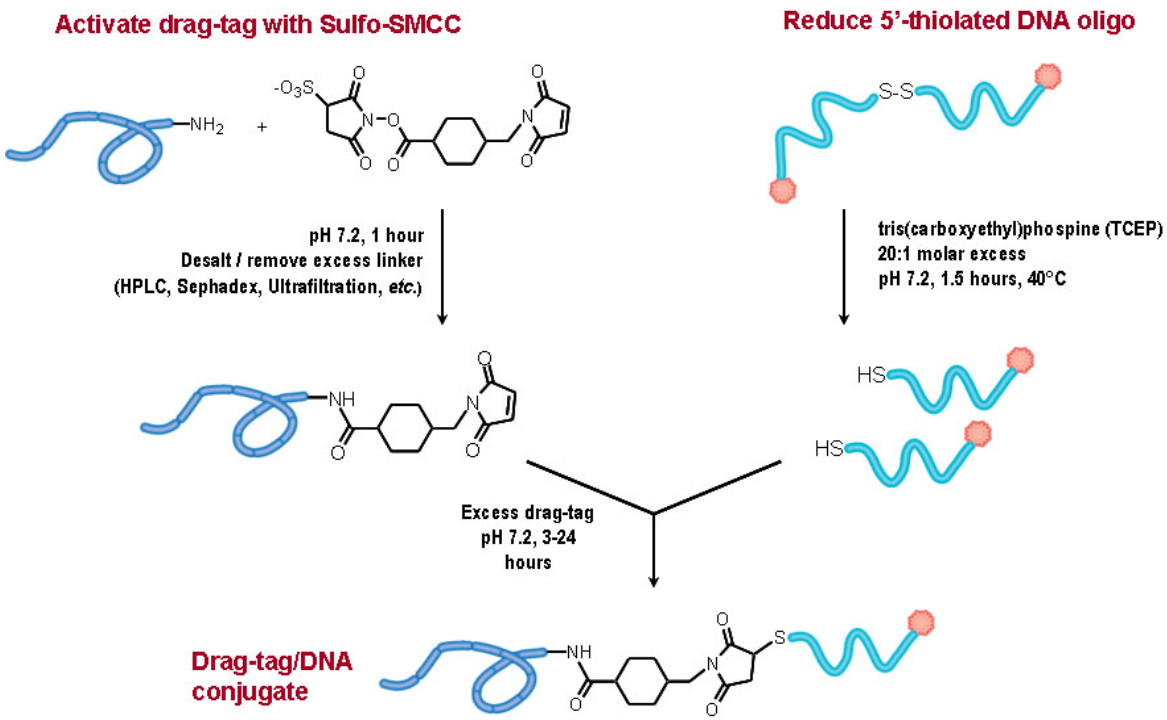


Figure 5-1: Conjugation strategy for attaching drag-tag to thiolated DNA primers. Note that for simple analysis of protein monodispersity, thiolated primers are purchased with an internal fluorescein label. For sequencing, the thiolated primers are not fluorescently labeled since the ddNTPs are already labeled with one of four fluorescent dyes.

5.2.1.1 Current conjugation protocol

Figure 5-1 illustrates the general bioconjugation strategy currently used to attach protein polymer drag-tags to DNA primers. First, oligonucleotides containing a thiol (-SH) functionality on the 5' terminus were purchased from IDT (Coralville, IA). For best conjugation results, the DNA must be reduced to prevent DNA-DNA disulfide bond formation. To reduce the DNA, 2 nmol of DNA primer are incubated with a 20:1 molar excess of Tris(2-carboxyethyl)phosphine (TCEP, Acros Organics, Morris Plains, NJ) at 40°C for 90 minutes in 20 μL of 70 mM sodium phosphate buffer, pH 7.2 [123]. Protein polymer drag-tags are activated at the N-terminus with a

maleimide by the addition of the heterobifunctional crosslinker sulfosuccinimidyl 4(*N*-maleimidomethyl)cyclohexane-1-carboxylate (sulfo-SMCC, Pierce Biotechnology, Rockford, IL). Sulfo-SMCC contains an amine-reactive *N*-hydroxysuccinimide (NHS ester) and a sulfhydryl-reactive maleimide group. A 10:1 molar excess of sulfo-SMCC is added to 1.2 mg of protein polymer in 80 μL of 100 mM sodium phosphate buffer, pH 7.2, and the mixture is vortexed for 1 hour at room temperature. Excess sulfo-SMCC is separated from the activated protein polymer drag-tag by gel filtration with a Centri-Sep column (Princeton Separations, Adelphia, NJ). The activated, purified protein polymer is frozen, lyophilized and then resuspended in water at 10 mg/mL concentration [123].

To conjugate the activated drag-tag to the reduced DNA, 90 pmol of DNA is mixed with 2.5 nmol of drag-tag to a final volume and concentration of 10 μL in 25 mM sodium phosphate buffer at pH 7.2. The mixture is then incubated at room temperature for 3-24 hours. A large excess of drag-tag to DNA (typically 100-fold) is necessary to ensure nearly complete (> 95%) conjugation of drag-tags to each DNA molecule [63, 123, 170].

5.2.1.2 Alternative bioconjugation strategies

One drawback to using maleimide/thiol chemistry for the conjugation is the potential for hydrolysis of the maleimide through a ring-opening reaction after thiol coupling [170, 171]. This creates an additional negative charge and would influence the free-solution mobility of the drag-tag-DNA conjugate. So far decomposition has not been a significant problem for sulfo-SMCC-activated drag-tags, likely due to enhanced stability of the maleimide ring in sulfo-SMCC [170, 171]. However, this behavior can become problematic in the future if extended thermal cycling

protocols are required for sequencing. Use of crosslinkers without maleimide was studied first by Dr. Robert Meagher (without success) and continues to be explored by graduate student Jennifer Coyne. "Click" chemistry is an alternative strategy under investigation. This method involves reacting alkyne groups with azides in a high-yielding linkage reaction favorable in water [172, 173]. One reactive group would be introduced into the protein and the other to the DNA oligomer.

An enzymatic linking strategy is also actively being researched. This technique was first investigated by Masters in Biotechnology student, Louisa Carr, and continues to be researched by Ph.D. candidate, Jennifer Coyne. This strategy is particularly appealing if ELFSE is to be eventually commercialized and reagents are to be sold in an easy-to-use kit. DNA ligase can be used to link, via a DNA "splint", a short oligonucleotide (already attached to drag-tag) to the sequencing primer. Preliminary proof-of-concept experiments demonstrate that this strategy has potential and the research is ongoing.

5.2.2 Electrophoresis of protein-DNA conjugates

Except where stated, all analyses were performed on an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) with a 16- capillary array of fused silica capillaries (50 μM inner diameter) and 4-color laser-induced fluorescence (LIF) detection using a 488 nm laser. Capillary electrophoresis separations of the conjugates were done in denaturing buffer consisting of 89 mM Tris(hydroxymethyl)aminomethane (Tris), 89 mM Tris(hydroxymethyl) methylaminopropanesulfonic acid (TAPS), 2 mM ethylenediaminetetraacetic acid

(EDTA), and 7 M urea. A 0.5–3% (v/v) POP-5 or POP-6 polymer solution (Applied Biosystems) was used for a dynamic wall coating agent to suppress electroosmotic flow and prevent adsorption to capillary walls. Capillaries with an effective length from inlet to detector of 36 cm were used for ELFSE separations. Typical electrophoresis conditions include electrokinetic injection with a potential of 1-2 kV applied for 5-30 seconds and running voltage of 15 kV, all at 55°C [63, 170, 174].

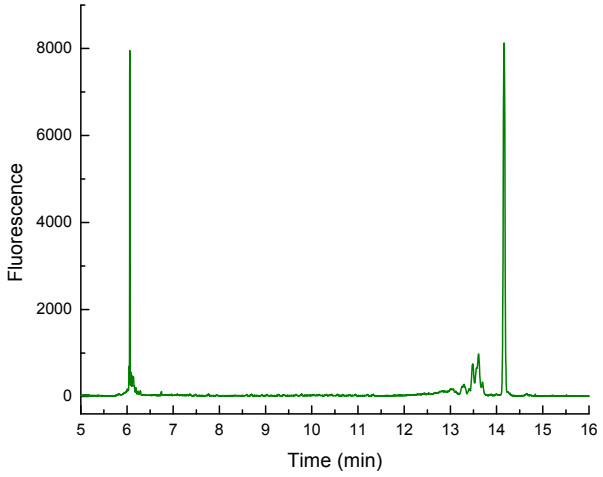


Figure 5-2: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for the PZm8-6 protein (127 amino acids, $\alpha \sim 25$) using 18-base primer. BioRad (Hercules, CA) BioFocus 3000, 44 cm capillary with 25mM ID, 1X TTE, 7M urea, 3%v/v POP5, 20 psi*s inject, 400 V/cm, 50°C

5.3 PZ8 series of protein polymers

The PZ8 protein polymer sequence (GAGTGSA) yielded a 127-amino acid drag-tag (6mer) that was nearly monodisperse (Figure 5-2) in contrast to earlier designs [120-122]. The peak on the far left of the electropherogram corresponds to the free (untagged) DNA whereas the

larger peak on the far right corresponds to the drag-tag-DNA conjugate, which eluted later due to attachment of the drag-tag. The smaller peaks adjacent to the major bioconjugate peak indicate low levels of impurities are present. The hydrodynamic drag, α, can be determined by rearranging Equation 1-3 into the form shown below:

$$\alpha = N\left(\frac{\mu_0}{\mu} - 1\right) \qquad (5\text{–}1)$$

where N is the # bases of the DNA primer, μ_0 and μ are the electrophoretic mobilities for the free DNA and the bioconjugate, respectively. Equation 5-1 can be further simplified into Equation 5-2 as the molecules are migrating the same distance (length of microchannel) under the same electric field strength. Therefore only the migration time of the two peaks and the DNA size is required.

$$\alpha = N\left(\frac{t}{t_0} - 1\right) \qquad (5\text{–}2)$$

A potentially more accurate measurement of α can be obtained from a mixture of different sizes of DNA (*e.g.,* sequencing electropherogram) by plotting the quantity (μ_0/μ - 1) versus 1/N where the slope of the resulting linear plot is equal to α. The calculated α value of the protein is 25. It was later determined that the actual drag-tag sequence was a variant (renamed PZm8) containing two Ser to Arg mutations (from *E. coli*). Nevertheless, this positively charged protein polymer did not demonstrate any of the detrimental effects expected to be caused by its two positive charges. Figure 5-3 compares the original PZ8 sequence to the PZm8 sequence where every 1 in 9 Ser is mutated into Arg. Specifically, the 5th and 13th Ser codons (out of 18) had a

single base mutation at the 3rd position of the codon, resulting in an AGG arginine codon in both cases. Not only did the successful sequencing prove ELFSE was possible with a protein polymer drag-tag, but also that a limited amount of positive charges could be beneficial for future sequence designs. The positive charges "pull" the protein in the opposite direction of the negatively charged DNA in an electric field, thereby increasing the hydrodynamic drag or α value of the protein. As a result, the arginine-containing PZm8 series has been the focus of extensive research due to their higher α values relative to similarly sized PZ8 proteins.

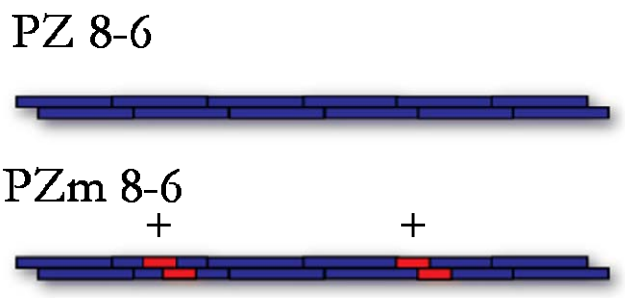


Figure 5-3: Illustration comparing uncharged PZ8 to mutated PZm8 sequence for a 6mer size

5.3.1 DNA sequencing using PZm8-6 as a drag-tag

A SNaPshot Multiplex Kit (Applied Biosystems), normally used for single-base extension (SBE) reactions, was used for the DNA sequencing reaction with the addition of dNTPs [123]. The SNaPshot kit premix includes a sequencing polymerase, reaction buffer and dichlororhodamine (dRhodamine) dye-labeled ddNTPs. For ELFSE sequencing, the following reaction mixture was used: 5 μL of SNaPshot premix, 8 nmol of dNTPs, 4.2 pmol of M13 primer (5'-$\mathbf{X_1}$GTTTTCCCA-GTCACGAC, where $\mathbf{X_1}$ is a 5'-C6 thiol linker) conjugated to a drag-tag, 0.16 μg of M13mp18 control DNA template (GE Healthcare, Piscataway, NJ), and sufficient water for a total volume of 10 μL. The mixture is then thermal cycled for 26 cycles of 96°C for 5 seconds (denaturation), 50°C for 5 seconds (annealing), and 60°C for 30 seconds (extension).

For purification, Centri-Sep gel filtration columns are used to remove excess dye terminators, buffer salts and dNTPs prior to electrophoresis [123]. Prior to analysis by electrophoresis, samples are denatured at 95°C for 30-60 seconds and snap-cooled on ice for 2-5 minutes.

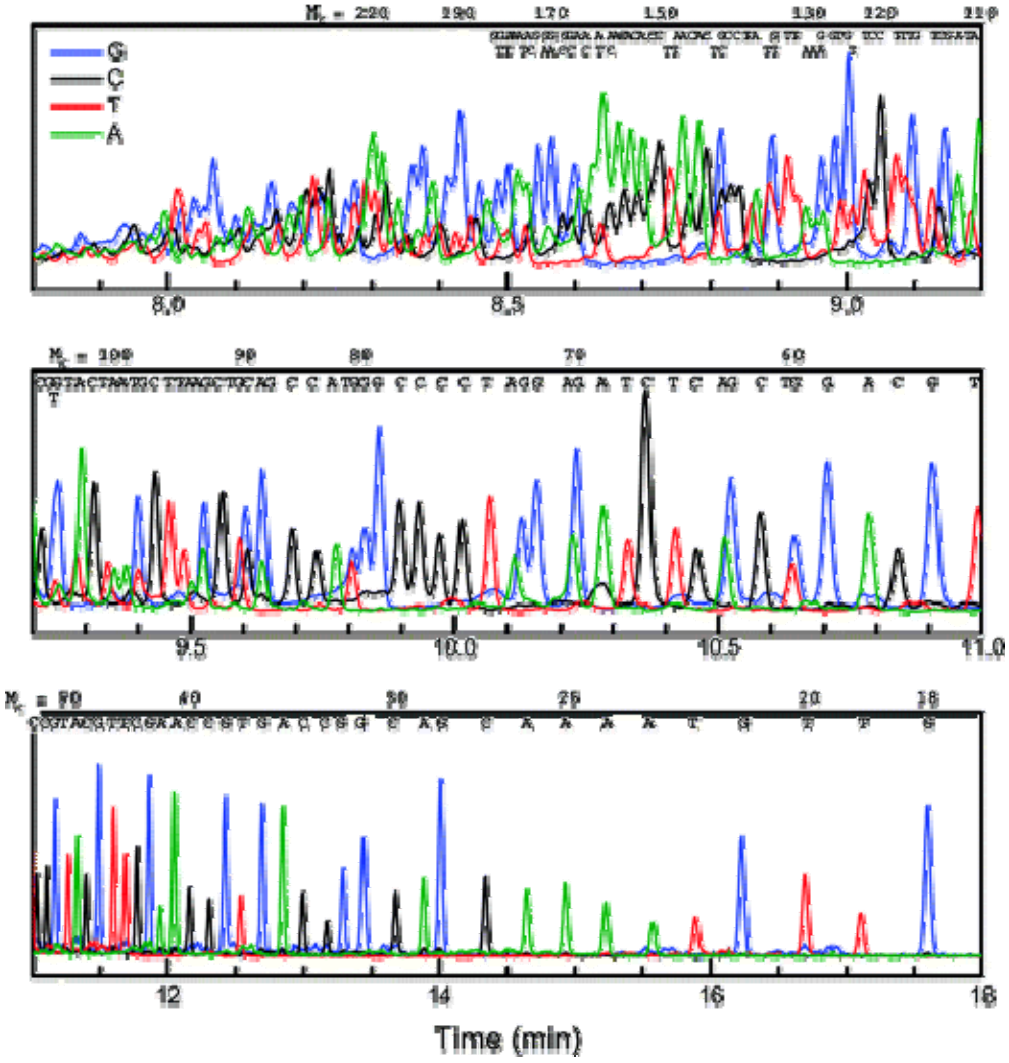


Figure 5-4: Four-color electropherogram for the products of a sequencing reaction carried out using the 127mer protein polymer drag-tag-labeled M13 primer and the SNaPshot SBE kit with dNTPs at a total concentration of 800 μM in the sequencing reaction. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP5, 1kV/20s injection, 312 V/cm, 55°C. Reproduced with permission from *Analytical Chemistry* 2008, *80*, 2842-2848. Copyright 2008, American Chemical Society.

The previously described protocol for electrophoretic analysis of protein-DNA conjugates in free solution was similarly used for analyzing the sequencing products. Figure 5-4 is a sequencing electropherogram using a 36 cm capillary and a field strength of 312 V/cm. A total concentration of 800 μM of dNTP generated sequencing products up to ~ 250 bases long [123]. Note that except for the correction for spectral overlap of the dyes (done automatically by the instrument), this electropherogram shows raw, unmodified data. Commonly performed techniques on matrix-based sequencing, such as peak height normalization and mobility shift correction, have not been applied. In addition, since large DNA fragments elute earlier, the sequence is read backwards from conventional sequencing from the lower right (smaller fragments) to the upper left. By comparing the experimentally observed sequence with the known M13mp18 sequence, we find that ~ 180 bp can be read successfully. With more advanced processing, a completely unknown template could be analyzed.

Cleaner, sharper peaks and significantly better resolution were some of the advances made over the earlier streptavidin sequencing result [55]. We were also able to conjugate our protein polymer to the primer and then take it through the thermal cycling reaction with no apparent negative effect. This was not possible for ELFSE sequencing with streptavidin which relied on streptavidin-biotin binding that had to be performed after the thermal cycling reaction. Our non-natural protein polymer (9 kDa) also had a similar effective drag to streptavidin (53 kDa) while being significantly smaller in size. In the process of performing large-scale expressions of the PZm8-6 protein, it was found that the protein could not be detected by the

Agilent 2100 Bioanalyzer. Consequently, future analyses of PZ8 and its variants were all performed on SDS-PAGE gels with Coomassie staining.

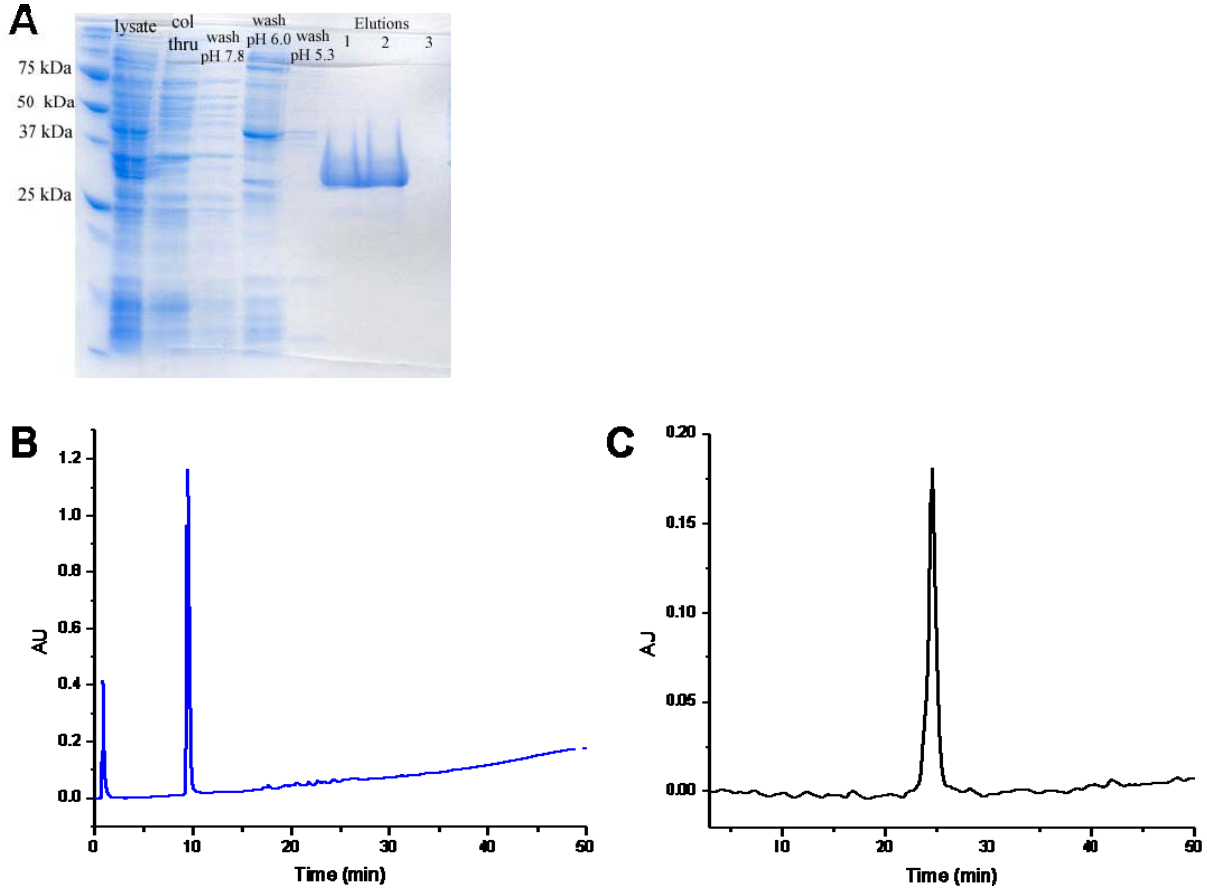


Figure 5-5: Results of PZm8-12 and PZm8-24 expression and purification A) 12% SDS-PAGE gel of PZm8-12 affinity chromatography fractions; B) RP-HPLC of PZm8-12 protein on C4 column 0-95% ACN gradient; C) RP-HPLC of PZm8-24 protein on C4 column 10-30% ACN gradient

5.3.2 Polydispersity in longer length protein polymers

With the success of the PZm8-6 drag-tag for DNA sequencing, we began doubling the gene, using the controlled cloning method, in order to make longer drag-tags with greater

hydrodynamic drag for improved sequencing read lengths. This led to the discovery that longer lengths of protein polymers (beyond 127 amino acids) all contained additional peaks of unknown origin when analyzed by ELFSE that increased in number with higher molecular weights.

5.3.2.1 PZm8-12 and PZm8-24

5.3.2.1.1 Expression results

PZm8-12 (253 amino acids) and PZm8-24 (505 amino acids) were generated via controlled cloning and inserted into the MpET-19b expression plasmid. SDS-PAGE of test expressions did not show a clearly distinguishable induced band but dot blot analysis confirmed that both proteins were expressed. These results were shown previously in Figure 2-7. Large-scale expressions of the proteins were purified by affinity chromatography (~ 20 mg/L yields) using a nickel resin (PZm8-24 had to be purified twice to remove impurities) and the His tag was removed from purified proteins by cyanogen bromide (CNBr). Figure 5-5A is an SDS-PAGE gel of the PZm8-12 purification. The purified proteins were identified as a single peak by RP-HPLC (Figure 5-5BC); however, the chromatogram for PZm8-12 showed an additional peak eluting at the start of the run of unknown origin. Both proteins were reacted with cyanogen bromide to remove the *N*-terminal histidine tag. MALDI-TOF confirmed the molecular mass of the proteins although the masses were slightly higher than expected. 18.6 kDa instead of the expected 18.4 kDa was observed for the 12mer while PZm8-24 was observed at 37.1 kDa instead

of the predicted 36.7 kDa.

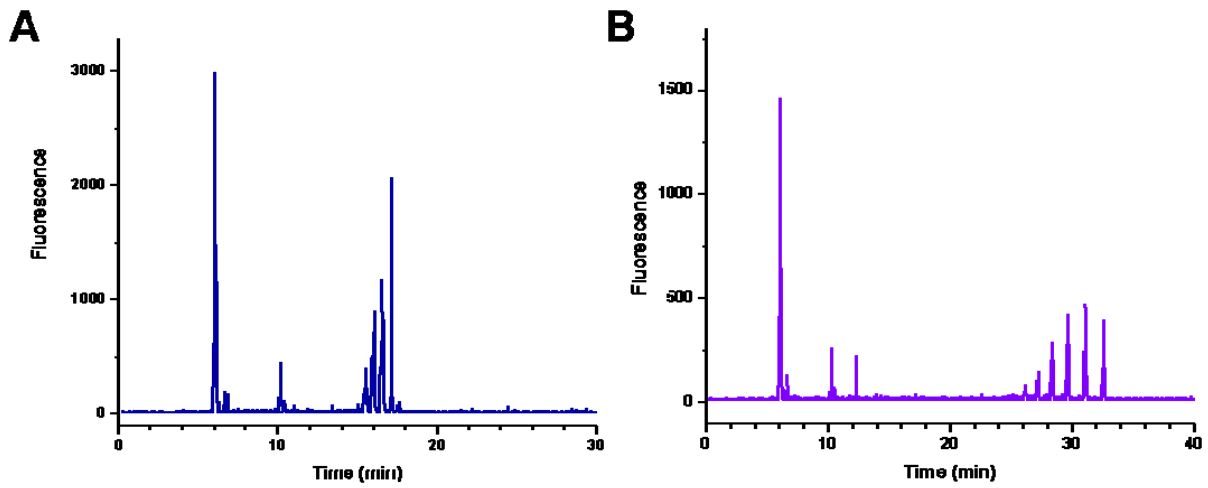


Figure 5-6: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for A) PZm8-12 (253 amino acids, α ~ 55 last peak) and B) PZm8-24 (505 amino acids, α ~ 130 last peak) using a 30-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 3%v/v POP5, 1kV/1s injection, 320 V/cm, 55°C

5.3.2.1.2 Analysis of the purified proteins

Both protein polymers were conjugated to DNA and analyzed in free solution. In contrast to the 6mer results, both proteins exhibited multiple, sharp peaks at regular intervals, meaning the proteins were not as monodisperse as originally thought (Figure 5-6). Mass spectrometry was done on both proteins by a different facility, the Biotechnology Center at University of Illinois at Urbana-Champaign, but results were the same as before (*i.e.,* single protein peak with slightly higher than expected mass). This confirmed that miscalibration or user error did not affect the previous MALDI-TOF results. Amino acid analysis was done on the 6mer, 12mer, and 24mer proteins by the Yale University W.M. Keck Facility (Section 2.6.5). The results are presented in Table 5-1 and confirm that these proteins are in fact our expressed

protein polymers and not another protein. Several unexpected amino acids are present at a very low percent, indicating the samples may contain a slight amount of impurities but not enough to generate the significant peaks seen by ELFSE.

Table 5-1: Amino acid analysis of PZm8-6, PZm8-12, and PZm8-24 proteins

Protein	**PZm 8-6**		**PZm 8-12**		**PZm 8-24**	
Amino acid (mol%)	expected	actual	expected	actual	expected	actual
Asx	0	0.2	0	0.3	0	1.1
Thr	14.2	13.3	14.2	13.2	14.3	12.5
Ser	12.6	11.2	12.6	10.9	12.7	10.8
Glx	0	0.8	0	1.5	0	1.4
Pro	0	0	0	0.3	0	0.4
Gly	43.3	43.3	43.1	42.1	43.0	40.3
Ala	28.3	29.3	28.5	29.2	28.5	27.7
Val	0	0.1	0	0.4	0	0.6
Met	0	0	0	0	0	0.2
Ile	0	0	0	0	0	0.2
Leu	0	0.1	0	0.2	0	0.2
Lys	0	0	0	0.1	0	0.3
His	0	0	0	0	0	2.3
Arg	1.6	1.7	1.6	1.9	1.6	1.9

Free-solution DNA sequencing was attempted using the PZm8-12 and PZm8-24 proteins as drag-tags despite their polydispersity. The Sanger cycle sequencing step was successful even with the larger drag-tags attached to the M13 primer as multiple fragment peaks were detected. Not surprisingly though, a sequence could not be read from the resulting electropherograms due to the many additional, overlapping peaks caused by the polydisperse drag-tags (results not shown).

Every analytical technique *except* for ELFSE seemed to indicate the proteins were pure, likely making it a challenge to separate and purify individual peaks seen in ELFSE by other available preparative techniques (*e.g.,* RP-HPLC). Dr. Robert Meagher attempted to purify the PZm8-12 protein by RP-HPLC on a C18 column but was unsuccessful (results not shown). Each fraction yielded the same profile when later analyzed by ELFSE. Therefore, if the cause of the polydispersity could be determined instead and easily prevented, then further purification steps would not be necessary.

5.3.2.2 PZ8-6, PZ8-12, PZ8-16, PZ8-21, PZ8-24 (no arginines)

This polydispersity cannot be the attributed to the previously reported glutamine deamidation [120]. The PZ8 sequence was specifically designed to exclude that amino acid. Nevertheless, this variant of the PZ8 gene contained arginines that were not part of the original sequence design. In fact, both of *E. coli* mutations in the PZm8-6 sequence resulted in the serines being converted to arginines using the least frequent (AGG) of six possible Arg codons.

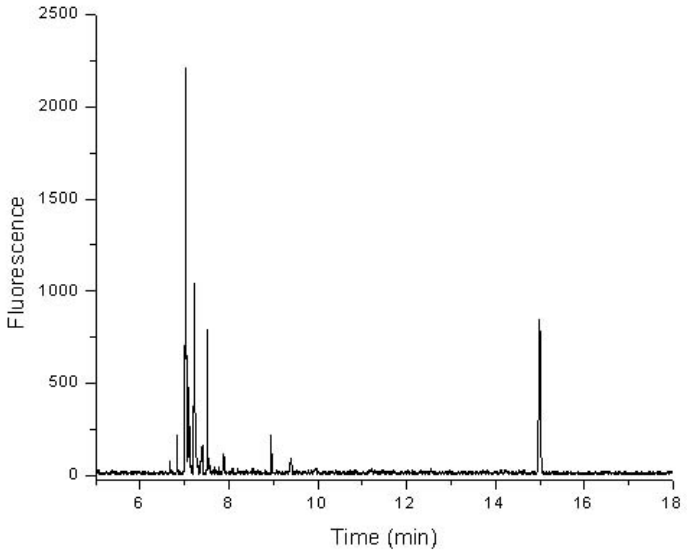


Figure 5-7: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for PZ8-6 (127 amino acids, $\alpha \sim 19$) using a 17-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 3%v/v POP5, 1kV/1s injection, 320 V/cm, 55°C

Protein polymers of various lengths were generated from an existing PZ8 trimer gene that did not contain the arginine mutations (confirmed by DNA sequencing). The PZ8-6 protein, the shortest length expressed and the same length as PZm8-6, was determined to be completely monodisperse by ELFSE (Figure 5-7).

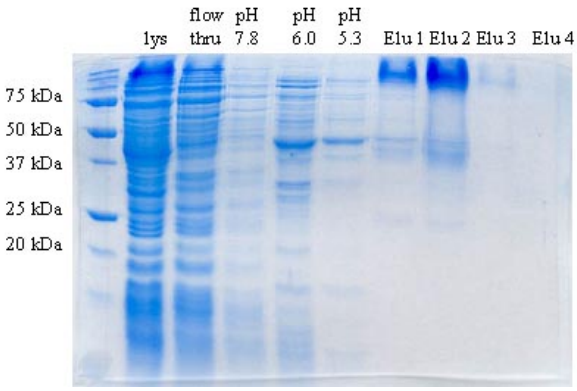


Figure 5-8: 12% SDS-PAGE gel of PZ8-12 purification

Longer length PZ8 proteins were generated by controlled cloning (PZ8-12 and PZ8-24). PZ8-16 and PZ8-21 were obtained by screening colonies after inserting the PZ8-12 gene into the MpET-

19b expression vector. All proteins were purified and the His tag was removed by cyanogen bromide cleavage. SDS-PAGE gels of the completely uncharged PZ8 sequences all showed abnormal migration patterns. Figure 5-8 is an SDS-PAGE gel of the first PZ8-12 purification. A second nickel column purification of PZ8-12 and PZ8-16 (using imidazole elution and not pH shift) yielded similar results. There appear to be multiple bands in the elution lanes up to an apparent molecular weight of 100 kDa. Most likely SDS is not associating with the uncharged protein polymers, resulting in little protein migration in the electric field. Additional bands may be due to the association of various amounts of SDS to a small fraction of the uncharged proteins.

All purified PZ8 proteins were of the correct size when analyzed by mass spectrometry and appeared to be a single peak. However, when the proteins were conjugated to DNA and analyzed in free solution, they still exhibited the same multiple peak pattern shown by the arginine-containing PZm8 sequences. Figure 5-9 compares the MALDI-TOF spectrum to the ELFSE electropherogram for the PZ8-16 protein. These results confirmed that it was not arginine that was causing the polydispersity but another not yet identified issue.

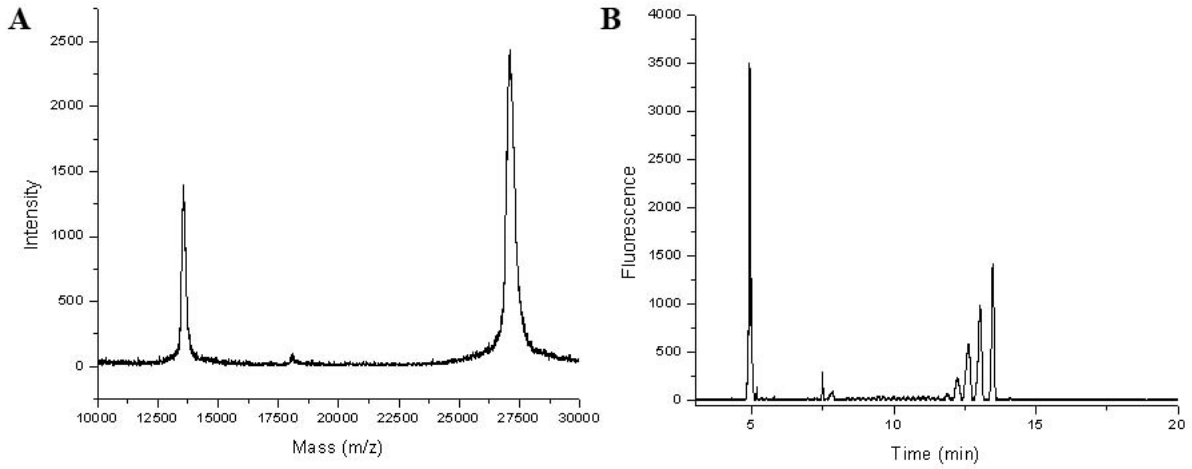


Figure 5-9: Comparison of PZ8-16 mass spectrometry and ELFSE results A) MALDI-TOF of uncleaved PZ8-16 protein (expected mass 27.06 kDa, actual 27.09 kDa) B) Free-solution capillary electrophoresis of drag-tag-DNA conjugates for PZ8-16 (337 amino acids, $\alpha \sim 52$ last peak) using a 30-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 1%v/v POP6, 1kV/2s injection, 320 V/cm, 55°C

5.3.2.3 PZ8-9, a length between 6mer and 12mer

So far no protein polymer lengths between the 6mer (monodisperse) and 12mer (polydisperse) sizes had been tested. Therefore PZ8-9 was constructed via controlled cloning from 6mer and trimer genes. However, PZ8-9 was also polydisperse, although slightly less so than the 12mer.

5.4 Alternative purification protocols

Since the arginines in the sequence were not the source of the polydispersity, we focused our efforts on the protein purification protocols. Each step in the purification process was examined to determine if improvements could be made to obtain higher purity protein polymers.

5.4.1 Affinity chromatography: resin and buffer conditions

Up to this point, all proteins had been purified using denaturing buffer conditions with nickel resin for the affinity chromatography step. However, an alternative resin based on cobalt, called Talon resin (Clontech, Mountain View, CA) is also available for purifying proteins with histidine tags. The manufacturer claims this resin is superior to nickel-based resins. There is no metal ion leakage even in the presence of strong denaturants and the resin possesses enhanced specificity for polyhistidine-tagged proteins over native proteins [138]. Elution at a higher pH or with a lower imidazole concentration is also possible compared to nickel resins.

A 2 L-expression of PZ8-21 was divided into four samples to test different purification conditions: nickel resin with native and denaturing conditions and cobalt resin with native and denaturing conditions. Comparison of the SDS-PAGE gels showed that Talon resin had fewer impurities in the elutions compared to ProBond nickel resin. Despite this improvement, the Talon-purified protein polymers had the same amount of polydispersity as the nickel-purified protein polymers when the protein-DNA conjugates were analyzed in free solution by capillary electrophoresis. There was also no apparent difference between native and denaturing buffer conditions.

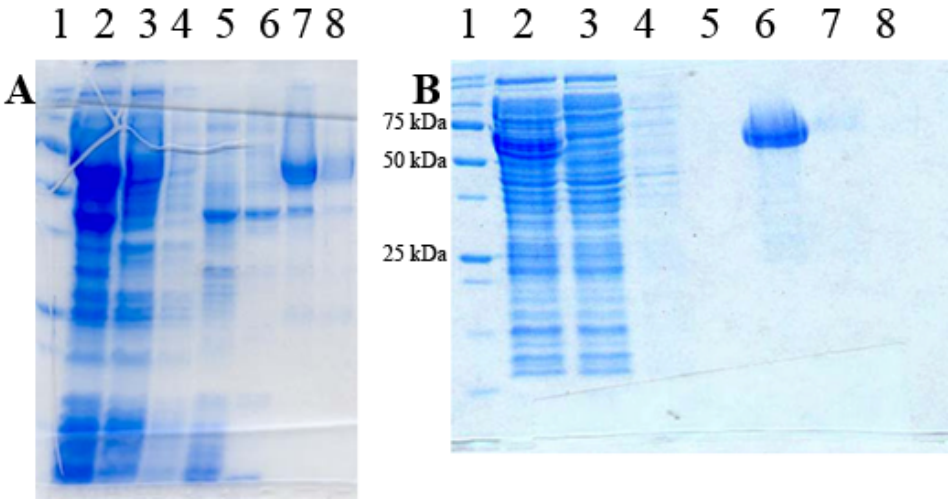


Figure 5-10: PZm8-24 purification using different metal ion resins A) ProBond nickel resin lane 1: ladder; lane 2: lysate; lane 3: flow through; lane 4: pH 7.8 wash; lane 5: pH 6.0 wash; lane 6: pH 5.3 wash; lanes 7-8: elutions B) Talon cobalt resin lane 1: ladder; lane 2: lysate; lane 3: flow through; lanes 4-5: washes; lanes 6-8: elutions

The SDS-PAGE results did show that Talon resin led to more pure protein in the affinity chromatography step, even though it did not solve the specific polydispersity issue. Another protein was purified using Talon resin, PZm8-24. As seen in Figure 5-10, there is significant reduction in nonspecific binding with the Talon resin compared to the nickel resin, thus avoiding the need to repeat the purification. Future protein purifications were performed using Talon

resin.

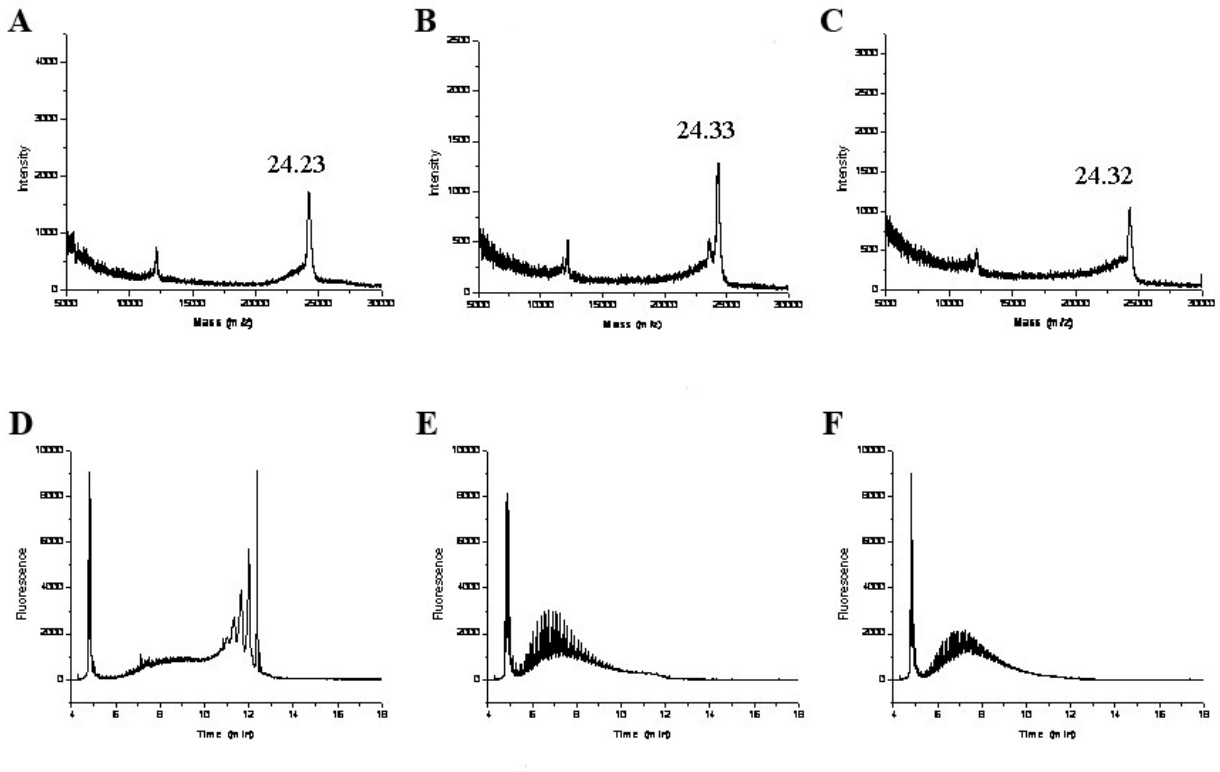


Figure 5-11: PZ8-16 extended cyanogen bromide cleavage (expected mass 24.15 kDa) over 2, 6, and 12 days. Comparison of MALDI-TOF results (top row) to free-solution capillary electrophoresis of drag-tag-DNA conjugates (bottom row). ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 3%v/v POP5, 1kV/15s injection, 320 V/cm, 55°C;. From left to right: 2 days, 6 days, 12 days reaction time

5.4.2 Varying cyanogen bromide cleavage reaction time

Previous work had shown that the cyanogen bromide cleavage reaction to remove the affinity tag could cause glutamine deamidation in protein polymers [120]. There was concern that, although these sequences did not contain any glutamines, other modifications or side

reactions were occurring during the cyanogen bromide step. This reaction is normally done for 24-48 hours.

5.4.2.1 Extended to several days

We have observed that after cyanogen bromide cleavage, some masses would be 50-300 Da higher than expected when analyzed by MALDI-TOF. The size difference varied between sequence, length, and even different batches of the same protein. The only consistent observation was that the largest protein (24mer with 505 amino acids) would have the largest mass difference.

Similar to the glutamine deamidation study using PZ6-16 [120], samples of PZ8-16 (no glutamines) were reacted with cyanogen bromide in 70% formic acid for two, six, and twelve days. Samples at each of these timepoints were analyzed by MALDI-TOF and in free solution by ELFSE (Figure 5-11). The overall trend is an increasing molecular mass and increasing polydispersity (and simultaneously decreasing α). These observations match the trends seen in the earlier study [175]. Since this protein does not contain glutamines, there is apparently an additional contributing factor to the polydispersity and mass increase. The mass increase may be resulting from formylation of the serine and/or threonine residues in the sequence (addition of 28 Da to the mass from -CH=O functionalization). For PZ8-16, there are 48 serines and 48 threonines in the sequence. This reaction is a reported potential side effect from performing the cyanogen bromide cleavage in formic acid [176, 177]. An alternative cleavage protocol would be to use 6 M guanidinium chloride in 0.1 M hydrochloric acid. It is unknown why formylation, which increases the molecular mass, would also lead to decreasing hydrodynamic drag. Perhaps

another reaction is involved which, like glutamine deamidation, generates an increasing number of negative charges over time.

5.4.2.2 Shortened to four hours

Clearly extended cleavage times are detrimental to the protein polymer. Purified samples of PZm8-12 and PZm8-24 were cleaved under a shortened reaction time of only four hours. As expected with the reduced reaction time, cleavage was mostly incomplete. After removing uncleaved protein with a second chromatography step, the samples were analyzed by ELFSE. There was no noticeable reduction in the height or number of peaks. Reducing the reaction time further would be impractical as almost no protein would be cleaved in such a short amount of time. It would be advantageous to use an alternate cleavage technique that is less harsh on the protein as even brief exposure to the chemicals involved in the cyanogen bromide reaction appear detrimental to the protein.

5.4.3 Enterokinase cleavage to remove affinity tag

The MpET-19b expression plasmid is encoded with an enterokinase recognition site between the *N*-terminal histidine tag and the target protein. Enterokinase is a site-specific protease that recognizes the DDDDK amino acid sequence, cleaving after the lysine residue, under mild conditions.

5.4.3.1 Test cleavages

An enterokinase cleavage/capture kit was purchased from Novagen (Madison, WI). As suggested in the manufacturer's protocol, different cleavage conditions were tested in small scale reactions to determine the minimum amount of enzyme for cleavage. All reactions were done at 20°C for 16 hours. Three enzyme to protein ratios (enzyme unit:μg protein) were tested (1:100, 1:50, and 1:20) in 25 and 50 μL reaction volumes. The results were analyzed on an SDS-PAGE gel for PZm8-12 and PZm8-24 (Figure 5-12). Cleavage appeared successful for all conditions tested based on the absence of the protein band in the reactions compared to controls. Therefore, the 1:100 ratio of protease to target protein in the 25 μL reaction volume was chosen for the scale-up conditions.

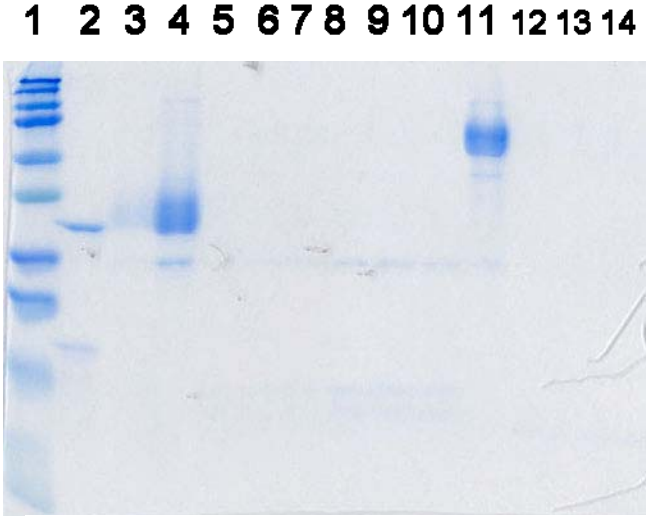


Figure 5-12: 12% SDS-PAGE of recombinant enterokinase (rEK) test cleavages; lane 1: ladder; lane 2: control protein 32 kDa and 16 kDa fragments when cleaved; lanes 3-4: PZm8-12 protein no protease added; lanes 5-7: PZm8-12 cleaved in 50 μL with 1:100, 1:50, 1:20 U rEK per μg protein; lanes 8-10: PZm8-12 cleaved in 25 μL with 1:100, 1:50, 1:20 U rEK per μg protein; lane 11: PZm8-24 protein no protease added; lanes 12-14: PZm8-24 cleaved in 50 μL with 1:100, 1:50, 1:20 U rEK per μg protein

5.4.3.2 Large-scale cleavage

Five milligrams of purified, but uncleaved, PZm8-12 and PZm8-24 protein were reacted with 50 units (U) of enterokinase in a total volume of 12.5 mL for 16 hours at 20°C. After the reaction, the samples were mixed with 3.5 mL of slurry of the provided EKapture™ agarose, which bound to the enterokinase after 30 minutes of mixing. The flow through was collected into Amicon Ultra-15 centrifugal filter devices (10K MWCO) from Millipore (Bedford, MA). Salts and the cleaved His tag were removed by ultrafiltration. Each sample was purified by IMAC to remove any uncleaved proteins. Mass spectrometry confirmed that enzymatic cleavage was successful. Approximately 4-5 mg of protein was recovered, indicating nearly complete cleavage occurred. Cleaved proteins appeared as a single peak with a trailing "tail" on MALDI-TOF (Figure 5-13A) but near the expected size (18.8 kDa instead of 18.7 kDa for the 12mer and 37.3 kDa instead of 37.0 kDa for the 24mer). Despite the gentle cleavage conditions, PZm8-12 did not demonstrate any improvement when analyzed by ELFSE as seen in Figure 5-13B (conjugation and analysis of the PZm8-24 protein was unsuccessful). Evidently, the cyanogen bromide reaction was not the main cause of the polydispersity.

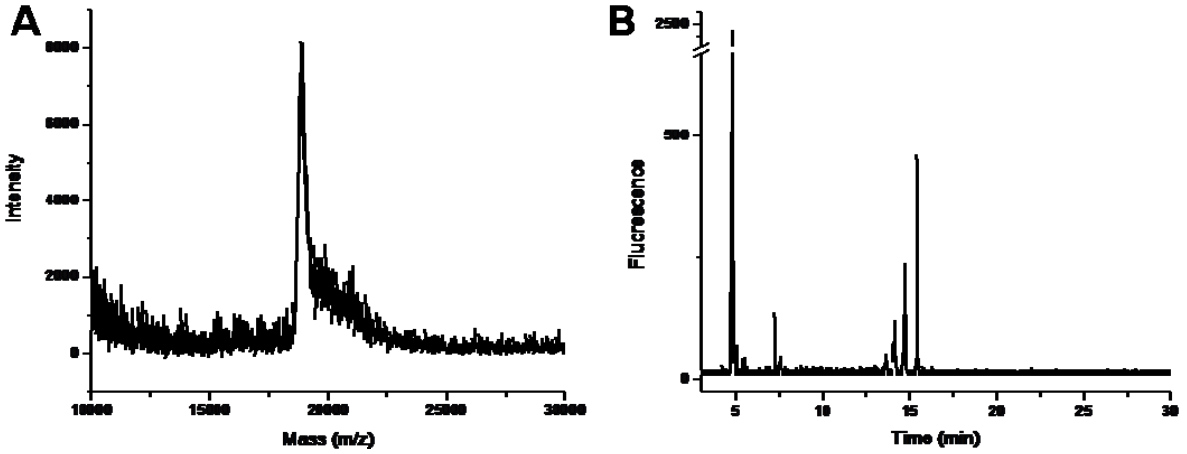


Figure 5-13: PZm8-12 after *N*-terminal His tag removal by recombinant enterokinase (rEK) A) MALDI-TOF spectra B) free-solution capillary electrophoresis of drag-tag-DNA conjugates using a 30-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 1%v/v POP6, 1kV/8s injection, 320 V/cm, 55°C

5.4.4 Cell lysis by sonication

Sonication has always been used when breaking apart the cells after harvesting (Chapter 2). We suspected our protein polymers might have been damaged or sheared into smaller pieces in the process, much like DNA can be sheared [124, 178]. Even a difference of a single amino acid in a drag-tag sample is distinguishable by ELFSE whereas the mass difference might be too small to be resolved by mass spectrometry. The typically broad peak we have observed for protein polymers may overlap with a neighboring peak. Protein polymer peak widths, measured at the base, can span upwards of ~ 2000 Da depending on protein molecular mass. By comparison, the amino acid with the highest molecular weight out of those used in these sequences has a mass of only 174 Da (Arg).

5.4.4.1 Detergent-based cell lysis

Several companies offer detergent-based formulations for cell lysis. Cells can be lysed by gentle mixing with these solutions which also claim to achieve higher yields than sonication. BugBuster reagent from Novagen (Madison, WI) was used to lyse the cells from a 2 L culture of PZm8-24 using the manufacturer's protocols. 40 mL of reagent was incubated with the cells at room temperature for 30 minutes with gentle shaking. Benzonase nuclease (provided with the lysis reagent) was added to break up chromosomal DNA. A low yield of 7 mg of protein was obtained (compared to an expected yield of around 40 mg). After purification and cleavage, the protein was analyzed by ELFSE. Not only did the polydispersity not improve, but also the baseline appeared to have gotten much worse (Figure 5-14) compared to previous analyses of PZm8-24. This may have been caused by unknown chemicals in the lysis reagent that were not removed during purification.

5.4.4.2 PZm8-12 purification comparison with protease inhibitors

A 2 L expression of PZm8-12 was split equally into four separate batches after harvesting.

One was lysed using only freeze/thaw techniques, the second with freeze/thaw and protease inhibitor (EDTA-free Halt Protease Inhibitor Cocktail from Pierce, Rockford, IL) added in afterwards, the third with freeze/thaw and sonication, and the fourth with freeze/thaw, protease inhibitor, and sonication. As seen in Figure 5-15, there is no significant difference between the four bioconjugates when analyzed by ELFSE (*i.e.,* all are polydisperse). This agreed with the previous results that sonication was not damaging the protein polymer. Also, the addition of a

broad spectrum protease inhibitor cocktail did not have any noticeable effect. Hence, the protein polymer was not being digested by native proteases.

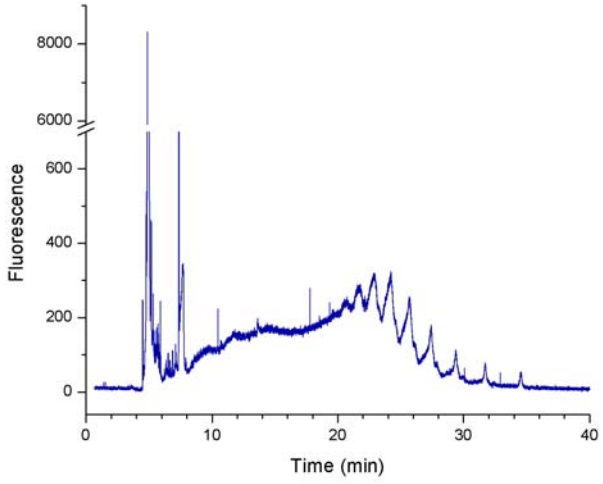


Figure 5-14: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for PZm8-24 (505 amino acids) using a 20-base primer. Cells lysed using BugBuster detergent with no sonication. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 1M urea, 0.5%v/v POP6, 1kV/15s injection, 320 V/cm, 55°C

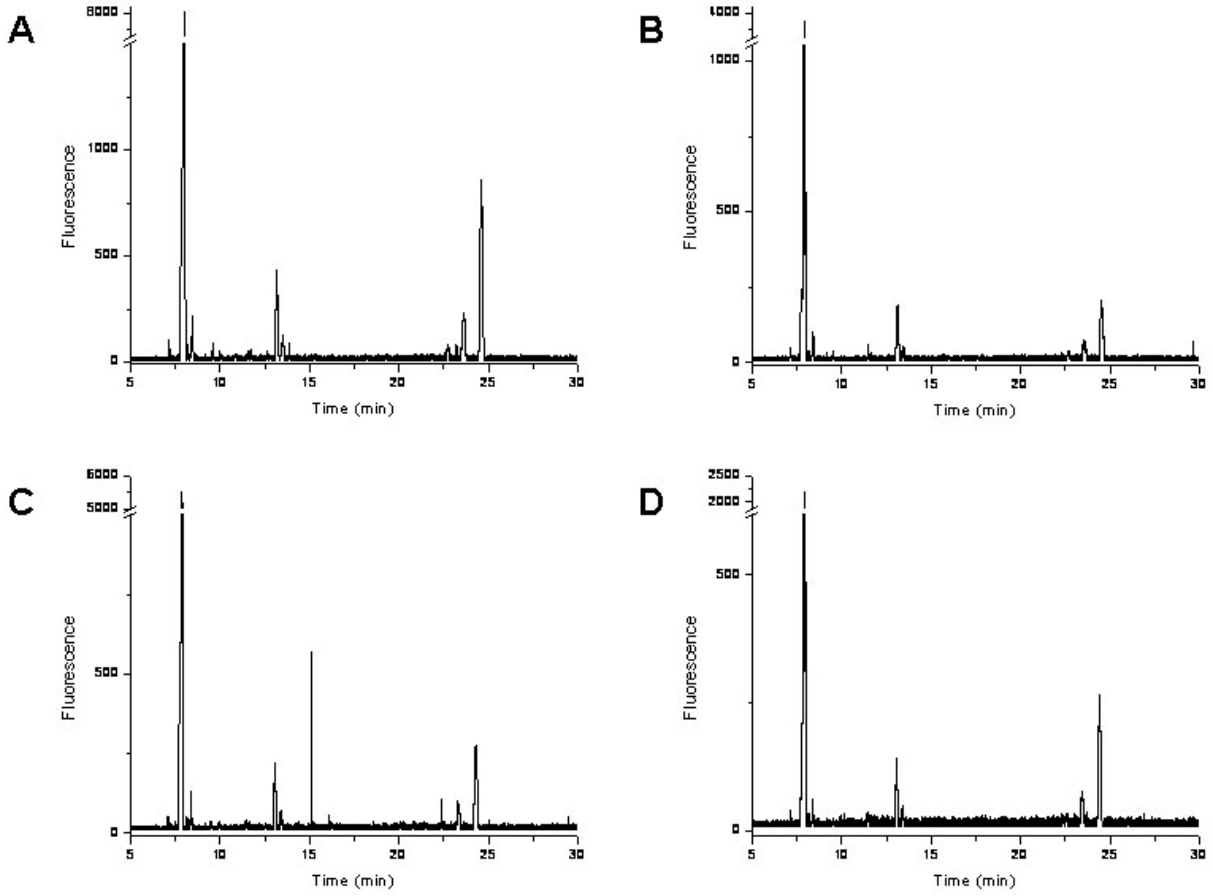


Figure 5-15: Free-solution capillary electrophoresis of drag-tag-DNA conjugates of PZm8-12 using a 30-base primer. Cells lysed under varying conditions A) freeze/thaw only B) freeze/thaw with protease inhibitor C) freeze/thaw, sonication, protease inhibitor D) freeze/thaw, sonication. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 3%v/v POP5, 1kV/1s injection, 320 V/cm, 55°C

5.5 Plasmid DNA verification

Since we are attempting to express long lengths of a highly repetitive sequence in *E. coli*, it is prudent to verify that the plasmid DNA was not becoming altered or mutated during large-scale expression. Samples were taken from large-scale expressions of PZ8-6 and PZ8-24 just

after the 3 hour induction period. The plasmid DNA was recovered by miniprep. Sequencing confirmed that there were no mutations in the sequences nor were there any abnormalities in the chromatograms (*i.e.,* another, overlapping sequence). Restriction enzyme digestion also confirmed the insert DNA was of the expected size and no other sizes (*e.g.,* from recombination events) were present at levels detectable by the Agilent 2100 Bioanalyzer.

5.6 Phosphorylation of serines and/or threonines during expression

We have also been investigating an alternative cause of the added molecular mass occasionally seen by mass spectrometry, which might contribute to the polydispersity observed by ELFSE as well. After analyzing a list summarizing potential post-translational modifications (http://ca.expasy.org/tools/findmod/PHOS.html), it was discovered that for prokaryotes Ser and Thr residues, in any position, can become phosphorylated. A single phosphorylation of either a S or T residue would produce a mass increase of 80 Da. A recombinant human protein expressed in *E. coli* was phosphorylated at serine, threonine, and tyrosine residues [179]. *In vivo* phosphorylation has not been investigated or reported for protein polymers; however, silk-based protein polymers have been intentionally enzymatically phosphorylated and dephosphorylated to control protein solubility [180].

5.6.1 CIP reaction

Calf intestinal phosphatase (CIP), which is used to dephosphorylate DNA during the controlled cloning step, can also be used to dephosphorylate proteins [180]. Five milligrams of uncleaved PZ8-16 was reacted with 100 units of CIP at 37°C for 16 hours. After the reaction, the protein polymer was separated from the CIP enzyme by affinity chromatography and then

cleaved by cyanogen bromide. There was no noticeable size change from untreated PZ8-16 when analyzed by MALDI-TOF (Figure 5-16A). A mass of 27.1 kDa was observed, matching the value obtained previously from Figure 5-9A. Analysis of the CIP-treated protein showed no improvement in the polydispersity (Figure 5-16B). In fact, the electropherogram appeared even worse than before compared to Figure 5-9B.

5.6.2 Phosphorylated protein purification columns

Qiagen (Valencia, CA) sells a PhosphoProtein Purification kit based on affinity chromatography and claims to be able to completely separate phosphorylated proteins from unphosphorylated proteins in cell lysates. Phosphorylated proteins bind to a column filled with resin while unphosphorylated proteins flow through. A few milligrams of cleaved PZm8-12 and PZm8-24 were purified using this kit. The flow through fractions, which would contain any unphosphorylated proteins, were dialyzed and lyophilized prior to ELFSE analysis. Even though the amount of protein recovered was less than the initial quantity, there was, again, no improvement in the resulting electropherograms. No protein was recovered in the phosphorylated fractions.

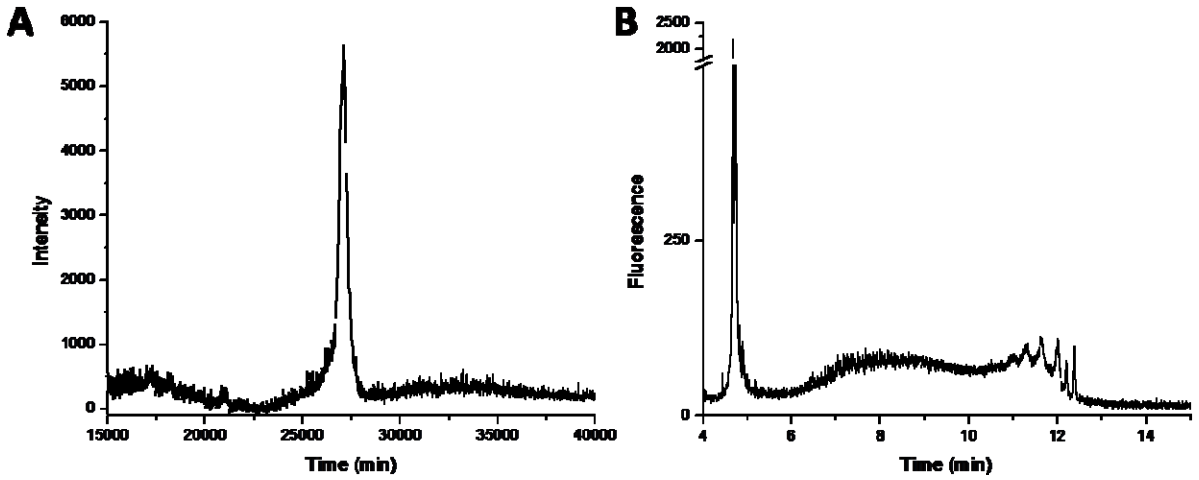


Figure 5-16: PZ8-16 protein after treatment with calf intestinal phosphatase (CIP) for dephosphorylation A) MALDI-TOF spectra B) free-solution capillary electrophoresis of drag-tag-DNA conjugates using a 30-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 2kV/10s injection, 320 V/cm, 55°C

5.6.3 Dot blot detection

Since efforts to dephosphorylate or separate potentially phosphorylated proteins were not successful, we attempted to simply confirm the presence of phosphorylated residues by using antibodies sensitive to phosphorylated serine and threonine residues in a dot blot. PhosphoSerine Antibody Q5 and PhosphoThreonine Antibody Q7 were purchased from Qiagen (Valencia, CA). The recommended rabbit anti-mouse IgG/IgM HRP-conjugate (secondary antibody) was obtained from Jackson ImmunoResearch (West Grove, PA). A phosphoprotein control set was purchased from Pierce (Rockford, IL) consisting of phosvitin (positive control) and soybean trypsin inhibitor (negative control).

Dot blot detection was done according to the provided primary antibody protocol. Bovine serum albumin (BSA) was used in the blocking buffer step as milk powder contains serine- and threonine-phosphorylated proteins. Tris-buffered saline was used instead of phosphate-buffered saline. As per the protocol, primary antibody incubation was done overnight at 4°C while the secondary antibody incubation was done at room temperature for 1 hour. One microgram (and later 5 μg on a second attempt) of purified samples of PZm8-6, PZm8-12, and PZm8-24 were dotted onto each membrane along with the two control proteins. Despite using more than enough sample for detection and 10 minutes of film exposure, no proteins, not even the positive control, were detected using either primary antibody. The reason for the failed antibody detection of phosphorylated residues could not be determined.

5.7 Conclusions and Recommendations

We have shown that ELFSE sequencing is possible using a small, nearly monodisperse protein polymer as the drag-tag [123]. The read length of 180 bases is not competitive with current Sanger sequencing read lengths of 500+ bases, requiring the production of larger drag-tags. However, while longer-length protein polymers were successfully created, they all exhibited multiple peaks when protein-DNA conjugates were analyzed in free solution, making the proteins unsuitable as drag-tags. This was in contrast to MALDI-TOF, SDS-PAGE, and RP-HPLC data which suggested the protein samples were pure. Many different theories were tested as the source of the polydispersity such as the affinity tag cleavage step and cell lysis by sonication. Nevertheless, none of these theories could be proven as the cause of the additional peaks seen in the electropherograms.

While the above mentioned experiments were ongoing, work was also in progress to adapt an expression plasmid containing different *N*- and *C*-terminal affinity tags for protein polymer expression and purification. Using the same His tag but with a pH gradient purification followed by a second imidazole-based purification was not sufficient to obtain completely pure protein. However, perhaps use of orthogonal affinity tags, based on two completely separate binding principles, would yield absolutely pure protein. In contrast to other applications of protein polymers (*e.g.,* biomaterials), obtaining a totally monodisperse protein is key to a successful drag-tag and, consequently, ELFSE sequencing. For longer read lengths, the protein will need a larger α than 25 in addition to being monodisperse.

Chapter Six

Protein Expression with a *C*-terminal Affinity Tag: Obtaining Truly Monodisperse Protein Polymers

6.1 Introduction

Monodispersity is an absolute necessity for ELFSE drag-tags. This chapter describes a new approach to ensure monodisperse production of protein polymer drag-tags by using *C*-terminal His tags. There is a possibility that expression of such highly repetitive sequences places additional stress on the cell biomachinery during protein production, leading to truncated versions of the desired protein [181-183]. Obtaining a truly monodisperse protein polymer with a large hydrodynamic drag would allow us to generate longer sequencing reads by ELFSE.

Insufficient tRNA pools can lead to translational stalling, premature translation termination, translation frameshifting, and amino acid misincorporation [125]. In addition, it is possible that even relatively abundant species of tRNA become depleted when expressing highly repetitive protein polymers. Using an N-terminally expressed His tag can lead to co-purification of truncated products along with the desired full-length protein, and these prematurely truncated proteins may account for the extra peaks observed in free-solution capillary electrophoresis of drag-tag-DNA conjugates using protein polymers greater than 127 amino acids in length (Chapter 5). In order to ensure that only the full-length protein is recovered in the final purified product, a *C*-terminal affinity tag must be used. Only complete proteins would have the affinity tag and therefore bind to the resin during purification. Obtaining a large, monodisperse protein polymer additionally required removal of the *C*-terminal His tag. Note that as in Chapter 5, the

electrophoretic analyses using ELFSE were performed by either Dr. Robert Meagher or graduate student Jennifer Coyne.

6.2 GST-His double tag vector (pET-41a)

A commercially available vector, pET-41a, was purchased from Novagen (Madison, WI). Figure 6-1A is a vector map of the pET-41a plasmid while Figure 6-1B represents the sequence of the multiple cloning site (MCS) in the vector. *Sap* I and *Ear* I sites are marked along with select unique restriction enzyme sites. The original pET-41a vector contains regions coding for kanamycin resistance (30 μg/mL), a GST (glutathione-S-transferase) tag, S-tag, and a 6X histidine tag at the N-terminus and a *C*-terminal 8X histidine tag. It was modified in a similar manner to MpET-19b (Section 2.5 and [133]) with a few additional steps.

6.2.1 Producing the MpET-41a expression vector

6.2.1.1 Site-directed mutagenesis to remove existing *Sap* I sites

Site-directed mutagenesis (QuikChange Kit, Stratagene, La Jolla, CA) was used to alter the two existing *Sap* I sites of pET-41a into *Ear* I recognition sites. Primer sequences 5'-CTT GAA GAA AAA TAT GAG GAG CAT TTG TAT GAG CGC GAT G-3' and 5'-GAG GAA GCG GAA GAG AGC CTG ATG CGG-3' along with the reverse complementary sequences (four primers total) were designed according to the manufacturer's guidelines and purchased as PAGE-purified DNA oligomers from IDT (Coralville, IA).

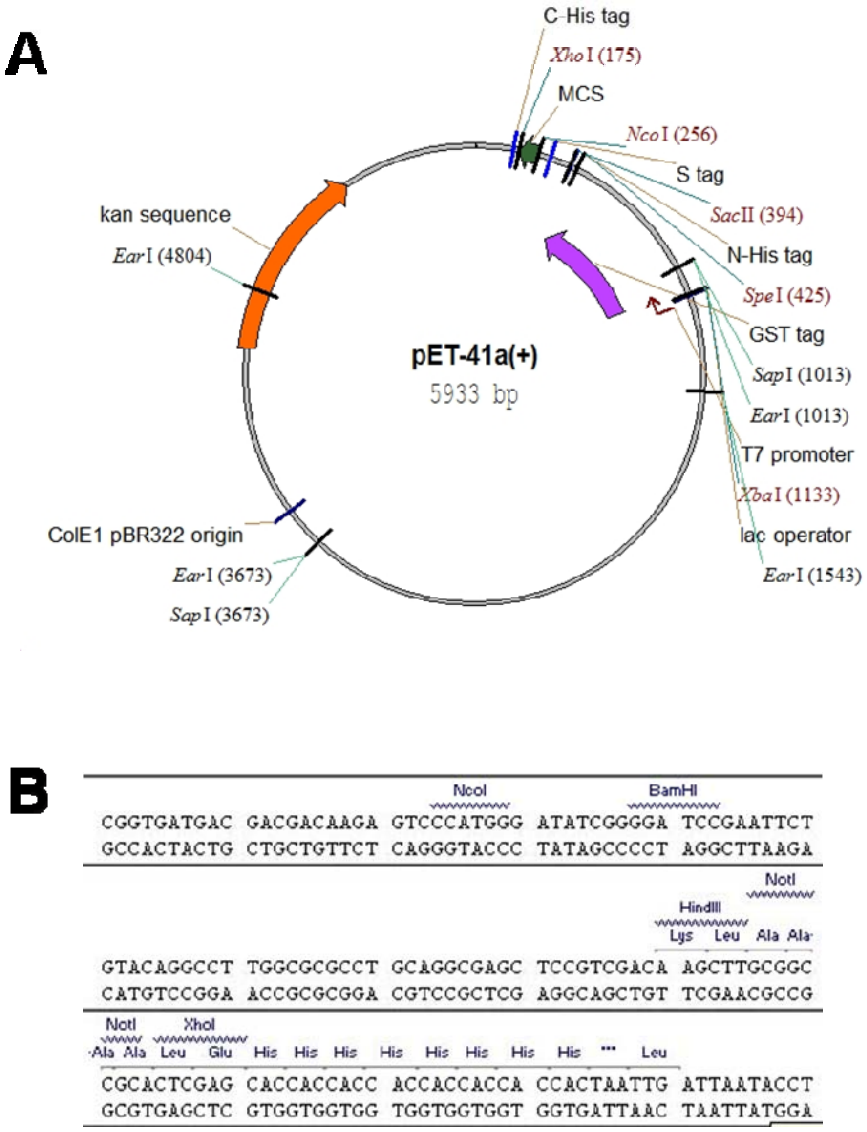


Figure 6-1: A) vector map of pET-41a expression plasmid B) expanded view of multiple cloning site (MCS)

Two rounds of mutagenesis were performed according to the suggested manufacturer's protocol. The thermal cycling program used consisted of an initial 30 second denaturing step at 95°C followed by 16 cycles of 30 second denaturation at 95°C, 1 minute annealing at 55°C, and

6 minute extension at 68°C. The methylated (by *E. coli*) parental DNA was digested by *Dpn* I allowing only the mutated DNA to be transformed into *E. coli* cells. *Sap* I digestion of the recovered plasmid DNA confirmed the modifications were successful.

6.2.1.2 Short sequence replacement

A short oligonucleotide sequence, 5'-GTT CAA CTA GTG GTT CTG GTC GCT CTG GTA CCT CTG GCT CCG CGG GTC TG-3', was designed to replace the existing *N*-terminal His tag sequence. This oligonucleotide codes for the amino acid sequence, STSGSGRSGTSGSAGL, which consists mostly of amino acids already used in PZ8 or PZm8 and, therefore, are unlikely to cause complications during protein expression. The His tag is flanked by unique *Spe* I and *Sac* II restriction enzyme sites in the vector. The replacement sequence was constructed with identical, flanking recognition sites. Double enzyme digestion of the plasmid DNA and the PCR-amplified oligonucleotide yielded the necessary cohesive ends for ligation. DNA sequencing verified the substitution. GST was left at the N-terminus to aid in protein expression by acting as a leader sequence known to promote solubility in some expressed proteins and as a second affinity tag for purification.

6.2.1.3 Dangled primer PCR to generate modified ends

The dangled primers 5'-AGT TAG CTC TTC AGG TAT GAA GCT TGC GGC CGC ACT CGA-'3 and 5'-AGT TAG CTC TTC AAC CCA TGG GAC TCT TGT CGT CGT C-3' were used to PCR-amplify the plasmid DNA and generate the necessary adapter ends to accept any insert from pUC18 [133]. *Sap* I sites are colored in red. Instead of inserting a stop codon at the end of the sequence, as was done for the MpET-19b plasmid, a Met residue (green) was

incorporated to allow for cyanogen bromide (CNBr) cleavage of the *C*-terminal His tag. The same thermal cycling protocol and *Pfu* polymerase used to amplify the MpET-19b plasmid with dangled primers [133] was used here as well and is reproduced below.

a) 95°C for 4 min
b) 58°C for 4 min
c) 72°C for 12 min
d) 95°C for 45s
e) 58°C for 35s
f) 72°C for 12 min
g) repeat d through f 25 more times
h) 72°C for 8 min

The *C*-terminal His tag sequence is G-MKLAAALE$(H)_8$ where the affinity tag is cleaved after the methionine residue by the cyanogen bromide reaction. Following PCR the vector was digested with *Sap* I to generate the necessary cohesive ends for the insert DNA (from pUC18) and then dephosphorylated by CIP to prevent circularization of the recipient vector. This new, modified vector was named MpET-41a. This strategy differs from the approach discussed in Chapter 4 for inteins by allowing direct transfer of genes from the cloning vector (via enzyme digestion and ligation) instead of from the pET-19b expression vector (via PCR amplification followed by enzymatic digestion and ligation).

6.2.2 PZ8-24 large scale expression, purification, and tag cleavage in both systems

PZ8-24 was tested concurrently in MpET-19b and the newly developed expression vector, MpET-41a. Both proteins were expressed in 2 L volumes under the same conditions and both had similar yields of 25 mg/L. However, GST is 35 kDa in mass so the actual yield of PZ8-24 (36.18 kDa) in MpET-41a was only half that amount. The MpET-19b protein was purified once by IMAC using denaturing conditions (8 M urea). The MpET-41a protein (GST-PZ8-24-

GMKLAAALE$(H)_8$) was purified first by IMAC under denaturing conditions and then the elutions were dialyzed and lyophilized. The recovered protein was then purified under native conditions on glutathione resin (native conditions are required for proper binding to the resin). Approximately 70% of the starting material was recovered after the second purification step.

Both purified proteins were analyzed by MALDI-TOF mass spectrometry before and after cyanogen bromide cleavage (Figure 6-2). The MpET-19b-expressed protein yielded the typical spectra for a purified protein polymer, consisting of a single peak at the expected masses before (39.1 kDa) and after cleavage (36.18 kDa) and an additional doubly charged ion peak at half the mass (m/z). However, the spectrum of the MpET-41a expressed protein contained several unexpected peaks in addition to the desired fusion protein peak (70.2 kDa). The CNBr-cleaved protein showed multiple peaks at lower masses than before and a small peak near the expected mass of a fully cleaved PZ8-24mer.

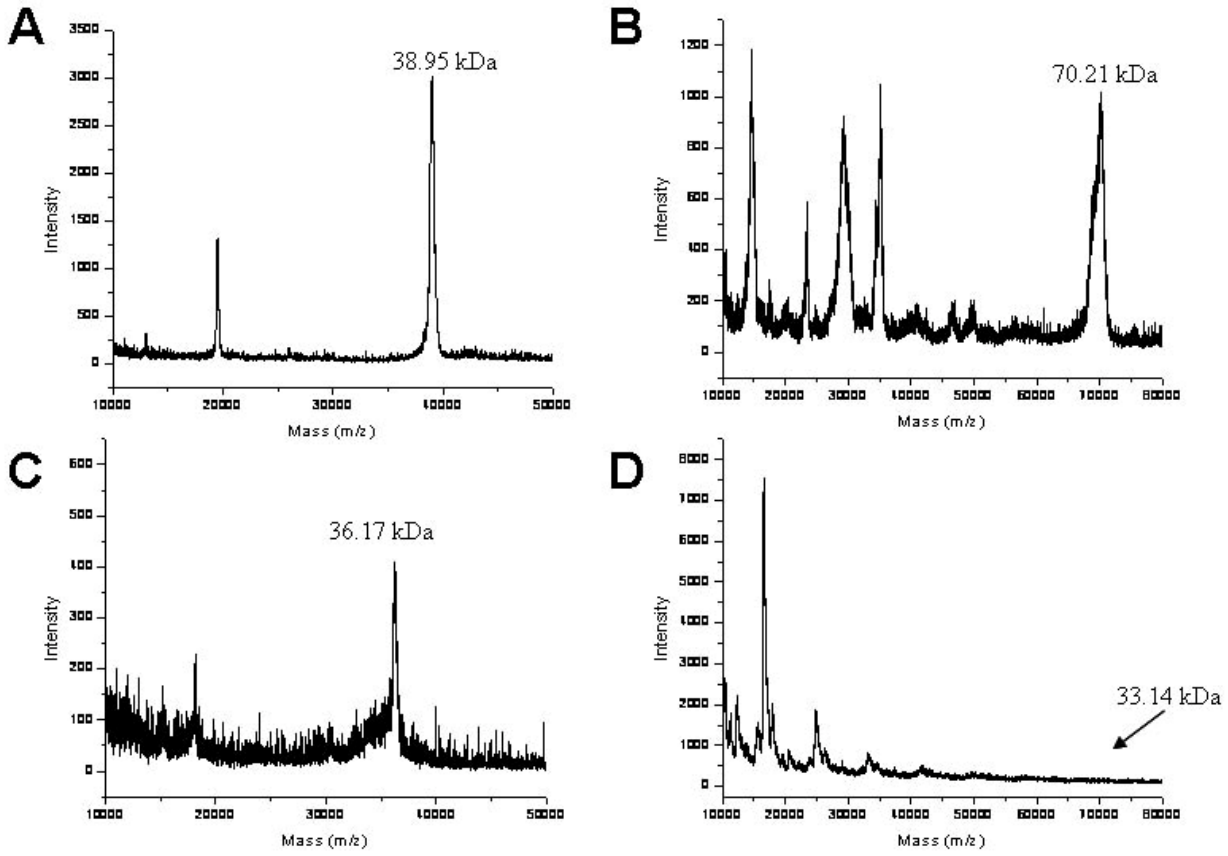


Figure 6-2: MALDI-TOF spectra for PZ8-24 before and after CNBr cleavage in different expression vectors A) in MpET-19b B) in MpET-41a C) in MpET-19b after cleavage D) in MpET-41a after cleavage

The *N*-terminal GST tag did not improve protein yield and the expressed protein was less pure by MALDI-TOF. Free-solution capillary electrophoresis of the PZ8-24 conjugates showed the same multiple peak pattern as other large protein polymers produced using an *N*-terminal His tag (Figure 6-3). The MpET-41a protein sample was not analyzed by ELFSE as it clearly was not pure. We later realized that one problem with this method is that the GST protein does in fact contain several methionine residues (unlike the *N*-terminal affinity tag in MpET-19b),

complicating cyanogen bromide cleavage and purification of uncleaved protein. It is also possible that the GST protein is highly susceptible to protease degradation and that this resulted in the multiple peaks seen in the purified protein sample.

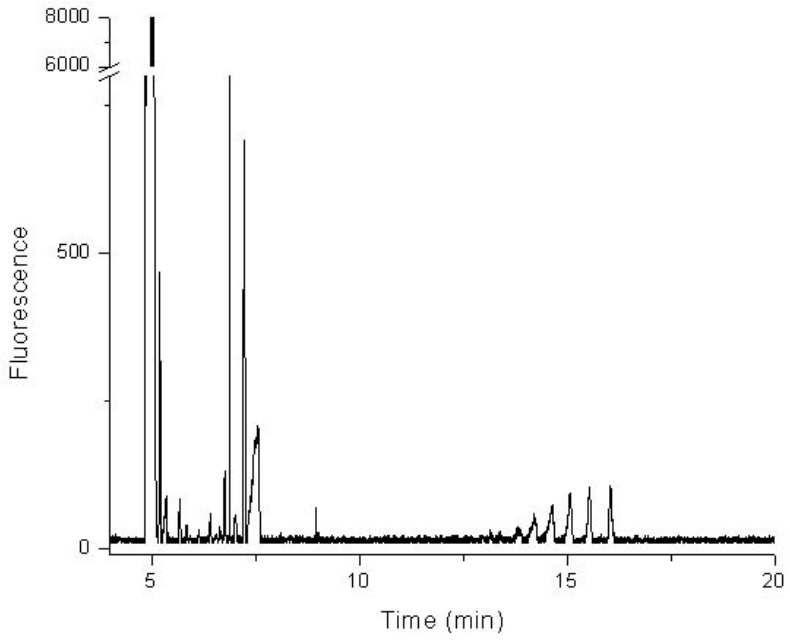


Figure 6-3: Free-solution capillary electrophoresis of drag-tag-DNA conjugates of PZ8-24 expressed in MpET-19b using a 26-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 3kV/20s injection, 320 V/cm, 55°C

6.3 MCHis41a expression vector variant

The MpET-41a vector was modified further by eliminating the entire *N*-terminal GST (and S-tag region). Removal of the GST tag would likewise eliminate the additional methionine reactive sites.

6.3.1 Removal of *N*-terminal GST tag

Unique *Xba* I and *Nco* I restriction enzyme sites flank the GST region in the plasmid that will be excised. This region was substituted with a sequence copied directly from a region in the pET-19b vector between its *Xba* I and *Nco* I sites (5'-ATT CCC CTC TAG AAA TAA TTT TGT TTA ACT TTA AGA AGG AGA TAT ACC ATG GAT ATC-3'). This ensured the sequence would be compatible for protein expression (*i.e.,* ribosome binding site and in frame). PZ8-6 and PZ8-24 were tested in this new version of the vector, MCHis41a, that contained only the *C*-terminal His tag. Both sequences had been inserted previously into the MpET-41a double tag vector. Therefore, only removal and replacement of the *N*-terminal affinity tag region was required. DNA sequencing confirmed the small oligonucleotide had successfully replaced the *N*-terminal affinity tag in the two plasmids.

6.3.2 PZ8-6 expression and purification

A 2 L culture of PZ8-6 with only a *C*-terminal His tag was expressed with a final yield of 6.5 mg/L. This value was approximately 25% of the yield obtained from expression of the same protein with an *N*-terminal His tag. This drastic reduction in yield may be due to the removal of a favorably expressed leader sequence such as the GST protein or the His tag at the N-terminus of the protein, the exclusion from the final product of suspected truncated proteins, or a combination of both factors.

6.3.3 PZ8-24 expressions with varied inducer concentrations

For the expression of PZ8-24, two concentrations of the inducer, IPTG, were compared. 2 L was induced at the standard 1 mM IPTG concentration used for past expressions and 2 L was

induced at 0.1 mM. After purification, it was found that more protein was obtained using the lower IPTG concentration (17.2 mg compared to 6.5 mg for 2 L expression). This was still less than the 25 mg/L expression levels typically obtained for protein expressed with an *N*-terminal affinity tag.

6.3.4 PZ8-6 and PZ8-24 cyanogen bromide cleavage

Both proteins were cleaved by cyanogen bromide and the masses were determined by MALDI-TOF and the protein-DNA conjugates were analyzed by ELFSE. The MALDI-TOF spectra of the protein polymers with a *C*-terminal His tag do not differ noticeably from spectra obtained for proteins with an *N*-terminal His tag. However, the ELFSE electropherograms were distinctly different for both proteins when compared to past results with an *N*-terminal His tag. As seen in Figure 6-4, both C-terminally tagged PZ8-6 and PZ8-24 exhibit twin peak patterns in sharp contrast to the single peak observed for PZ8-6 (Figure 5-7) or the six peaks seen previously for PZ8-24 (Figure 6-3), both produced with an *N*-terminal His tag later removed by cyanogen bromide. The small pair of peaks offset from the main twin peaks for the PZ8-24 protein is likely due to incomplete removal by chromatography of uncleaved protein (confirmed by MALDI-TOF). The amount of IPTG used had no effect on the electrophoretic analyses. These results were the first evidence of significant alterations to the protein-DNA conjugate profile in ELFSE (in contrast to the extensive testing done in Chapter 5). Only a single peak was detected when the PZ8-24 protein was analyzed by RP-HPLC on the C4 column. Therefore this second peak detected by ELFSE could not be easily isolated and purified using RP-HPLC.

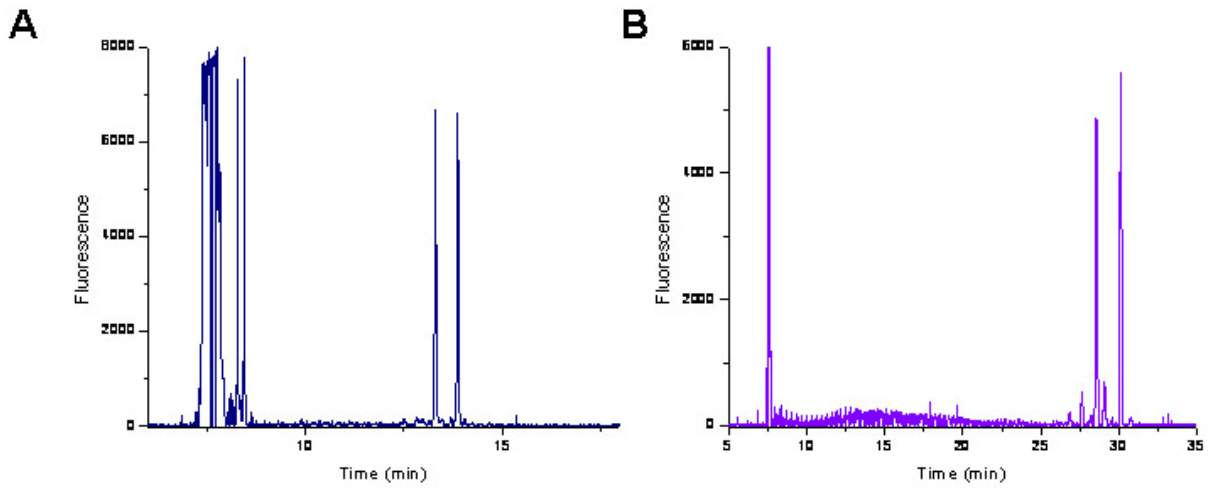


Figure 6-4: Free-solution capillary electrophoresis of drag-tag-DNA conjugates using a 30-base primer for PZ8 proteins expressed in MCHis41a with a *C*-terminal His tag that was removed by CNBr A) PZ8-6 α ~ 22, 25 B) PZ8-24 α ~ 85, 90. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 1kV/20s injection, 320 V/cm, 55°C

6.3.5 Consequences of cyanogen bromide cleavage

We hypothesized that the double peak results are due to the cyanogen bromide reaction that is used to remove the *C*-terminal affinity tag. No other procedures have changed except the positioning of the affinity tag and the cleavage point. As a byproduct of the reaction, methionine (the new C-terminus of the cleaved protein), is converted into homoserine lactone in acidic conditions (such as 70% formic acid) but under

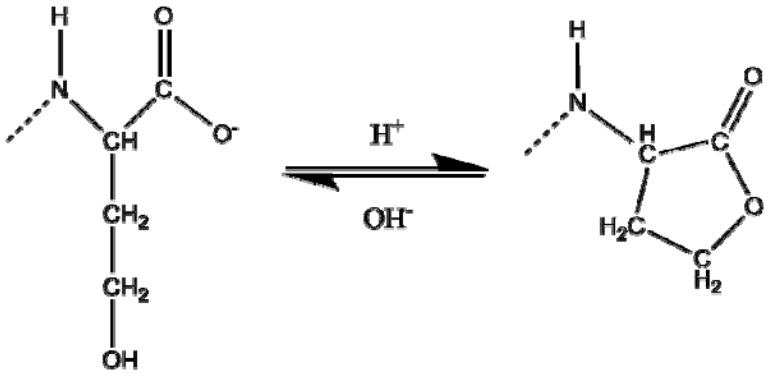


Figure 6-5: Homoserine (left) and homoserine lactone (right)

neutral conditions there exists an equilibrium between homoserine (linear form) and homoserine lactone (ring form), differing by 18 Da (Figure 6-5) [139, 184-187]. The conjugation reaction and free-solution electrophoresis are performed in buffer at pH 7.2, near neutral conditions. The structural differences between these two forms may be enough to cause the two peak results seen by ELFSE. In addition, *C*-terminal homoserine side chains (but not homoserine lactone) can be formylated during cleavage [177], further distinguishing the two variants. Cyanogen bromide cleavage of the *N*-terminal affinity tag had not been problematic as the homoserine lactone residue would have been part of the cleaved affinity tag and not the protein polymer itself.

6.4 Opting to leave the His tag attached

An alternative method is to enzymatically remove the affinity tag using a site-specific protease. However, this strategy adds several undesirable (*e.g.,* hydrophobic, negatively charged) amino acids from the recognition site to the C-terminus of the drag-tag. Table 6-1 lists the commercially available site-specific proteases and their respective recognition sites.

Table 6-1: Recognition/cleavage sites of site-specific proteases

Site-specific Protease	Recognition/Cleavage Site(*)
Enterokinase	DDDDK*
Factor Xa protease	IEGR*
Thrombin	LVPR*GS
TEV protease	ENLYFQG*
PreScission™ protease	LEVLFQ*GP
Genenase I	PGAAHY*

Additionally, no enzyme was found that could cleave selectively at the Met residue like cyanogen bromide. We opted to leave the His tag attached to the protein instead. The *N*-terminal His tag was always removed due to the presence of the enterokinase recognition site (DDDDK). If the *C*-terminal His tag is to be left attached, the existing lysine in the affinity tag will need to be removed. The primary amine would act as a reactive site during the maleimide activation step, competing with the amino terminus for sulfo-SMCC. In addition the methionine that was inserted into the sequence for cyanogen bromide cleavage was no longer needed. By leaving the affinity tag attached, a second purification step is also avoided.

Previous sequences in MCHis41a (PZ8-6 and PZ8-24) were obtained by replacing the *N*-terminal affinity tag region from an existing MpET-41a plasmid (already containing a gene insert). Now that there is evidence that a *C*-terminal affinity tag can improve the polydispersity of a protein when analyzed by ELFSE, a more practical method to test other sequences would be to generate a recipient vector, like the MpET-19b and MpET-41a plasmids, that can accept any gene from the pUC18 cloning vector.

6.4.1 Removal of Met and Lys residues using dangled primer PCR

The Met and Lys residues were excluded from the revised dangled primers (5'- AGT TAG CTC TTC AGG TCT TGC GGC CGC ACT CGA-3' and 5'-AGT TAG CTC TTC AAC CCA TGG TAT ATC TCC TTC TTA A-3'). Again, *Sap* I sites are in red. In addition, 16 bases in one primer had to be changed so that it could anneal to the new *N*-terminal region as the GST/S-tag portion had been removed. Template DNA was produced by substitution of the *N*-terminal region in a sample of the MpET-41a plasmid. Attempts to generate the correctly

amplified plasmid were unsuccessful. It was determined that primer amplification without a template was occurring, likely due to the large difference in melting temperatures between the two primers, (73.0°C versus 60.8°C). The difference is unavoidable since the primers must anneal to specific regions. Site-directed mutagenesis was also unsuccessful possibly due to strong secondary structure interactions between the required primer sequences.

The original approach (from MCHis41a) was used to generate a few sequences without the Met and Lys residues in the affinity tag. The MpET-41a vector was amplified using the dangled primer designed to anneal to the GST/S-tag region (5'-AGT TAG CTC TTC AAC CCA TGG GAC TCT TGT CGT CGT C-3') and the revised primer designed to exclude Met and Lys (5'- AGT TAG CTC TTC AGG TCT TGC GGC CGC ACT CGA-3'). Amplification was successful with this combination of primers (one old and one new) and the DNA was reacted with *Sap* I and CIP.

PZ8-24 was inserted into the prepared recipient vector. The *N*-terminal affinity tag region in the modified PZ8-24 MpET41a plasmid was replaced with the short oligonucleotide as before. DNA sequencing confirmed that the GST/S-tag region was successfully replaced. A 4 L culture of PZ8-24 was expressed and purified. The MALDI-TOF result is shown in Figure 6-6. The protein mass was 37.77 kDa, slightly below the expected mass of 37.85 kDa. However, DNA conjugation and ELFSE analysis were not successful.

Attempts to insert previously expressed PZ8 and PZm8 sequences into the newest *C*-terminal His tag vector using the same method described above were problematic. Despite ample concentrations of insert DNA, few colonies were obtained after transformation of the

insert-vector ligation reactions. DNA sequencing of these few colonies revealed various mutations and were not usable.

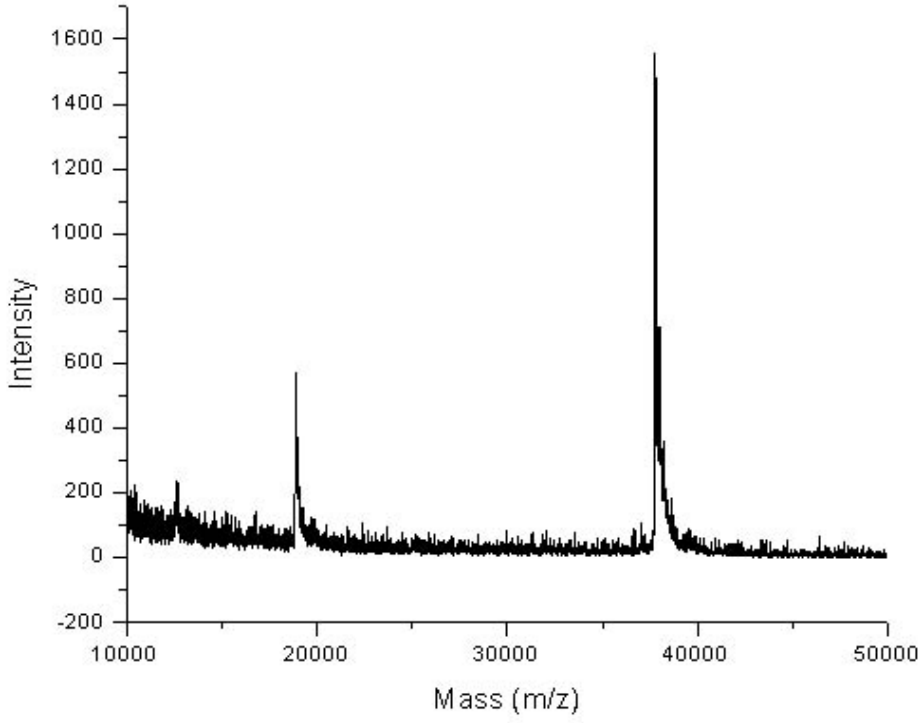


Figure 6-6: MALDI-TOF spectrum of PZ8-24 expressed with a *C*-terminal His tag containing no Lys or Met residues.

6.4.2 Further customization of *C*-terminal His tag

The procedure described above to insert a sequence into the *C*-terminal His tag expression vector is lengthy and not always successful. First, the gene must be inserted into the MpET-41a vector that has been PCR-amplified using dangled primers. Afterwards, the *N*-terminal affinity tag region is excised by double enzyme digestion and replaced with a short

oligonucleotide. Hence, the plasmid is converted into a *C*-terminal His tag expression vector with the accompanying gene insert.

An improved protocol that is both quicker and more reliable would be preferred. In addition, the presence of glutamic acid in the *C*-terminal His tag is unnecessary and results in a decreased α of the drag-tag. However, the glutamic acid is adjacent to the histidine residues. It is unlikely that the dangled primer PCR strategy could successfully remove the amino acid. Longer primers approaching 40 bp in length would be required, leading to an increased difference in melting temperatures and likely PCR complications. There was also concern that use of the lower frequency of usage of the two histidine codons in the affinity tag was causing the slightly smaller than expected molecular mass of PZ8-24 due to truncation within the His tag.

6.4.3 Custom-designed oligo linker to replace existing cloning region

An entirely different approach, inspired in part by cloning work done by the Kiick group [98, 100], was devised to generate the new recipient vector. Instead of making small changes to the original plasmid sequence through dangled primer PCR, as was done for recent versions of the *C*-terminal His tag vector as well as the MpET-19b vector, the whole cloning region will be excised from the plasmid and replaced with an oligonucleotide sequence entirely of our own design. This "adapter" oligonucleotide would contain the two *Sap* I sites necessary for controlled cloning (previously incorporated into the dangled primers) as well as two restriction enzyme sites to allow insertion into an expression vector. A drawback of the dangled primer PCR method is that even if amplification is successful, the DNA yield is often poor. In addition, PCR needs to be performed every time a new batch of recipient vector is needed. This new

method allows us to maintain a circular version of the vector that can be easily propagated in *E. coli* and later obtained by miniprep when needed. This was not possible for recipient vectors made by dangled primer PCR as digestion with *Sap* I also removes the recognition sites from the vector. Hence, if the plasmid were to recircularize, there would be no *Sap* I site to cut the vector again. Additionally, a problem with *Sap* I itself is that enzymatic digestion is never complete due to its low activity. It is impossible to detect a size shift from successful *Sap* I digestion since the plasmid is 6000 bp and the DNA that is removed is only 10 bp. Therefore the extent of digestion cannot be determined. Frequently, the *Sap* I digestion would be repeated before the CIP reaction in an effort obtain a greater number of cleaved product.

6.4.3.1 Sequence design choices

Originally the adapter oligonucleotide was designed for insertion into the MpET-19b using the *Nco* I and *Bam* HI enzymes. However, for unknown reasons, neither double nor sequential digestion of the MpET-19b vector successfully yielded the 80 bp excised region. A revised 152-bp adapter sequence (Figure 6-7) coding for an affinity tag sequence of G-LAAAHHHHHHHH was designed for incorporation into the MpET-41a vector instead. It includes the *Xba* I and *Xho* I sites needed for insertion into the plasmid and the two *Sap* I sites for controlled cloning. The glutamic acid and its adjacent leucine were removed and the codon for the remaining leucine was changed to a higher frequency of usage codon. The histidine codons also were changed to the higher preference CAT codon. The *Nde* I site was added between the two *Sap* I sites so that any circular plasmid DNA present after incomplete *Sap* I digestion could still be linearized by *Nde* I. Since *Sap* I digestion is never complete, some

portion of the plasmid DNA will still be circular. If the vector is not linearized then it will be carried through the ligation and transformation steps, essentially yielding a false positive colony that does not contain the desired insert DNA.

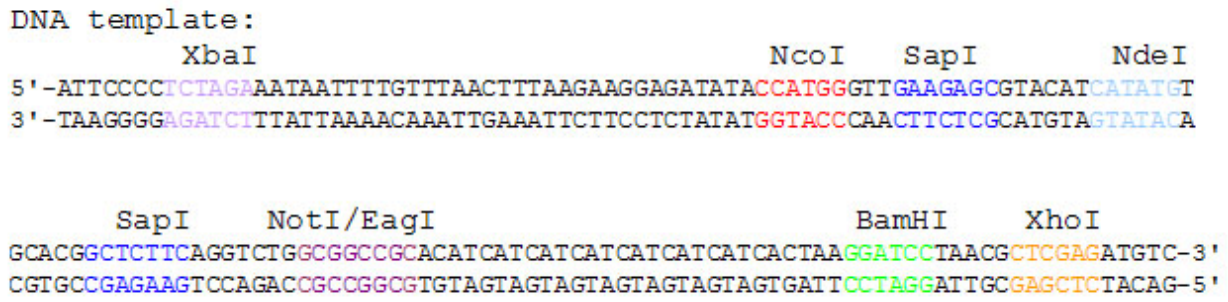


Figure 6-7: Adapter oligonucleotide for replacement of cloning region in MpET-41a to generate a *C*-terminal His tag

6.4.3.2 Assembly PCR and replacement

Due to the size of the oligonucleotide, assembly PCR was used to create the full-length DNA as it would be difficult and expensive to synthesize a single oligonucleotide of 152 bases. Assembly PCR involves using shorter, overlapping fragments of the full sequence in a two-step amplification process to generate the full-length oligonucleotide. The adapter sequence was split into four fragments ranging in size from 51 bp to 63 bp and two flanking primers were designed, shown in Figure 6-8. All DNA had a similar melting temperature (55°C) and were generated using a web-based program with the adapter sequence as the input, available at http://publish.yorku.ca/~pjohnson/AssemblyPCRoligomaker.html [188]. Figure 6-8 also illustrates how the four different oligonucleotides overlap to cover the complete length of the sequence.

Assembly PCR was performed using the provided example conditions and concentrations. For the first thermal cycling step, all four oligonucleotides (resuspended at 12.5 μg/μL in water) were combined into one 50 μL reaction and amplified using the protocol below.

1. 0.25 μL Go*Taq* polymerase (Promega)
2. 0.4 μL 25 mM dNTP
3. 10 μL 5X Go*Taq* buffer
4. 2 μL of each oligonucleotide (8 μL)
5. 31.35 μL water

a) 94°C for 7 min
b) 54°C for 2 min
c) 72°C for 3 min
d) 94°C for 1.5 min
e) 54°C for 2 min
f) 72°C for 3 min
g) Repeat d through f six more times
h) 72°C for 5 min

Figure 6-8: Complementary oligonucleotides and flanking primers for assembly PCR

During the thermal cycling, the oligonucleotides anneal to complementary fragments and the gaps are then filled in by the *Taq* polymerase. Various fragments are randomly extended during each cycle depending on which oligonucleotides encounter each other.

This resulting reaction mixture is then used as the template for a standard PCR amplification using the two flanking primers. These primers ensure that only the full length, correct sequence is amplified and not the various incomplete fragments that were also generated in the first reaction. The primers were resuspended in water to a concentration of 0.25 μg/μL and used in the following protocol for a 100 μL reaction.

1. 0.5 μL Go*Taq* polymerase (Promega)
2. 0.8 μL 25 mM dNTP
3. 20 μL 5X Go*Taq* buffer
4. 4 μL of each primer (8 μL)
5. 1 μL from 1st reaction
6. 69.7 μL water

a) 94°C for 5 min
b) 94°C for 30s
c) 54°C for 2 min
d) 72°C for 1.5 min
e) Repeat b through d 24 more times
f) 72°C for 5 min

Figure 6-9: Agarose gel of assembly PCR products lane 1: 25 bp ladder; lane 2: first reaction; lane 3: second reaction; lane 4: double enzyme digestion of PCR product to generate insert

The assembly PCR product was digested with *Xba* I and *Xho* I to generate the required cohesive ends for insertion into the MpET-41a plasmid. Figure 6-9 is an agarose gel of the products of the two thermal cycling steps and the double-digested PCR product. DNA

sequencing confirmed that the adapter oligonucleotide successfully replaced the cloning region in the MpET-41a vector. The new vector is designated MpET-41a-CHis2 (for version 2 of the adapter) and has G-LAAAHHHHHHHH as the affinity tag sequence.

6.4.3.3 Large-scale expressions of previous PZ8 sequences

The recipient vector was prepared by *Sap* I digestion of the MpET-41a-CHis2 plasmid followed by *Nde* I digestion to linearize any incompletely digested DNA. This was followed by reaction with CIP to prevent re-ligation. PZ8-6, PZm8-6, PZ8-9, PZ8-12, PZm8-12, and PZm8-24 genes were all inserted into the new vector. PZ8-9 and PZm8-24 were transferred from an earlier vector version using *Nco* I/*Not* I. 2 L cultures were expressed for each sequence. MALDI-TOF of each protein showed the presence of an unknown contaminant protein in both 12mer samples as well as PZ8-6 at a mass of 20.8 kDa.

This peak has been observed before in the expression of other protein polymers in our lab but, until now, was absent from drag-tag purifications (all performed on Talon resin). Others have found that filtration (0.2 μM pore size) of dissolved protein samples removed most of the 20.8 kDa peak (results not shown). A native *E. coli* protein, SlyD, is a known potential contaminant of IMAC purifications due to its histidine rich C-terminus (involved in metal binding) [189, 190]; however, this protein has a molecular mass of 27 kDa and has also been shown to have a substantially lower affinity for Talon (cobalt) resin compared to nickel resins [191]. Large amounts of a histidine-tagged protein can outcompete native SlyD for resin binding on nickel resins but our *C*-terminal His-tagged proteins all have low expression levels. Another possible cause is an existing protein polymer sequence in the lab contaminated the cultures, yet

no one has been working with a protein polymer of that size. Samples of PZm8-12 and PZ8-12 were dissolved at ~ 1 mg/mL concentration in water for analysis by RP-HPLC. Despite rigorous mixing, both samples could not be completely dissolved and had to be filtered prior to loading onto the HPLC columns. Samples were analyzed on the C4 and C18 columns but only the solvent peak could be detected. It is possible that most of the purified protein consisted of the 20.8 kDa protein which was insoluble in water and removed by sterile filtering.

Based on suggestions by Clontech (Mountain View, CA) technical support regarding improving protein purity with their Talon resin, the binding time was shortened to 30 minutes (from overnight) and only the minimum amount of resin was used, for better competitive binding, based on expected protein yields. Expression of PZm8-12 was repeated on the 8 L scale. Whereas for the 2 L expression only the 20.8 kDa peak could be detected, for the 8 L expression, a second, smaller peak was observed at the expected molecular weight of the protein, in addition to the still prominent 20.8 kDa peak. Figure 6-10 is a comparison of the MALDI-TOF spectra for the PZm8-12 proteins 2 L and 8 L expression and purifications (expected mass 19.829 kDa). Only 4 mg of protein was recovered for the entire 8 L expression although protein was clearly observed in the elution fractions using SDS-PAGE. Initially it was suspected that the binding time was too short but re-purification of the column flow through fraction did not yield any additional protein. Some precipitate was found after dialysis of elution fractions and protein may have been lost during the transfer into a new tube for lyophilization. The cause of the low yield could not be definitively determined.

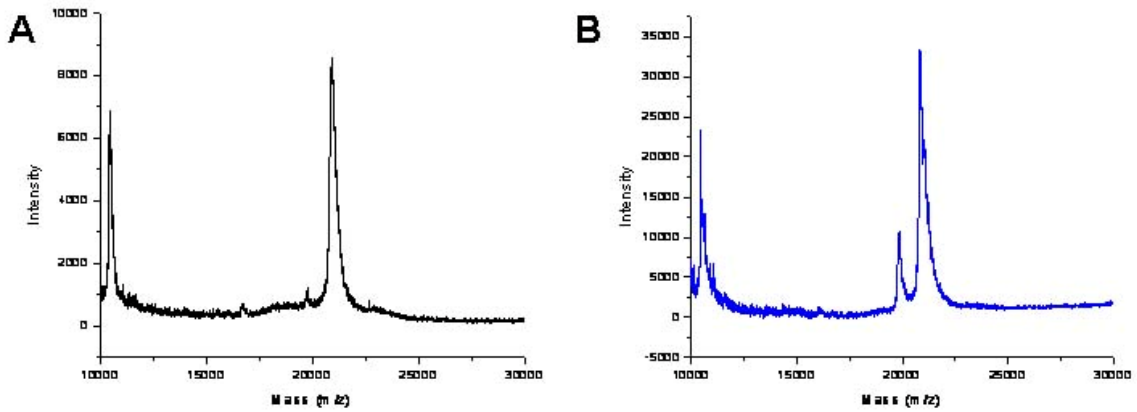


Figure 6-10: MALDI-TOF spectra for PZm8-12 with a *C*-terminal His tag A) 2 L expression B) 8 L expression

Most of the ELFSE analyses of drag-tag-DNA conjugates using these new drag-tags were not successful and no conjugate peaks were observed. However, conjugation was observed for PZ8-12 and PZm8-12 (Figure 6-11). These samples are still not completely monodisperse but instead have a different peak distribution than the *N*-terminal His-tagged proteins (*i.e.,* likely a different source of the additional peaks). Also important to note is that the conjugation efficiency is poor, resulting in significant free DNA peaks and small bioconjugate peaks. The graph has been rescaled so that the smaller conjugate peaks can be better visualized. The α of the major peak in Figure 6-11B is 60 whereas the smaller peak which eluted later has an α of 62.

The histidines appear to be mostly (if not completely) uncharged in these buffer conditions (pH 7.2) otherwise there would be a noticeable jump in α value. This is not surprising since pH 7.0 buffer is used for binding fusion proteins to the positively charged cobalt resin. Additionally, with a pK_a of 6.0 for histidine (imidazole) approximately 6% of the

histidines would be charged at pH 7.2. The Henderson-Hasselbalch equation (pH = pK_a + log {[A^-]/[HA]}) [134] can be used to generate a titration curve, f(HA), where [A^-] is the concentration of conjugate base, [HA] is the concentration of the acid and f(HA) = [HA]/([HA] + [A^-]). At pH 7.2 with a pK_a of 6.0, f(HA) = 0.059 or 6% charged.

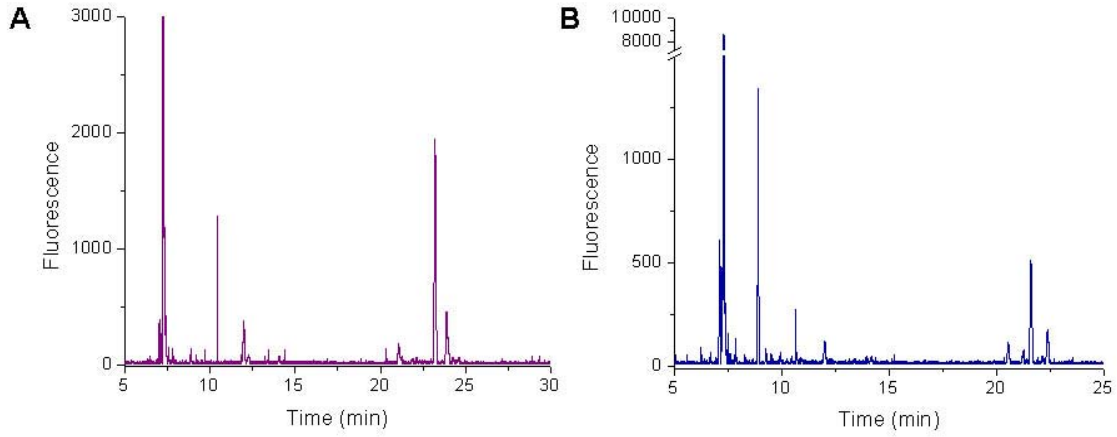


Figure 6-11: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for A) PZ8-12 using a 20-base primer and B) PZm8-12 using a 30-base primer. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 1kV/20s injection, 320 V/cm, 55°C

6.4.3.4 Overnight expression of PZm8-12

Professor David Wood at Princeton University suggested to us that protein polymer expression and/or purity may improve if no IPTG is used for induction and the cells are simply allowed to grow overnight or longer at a reduced temperature. The "leaky" T7 promoter (*i.e.,* basal level protein expression) may yield more protein if grown at a slower rate instead of quickly in a few hours. There is also commercially available media from Novagen (Madison, WI) that is designed for overnight culturing. These two methods were compared to the

traditional growth method for any differences in yield and purity. PZm8-12 with the *C*-terminal His tag was chosen as the test protein.

6.4.3.4.1 David Wood's protocol

According to the protocol provided by David Wood, the culture is grown in TB media at 37°C until the OD_{600} = 0.8-1.0, then it is transferred to a room temperature incubator and grown an additional 24-48 hours. A 1:100 ratio is used to inoculate the day culture from the overnight starter culture. A 25 mL LB starter culture was grown overnight. Ten milliliters of the culture were used to inoculate two 1 L flasks of TB media. After 4 hours the OD_{600} value reached 0.87. Both flasks were then incubated at room temperature for either 24 or 48 hours. Note that the incubators were not refrigerated designs and over time the temperature increased to 30°C from the constant shaking.

6.4.3.4.2 Novagen's Overnight Express™ Autoinduction System

Instant TB media was purchased from Novagen that is specially formulated for long-term cell growth. As per the manufacturer's recommendation, 5% v/v staged inoculations were used to increase the culture size. One colony was grown in 2.5 mL of instant TB media to an OD_{600} of 0.5 (3.5 hours). The entire 2.5 mL culture was transferred to a 60 mL culture of new instant TB and grown again to OD_{600} = 0.5 (3.5 hours). Forty milliliters of the culture was added to a 1 L flask of instant TB media while 20 mL was added to 500 mL instant TB media. The final cultures were grown for 16 hours at 37°C.

6.4.3.4.3 Results of expressions

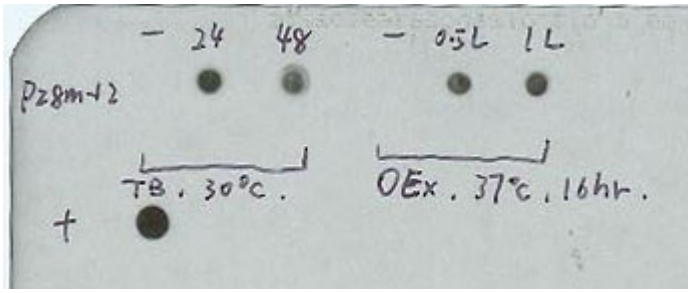


Figure 6-12: T7 "leaky" promoter overnight expression of PZm8-12 where TB corresponds to D. Wood's protocols and OEx is the Novagen Overnight Express™ Autoinduction System

Figure 6-12 is a dot blot of samples from each of the four cultures. The negative controls are from samples obtained prior to overnight incubation of the cultures. The positive control is from a well-expressing protein polymer (50 mg/L). SDS-PAGE of the subsequent column purifications confirmed that protein was successfully expressed, although a smaller molecular weight band was also detected below each eluted protein band but at a lower concentration. As with the large-scale 8 L expression under standard conditions, only a small amount of protein was recovered. Conjugate peaks could not be detected by ELFSE to determine whether or not purity had improved with overnight expression. Nor was there sufficient protein to properly compare which of the four culture conditions had the best protein yield (all appeared similar on the dot blot). Interestingly, MALDI-TOF of the proteins showed only the expected peak and not the 20.8 kDa contaminant.

6.4.3.5 Increasing the protein yield and solubility

It has been observed that water solubility was significantly lower with these recent proteins than past *C*-terminal His tag proteins and the *N*-terminal His tag proteins. This caused

difficulties with the DNA conjugation and possibly contributed to the overall poor results and low yields. Specifically, insoluble protein is filtered out with the excess sulfo-SMCC after the maleimide-activation reaction with approximately ~ 10% protein recovered (which is then resuspended at the 10 mg/mL concentration). Poor solubility may have contributed to the unsuccessful HPLC analyses as well. Elimination of the hydrophilic glutamic acid but not the hydrophobic leucine may have led to the observed reduction in protein polymer solubility.

A new version of the *C*-terminal His tag would eliminate the leucine and simplify the sequence to just G-AAAHHHHHHHH but there is still the existing problem of the very low protein yields. A possible solution would be to introduce a small leader peptide at the N-terminus of the protein that would aid in expression. A T7 tag is used by the Tirrell group for some of their protein polymer expressions [192]. This T7 tag is placed at the N-terminus of the protein and consists of 11 amino acids from the N-terminus of the most abundant gene in the T7 phage and is believed to aid in protein expressibility. The sequence is MASMTGGQQMG and it can also be used for antibody-based affinity chromatography. It is small enough that it can be left attached to the protein like the histidine tag, assuming methionine and glutamine do not cause complications downstream.

Two variations of the *C*-terminal His tag expression vector (MpET-41a-CHis3) were generated. One version simply eliminated the leucine in the affinity tag sequence whereas the other version also introduced a T7 tag into the vector at the *N*-terminal position. Both sequences were designed and produced in the same manner as the previous version of the expression vector

by using assembly PCR and restriction enzyme digestion to replace the cloning region in the plasmid.

6.4.3.6 Revised linker sequence with no T7 tag

Elimination of only three bases (leucine codon) from the original design was not likely to dramatically affect the melting temperature of the oligonucleotide. Hence, only the third oligonucleotide, which overlaps the region where leucine is present, was replaced with a shorter version, 5'-CAT CAT ATG TGC ACG GCT CTT CAG GTG CGG CCG CAC ATC ATC ATC ATC ATC ATC ATC AC-3'. The other three oligonucleotides and the two flanking primers from Figure 6-8 were used as before in assembly PCR to generate the adapter sequence.

6.4.3.7 Revised linker sequence with T7 tag

Inclusion of the T7 tag required that a new adapter sequence, 179 bases long, be constructed, as shown in Figure 6-13 along with the six overlapping oligonucleotides. The same flanking primers were used for the second thermal cycling reaction as in previous assembly PCR protocols. Note that the beginning Met residue of the T7 tag will be removed during processing of the protein as it is acting as the ATG start codon. Fortuitously, the final glycine of the T7 tag uses the GGT codon, thus it was used as the first glycine of the protein polymer sequence.

6.4.3.8 Test expression of PZm8-12 in both vector versions

PZm8-12 was inserted into both recipient vectors as the test protein for the new expression vectors. A test expression, using TB media and 3 hour induction with 1 mM IPTG, was started using three colonies from each vector version but one of the selected T7 tag colonies did not grow overnight. Figure 6-14 is a dot blot comparing the two T7 tag colonies to three

colonies of the vector consisting of only a *C*-terminal His tag. The dots for the T7 tag proteins are much greater in intensity compared to proteins without the T7 tag, indicating that addition of the T7 tag was leading to higher protein expression levels as intended.

```
DNA template:

5'-
ATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGCTAGCATGACTGGTGGACAGCAAATGGGTTGAAGAGCGTACATCATATGTG
CACGGCTCTTCAGGTCTGGCGGCCGCACATCATCATCATCATCATCATCACTAAGGATCCTAACGCTCGAGATGTC-3'
length: 179

Output:

6 assembly oligos:

5'-ATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACC-3'
length: 48

5'-CACCAGTCATGCTAGCCATGGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTC-3'
length: 56

5'-GGCTAGCATGACTGGTGGACAGCAAATGGGTTGAAGAGCGTACATCA-3'
length: 47

5'-CCGCCAGACCTGAAGAGCCGTGCACATATGATGTACGCTCTTCAACCC-3'
length: 48

5'-GCTCTTCAGGTCTGGCGGCCGCACATCATCATCATCATCATCACTAAGGA-3'
length: 53

5'-GACATCTCGAGCGTTAGGATCCTTAGTGATGATGATGATGATGATG-3'
length: 49
```

Figure 6-13: Adapter oligonucleotide for replacement of cloning region in MpET-41a to generate an *N*-terminal T7 tag and a *C*-terminal His tag using six overlapping oligonucleotides in assembly PCR

6.4.3.9 Large-scale expression results

PZm8-12 was expressed under standard growth and induction conditions (TB media, 3 hour induction at 1 mM IPTG) and also overnight without any IPTG added.

6.4.3.9.1 Under standard growth and induction conditions

The increased protein yield with a T7 tag was confirmed with large scale (4 L) expressions of the two variants. 21.0 mg of purified T7 tag protein was recovered compared to 7.5 mg of the protein with only a *C*-terminal His tag. In addition, it was observed by SDS-PAGE (Figure 6-15) that a fraction of the T7 tag protein had apparently eluted prematurely in the second wash (25 mM imidazole buffer).

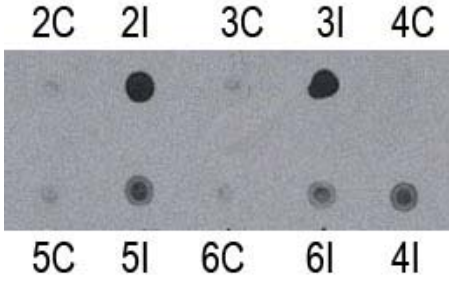


Figure 6-14: Dot blot of PZm8-12 test expression colonies where C = control, I = induced. Colonies #2, 3 have both T7 tag/CHis tag whereas #4, 5, 6 have CHis tag only

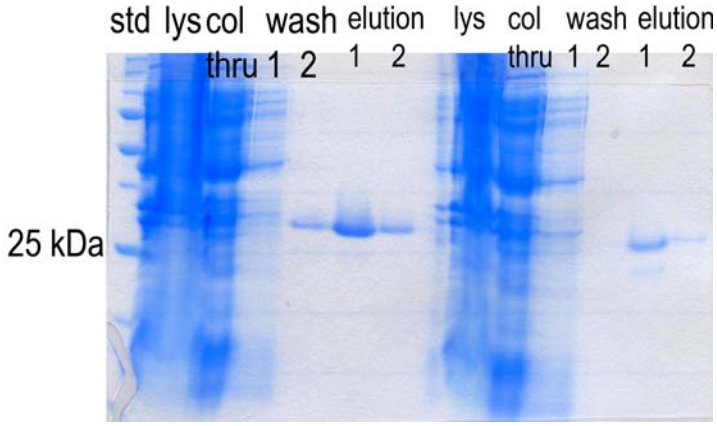


Figure 6-15: 12% SDS-PAGE gel of PZm8-12 purification. Left: Protein with T7 and CHis tags Right: protein with CHis tag only

Approximately, 24.4 mg of protein was recovered from the second wash which appeared to contain only the T7 tag protein and not other wash contaminants based on the gel analysis. The recovered 24.4 mg of protein was rerun through column purification using a lower, 10 mM imidazole second wash. However, we observed that a significant portion of the protein still eluted early in the second wash for unknown reasons. The amount of resin used for the column purification (20 mL) is enough to bind over 100 mg of protein. Therefore, the resin was not

close to maximum capacity. MALDI-TOF mass spectrometry confirmed that the protein in the wash and elutions is the same protein. Both proteins were observed to have improved solubility (with the T7 tag imparting greater improvement compared to just elimination of the leucine in the *C*-terminal His tag) and the conjugation reactions were analyzed by ELFSE. However, despite improvements in water solubility, poor drag-tag conjugation was still observed and the electropherograms were similar to those in Figure 6-11.

6.4.3.9.2 T7 leaky promoter overnight expression

The overnight expressions of PZm8-12 with the Leu-containing affinity tag demonstrated that the protein polymer could be successfully expressed (according to the dot blot of cell lysates). The low purified protein yields were likely due to the poor solubility of the protein and not the expression method itself. Therefore overnight expression using David Wood's protocol was repeated but with the newest *C*-terminal affinity tag and the additional T7 tag variant. Refrigerated incubators were also available for these expressions, allowing the temperature to be maintained at a constant 25°C.

Four different 1 L cultures were expressed using the two variants of PZm8-12 and the different 24 hour and 48 hour incubation times. All four cultures were purified by IMAC and the fractions were visualized by SDS-PAGE. The 10 mM imidazole concentration in the wash buffer was reduced to 5 mM in an effort to minimize any premature elution of the T7 tag protein.

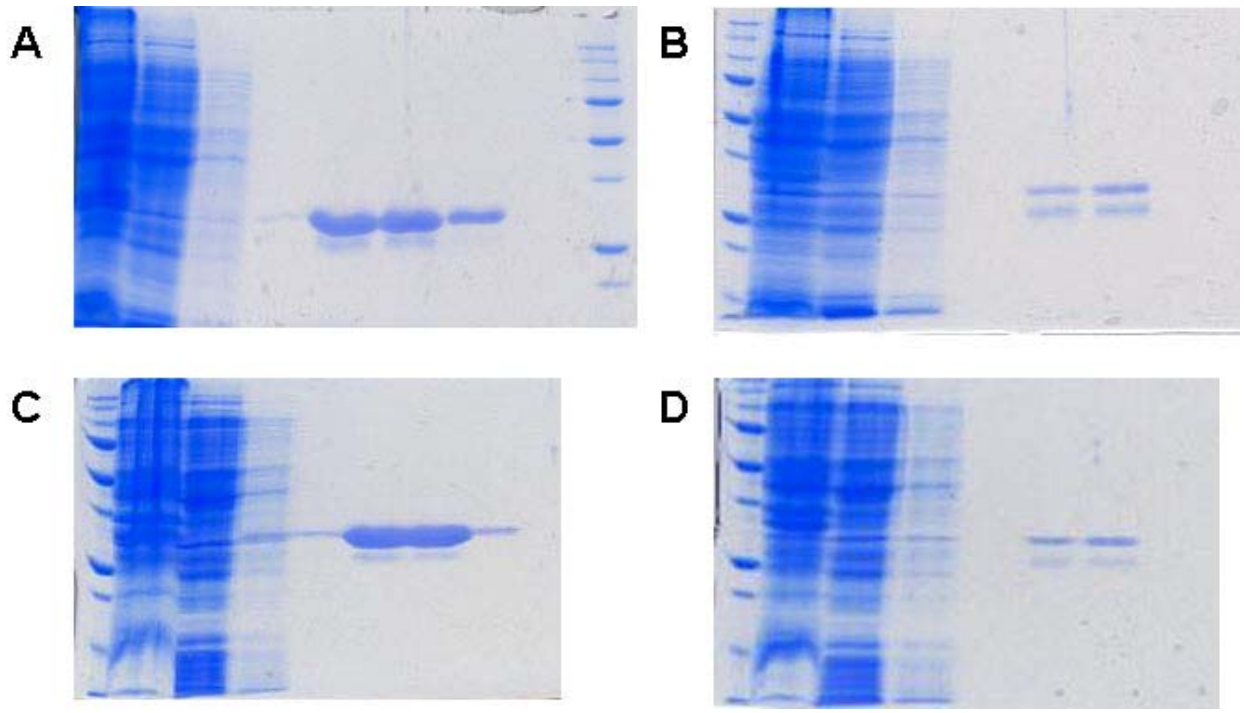


Figure 6-16: 12% SDS-PAGE gels of PZm8-12 overnight expression using the "leaky" T7 promoter. From left to right, fractions are lysate, column flow through, wash 1, wash 2, elution 1, elution 2, elution 3 A) T7 tag protein 24 hours B) CHis tag only protein 24 hours C) T7 tag protein 48 hours D) CHis tag only protein 48 hours

However, as seen in the SDS-PAGE gels in Figure 6-16, premature elution of the T7 tag proteins still occurred in the wash buffer. As expected, the T7 tag proteins had much higher levels of protein expression, but another band at a lower molecular weight was observed in all four purifications. The final yield of the proteins were 13 mg/L and 10 mg/L for the T7 tag 24 hour and T7 tag 48 hour expressions, respectively. The yield of the 24 hour and 48 hour expressions with no T7 tag were 2 mg/L and 1 mg/L, respectively. Twenty-nine milligrams of protein were recovered in the second wash of the T7 tag 24 hour expression and 44 mg for the

T7 tag 48 hour expression. However, mass spectrometry on both samples showed the presence of impurities along with the target protein.

Table 6-2 compares the yields from the two variants under standard and overnight expression conditions to PZm8-12 expressed with an *N*-terminal His tag. Overnight expression appeared to result in better yields for protein with the T7 tag (if the prematurely eluted proteins from the second wash of the T7 tag purifications are all omitted) but no noticeable difference for the *C*-terminal His tag only version.

MALDI-TOF analysis of the proteins confirmed the molecular mass of the purified proteins. An additional peak was detected at 11 kDa in the T7 tag 24 hour spectrum, which cannot be attributed to a doubly charged ion. This peak may correspond to the unidentified band observed on the SDS-PAGE gels. There was no discernable difference between drag-tag-DNA conjugates using protein expressed under standard conditions compared to protein expressed overnight at 25°C.

Table 6-2: PZm8-12 yields under different vectors and induction conditions. Number in parenthesis indicates yield when elution combined with protein recovered in second wash.

PZm8-12 version and expression conditions	**Yield (mg/L)**
N-terminal His tag for 3 hours with inducer	20
T7 tag & CHis tag for 3 hours with inducer	5.3 (11.4)
CHis tag for 3 hours with inducer	2
T7 tag & CHis tag for 24 hours	13
T7 tag & CHis tag for 48 hours	10
CHis tag for 24 hours	2
CHis tag for 48 hours	1

6.5 Protease cleavage of affinity tag to improve conjugation reaction efficiency

A reduction in the efficiency of the conjugation reaction between drag-tag and DNA has been observed when the *C*-terminal His tag was left attached to the drag-tag. While this issue is not a major problem when simply checking the monodispersity of a potential drag-tag, the low conjugation efficiency becomes problematic if the protein polymer is to be used for actual DNA sequencing. First, with a 100-fold molar excess of drag-tag being added to DNA per reaction, essentially 1 mole of drag-tag would be conjugated to DNA out of 1000 moles of drag-tag (assuming 10% successful conjugation). Additionally, with poor conjugation, peak heights in a sequencing electropherogram will not be large enough to detect. If 10% of the DNA is

conjugated to drag-tag, then only 10% of the generated Sanger fragments will have drag-tag attached. When this quantity is divided by the number of bases or peaks in the electropherogram (*i.e.,* 180 bases or more, ideally), the peaks would be impossibly small to read. To examine recent conjugates of drag-tag to a *single* DNA size (*e.g.,* 30mer primer), significant magnification is required for visualization. Separation/filtration after conjugation of drag-tag to DNA primer is typically not performed but most likely such a purification step would lead to further loss of the drag-tagged primers. Enzymatic ligation or another post-PCR conjugation method are appealing here as ways to circumvent this problem.

It was initially believed that the low conjugation reaction efficiency was solely the result of the poor solubility of the protein polymer. However, additional modifications to the affinity tag along with attachment of the T7 tag both improved protein polymer water solubility while seemingly having no effect on the conjugation reaction efficiency. We theorized that the histidines in the affinity tag are somehow interfering with the conjugation reaction. Perhaps the histidines are interacting electrostatically with the DNA primer. Even though the affinity tag is at the C-terminus it may be blocking access to the N-terminus by sulfo-SMCC. Additional peaks seen in the ELFSE electropherogram may be due to formation of side products.

A plausible explanation is that histidine is reacting with the sulfo-SMCC reagent (although it is a common protein crosslinker). We have since discovered that histidine can react with *N*-hydroxysuccinimide (NHS) esters, effectively accelerating the rate of hydrolysis of the NHS groups in solution (unstable reaction product rapidly hydrolyzes) [193, 194]. The NHS-ester reaction is performed first (*i.e.,* drag-tag activation) to minimize hydrolysis as it is less

resistant to hydrolysis than the maleimide group in sulfo-SMCC [193]. The histidines on the affinity tag may essentially be accelerating the hydrolysis of the reagent. Unlike a natural protein, there is only a single primary amine at the N-terminus of the protein polymer which may not be a strong enough nucleophile compared to the eight adjacent histidines at the C-terminus. Thus the sulfo-SMCC reagent preferentially reacts with the histidines, accelerating hydrolysis of the crosslinker and thus rendering it ineffective for conjugation as the crosslinker is now two separate molecules. Higher concentrations of sulfo-SMCC would be required to overcome this behavior. However, recently a 100-fold excess of sulfo-SMCC reagent has been tested as opposed to the standard 10-fold molar excess. There was no noticeable improvement in conjugation efficiency.

As previously reported in Section 6.3.5, cyanogen bromide cannot be used to remove the *C*-terminal affinity tag as the cleavage reaction will lead to the creation of at least two different forms of the protein polymer. The other option is to use a site-specific protease to remove the affinity tag. This option was formerly discounted given that it would inevitably lead to the addition of extra amino acids at the C-terminus of the protein from the protease recognition site. The commonly used proteases for affinity tag removal were shown in Table 6-1 along with their recognition and cleavage sites.

6.5.1 Factor Xa protease

As no better option currently exists, a site-specific protease recognition site was incorporated into the affinity tag between the protein polymer and the *C*-terminal affinity tag. This change allows for removal of the histidines after purification, but also adds several extra

amino acids to the C-terminus of the protein. Factor Xa was chosen since it adds only four additional amino acids (IEGR) for its recognition site. Three amino acids have already been used in past or present protein polymer designs and are not expected to cause complications. Only one hydrophobic residue (isoleucine) is added to the protein. The negative charge of the glutamic acid is counteracted by the addition of a positively charged arginine. Adding two charged residues may also balance out the hydrophobicity of the isoleucine.

6.5.1.1 Insertion of the IEGR sequence into existing PZm8-12 plasmids

The IEGR recognition site was inserted into existing PZm8-12 MpET-41a-T7tag/CHis3 vector and PZm8-12 MpET-41a-CHis3 vector by replacement of the *C*-terminal His tag with a 76-bp oligonucleotide sequence, 5'-CAG GTG CGG CCG CA**A TCG AGG GAA GG**C ATC ATC ATC ATC ATC ATC ACT AAG GAT CCT AAC GCT CGA GCA CCA C – 3'. This sequence includes IEGR (bold) and eight histidines with flanking *Not* I and *Xho* I restriction enzyme sites (underlined) for insertion into the plasmid. The existing *C*-terminal His tag was removed from the plasmid by double digestion then replaced with the new sequence. Using this method, sequences already existing in the expression vector can be quickly modified to test the cleavage strategy before making final changes to the adapter design for generating the new recipient vector. The PCR-amplified oligonucleotide was successfully digested and inserted into the two PZm8-12 vectors, as confirmed by DNA sequencing. The revised affinity tag sequence is now G-AAAIEGRHHHHHHHH.

6.5.1.2 Large-scale expression and purification

Three liters of each protein were expressed under standard conditions. For the IMAC step, the column was washed with double the volume of the first buffer (no imidazole) instead of one wash each of the zero imidazole buffer and the 5-10 mM imidazole buffer. A thin band of pre-eluted protein was still detected in the second wash but this time for both proteins even without the T7 tag. As before with other recent PZm8-12 expressions, a faint band was observed at a lower molecular weight in addition to the expressed protein in the elution fractions. 10 mg (3.3 mg/L) of the T7 tag protein was recovered compared to 5 mg (1.7 mg/L) of the *C*-terminal His tag-only protein. The yields were comparable to previous expressions of the protein without the IEGR insertion for standard conditions (Table 6-2).

6.5.1.3 Assay of cleavage reaction conditions

A Factor Xa cleavage/capture kit was purchased from Novagen (Madison, WI) which consists of Factor Xa protease, Xarrest agarose for protease removal, a control protein, and a concentrated buffer solution. Factor Xa : target protein ratios (unit : μg) of 1:100, 1:50, and 1:20 were tested initially. Ten micrograms of protein were digested by varying amounts of enzyme (0, 0.1, 0.2, 0.5 units) in a 50 μL reaction at 20°C. Ten microliters of sample were taken at 2, 4, 8, and 16 hour time intervals and immediately mixed with 10 μL of SDS-containing sample buffer for future SDS-PAGE analysis and to halt the cleavage reaction. Both the T7 tag protein and the *C*-terminal His tag-only protein were tested (with IEGR insert). The control protein was digested with 0.1 units of enzyme for 2 μg protein for 16 hours.

All samples were analyzed by SDS-PAGE. Digestion of the control protein confirmed that Factor Xa was active. However, no other protein bands could be detected. It is likely that the PZm8-12 proteins do not stain well at the amount present on the gel (~ 0.5 μg). Therefore a Western blot was performed to detect the presence of any protein with a His tag. A Western blot is similar to a dot blot except that the samples are first separated by size using SDS-PAGE and then transferred to the membrane. Then the same protocol is used as for the dot blot for detection of proteins with a His tag by antibodies and chemiluminescence.

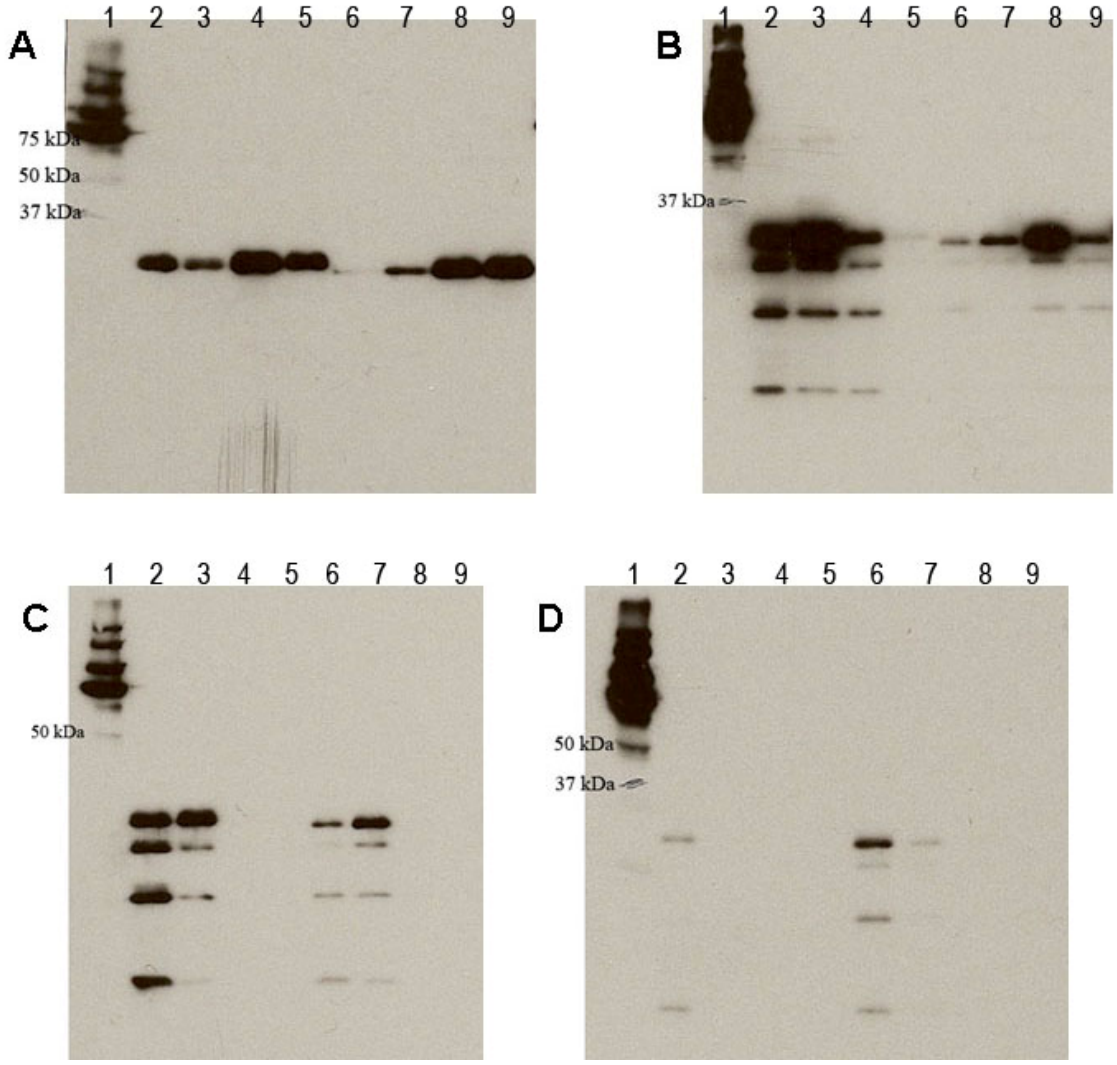


Figure 6-17: Western blot of Factor Xa digestion of PZm8-12 protein. Lane 1: ladder; lanes 2-5: T7 tag protein after 2, 4, 8, 16 hrs; lanes 6-9: CHis tag only protein after 2, 4, 8, 16 hours. A) no protease added B) 1:100 unit:µg protein C) 1:50 unit:µg protein D) 1:20 unit:µg protein

Figure 6-17 shows the results for the four different Factor Xa concentrations tested on the two versions of the protein. The control blot shows a single band in each lane, corresponding to

the uncleaved proteins. However, in each of the other three blots where varying amounts of Factor Xa were added, four bands (including the uncleaved protein) are observed. Eventually the His tag is removed from all of the target protein in the sample and the protein can no longer be visualized as indicated by a blank sample lane. It is likely that Factor Xa is cleaving at a secondary recognition site, GR, which is present in four locations in the PZm8-12 protein. The expected molecular weights produced by these secondary cleavages are presented in Figure 6-18 for fragment sizes with the His tag attached (*i.e.,* visible by Western blot detection). The lower molecular weight bands most likely migrated off the gel during electrophoresis, explaining their absence from the Western blot.

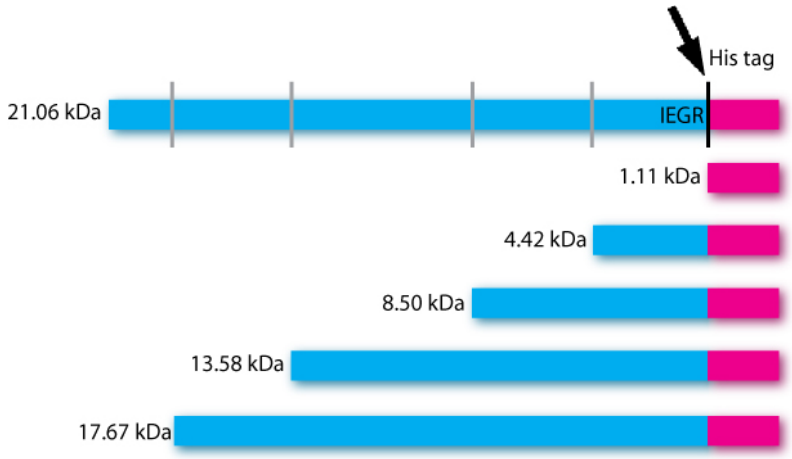


Figure 6-18: Factor Xa potential cleavage sites in PZm8-12 protein containing T7 tag and *C*-terminal His tag. Preferred cleavage site is marked in black, secondary in gray.

The reaction conditions were modified in an effort to minimize cleavage at secondary sites. The second set of reactions used ten times diluted protease concentrations of 0.01 and 0.05 units per 10 μg target protein at 25°C and 4°C. Enzyme concentrations of 0.1 and 0.5 units per

10 μg target protein were also tested at 4°C. The extra bands are present in all the reactions even after only 2 hours of enzyme addition as seen in Figure 6-19. Moreover, a large amount of uncleaved protein is still present even after 16 hours. Analysis of the protease confirms that it does not possess a His tag and is therefore not contributing to one or more of the bands seen. Analysis of just the T7 tag protein alone suggests that it is not as pure as originally thought (previous experiment showed a single band). In addition to a smaller, lower molecular weight band that had earlier been assumed to be from secondary protease cleavage, there is a larger band at ~ 60 kDa which is present in all the samples, including the pure T7 tag protein. Regardless of the original sample purity, the change in reaction conditions reduced but did not prevent enzymatic cleavage at both the preferred IEGR site and at secondary GR regions. Clearly it will be challenging to use Factor Xa to cleave the affinity tag from a protein polymer containing arginine residues.

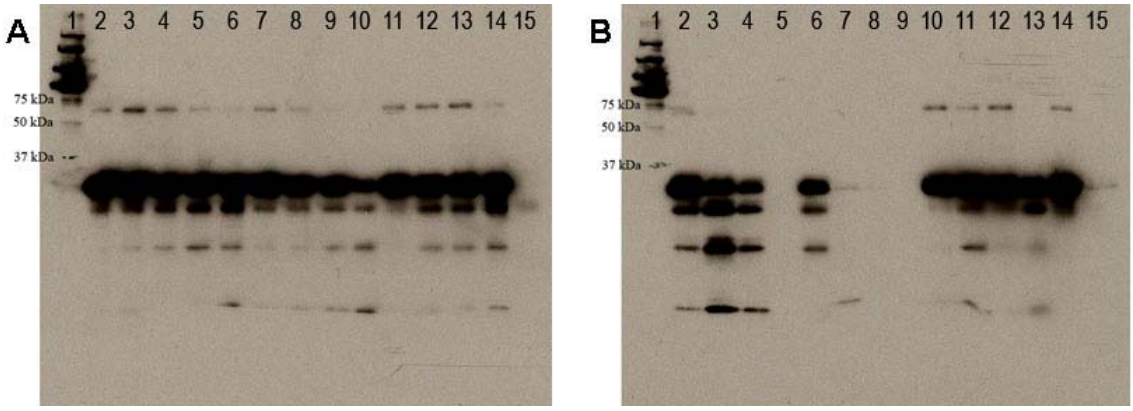


Figure 6-19: Western blot of Factor Xa digestion of PZm8-12 protein with T7 tag/CHis tag. A) lane 1: ladder; lane 2: pure PZm8-12 in water; lanes 3-6: 1:1000 unit:µg protein after 2, 4, 8, 16 hrs at 25°C; lanes 7-10: 1:200 unit:µg protein after 2, 4, 8, 16 hrs at 25°C; lanes 11-14: 1:1000 unit:µg protein after 2, 4, 8, 16 hrs at 4°C; lane 15: Factor Xa B) lane 1: ladder; lanes 2-5: 1:100 unit:µg protein after 2, 4, 8, 16 hrs at 4°C; lanes 6-9: 1:20 unit:µg protein after 2, 4, 8, 16 hrs at 4°C; lanes 10-13: 1:200 unit:µg protein after 2, 4, 8, 16 hrs at 4°C; lane 14: pure PZm8-12 in water; lane 15: Factor Xa

6.5.2 Affinity tag removal by endoproteinase-GluC

The insertion of the IEGR Factor Xa recognition site into the plasmid introduced a single glutamic acid into the *C*-terminal affinity tag. Endoproteinase GluC is a serine protease that can cleave specifically after Glu residues (and at Asp residues at 100-300 times slower rate but no aspartic acid residues are present in our sequences). This enzyme is typically used for peptide digestion and identification using mass spectrometry and not for affinity tag cleavage. Therefore there is no resin available specifically designed to "capture" the protein after the cleavage reaction. It is sold by various vendors but the version sold by New England BioLabs (Ipswich, MA) includes a histidine tag at its C-terminus. Consequently, after protease digestion, the cleaved His tag, uncleaved protein, and the protease can all be removed in a single chromatographic step.

6.5.2.1 Assay of cleavage reaction conditions

Digestion was done at 25°C in the provided reaction buffer using the protease : target protein (μg:μg) ratios of 1:100, 1:50, and 1:20. These reactions were monitored over the course of 8 hours and a second reaction was setup to run for 16 hours. Western blot was done for all the reactions. Due to the limited quantity of purified *C*-terminal His tag-only PZm8-12 protein, only the T7 tag version was tested with this protease. Results show that cleavage was successful in less than 8 hours for all protease concentrations tested (Figure 6-20). The endoproteinase GluC has a different mass than the PZm8-12 protein and thus they are easily distinguishable on the blot. Curiously, faint target protein bands are detected in all three 16 hour reactions. These unexpected bands may be due to the resuspended enzyme rapidly losing activity during the 8 hours the protease was stored frozen in solution before it was used again for the 16 hour reactions. Another explanation is that long term incubation with the protease may lead to the formation of side products.

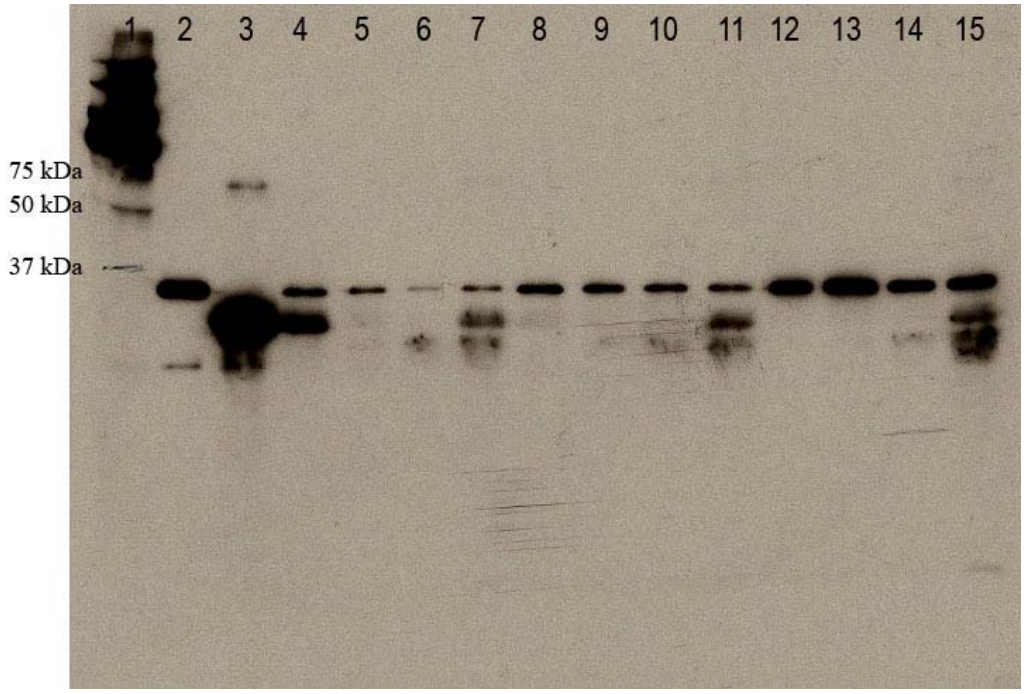


Figure 6-20: Western blot of endoproteinase GluC digestion of PZm8-12 protein containing T7 tag/CHis tag. Lane 1: ladder; lane 2: endoproteinase GluC; lane 3: PZm8-12 protein; lanes 4-7: 1:100 μg:μg protein for 2, 4, 8, 16 hrs; lanes 8-11: 1:50 μg:μg protein for 2, 4, 8, 16 hrs; lanes 12-15: 1:20 μg:μg protein for 2, 4, 8, 16 hrs

6.5.2.2 Cleavage of the affinity tag and protease removal

New lyophilized endoproteinase GluC was purchased for the large-scale reactions. Approximately 4-5 mg of each protein (with and without T7 tag) were cleaved by 50 μg of enzyme for 6 hours at room temperature in a final buffer volume of 12.5 mL. The reactions were dialyzed overnight using 3500 MWCO membrane to remove salts and the cleaved His tag. The samples were then purified by IMAC using 3 mL of Talon resin. Enough wash buffer was added so the total volume of the flow through and wash fractions, containing protein with the His tag removed, would be approximately 30 mL. Uncleaved protein and the His-tagged protease were

collected in 15 mL of elution buffer. Five milligrams of the T7 tag protein were recovered after protease cleavage as well as after removal of any His tag-containing proteins by IMAC. However, only 1.3 mg was recovered in the flow through for the other protein following IMAC while 2.6 mg was obtained in the elution fraction.

MALDI-TOF was used to analyze the protein molecular weights post-cleavage. The measured masses correspond to the expected sizes of the cleaved proteins, confirming successful removal of the affinity tag. For PZm8-12 with only a *C*-terminal His tag, MALDI-TOF showed that an impurity (10.8 kDa) was present in the sample during the cleavage reaction. This impurity was later removed during IMAC where it was detected solely in the elution fraction by mass spectrometry.

6.5.2.3 ELFSE analysis of cleaved proteins

The cleaved drag-tags were then conjugated to DNA and analyzed by free solution capillary electrophoresis. The results are presented in Figure 6-21. Removal of the *C*-terminal affinity tag not only resulted in improved conjugation efficiency (as evidenced by the higher signal of the conjugate peaks relative to free DNA) but also in nearly monodisperse protein. These results are in sharp contrast to the electrophoretic analysis of the same protein sequence except with the His tag still attached. A couple minor peaks of unknown origin can be observed in the electropherogram for protein expressed with only the *C*-terminal affinity tag (less pronounced in the other electropherogram with T7 tag). These may be due to protease cleavage at other sites along the affinity tag such as the G or R residues adjacent to the glutamic acid.

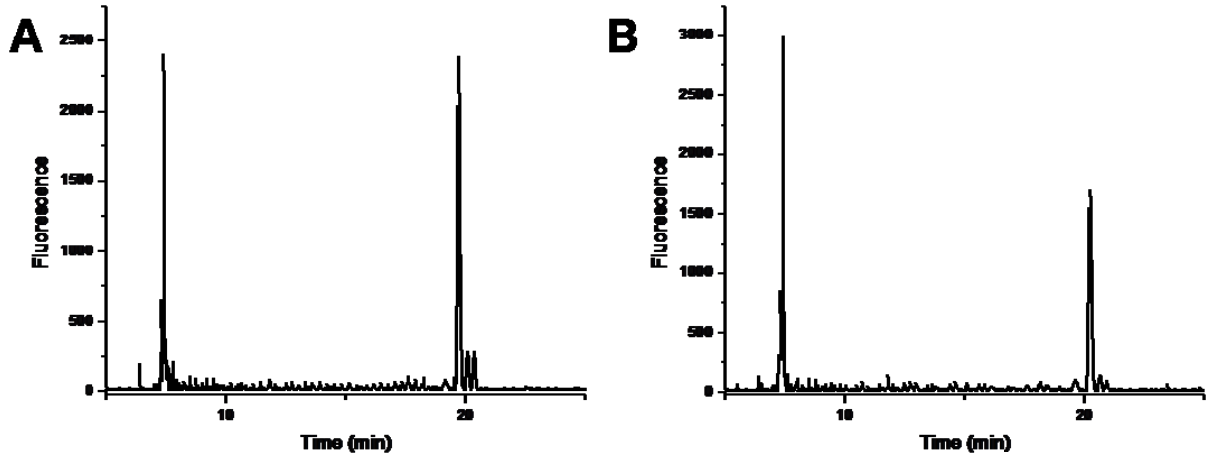


Figure 6-21: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for PZm8-12 using a 30-base primer A) expressed with only *C*-terminal His tag later removed by protease B) expressed with T7 tag and *C*-terminal His tag with *C*-terminal affinity tag removed by protease. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 1kV/20s injection, 320 V/cm, 55°C

The α values for the two versions are 49 for the protein expressed with the *C*-terminal affinity tag and 51 for the protein expressed with both tags (the additional mass from the T7 tag leads to slightly higher α). As expected, the presence of the negatively charged glutamic acid at the C-terminus led to a reduced hydrodynamic drag compared to PZm8-12 expressed with an *N*-terminal His tag (~ 55). Nevertheless, these proteins are of comparable purity to PZm8-6, which was used successfully for DNA sequencing, but with nearly double the hydrodynamic drag. DNA sequencing with these new drag-tags is currently underway.

We can calculate the expected sequencing read length using the equation for peak resolution shown below [121] and setting R = 1 for single peak resolution.

$$R = 4\left[\frac{\ln(2)D_1}{\mu_0 EL}\right]^{0.5}\left[\frac{M_c{}^{0.5}(M_c+\alpha)}{\alpha}\right]^{1.25} \qquad (6\text{-}1)$$

where D_1 is 7 x 10^{-6} cm^2/s [121], μ_0 = 2.6 x 10^{-4} cm^2/Vs, E = 312.8 V/cm, and L = 36 cm, based on experimental conditions. Solving for M_c when R = 1 and α = 51, we obtain M_c = 157 bases. However, the human eye or computer software would likely be able to identify peaks past this point of single peak resolution. For the previous sequencing analysis of M13mp18 using PZm8-6 [123], this equation fairly accurately predicts R = 1 between 100 and 120 bases, although bases past this point could be easily identified with knowledge of the M13mp18 sequence.

6.6 Conclusions and Recommendations

Introduction of a *C*-terminal His tag for protein expression and purification has yielded protein polymers that are more pure than large proteins (> 127 amino acids) expressed previously with an *N*-terminal His tag. A survey of the major protein polymer research groups presented in Table 1-1 revealed that only one group, Cappello with Protein Polymer Technologies, exclusively uses *C*-terminal histidine tags over *N*-terminal histidine tags for protein expression. Elastin tags used for inverse transition cycling purification are typically placed at the C-terminus as well. Kaplan used an *N*-terminal T7 tag coupled with a *C*-terminal histidine tag for expression of serine-rich protein polymers [92]. In general, for most applications of protein polymers, absolute purity is not as crucial as it is for drag-tags for ELFSE. Obtaining a completely monodisperse protein polymer was undoubtedly a challenging task.

Cyanogen bromide cleavage of the affinity tag has been shown to generate two similar drag-tag species. Similarly, not removing the His tag has led to problems with poor protein-

DNA conjugation. An IEGR recognition site for Factor Xa was introduced for affinity tag removal but it was later discovered that the protease would cleave at one or more GR sites in the protein instead of solely the preferred recognition site. We have shown that endoproteinase GluC can selectively cleave after glutamic acid residues and be used as an alternative method to enzymatically remove the affinity tag. ELFSE analysis of the drag-tag-DNA conjugates has shown that this protein (PZm8-12), expressed with a *C*-terminal affinity tag that is later removed by endoproteinase GluC, is nearly monodisperse. These drag-tags (PZm8-12, with and without the T7 tag) are currently being tested for DNA sequencing. Larger proteins such as PZm8-24 which has an α of 130 (but was polydisperse when expressed with an *N*-terminal affinity tag) will need to be tested to determine if a longer length, pure drag-tag is possible using the same strategy.

Incorporation of the short T7 tag sequence into the N-terminus of the protein has resulted in improved yields with a *C*-terminal His tag, although not to the same level of proteins expressed with an *N*-terminal His tag. Overnight expression without an inducer did not lead to any significant change in protein yields compared to a 3 hour induction with IPTG. Presently, addition of the T7 tag does not appear to be detrimental to the drag-tag although further testing (*i.e.,* thermal cycling and DNA sequencing) would need to be performed. Nevertheless, the promising yields have led to the expression of newer sequences with the T7 tag, to be discussed in Chapter 7.

Future designs of the expression vector and, hence, the adapter oligonucleotide may incorporate an enzymatic cleavage site between the T7 tag and the protein so the short tag can be

removed after protein expression if desired. Additionally, since the affinity tag is again being removed, other small affinity tags can be reconsidered that were previously discounted due to the presence of undesirable amino acids such as lysine. For example, *Strep*-tag II (WSHPQFEK) or FLAG (DYKDDDDK) can be combined with the His tag for two orthogonal purifications for improving protein purity. The T7 tag itself can also be used for antibody-based affinity chromatography.

Ultimately, self-cleaving intein affinity tags may be the best choice. As discussed in Chapter 4, self-cleaving tags avoid the expense and cleanup of protease cleavage, particularly when processing larger amounts of protein. In addition, site-specific proteases require a recognition site for cleavage which adds several, potentially undesirable amino acids to the C-terminus of the protein polymer, whereas intein cleavage does not. Factor Xa can be used to cleave only non-arginine containing sequences while endoproteinase GluC can only be used to remove the affinity tag from proteins without Glu residues in the repetitive sequence. Additionally, introduction of the negatively charged glutamic acid for the sole purpose of enzymatic cleavage leads to a decrease in the hydrodynamic drag of the drag-tag. Applying self-cleaving *C*-terminal affinity tags to drag-tag expression and purification is being investigated by another Ph.D. student in the lab, Xiaoxiao Wang.

Chapter Seven

Expanding Upon the PZ8 Series Protein Polymers: New Variations

7.1 Introduction

New sequence designs are in development based on our experiences with the PZ8 and PZm8 protein polymer series. A 127-amino acid protein polymer (PZm8) was successfully used as a drag-tag for ELFSE DNA sequencing (Section 5.3). Originally based on the PZ8 sequence (GAGTGSA), PZm8 was discovered to have two serine to arginine mutations, resulting in an α value of ~ 25 (an improvement over the α of 20 for the PZ8-6 uncharged version). The addition of two positively charged arginines did not result in any of the expected detrimental interactions with the negatively charged DNA or microchannel walls. These results required us to revise the anticipated ideal drag-tag properties, previously listed in Section 1.3.2, which included a preference that the proteins be uncharged. It became apparent that a few arginines in a sequence not only can benefit the drag-tag by increasing water solubility, but also can boost the hydrodynamic drag by "pulling" the protein in the opposite direction of the DNA in an electric field. Indeed this was seen for the larger PZm8 proteins that included 4 or 8 arginine mutations out of 253 and 505 amino acids, respectively, compared to their uncharged counterparts.

It has also become apparent from the extensive testing done on longer-length protein polymers and the use of a *C*-terminal His tag (Chapters 5 and 6) that truncation of the highly repetitive sequences was occurring. These truncated proteins were detected as a series of distinct, regularly repeating peaks by analyzing protein-DNA conjugates in free-solution

capillary electrophoresis. Even though the switch to the *C*-terminal His tag has resulted in mostly pure protein, the protein yields are very low especially in comparison to expressions using an *N*-terminal His tag. Thus if the cause of the truncation could be determined and then minimized or eliminated, protein yields should improve with less protein lost as incomplete fragments. Since the truncation was occurring at regular intervals (generating a series of evenly spaced peaks by ELFSE), attention was focused on the codon choices in the original PZ8 design.

7.2 Variable arginine sequences

As stated previously, addition of a limited amount of arginines can be beneficial to the properties of the drag-tag. Research is currently being undertaken by graduate student Xiaoxiao Wang, to determine the maximum amount of arginines that can be incorporated into a sequence before we begin to see unfavorable interactions. These designs consist of 1 Arg in 18 up to 1 Arg in 8 amino acids. In comparison, the PZm8 sequences have 1 Arg in 63 amino acids. Another approach that will be discussed here, is to use site-directed mutagenesis to deliberately introduce additional serine to arginine mutations in the PZ8 sequence to obtain a sequence with greater hydrodynamic drag per length of protein.

7.2.1 Site-directed mutagenesis to introduce additional Arg residues

Site-directed mutagenesis using the QuikChange Kit (Stratagene, La Jolla, CA) has successfully been used in the past for creating single-base changes in the pET-41a plasmid (Section 6.2.1.1) at unique locations. Primers were designed to insert a mutation into the middle serine out of the three in the PZ8 monomer sequence, converting that serine (AGC) into an arginine with a highly preferred codon (CGC), in contrast to the AGG *E. coli* mutation in PZm8.

A PAGE-purified oligonucleotide, 5'-CTG GAA CGG GCC GCG CAG GAG CTG-3', and its reverse complement were obtained from IDT (Coralville, IA). The same mutagenesis protocol described in Section 6.2.1.1 was applied here using PZ8-6 (no arginines) in the pUC18 cloning vector as the DNA template.

7.2.1.1 Colony screening, sequencing, and controlled cloning

Many colonies were obtained after mutagenesis and their plasmid DNA was sequenced. However, the majority of colonies tested had either frameshifts or deletions in their inserts and were unusable. The largest correct sequence that was obtained was a trimer that contained a single mutation out of the three potential primer binding locations.

This trimer was doubled into a 6mer gene through controlled cloning [133] and then inserted into the cloning vector. Additional mutations were discovered among the colonies that were sequenced in addition to the correctly doubled gene. Figure 7-1 is a diagram of the two previously studied PZ8 variants and the three new PZ8-6 variants that were obtained, designated PZ8+1, PZ8+2, and PZ8+3 based on the number of positively charged arginine residues in the sequence. PZ8+2 is the correctly doubled trimer containing the expected two arginines. PZ8+1 had an additional mutation that resulted in an arginine codon mutating back into a serine whereas PZ8+3 was the result of an additional serine to arginine mutation.

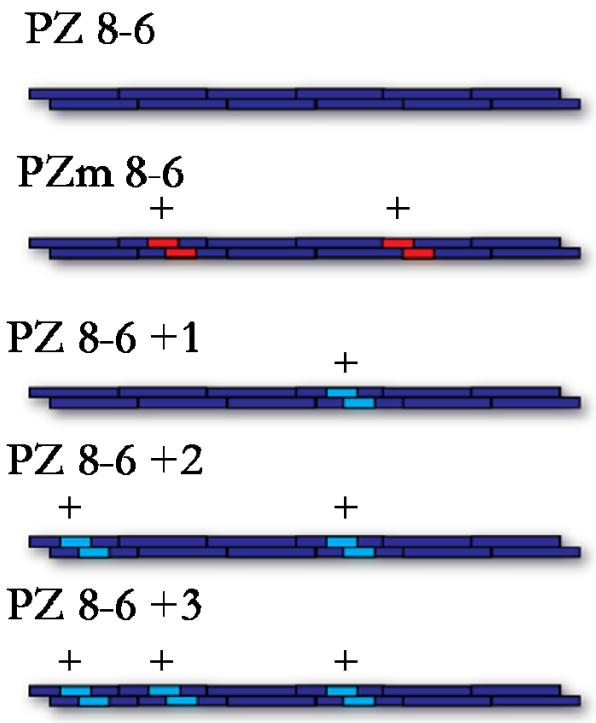


Figure 7-1: The two earlier PZ8 variants in comparison to the three new variants created by site-directed mutagenesis and controlled cloning. Position of the charged residues are marked for a 127-amino acid 6mer protein. Each segment or "monomer" consists of three repeats of the seven amino acid sequence. The repeat with an arginine mutation is marked (red or light blue).

These genes were doubled again via controlled cloning to generate 12mer versions of each of the three variants. All genes were successfully inserted into the recently developed T7 tag/*C*-terminal His tag vector (Section 6.4.3.7) except for the PZ8+1 6mer gene for which no colonies could be obtained for unknown reasons.

7.2.1.2 Test expression of the new sequences

The five remaining sequences were all transformed into BLR(DE3) cells. Three colonies from each of the five proteins were grown on the small scale for test expression and then the cell

lysates were analyzed on a dot blot (Figure 7-2). For sample identification purposes, each sequence was given a number designation and the three different colonies chosen for each sequence were designated A, B, and C. As seen in the blot, all colonies expressed well with some basal level expression present in most of the uninduced control samples as well. Colonies designated 1C, 2A, 3A, 4A, and 5C were selected to represent each sequence based on intensity of the expressed protein signal on the blot and rate of cell growth during culturing.

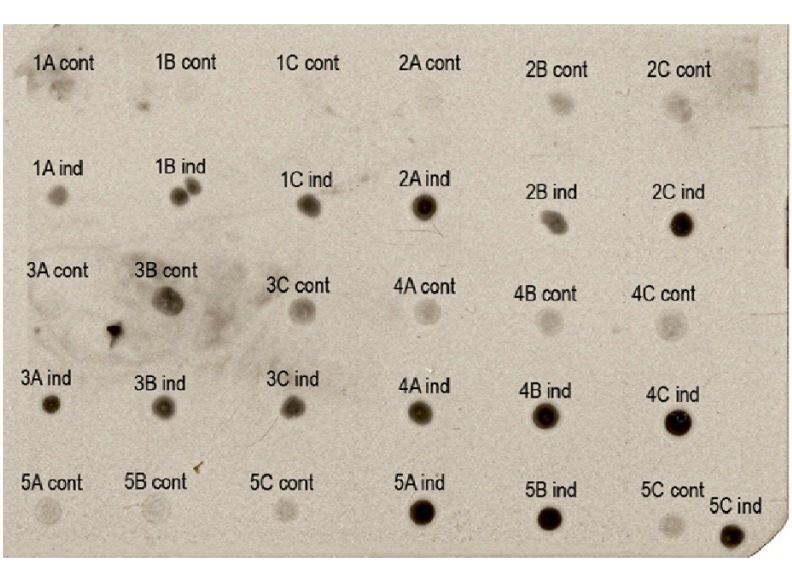


Figure 7-2: Dot blot of variable arginine test expressions. Cont = control; ind = induced.

7.2.1.3 Large-scale expressions

To date, only large-scale (4 L) expressions of the two PZ8+3 sizes have been performed although all the remaining sequences will be expressed in the near future. The affinity chromatography, MALDI-TOF, and ELFSE results for the 6mer and 12mer expressions are presented in Figure 7-3. For the PZ8+3 6mer purification, the reduced 5 mM imidazole buffer

was used for the second wash; however, as seen in Figure 7-3A, there is significant loss of protein in the second wash (200 mL total wash buffer volume whereas elutions are 25 mL each). In an effort to mitigate this loss of protein, for the 12mer purification three washes were done prior to the elution step. The resin was washed twice with the first buffer containing no imidazole followed by a single wash with the 5 mM imidazole buffer. This appeared to lessen the amount of protein lost in the 5 mM imidazole wash based on the lack of a strong protein band in the 3rd wash lane (although the two proteins had different expression levels as well). Each of the final washes were dialyzed, lyophilized, and analyzed by MALDI-TOF. Approximately 45 mg of protein was recovered from the final washes of the 6mer purification and 0.3 mg from the 12mer purification. Mass spectrometry detected the presence of the target protein as well as other contaminant proteins in the wash samples. MALDI-TOF confirmed the elutions contained the correct and pure target proteins (expected 11.512 kDa, actual 11.518 kDa for the 6mer and expected 20.746 kDa, actual 20.751 kDa for the 12mer). Only 6 mg of PZ8+3 6mer and 42 mg of PZ8+3 12mer were recovered from the elution fractions.

Unlike the previously tested PZ8-6 protein discussed in Chapter 5, the PZ8+3 version of the same size exhibited multiple peaks in ELFSE (Figure 7-3E) reminiscent of the truncated peaks seen for longer proteins expressed with an *N*-terminal His tag. In contrast, to the 6mer result, the 12mer protein is devoid of these peaks and is nearly monodisperse (Figure 7-3F). The α values for the 6mer and 12mer were 31 and 71, respectively. These values are slightly higher than the corresponding PZm8 proteins (containing two Arg per 6mer length) once the added mass (and hydrodynamic drag) of the two affinity tags is included.

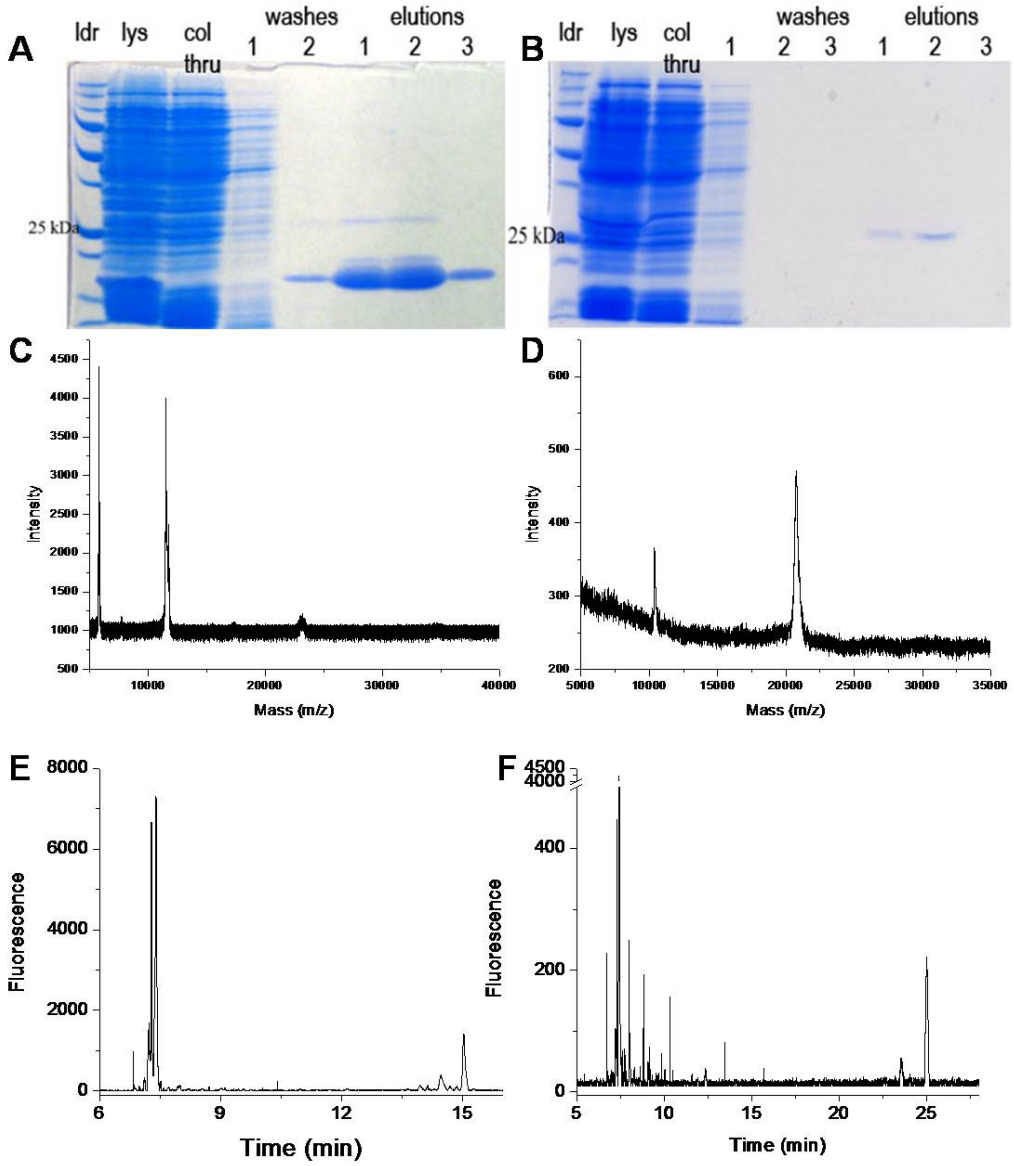


Figure 7-3: PZ8+3 6mer and 12mer proteins A) SDS-PAGE of 6mer B) SDS-PAGE of 12mer C) MALDI-TOF of 6mer D) MALDI-TOF of 12mer. Free-solution capillary electrophoresis of drag-tag-DNA conjugates for E) 6mer and F) 12mer sizes using a 30-base primer. ABI 3100, 36 cm array with 50µM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 1kV/5s injection, 320 V/cm, 55°C

7.2.1.4 Test of media and induction conditions

We have shown previously that a reduced inducer concentration (0.1 mM compared to 1 mM of IPTG) may lead to better yields for proteins expressed with only a *C*-terminal His tag (Section 6.3.3). Further testing was done using the PZ8+3 6mer protein to determine what effect, if any, the T7 tag had on these previous results. Note that expressions with the T7 tag/CHis tag vector and the related vector with only the *C*-terminal His tag have been induced with 1 mM IPTG for better comparison to past results with an *N*-terminal His tag. Testing of protein expression levels in LB, TB, and 2XYT (another rich media like Terrific Broth) was performed concurrently at a fixed IPTG concentration of 1 mM. All samples were induced for 3 hours.

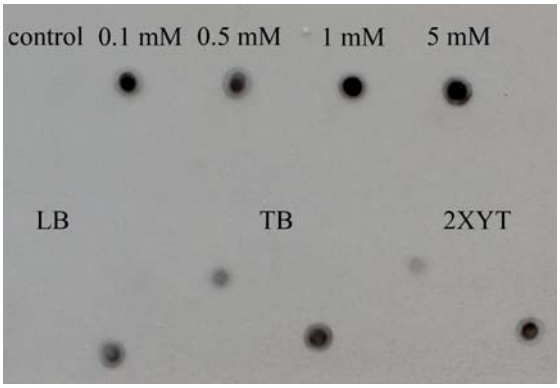


Figure 7-4: Dot blot of IPTG and media test expressions of PZ8+3 6mer. For media test, upper left is control and bottom right is induced sample.

The results of these test expressions were analyzed on a dot blot (Figure 7-4). For changes in IPTG concentration, the protein expressed nearly equally well for all four inducer concentrations tested (0.1, 0.5, 1, and 5 mM), indicating that the same amount of protein could be obtained using a tenth of the current 1 mM inducer concentration. For the media test, expression appeared better in both rich media compared to LB although LB did have less basal level protein expression in the control sample.

7.3 Truncation and the *C*-terminal affinity tag

We have theorized that the polydispersity observed for protein polymers greater than 127 amino acids that were produced with an *N*-terminal affinity tag and analyzed by ELFSE, is due to truncation of the protein during expression. This truncation appeared to occur at specific points in the protein resulting in the regularly spaced peaks. Even though use of the *C*-terminal His tag has mostly eliminated these impurities, the protein yields are much lower, possibly due a significant amount of incomplete protein fragments.

7.3.1 Analysis of current PZ8 codon choices

The original PZ8 gene sequence was designed by Dr. Jong-In Won based on his work with the early drag-tag designs. A simple *E. coli* codon usage analyzer (http://www.faculty.ucr.edu/~mmaduro/codonusage/usage.htm) was used to reevaluate the codon choices in this gene. The data is based on Class II gene data for E. coli [195]. Class II genes correspond to genes highly and continuously expressed during exponential growth and are biased against codons ending in A (adenine).

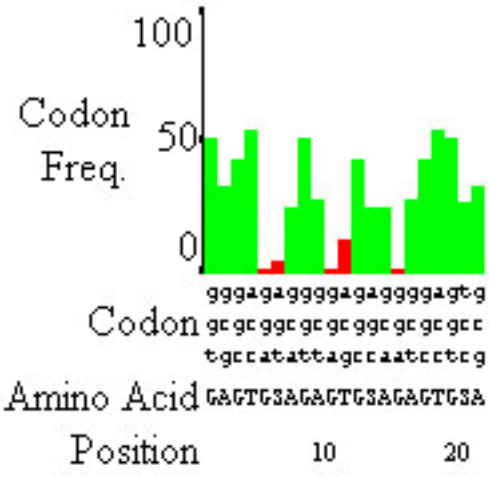


Figure 7-5: Codon analysis of original PZ8 "monomer" sequence. Red = less than 15% of codons for same amino acid.

Figure 7-5 is the codon analysis result for the original PZ8 gene sequence. Five of the codons used in this gene (red) are below the arbitrarily set 15% usage threshold. The GGA

codon (lowest preference out of four glycine codons) was used unnecessarily for three out of the total ten glycines. Additionally, one of the serines is also a low usage codon despite there being only three serines total in the sequence (and six possible codons to choose). The fifth "red" codon is for one of the three threonines. More importantly, two of these "red" codons (glycine and serine) are adjacent to each other. This point is most likely where the highly repetitive protein polymer becomes truncated during expression.

7.3.2 Analysis of PZ8-9

PZ8-9 is the smallest, uncharged, polydisperse protein we have expressed with the *N*-terminal His tag. The free-solution electrophoretic analysis of the protein-DNA conjugate is presented in Figure 7-6. Since it is assumed that the last peak to elute is the full-length protein and the lesser peaks are truncated proteins, then based on the α values for each of these peaks, we should be able to determine the molecular weight of each fragment. Hence, the approximate breakage point in the sequence can be determined and compared to locations of the "red" low usage codons.

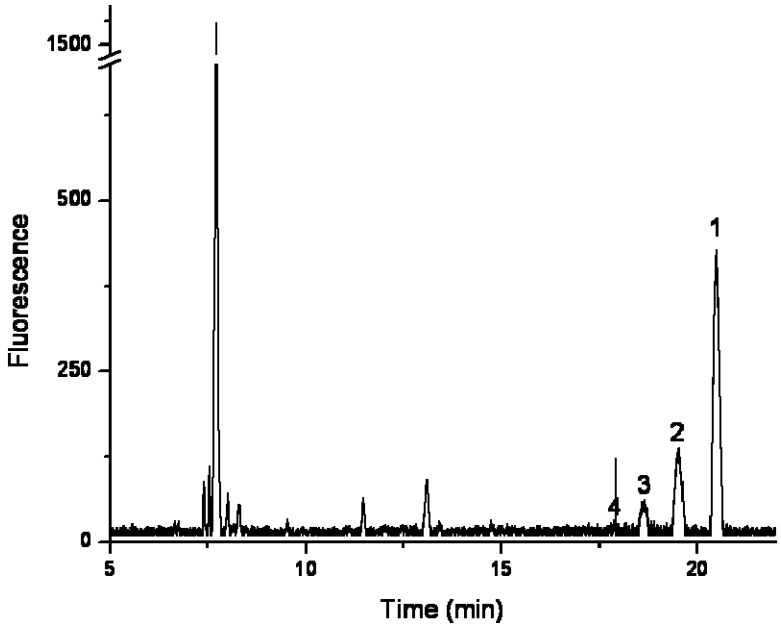


Figure 7-6: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for PZ8-9 using a 20-base primer with the four visible peaks numbered from largest to smallest. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 1kV/5s injection, 320 V/cm, 55°C

Table 7-1 presents the calculated α values for each of the four visible peaks in the electropherogram and the estimated molecular weight of the corresponding protein fragment. The three estimated molecular weights (12.636 kDa, 11.694 kDa, and 10.874 kDa) were then compared to the masses of potential truncated PZ8 proteins. All these masses closely correspond to the position of at least one low usage codon within 142 Da or less. These results further support the theory that truncation is occurring at (or very near) locations containing a low usage codon. The estimated difference in mass between each fragment is all within 820 and 980 Da.

By comparison, the PZ8-9 peak in the MALDI-TOF spectrum has a width, at its base, of ~ 1000 Da.

Table 7-1: Estimated molecular weights from PZ8-9 electropherogram and comparison to expected points of protein truncation

Peak #	α	Calculated amino acid length	Mass (Da)	Amino acid sequence of ending "monomer" segment (assuming truncation at low use codon)	Est. Mass (Da)	Est. amino acid length
1	33.1	190	13615			
2	30.8	176.3	12636	GAGTGSAGA	12684	177
3	28.5	163.2	11694	GAGTGSAGAGTGSA	11552	161
4	26.5	151.7	10874	GAGT	10893	151

7.3.3 Codon substitution

A series of three different sequences were designed to incorporate more of the higher usage codons into the sequence. In other words, these sequences have more occurrences of the higher usage codons but overall have less variety in the codons chosen. This change does not guarantee that truncation will not occur as the larger tRNA pools for more frequently used *E. coli* codons may still become depleted with expression of a highly repetitive sequence.

The first sequence, designated PZc8, is a simple revision of the original PZ8 gene with the low use codons replaced by higher usage codons. PZ9 is a scrambled version of PZ8 (GSGGATA) that also incorporates the codon changes. Hydrodynamic drag is not expected to change as no charged residues are present; however, protein expression levels may be altered, possibly improved, due to the rearrangement. Additionally, with the amino acids repositioned, the truncation peak pattern may also change. Finally, for PZ10 an additional serine was

incorporated into the sequence to aid in water solubility, creating an 8-amino acid repeating sequence (GASGTGSA). PZ10 also incorporated the codon changes. The genes for these sequences as well as their codon usage analyses are presented in Figure 7-7.

PZc8

ATATAGAATTCCTCTTCAGGTGCGGGCACCGGTAGCGCAGGTGCTGGCACCGGCAGCGCAGGTGCTGGCA
CTGGTTCCGCGGGTAGAAGAGGAATTCATATA

PZ9

ATATAGAATTCCTCTTCAGGTAGCGGCGGTGCGACTGCAGGTAGCGGCGGTGCTACCGCAGGCTCCGGTG
GCGCTACCGCGGGTAGAAGAGGAATTCATATA

PZ10

ATATAGAATTCCTCTTCAGGTGCGTCTGGCACCGGTAGCGCAGGCGCTAGCGGTACCGGCTCTGCAGGTG
CTTCCGGCACTGGTTCCGCGGGTAGAAGAGGAATTCATATA

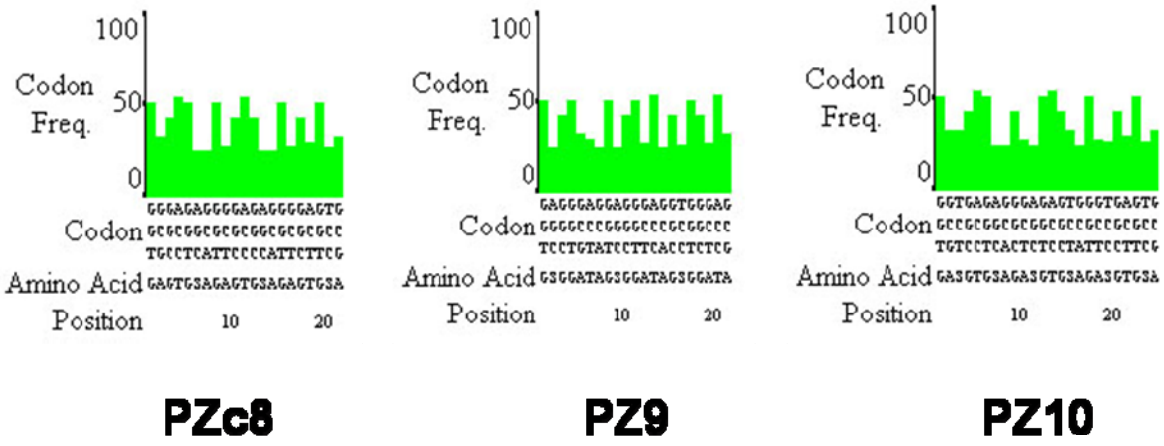


Figure 7-7: Gene monomer sequences and codon usage analyses for PZc8, PZ9, and PZ10 designs

7.3.4 Expression of PZc8, 9, 10

All oligonucleotides were purchased from IDT (Coralville, IA), PCR-amplified, and concatemerized into a ladder of multimers in the same manner as the RZ genes (Chapter 3) and then inserted into the pUC18 cloning vector. Colony screening yielded plasmids with inserts ranging from 3-5mers but no 6mers. Consequently, the trimers were doubled into 6mers using controlled cloning and then doubled again to create 12mers of each sequence. These 12mers were then inserted into the T7 tag/*C*-terminal His tag expression vector (MpET-41a-T7/CHis3) and then transformed into BLR(DE3) cells. Test expressions showed that all three sequences expressed well in the cells.

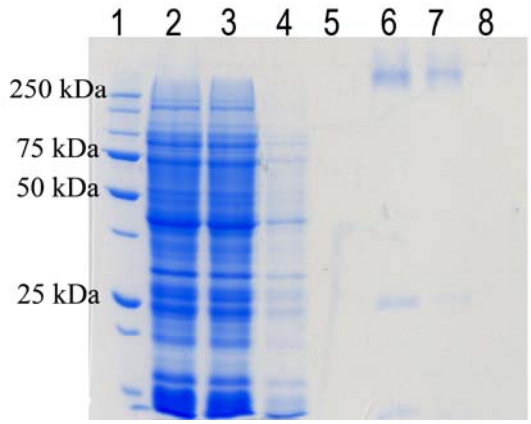


Figure 7-8: 12% SDS-PAGE of PZ10-12 protein purification. Lane 1: ladder, lane 2: lysate; lane 3: flow through; lanes 4-5: washes; lanes 6-8: elutions

3 L to 4 L large-scale expressions were performed for PZ8-12, PZc8-12, and PZ10-12. The proteins were purified by affinity chromatography using two washes of buffer containing no imidazole to minimize premature elution of the protein. All three uncharged proteins exhibited the same unusual migration in the gel as previous expressions of the PZ8 gene (Section 5.3.2.2). Figure 7-8 is the SDS-PAGE result for PZ10-12.

7.3.5 Varied inducer concentrations

PZ9-12 was expressed at four different IPTG concentrations (0.1, 0.5, 1, and 5 mM IPTG) in 1 L cultures each. As with the other uncharged sequences, the protein migrated

unusually in the gel. After dialysis, white solid was observed in all four samples, most likely precipitated protein. The precipitate was lyophilized together with the soluble fraction.

7.4 Comparison to PZ8/PZm8 sequences

7.4.1 Yield

The yields for the large-scale expressions are presented in Table 7-2 for all 12mer (253 amino acid) protein polymers expressed with both a T7 tag and a *C*-terminal His tag. Instead of improving protein polymer yields with the replacement of lower usage codons, the yields for PZc8 were found to be slightly lower than the yield for the original PZ8 sequence. PZ9, which was a rearrangement of the same amino acids in the PZ8 repeating sequence, expressed at yields comparable to the original PZ8 sequence and better than the PZc8 sequence, which incorporated similar codon changes. IPTG concentration had no significant affect on the yield of PZ9-12.

Table 7-2: Summary of 12mer (253 amino acid) protein polymer yields using T7 tag and *C*-terminal His tag for expression

Protein	**Yield (mg/L)**
PZ8-12	21.7
PZc8-12	16.6
PZ9-12 (0.1 mM IPTG)	19.0
PZ9-12 (0.5 mM IPTG)	18.8
PZ9-12 (1 mM IPTG)	21.7
PZ9-12 (5 mM IPTG)	18.4
PZ10-12	14.1

7.4.2 Solubility

PZ9 and PZ10 were observed to have reduced solubility in comparison to the original PZ8 sequence. Both required the addition of urea in an attempt to solubilize the proteins during the conjugation steps. GOR IV [128] predicts the PZ9 sequence to have a lower percent of random-coil secondary structure (70.1% compared to 96.4% for PZ8), which may account for the difference in solubilities. The remaining 29.9% is predicted to be extended strand/β-sheet conformation. However, GOR IV also predicts that PZ10 would have slightly more random-coil structure than PZ8 (97.2%) yet PZ10 was less soluble. The inclusion of the additional serine was intended to generate a more soluble protein polymer but this appears to not be the case.

7.4.3 Hydrodynamic drag

Attempts to conjugate PZ9-12 and PZ10-12 to DNA and analyze the bioconjugates by ELFSE were unsuccessful. Most likely this was due to their poor solubility coupled with the presence of the *C*-terminal His tag. As expected, PZ8-12 and PZc8-12 exhibited similar conjugate profiles as their amino acid sequences are identical (Figure 7-9). However, these electropherograms were unexpectedly similar to the profiles of drag-tags expressed with an *N*-terminal His tag. A glutamic acid site will need to be introduced for enzymatic removal of the *C*-terminal His tag and elimination of any associated side products.

Even though the larger protein polymers are not monodisperse when expressed with an *N*-terminal affinity tag, a graph of molecular weight versus hydrodynamic drag can be constructed based on the "effective" α value of the protein. The last peak in the electropherogram is assumed to be the full-length, uncharged drag-tag. The "effective" α term is used here because we are assuming the drag-tags are neutral ($\beta = 0$) even though some sequences contain arginines.

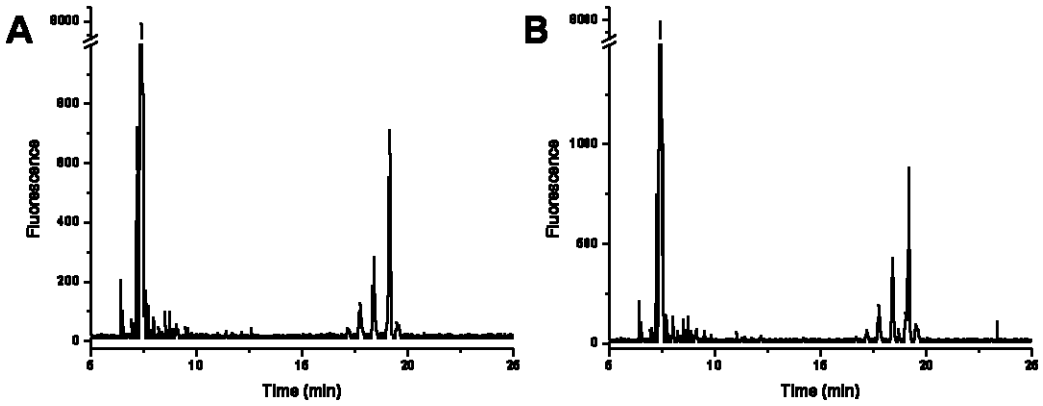


Figure 7-9: Free-solution capillary electrophoresis of drag-tag-DNA conjugates for A) PZ8-12 and B) PZc8-12 using a 30-base primer. Proteins were both expressed with a T7 tag and *C*-terminal His tag which were not removed. ABI 3100, 36 cm array with 50μM ID, 1X TTE, 7M urea, 0.5%v/v POP6, 1kV/5s injection, 320 V/cm, 55°C

All lengths of PZ8 and PZm8 expressed previously with an *N*-terminal His tag (removed by CNBr) are included for comparison. Data for PZ8-12, PZm8-12, and PZ8+3 6mer and 12mer expressions with a T7 tag and *C*-terminal His tag (not removed) are included as well. Molecular weights are derived from MALDI-TOF results and not the expected mass of the protein (< 100 Da difference for most proteins). "Effective" α values are based on the use of a 30mer DNA primer for conjugation.

From the plot in Figure 7-10, "effective" α increases linearly with chain length for both uncharged PZ8 and charged PZm8 proteins. A significant difference in hydrodynamic drag is seen for the 505-amino acid 24mer proteins. The addition of eight arginines to PZm8-24 has more than doubled its "effective" α compared to PZ8-24. However, the addition of a third

arginine to the 6mer sequence did not yield significant improvements in hydrodynamic drag (PZm8 compared to PZ8+3) once the additional mass from its two tags was accounted for. Inclusion of the T7 tag and *C*-terminal His tag does not appear to affect the hydrodynamic drag beyond the increased chain length and mass as evidenced by the position near their respective trendlines of PZm8-12 and PZ8-12 when expressed with both tags.

7.5 Conclusions and Recommendations

Several new protein polymers have been expressed and tested as drag-tags for ELFSE with mixed results. These designs, based on previous work with the PZ8 and PZm8 sequences, investigated increasing the number of arginines for greater hydrodynamic drag per length of protein. Other designs altered the original PZ8 codon selection in an effort to boost protein yields with less truncation during expression. Both PZ9 and PZ10 had lower water solubility than the original PZ8 design and DNA conjugation with these proteins as drag-tags was not successful. The yield of PZc8-12 was slightly lower compared to the yield of PZ8-12, indicating that simply changing all the codons to higher usage codons (> 15%) does not necessarily mean less truncated proteins will be made, although in some cases protein yields have increased as a result [181]. Further balancing will need to be done to ensure that the tRNA pool for higher usage codons is not rapidly depleted with the alterations while simultaneously ensuring that the inclusion of a few low usage codons is likewise not limiting. A revised PZ8 sequence might include only 2-3 codon changes instead of the 5 that were implemented in the PZc8 design. Interestingly, PZ9, which also incorporated the codon changes but in a rearranged sequence, had comparable expression levels to the original PZ8 gene.

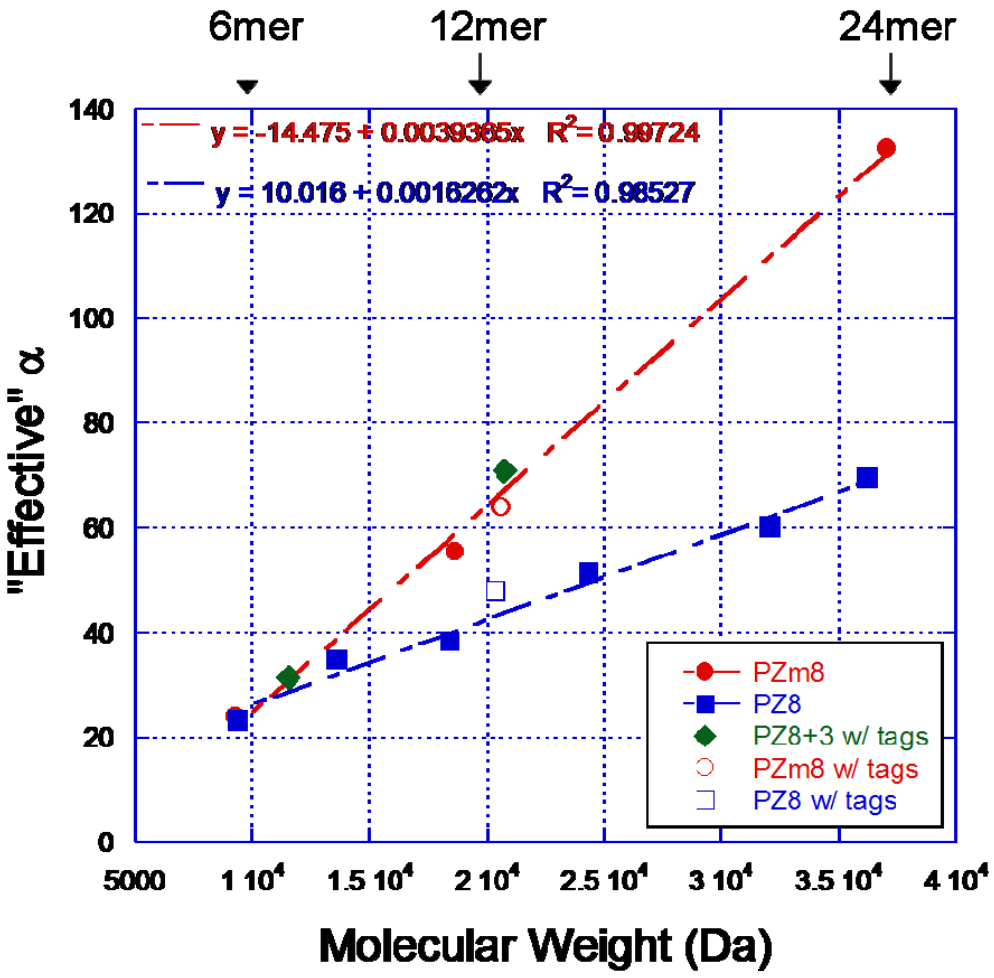


Figure 7-10: "Effective" α of protein polymers of varying length and charge as determined by ELFSE using a 30 bp DNA primer for conjugation. Proteins expressed with a T7 tag and *C*-terminal His tag (not removed) are compared to proteins expressed previously with an *N*-terminal His tag (removed by CNBr).

Use of the PZc8, PZ9, and PZ10 sequences with an *N*-terminal His tag should result in a different truncation peak pattern (in ELFSE) compared to the PZ8 and PZm8 sequences due to the change in codons or rearrangement of amino acids (PZ9). With the reduced yield of PZc8 compared to the original gene, we would expect there to be a higher number of truncation peaks.

Overall the yields of the uncharged protein polymers were greater than the yields of the charged sequences.

The PZ8+3 sequence, consisting of one additional arginine per 6mer length (or 126 amino acids) than PZm8, resulted in only slightly greater hydrodynamic drag for the two lengths tested (6mer and 12mer lengths) once the added mass of the T7 tag and *C*-terminal His tag were accounted for. If the masses of PZ8+3 (with tags) are used to calculate the "effective" α using the corresponding linear trendline equation for the PZm8 proteins from Figure 7-10 ($y = -14.475 + 0.0039365x$), we find that the difference between PZ8+3 and PZm8 proteins of the same size for the 6mer is negligible (31 for PZ8+3 versus 30.9 for PZm8). For the 12mer protein the difference is 71 versus 67.3. A 24mer version of PZ8+3 would need to be produced to better compare the new sequence to PZ8 and PZm8. A 18mer protein can also be constructed by controlled cloning by combining the genes for the 6mer and 12mer proteins. Even though PZ8+3 has more arginines than PZm8, as seen in Figure 7-1, the positioning of the charges in the sequence are different, with more of the PZ8+3 positive charges situated closer to the N-terminus of the protein.

7.5.1 End effects theory

Recent theoretical work has examined the contribution of end effects to the theory of ELFSE. The electrophoretic mobility of a drag-tag-DNA conjugate, μ, is determined by the weighted average of the electrophoretic mobilities of charged and uncharged monomers. The weighting of individual monomer units was reexamined to account for end effects [196]. It was theorized that monomer units near either end of the polymer chain have greater influence on the

electrophoretic mobility of the conjugate [197]. Therefore, addition of drag-tags to both ends of the DNA molecule yields more than double the drag of using a single drag-tag of the same size at one end of the DNA. Experiments done by Dr. Robert Meagher using drag-tags of different lengths supported this assertion [174].

The arginines in PZm8 are positioned closer to the C-terminus of the protein than the arginines in the PZ8+2 sequence. Therefore, despite having the same net charge, PZm8 should exhibit greater hydrodynamic drag. Further testing on the effects of charge location and density on α can be made by generating block copolymers from the different charged and uncharged versions of PZ8 through controlled cloning. Concentrating the positive charges at the end of the protein polymer would theoretically boost the hydrodynamic drag, assuming no electrostatic interactions with the DNA and microchannel walls.

The end effects theory may also explain why PZ8+3 is, at least for the sizes tested so far, only a slight improvement over the PZm8 sequence. PZm8 has one of its two arginines located seven amino acid residues closer to the C-terminus of the protein. This slight difference in positioning may counteract the addition of a third arginine nearer the N-terminus for the PZ8+3 sequence. Not all sequences have been expressed yet on the large scale (*i.e.,* PZ8+1 and PZ8+2). PZ8+2 includes two arginines (like PZm8) but the positioning of the two arginines is different (similar to PZ8+3). According to the end effects theory, this protein would have a reduced hydrodynamic drag compared to PZm8 despite having the same net charge.

It would be useful to generate one or more arginine mutations close to the end of the protein to maximize the potential of the positive charge. Further site-directed mutagenesis and

screening would be needed to identify plasmids containing these mutations. Work is also ongoing to explore the inclusion of Arg directly into the repeating monomer sequence by graduate student Xiaoxiao Wang.

Chapter Eight

Biophysical Studies of the Protein Polymers

8.1 Introduction

In order to determine the suitability of the designed protein polymers as drag-tags, the expressed proteins have been analyzed for molecular mass (MALDI-TOF) and purity (RP-HPLC, ELFSE) (Chapters 5 and 6). However, the drag-tags must also be water soluble at elevated temperatures (55°C for capillary electrophoresis and up to 95°C if the drag-tag is conjugated to the DNA primer prior to the thermal cycling steps of the Sanger reaction). Additionally, the protein polymers were designed for a random-coil secondary structure (Section 2.3.2). Since protein polymers have now been expressed and purified, their conformations can be verified by circular dichroism spectroscopy.

8.2 Temperature-dependent Vis spectroscopy for solubility investigation

Polypeptide solubility can be measured as a function of temperature, in order to determine if the protein polymer exhibits LCST (lower critical solution temperature) behavior. The LCST is the transition temperature at which a polymer solution undergoes a solubility-to-insolubility transition, and separates into two immiscible phases. Elastin and elastin-like protein polymers (ELPs) are examples of proteins that exhibit this phase transition behavior [72, 152, 198]. This behavior is the basis of the inverse transition cycling purification strategy discussed in Section 2.8.4 using elastin tags. A large increase in solution absorbance at 500 nm (*i.e.,* within visible spectrum) occurs when the LCST is reached [199].

The relative solubility of the PZm8 proteins (127, 253, and 505 amino acids) were measured by a Cary 500 UV-Vis spectrophotometer equipped with temperature control (Varian, Inc., Palo Alto, CA) at the Northwestern Keck Facility. Sample absorbance was observed at 500 nm for a temperature range of 20°C to 95°C for protein dissolved at 100 μM in water. A temperature ramping rate of 1.5ºC/min was used. The 100 μM concentration was chosen to be within the range used for elastin-like protein polymers. Figure 8-1 compares a representation of LCST behavior to the experimental results for the three lengths of PZm8 protein that were tested. No solution absorbance increase was observed indicating no phase change was occurring for the proteins and concentration tested. Water solubility is an important drag-tag property, particularly in light of the difficulties with the earlier versions of the *C*-terminal His tag proteins (Chapter 6).

8.3 Circular dichroism spectroscopy for secondary structure determination

Three sizes of the PZ8 and PZm8 sequences were analyzed using circular dichroism (CD) spectroscopy. Circular dichroism is a technique that can assess the average content secondary structure in a protein (*e.g.,* random coil, α-helix, or β-sheet) [200]. CD spectrophotometers measure the difference in absorbance for left and right circularly polarized light or the molecular asymmetry (ellipticity) of a material [141]. Absorbance is measured in the far UV range corresponding to the amide chromophore. Different secondary structures yield different CD characteristics. A random-coil conformation exhibits a strong negative band between 192-201 nm [141].

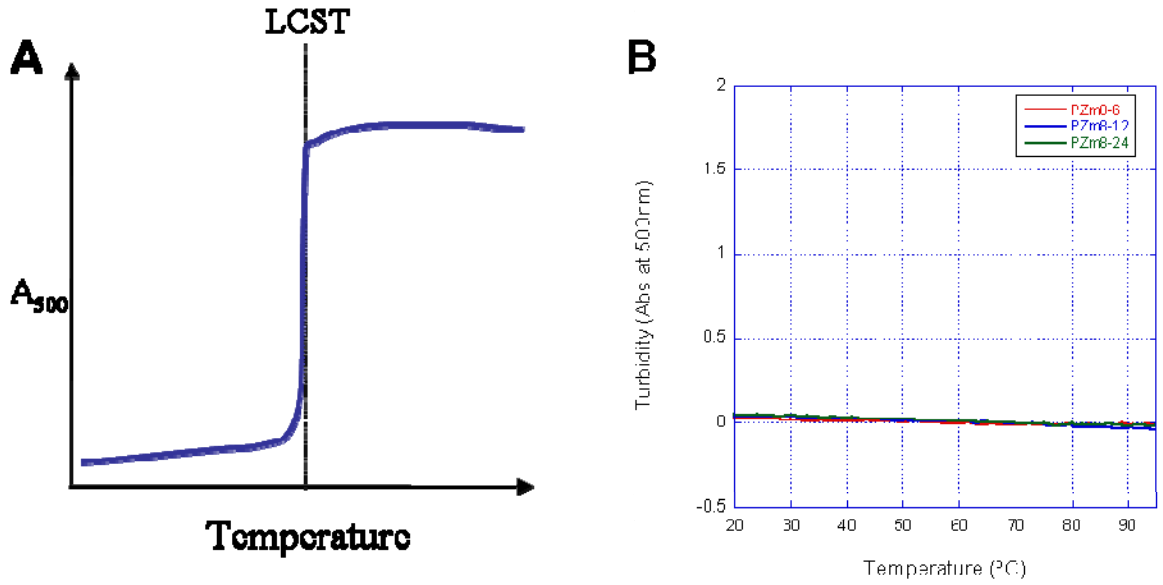


Figure 8-1: A) Diagram of LCST behavior as indicated by a sharp increase in absorbance B) Behavior of three different sizes of PZm8 protein in water at 100 μM concentration between 20°C and 95°C

All six proteins that were tested, ranging from 127 to 505 amino acids in size, exhibited the desired random-coil secondary structure at 25°C in water (no urea) using concentrations ranging from 10 μM to 50 μM (Figure 8-2). A Jasco (Easton, MD) J-715 spectrophotometer (Northwestern Keck Biophysics Facility) was used to collect the data between 185-280 nm using a cuvette with a 0.02 cm path length.

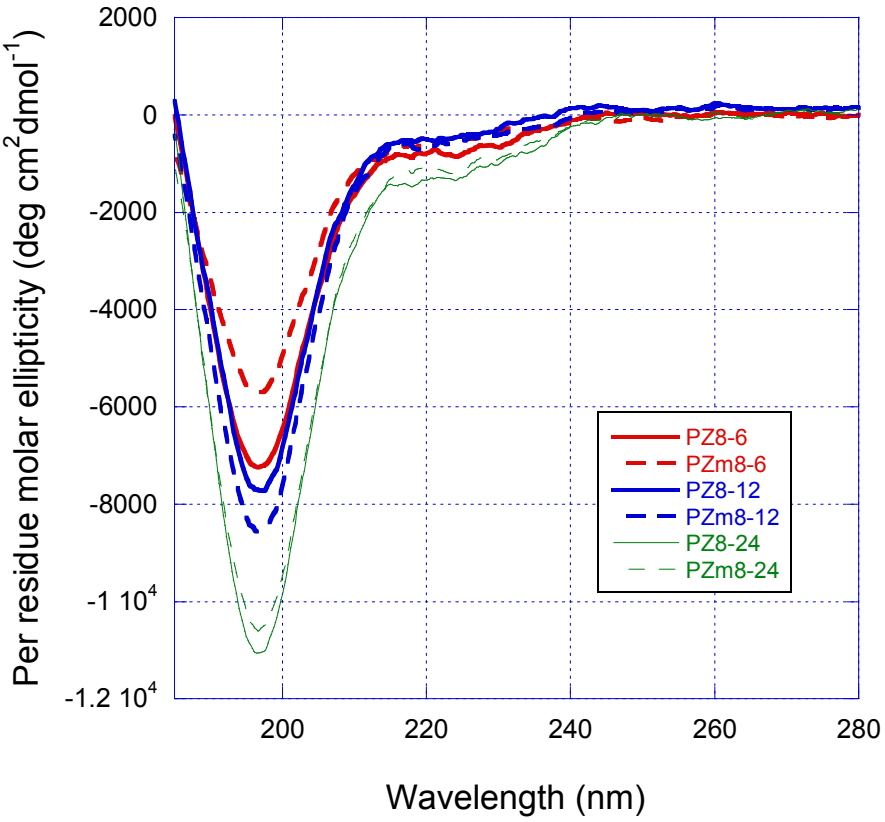


Figure 8-2: CD spectra for various lengths of PZ8 and PZm8 protein in water at 25°C

PZ8-12 was also tested at elevated temperatures using a water bath to control the sample temperature. Readings were taken at 25°C, 55°C, 75°C, and 90°C, the maximum that could be reached with the water bath setup (Figure 8-3). The protein maintains its random-coil structure for the range tested. Additionally, there appears to be an isodichroic point at approximately 201 nm, possibly indicating a transition into another conformation. A method called SOMCD [201] was used to estimate the protein secondary structure content from the CD data in the 190-240 nm range. This program estimates that PZ8-12 is approximately 95% random coil at 25°C and 39%

at 90°C with the α-helical, β-sheet, and turn structure content increasing to 18%, 27%, and 16%, respectively.

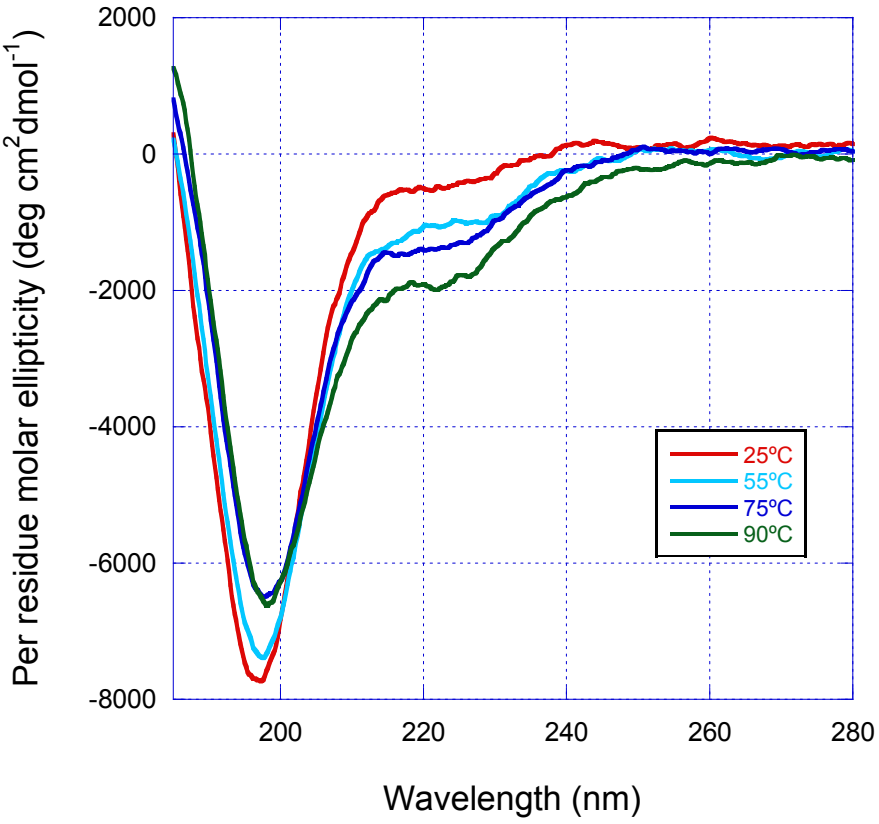


Figure 8-3: CD spectra of PZ8-12 protein in water at various temperatures

8.4 Conclusions and Recommendations

At the 100 μM concentration tested, the three PZm8 proteins did not exhibit LCST behavior between 20-95°C in water. Additional proteins (*e.g.,* uncharged PZ8) would need to be tested as well as other concentrations and solvent conditions. For example, the protein concentration during the sulfo-SMCC activation step in sodium phosphate buffer is 15 mg/mL while the concentration during the DNA conjugation step is 250 μM, all done at room

temperature. For the Sanger cycle sequencing reaction, the drag-tag concentration is only 0.42 μM but the protein undergoes several temperature cycles. Temperature-controlled Vis spectroscopy can be used to better quantify the solubility of the drag-tags since currently the relative solubility of a protein is based on empirical observations by Jennifer Coyne when performing the conjugation reaction.

CD spectroscopy was used to confirm that the PZ8 and PZm8 proteins adopt random-coil conformations as per their intended design. PZ8-12 (and likely the other variants of PZ8) also maintained this random-coil secondary structure up to 90°C. However, PZ9-12 was observed to have reduced water solubility and was predicted to have more extended sheet/β-strand structure than PZ8. Therefore, a CD spectrum of PZ9 may include a noticeable fraction of β-sheet structure along with the expected random-coil conformation. Analysis of PZ10-12 would likewise provide insight into what may be causing the reduced solubility of the protein with the addition of a serine (*e.g.,* perhaps an unexpectedly large proportion of β-sheet structure).

8.4.1 Dynamic light scattering

Dynamic light scattering can yield information on the size, distribution, and monodispersity of the proteins in solution and identify possible aggregation. Only dilute protein concentrations ~ 1 mg/mL and approximately 35 μL of sample are needed for analysis using a Wyatt Technologies (Santa Barbara, CA) DynaPro™ Plate Reader. Using this instrument we can monitor proteins in solution at varying temperatures and concentrations. Particle size is determined by assessing the movement of particles over time (Brownian motion) where their rate of motion is a factor of their size. Rate of motion is measured by using laser light to determine

the rate at which light scattered or reflected by the particles changes over time. Analysis of the correlation function (rate of time intensity-fluctuations from particle movement) using numerical methods leads to a size distribution. Models that predict molecular weight as a function of radius for several polymer types can be evaluated by comparing predicted molecular weight versus known molecular weight. This can help evaluate how our random-coil protein polymers compare to other polymers and proteins in solution.

Chapter Nine

Conclusions and Future Research Directions

9.1 General conclusions

End-labeled free-solution electrophoresis (ELFSE) is a novel method of DNA sequencing based on the Sanger reaction that allows for size-based DNA separation without a sieving matrix. This technique has the potential to yield rapid, long read-length DNA sequencing without the use of a viscous polymer solution. Read lengths of ~ 180 bases were previously obtained using a small, nearly monodisperse, 127-amino acid protein polymer as the drag-tag. Much of the research effort has been focused on overcoming the obstacles of obtaining larger drag-tags with greater hydrodynamic drag. Higher values of α would lead to longer DNA sequencing read lengths approaching that obtained by standard DNA sequencing techniques (500+ bases), but with the advantage of requiring no polymer solution for DNA separation.

A nearly monodisperse protein polymer with double the hydrodynamic drag of the 127-amino acid protein polymer has finally been achieved, proving that larger monodisperse protein polymer drag-tags are possible. This drag-tag is currently being tested for use in ELFSE-based DNA sequencing. Future research efforts will focus on generating even larger, monodisperse drag-tags using the same techniques. The monodispersity requirement for ELFSE drag-tags is strict yet still achievable.

9.2 Drag-tag design

Not all sequence designs have been as successful as the PZ8 series, illustrating the challenges of creating new designs *de novo*. PZ8 was the result of work on seven previous designs by Dr. Jong-In Won and PZm8 with its arginine mutations was actually serendipitously created by the *E. coli* cells. The success of PZm8 has led to further investigation of proteins that deliberately include the positively charged arginines (*i.e.,* PZ8+1, PZ8+2, and PZ8+3 sequences) when previously positive charges were specifically avoided. The inclusion of a few arginines has greatly boosted the hydrodynamic drag of the protein polymers compared to uncharged proteins (Figure 7-10). The results for the RZ series (Chapter 3) and PZc8, PZ9 and PZ10 (Chapter 7) show that there is currently no way to predict with 100% certainty if a protein will express in *E. coli*, and if so with what yields and solubility.

Block copolymer designs can be easily constructed using controlled cloning although this has not yet been done for protein polymer drag-tag sequences. The newer designs being investigated by graduate student Xiaoxiao Wang, which incorporate Arg into the repetitive sequence itself, may prove to contain too many arginines in the sequence (*e.g.,* detrimental interaction with negatively charged DNA or microchannel walls). In such a case, a block copolymer design can be assembled by combining a long, uncharged PZ8 gene with an Arg-repeating segment. This Arg segment would be placed towards the end of the sequence to maximize the benefit of the positive charges based on the end effects theory [174, 197].

9.3 *C*-terminal affinity tag and protein polymer monodispersity

The creation of longer protein polymers that were monodisperse enough to be used as drag-tags for ELFSE sequencing proved to be challenging. After extensive research, detailed in Chapter 5, it was determined that a *C*-terminal affinity tag was required to exclude truncated protein polymers from the final product (Chapter 6). Additionally, the presence of the *C*-terminal His tag and not solely poor water solubility caused the low efficiency of drag-tag to DNA conjugation reactions. Cyanogen bromide could not be used to remove the affinity tag as this leads to the generation of two species of proteins with either a homoserine or homoserine lactone C-terminus. Site-specific protease cleavage was the other option. However, Factor Xa, which adds IEGR to the C-terminus of the protein, proved to favor secondary cleavage sites of GR as much as the primary recognition sequence, IEGR.

Alternatively, endoproteinase GluC, which cleaves specifically at Glu or Asp residues, can also be used since no protein polymer design currently in use contains these two amino acids in its repeating sequence. This protease is later removed along with any uncleaved protein by IMAC since the version sold by New England BioLabs (Ipswich, MA) has a His tag at its C-terminus. The combination of a *C*-terminal affinity tag for expression that was removed by endoproteinase GluC was what led to a nearly monodisperse protein polymer with double the α of the previous drag-tag used for DNA sequencing. Only 50 μg was used to cleave 5 mg of protein; however, it may be possible to cleave larger quantities with the same amount of enzyme. No uncleaved protein was detected in the IMAC elutions after cleavage of 5 mg of protein, suggesting that cleavage of larger quantities may be possible.

Ultimately, a self-cleaving *C*-terminal affinity tag that incorporates an intein may be the best option. No glutamic acid residue would be required, which reduces the hydrodynamic drag of the protein slightly with its negative charge. Additionally the expenses associated with protease cleavage followed by another IMAC purification are avoided if the tag can simply be cleaved on-column during the initial purification from the cell lysate. Intein tags were previously explored in Chapter 4 for an *N*-terminal affinity tag before we knew that a *C*-terminal affinity tag would be necessary for achieving monodisperse protein polymers. This direction of research will be pursued by Ph.D. candidate Xiaoxiao Wang.

9.4 Biophysical studies

Our protein polymers are unique in that they have random-coil structures in pure water (no denaturant) as confirmed by circular dichroism spectroscopy (Chapter 8). These protein sequences have not been created or studied before by others. A comprehensive solubility study of the different protein polymers at various temperatures and concentrations will be performed using Vis spectroscopy. Similarly dynamic light scattering will be used to investigate particle size in solution at varying temperatures. The study of unfolded proteins is gaining interest as more research uncovers the importance of these unstructured regions to the function of natural proteins. Early work on denatured proteins showed reasonable agreement with theoretical estimates of the hydrodynamic radius of random coils; however, more recent results have shown that even denatured proteins possess residual structure [202, 203]. Our protein polymers, on the other hand, are random coil by design and require no denaturants to achieve this conformation.

The study of polymer physical properties using truly monodisperse polymers of tailored molecular weights and charges may be possible using our protein polymers, assuming they are large enough (largest size produced to date is 36 kDa). In particular, the overlap threshold concentration (c*) can be determined as a function of protein polymer chain length and compared to theory. The overlap threshold concentration c* is defined as the concentration at which polymer molecules begin to overlap and interact strongly in solution [204]. This value can be estimated experimentally by identifying the point of departure from linearity of a log-log plot of specific viscosity versus concentration. Theoretically, c* can be calculated from molecular weight and the radius of gyration or from intrinsic viscosity, defined as the volume occupied by a polymer per unit mass [205]. Attempts to correlate theoretical predictions with experimental data are complicated by the polydispersity of the synthetic polymers used for analysis [206, 207].

9.5 ELFSE on microfluidic devices

The greatest benefit of ELFSE sequencing, which uses no polymer solution for the separation but simply buffer, would be realized on microfluidic devices. A related technique that separates single-base extension products in free solution for genotyping was successfully applied to a microfluidic device after the technique was initially developed using capillary electrophoresis [61]. ELFSE sequencing using the PZm8-6 drag tag has been attempted on a glass microchip but the results show that additional challenges will need to be overcome to successfully apply it to microchip separations. Approximately 20 bases were observed in less

than 110 seconds (results not shown). These bases correspond to the smallest DNA fragments which elute last in ELFSE sequencing. Hence, 180 bases of sequence should be obtainable in less than 2 minutes once these complications are overcome. Transitioning ELFSE sequencing from capillary to microchip will be undertaken by Ph.D. candidate Jennifer Coyne. Ultimately we plan to use ELFSE in integrated "lab-on-a-chip" devices to quickly and economically perform genotyping and DNA sequencing.

References

[1] Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H. M.; Yu, J.; Wang, J.; Huang, G. Y.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S. Z.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H. Q.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G. R.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W. H.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J. R.; Slater, G.; Smit, A. F. A.; Stupka, E.; Szustakowki, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.;

Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J., Initial sequencing and analysis of the human genome. *Nature* 2001**,** 409, (6822), 860-921.

[2] Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. Q. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J. H.; Miklos, G. L. G.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, C.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z. M.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W. M.; Gong, F. C.; Gu, Z. P.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z. X.; Ketchum, K. A.; Lai, Z. W.; Lei, Y. D.; Li, Z. Y.; Li, J. Y.; Liang, Y.; Lin, X. Y.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B. X.; Sun, J. T.; Wang, Z. Y.; Wang, A. H.; Wang, X.; Wang, J.; Wei, M. H.; Wides, R.; Xiao, C. L.; Yan, C. H.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W. Q.; Zhang, H. Y.; Zhao, Q.; Zheng, L. S.; Zhong, F.; Zhong, W. Y.; Zhu, S. P. C.; Zhao, S. Y.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H. J.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H. Y.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A. D.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X. J.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.;

Wen, M. Y.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. H., The sequence of the human genome. *Science* 2001**,** 291, (5507), 1304-+.

[3] Collins, F. S.; Green, E. D.; Guttmacher, A. E.; Guyer, M. S., A vision for the future of genomics research. *Nature* 2003**,** 422, (6934), 835-847.

[4] Collins, F. S.; Lander, E. S.; Rogers, J.; Waterston, R. H., Finishing the euchromatic sequence of the human genome. *Nature* 2004**,** 431, (7011), 931-945.

[5] Kling, J., The search for a sequencing thoroughbred. *Nat. Biotechnol.* 2005**,** 23, (11), 1333-1335.

[6] Service, R. F., Gene sequencing - The race for the $1000 Genome. *Science* 2006**,** 311, (5767), 1544-1546.

[7] NIH News Release: NHGRI Seeks Next Generation of Sequencing Technologies, http://www.genome.gov/12513210. 2004.

[8] Recommendation for a Human Cancer Genome Project: Report of the Working Group on Biomedical Technology. http://www.genome.gov/15015123. 2005.

[9] Fredlake, C. P.; Hert, D. G.; Mardis, E. R.; Barron, A. E., What is the future of electrophoresis in large-scale genomic sequencing? *Electrophoresis* 2006**,** 27, (19), 3689-3702.

[10] Ledford, H., Kudos, not cash, is the real X-factor. *Nature* 2006**,** 443, 733.

[11] Pennisi, E., GENOMICS: On Your Mark. Get Set. Sequence! *Science* 2006**,** 314, (5797), 232-.

[12] Lazarou, J.; Pomeranz, B. H.; Corey, P. N., Incidence of adverse drug reactions in hospitalized patients - A meta-analysis of prospective studies. *Jama-Journal of the American Medical Association* 1998**,** 279, (15), 1200-1205.

[13] Special report: genotyping for cytochrome P450 polymorphisms to determine drug-metabolizer status. *Technol Eval Cent Asses Program Exec Summ* 2004**,** 19, (9), 1-2.

[14] Singer, E., Amplichip CYP450. *Technol. Rev.* 2006**,** 109, (5), 76-77.

[15] Jain, K. K., Applications of AmpliChip CYP450. *Mol. Diagn.* 2005**,** 9, (3), 119-27.

[16] Hutchison, C. A., DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* 2007**,** 35, (18), 6227-6237.

[17] Roach, J. C.; Boysen, C.; Wang, K.; Hood, L., Pairwise end sequencing - A unified approach to genomic mapping and sequencing. *Genomics* 1995**,** 26, (2), 345-353.

[18] Chaisson, M.; Pevzner, P.; Tang, H. X., Fragment assembly with short reads. *Bioinformatics* 2004**,** 20, (13), 2067-2074.

[19] Lerman, L. S.; Frisch, H. L., Why does the electrophoretic mobility of DNA in gels vary with the length of the molecule. *Biopolymers* 1982**,** 21, (5), 995-997.

[20] Olivera, B. M.; Baine, P.; Davidson, N., Electrophoresis of the nucleic acids. *Biopolymers* 1964**,** 2, 245-257.

[21] Slater, G. W.; Drouin, G., Why can we not sequence thousands of DNA bases on a polyacrylamide-gel. *Electrophoresis* 1992**,** 13, (8), 574-582.

[22] Weinberger, R., *Practical Capillary Electrophoresis*. 2nd ed.; Academic Press: New York, 2000.

[23] Slater, G. W.; Desrulsseaux, C.; Hubert, S. J.; Mercier, J. F.; Labrie, J.; Boileau, J.; Tessier, F.; Pepin, M. P., Theory of DNA electrophoresis: A look at some current challenges. *Electrophoresis* 2000**,** 21, (18), 3873-3887.

[24] Viovy, J. L., Electrophoresis of DNA and other polyelectrolytes: Physical mechanisms. *Reviews of Modern Physics* 2000**,** 72, (3), 813-872.

[25] Zhou, H. H.; Miller, A. W.; Sosic, Z.; Buchholz, B.; Barron, A. E.; Kotler, L.; Karger, B. L., DNA sequencing up to 1300 bases in two hours by capillary electrophoresis with mixed replaceable linear polyacrylamide solutions. *Anal. Chem.* 2000**,** 72, (5), 1045-1052.

[26] Sanger, F.; Nicklen, S.; Coulson, A. R., DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 1977**,** 74, (12), 5463-5467.

[27] Yang, D. L., Sauvageot, Rachel, Pentoney, Jr., Stephen, DNA sequencing by capillary electrophoresis. In *Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques*, third ed.; Landers, J. P., Ed. CRC Press: New York, 2008; p 1567.

[28] Jacobson, S. C.; Ramsey, J. M., Microchip Electrophoresis With Sample Stacking. *Electrophoresis* 1995**,** 16, (4), 481-486.

[29] Jacobson, S. C.; Hergenroder, R.; Koutny, L. B.; Warmack, R. J.; Ramsey, J. M., Effects of Injection Schemes and Column Geometry On the Performance of Microchip Electrophoresis Devices. *Analytical Chemistry* 1994**,** 66, (7), 1107-1113.

[30] Zhang, C. X.; Manz, A., Narrow sample channel injectors for capillary electrophoresis on microchips. *Analytical Chemistry* 2001**,** 73, (11), 2656-2662.

[31] Woolley, A. T.; Mathies, R. A., Ultra-high speed DNA-sequencing using capillary electrophoresis chips. *Anal. Chem.* 1995**,** 67, (20), 3676-3680.

[32] Fredlake, C. P.; Hert, D. G.; Kan, C. W.; Chiesl, T. N.; Root, B. E.; Forster, R. E.; Barron, A. E., Ultrafast DNA sequencing on a microchip by a hybrid separation mechanism that gives 600 bases in 6.5 minutes. *Proceedings of the National Academy of Sciences of the United States of America* 2008**,** 105, (2), 476-481.

[33] McDonald, J. C.; Duffy, D. C.; Anderson, J. R.; Chiu, D. T.; Wu, H. K.; Schueller, O. J. A.; Whitesides, G. M., Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis* 2000**,** 21, (1), 27-40.

[34] Ronaghi, M.; Sbokralla, S.; Gharizadeh, B., Pyrosequencing for discovery and analysis of DNA sequence variations. *Pharmacogenomics* 2007**,** 8, (10), 1437-1441.

[35] Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z. T.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L. I.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P. G.; Begley, R. F.; Rothberg, J. M., Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005**,** 437, (7057), 376-380.

[36] Mardis, E. R., The impact of next-generation sequencing technology on genetics. *Trends Genet.* 2008**,** 24, (3), 133-141.

[37] Pop, M.; Salzberg, S. L., Bioinformatics challenges of new sequencing technology. *Trends Genet.* 2008**,** 24, (3), 142-149.

[38] Rogers, Y. H.; Venter, J. C., Genomics - Massively parallel sequencing. *Nature* 2005**,** 437, (7057), 326-327.

[39] Wheeler, D. A.; Srinivasan, M.; Egholm, M.; Shen, Y.; chen, l.; mcguire, a.; he, w.; chen, y.-j.; makhijani, v., The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008**,** 452, (7189), 872-877.

[40] Levy, S.; Sutton, G.; Ng, P. C.; Feuk, L.; Halpern, A. L.; Walenz, B. P.; Axelrod, N.; Huang, J.; Kirkness, E. F.; Denisov, G.; Lin, Y.; MacDonald, J. R.; Pang, A. W. C.; Shago, M.; Stockwell, T. B.; Tsiamouri, A.; Bafna, V.; Bansal, V.; Kravitz, S. A.; Busam, D. A.; Beeson, K. Y.; McLntosh, T. C.; Remington, K. A.; Abril, J. F.; Gill, J.; Borman, J.; Rogers, Y. H.; Frazier, M. E.; Scherer, S. W.; Strausberg, R. L.; Venter, J. C., The diploid genome sequence of an individual human. *PLoS Biol.* 2007**,** 5, (10), 2113-2144.

[41] Wadman, M., James Watson's genome sequenced at high speed. *Nature* 2008**,** 452, (7189), 788.

[42] Shendure, J.; Porreca, G. J.; Reppas, N. B.; Lin, X. X.; McCutcheon, J. P.; Rosenbaum, A. M.; Wang, M. D.; Zhang, K.; Mitra, R. D.; Church, G. M., Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005**,** 309, (5741), 1728-1732.

[43] Blow, N., Genomics: The personal side of genomics. *Nature* 2007**,** 449, (7162), 627-632.

[44] Hutchison, C. A., DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research* 2007**,** 35, 6227-6237.

[45] Hillier, L. W.; Marth, G. T.; Quinlan, A. R.; Dooling, D.; Fewell, G.; Barnett, D.; Fox, P.; Glasscock, J. I.; Hickenbotham, M.; Huang, W. C.; Magrini, V. J.; Richt, R. J.; Sander, S. N.; Stewart, D. A.; Stromberg, M.; Tsung, E. F.; Wylie, T.; Schedl, T.; Wilson, R. K.; Mardis, E. R., Whole-genome sequencing and variant discovery in C-elegans. *Nat. Methods* 2008**,** 5, 183-188.

[46] Harris, T. D.; Buzby, P. R.; Babcock, H.; Beer, E.; Bowers, J.; Braslavsky, I.; Causey, M.; Colonell, J.; Dimeo, J.; Efcavitch, J. W.; Giladi, E.; Gill, J.; Healy, J.; Jarosz, M.; Lapen, D.; Moulton, K.; Quake, S. R.; Steinmann, K.; Thayer, E.; Tyurina, A.; Ward, R.; Weiss, H.; Xie, Z., Single-molecule DNA sequencing of a viral genome. *Science* 2008**,** 320, (5872), 106-109.

[47] Bayley, H., Sequencing single molecules of DNA. *Curr. Opin. Chem. Biol.* 2006**,** 10, (6), 628-637.

[48] Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W., Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* 1996**,** 93, (24), 13770-13773.

[49] Noolandi, J., A new concept for sequencing DNA by capillary electrophoresis. *Electrophoresis* 1992**,** 13, (6), 394-395.

[50] Mayer, P.; Slater, G. W.; Drouin, G., Theory of DNA-sequencing using free-solution electrophoresis of protein-DNA complexes. *Anal. Chem.* 1994**,** 66, (10), 1777-1780.

[51] Noolandi, J., A new concept for separating nucleic-acids by electrophoresis in solution using hybrid synthetic end labeled nucleic-acid molecules. *Electrophoresis* 1993**,** 14, (8), 680-681.

[52] Slater, G. W., Personal communication. In 2006.

[53] Vreeland, W. N.; Desruisseaux, C.; Karger, A. E.; Drouin, G.; Slater, G. W.; Barron, A. E., Molar mass profiling of synthetic polymers by free-solution capillary electrophoresis of DNA-polymer conjugates. *Anal. Chem.* 2001**,** 73, (8), 1795-1803.

[54] Heller, C.; Slater, G. W.; Mayer, P.; Dovichi, N.; Pinto, D.; Viovy, J. L.; Drouin, G., Free-solution electrophoresis of DNA. *J. Chromatogr.* 1998**,** 806, (1), 113-121.

[55] Ren, H.; Karger, A. E.; Oaks, F.; Menchen, S.; Slater, G. W.; Drouin, G., Separating DNA sequencing fragments without a sieving matrix. *Electrophoresis* 1999**,** 20, (12), 2501-2509.

[56] Zuckermann, R. N.; Kerr, J. M.; Kent, S. B. H.; Moos, W. H., Efficient method for the preparation of peptoids [oligo(N-substituted glycines)] by submonomer solid-phase synthesis. *J. Am. Chem. Soc.* 1992**,** 114, (26), 10646-10647.

[57] Kirshenbaum, K.; Barron, A. E.; Goldsmith, R. A.; Armand, P.; Bradley, E. K.; Truong, K. T. V.; Dill, K. A.; Cohen, F. E.; Zuckermann, R. N., Sequence-specific polypeptoids: A diverse family of heteropolymers with stable secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* 1998**,** 95, (8), 4303-4308.

[58] Simon, R. J.; Kania, R. S.; Zuckermann, R. N.; Huebner, V. D.; Jewell, D. A.; Banville, S.; Ng, S.; Wang, L.; Rosenberg, S.; Marlowe, C. K.; Spellmeyer, D. C.; Tan, R. Y.; Frankel, A. D.; Santi, D. V.; Cohen, F. E.; Bartlett, P. A., Peptoids - a modular approach to drug discovery. *Proc. Natl. Acad. Sci. U. S. A.* 1992**,** 89, (20), 9367-9371.

[59] Vreeland, W. N.; Barron, A. E., Free-solution capillary electrophoresis of polypeptoid-oligonucleotide conjugates. *Abstracts of Papers of the American Chemical Society* 2000**,** 219, U449-U449.

[60] Vreeland, W. N.; Meagher, R. J.; Barron, A. E., Multiplexed, high-throughput genotyping by single-base extension and end-labeled free-solution electrophoresis. *Anal. Chem.* 2002**,** 74, (17), 4328-4333.

[61] Meagher, R. J.; Coyne, J. A.; Hestekin, C. N.; Chiesl, T. N.; Haynes, R. D.; Won, J. I.; Barron, A. E., Multiplexed p53 mutation detection by free-solution conjugate microchannel electrophoresis with polyamide drag-tags. *Anal. Chem.* 2007**,** 79, (5), 1848-1854.

[62] Haynes, R. D.; Meagher, R. J.; Won, J. I.; Bogdan, F. M.; Barron, A. E., Comblike, monodisperse polypeptoid drag-tags for DNA separations by end-labeled free-solution electrophoresis (ELFSE). *Bioconj. Chem.* 2005**,** 16, (4), 929-938.

[63] Coyne, J. C., Lin, J. S., Barron, A. E., DNA sequencing and genotyping by free-solution conjugate electrophoresis. In *Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques*, third ed.; Landers, J. P., Ed. CRC Press: New York, 2008; p 1567.

[64] van Hest, J. C. M.; Tirrell, D. A., Protein-based materials, toward a new level of structural control. *Chem. Commun.* 2001, (19), 1897-1904.

[65] Kiick, K. L.; Saxon, E.; Tirrell, D. A.; Bertozzi, C. R., Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc. Natl. Acad. Sci. U. S. A.* 2002**,** 99, (1), 19-24.

[66] Connor, R. E.; Tirrell, D. A., Non-canonical amino acids in protein polymer design. *Polymer Reviews* 2007**,** 47, (1), 9-28.

[67] Kim, W.; McMillan, R. A.; Snyder, J. P.; Conticello, V. P., A stereoelectronic effect on turn formation due to proline substitution in elastin-mimetic polypeptides. *J. Am. Chem. Soc.* 2005**,** 127, (51), 18121-18132.

[68] Cappello, J.; Crissman, J.; Dorman, M.; Mikolajczak, M.; Textor, G.; Marquet, M.; Ferrari, F., Genetic-engineering of structural protein polymers. *Biotechnol. Prog.* 1990**,** 6, (3), 198-202.

[69] Nagarsekar, A.; Crissman, J.; Crissman, M.; Ferrari, F.; Cappello, J.; Ghandehari, H., Genetic engineering of stimuli-sensitive silkelastin-like protein block copolymers. *Biomacromolecules* 2003**,** 4, (3), 602-607.

[70] Megeed, Z.; Haider, M.; Li, D. Q.; O'Malley, B. W.; Cappello, J.; Ghandehari, H., In vitro and in vivo evaluation of recombinant silk-elastinlike hydrogels for cancer gene therapy. *J. Controlled Release* 2004**,** 94, (2-3), 433-445.

[71] Nagarsekar, A.; Crissman, J.; Crissman, M.; Ferrari, F.; Cappello, J.; Ghandehari, H., Genetic synthesis and characterization of pH- and temperature-sensitive silk-elastinlike protein block copolymers. *J. Biomed. Mater. Res.* 2002**,** 62, (2), 195-203.

[72] Chilkoti, A.; Dreher, M. R.; Meyer, D. E., Design of thermally responsive, recombinant polypeptide carriers for targeted drug delivery. *Adv. Drug Del. Rev.* 2002**,** 54, (8), 1093-1111.

[73] Chilkoti, A.; Dreher, M. R.; Meyer, D. E.; Raucher, D., Targeted drug delivery by thermally responsive polymers. *Adv. Drug Del. Rev.* 2002**,** 54, (5), 613-630.

[74] Meyer, D. E.; Chilkoti, A., Genetically encoded synthesis of protein-based polymers with precisely specified molecular weight and sequence by recursive directional ligation: Examples from the elastin-like polypeptide system. *Biomacromolecules* 2002**,** 3, (2), 357-367.

[75] McHale, M. K.; Setton, L. A.; Chilkoti, A., Synthesis and in vitro evaluation of enzymatically cross-linked elastin-like polypeptide gels for cartilaginous tissue repair. *Tissue Eng.* 2005**,** 11, (11-12), 1768-1779.

[76] Christensen, T.; Trabbic-Carlson, K.; Liu, W. G.; Chilkoti, A., Purification of recombinant proteins from Escherichia coli at low expression levels by inverse transition cycling. *Anal. Biochem.* 2007**,** 360, (1), 166-168.

[77] Lim, D. W.; Trabbic-Carlson, K.; MacKay, J. A.; Chilkoti, A., Improved non-chromatographic purification of a recombinant protein by cationic elastin-like polypeptides. *Biomacromolecules* 2007**,** 8, (5), 1417-1424.

[78] Shamji, M. F.; Betre, H.; Kraus, V. B.; Chen, J.; Chilkoti, A.; Pichika, R.; Masuda, K.; Setton, L. A., Development and characterization of a fusion protein between thermally responsive elastin-like polypeptide and interleukin-1 receptor antagonist - Sustained release of a local antiinflammatory therapeutic. *Arthritis Rheum.* 2007**,** 56, (11), 3650-3661.

[79] Simnick, A. J.; Lim, D. W.; Chow, D.; Chilkoti, A., Biomedical and biotechnological applications of elastin-like polypeptides. *Polymer Reviews* 2007**,** 47, (1), 121-154.

[80] McMillan, R. A.; Lee, T. A. T.; Conticello, V. P., Rapid assembly of synthetic genes encoding protein polymers. *Macromolecules* 1999**,** 32, (11), 3643-3648.

[81] McMillan, R. A.; Conticello, V. P., Synthesis and characterization of elastin-mimetic protein gels derived from a well-defined polypeptide precursor. *Macromolecules* 2000**,** 33, (13), 4809-4821.

[82] Qu, Y.; Payne, S. C.; Apkarian, R. P.; Conticello, V. P., Self-assembly of a polypeptide multi-block copolymer modeled on dragline silk proteins. *J. Am. Chem. Soc.* 2000**,** 122, (20), 5014-5015.

[83] Goeden-Wood, N. L.; Conticello, V. P.; Muller, S. J.; Keasling, J. D., Improved assembly of multimeric genes for the biosynthetic production of protein polymers. *Biomacromolecules* 2002**,** 3, (4), 874-879.

[84] Nagapudi, K.; Brinkman, W. T.; Thomas, B. S.; Park, J. O.; Srinivasarao, M.; Wright, E.; Conticello, V. P.; Chaikof, E. L., Viscoelastic and mechanical behavior of recombinant protein elastomers. *Biomaterials* 2005**,** 26, (23), 4695-4706.

[85] Kim, W.; Conticello, V. P., Protein engineering methods for investigation of structure-function relationships in protein-based elastomeric materials. *Polymer Reviews* 2007**,** 47, (1), 93-119.

[86] Mi, L. X., Molecular cloning of protein-based polymers. *Biomacromolecules* 2006**,** 7, (7), 2099-2107.

[87] Mi, L. X.; Fischer, S.; Chung, B.; Sundelacruz, S.; Harden, J. L., Self-assembling protein hydrogels with modular integrin binding domains. *Biomacromolecules* 2006**,** 7, (1), 38-47.

[88] Rizzi, S. C.; Hubbell, J. A., Recombinant protein-co-PEG networks as cell-adhesive and proteolytically degradable hydrogel matrixes. Part 1: Development and physicochernical characteristics. *Biomacromolecules* 2005**,** 6, (3), 1226-1238.

[89] Rizzi, S. C.; Ehrbar, M.; Halstenberg, S.; Raeber, G. P.; Schmoekel, H. G.; Hagenmuller, H.; Muller, R.; Weber, F. E.; Hubbell, J. A., Recombinant protein-co-PEG networks as cell-adhesive and proteolytically degradable hydrogel matrixes. Part II: Biofunctional characteristics. *Biomacromolecules* 2006**,** 7, (11), 3019-3029.

[90] Halstenberg, S., Biologically engineered protein-graft-poly(ethylene glycol) hydrogels: A cell adhesive and plasm in-degradable biosynthetic material for tissue repair. *Biomacromolecules* 2002**,** 3, (4), 710-723.

[91] Altman, G. H.; Diaz, F.; Jakuba, C.; Calabro, T.; Horan, R. L.; Chen, J. S.; Lu, H.; Richmond, J.; Kaplan, D. L., Silk-based biomaterials. *Biomaterials* 2003**,** 24, (3), 401-416.

[92] Huang, J.; Valluzzi, R.; Bini, E.; Vernaglia, B.; Kaplan, D. L., Cloning, expression, and assembly of sericin-like protein. *J. Biol. Chem.* 2003**,** 278, (46), 46117-46123.

[93] Bini, E.; Foo, C. W. P.; Huang, J.; Karageorgiou, V.; Kitchel, B.; Kaplan, D. L., RGD-functionalized bioengineered spider dragline silk biomaterial. *Biomacromolecules* 2006**,** 7, (11), 3139-3145.

[94] Huang, J.; Foo, C. W. P.; Kaplan, D. L., Biosynthesis and applications of silk-like and collagen-like proteins. *Polymer Reviews* 2007**,** 47, (1), 29-62.

[95] Vepari, C.; Kaplan, D. L., Silk as a biomaterial. *Prog. Polym. Sci.* 2007**,** 32, (8-9), 991-1007.

[96] Prince, J. T.; McGrath, K. P.; Digirolamo, C. M.; Kaplan, D. L., Construction, cloning, and expression of synthetic genes encoding spider dragline silk. *Biochemistry (Mosc).* 1995**,** 34, (34), 10879-10885.

[97] Asakura, T.; Nitta, K.; Yang, M. Y.; Yao, J. M.; Nakazawa, Y.; Kaplan, D. L., Synthesis and characterization of chimeric silkworm silk. *Biomacromolecules* 2003**,** 4, (3), 815-820.

[98] Farmer, R. S.; Kiick, K. L., Conformational behavior of chemically reactive alanine-rich repetitive protein polymers. *Biomacromolecules* 2005**,** 6, (3), 1531-1539.

[99] Wang, Y.; Kiick, K. L., Monodisperse protein-based glycopolymers via a combined biosynthetic and chemical approach. *J. Am. Chem. Soc.* 2005**,** 127, (47), 16392-16393.

[100] Farmer, R. S.; Argust, L. M.; Sharp, J. D.; Kiick, K. L., Conformational properties of helical protein polymers with varying densities of chemically reactive groups. *Macromolecules* 2006**,** 39, (1), 162-170.

[101] Kiick, K. L., Biosynthetic methods for the production of advanced protein-based materials. *Polymer Reviews* 2007**,** 47, (1), 1-7.

[102] Farmer, R. S.; Top, A.; Argust, L. M.; Liu, S.; Kiick, K. L., Evaluation of conformation and association behavior of multivalent alanine-rich polypeptides. *Pharm. Res.* 2008**,** 25, (3), 700-708.

[103] Liu, S.; Kiick, K. L., Architecture effects on the binding of cholera toxin by helical glycopolypeptides. *Macromolecules* 2008**,** 41, 764-772.

[104] Wang, C.; Stewart, R. J.; Kopecek, J., Hybrid hydrogels assembled from synthetic polymers and coiled-coil protein domains. *Nature* 1999**,** 397, (6718), 417-420.

[105] Wang, C.; Kopecek, J.; Stewart, R. J., Hybrid hydrogels cross-linked by genetically engineered coiled-coil block proteins. *Biomacromolecules* 2001**,** 2, (3), 912-920.

[106] Xu, C. Y.; Breedveld, V.; Kopecek, J., Reversible hydrogels from self-assembling genetically engineered protein block copolymers. *Biomacromolecules* 2005**,** 6, (3), 1739-1749.

[107] Kopecek, J., Hydrogel biomaterials: A smart future? *Biomaterials* 2007**,** 28, (34), 5185-5192.

[108] McGrath, K. P.; Tirrell, D. A.; Kawai, M.; Mason, T. L.; Fournier, M. J., Chemical and biosynthetic spproaches to the production of novel polypeptide materials. *Biotechnol. Prog.* 1990**,** 6, (3), 188-192.

[109] McGrath, K. P.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A., Genetically directed syntheses of new polymeric materials - expression of artificial genes encoding proteins with repeating (alagly)3proglugly elements. *J. Am. Chem. Soc.* 1992**,** 114, (2), 727-733.

[110] Zhang, G. H.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A., Biological synthesis of monodisperse derivatives of poly(alpha,L-glutamic Acid) - model rodlike polymers. *Macromolecules* 1992**,** 25, (13), 3601-3603.

[111] Petka, W. A.; Harden, J. L.; McGrath, K. P.; Wirtz, D.; Tirrell, D. A., Reversible hydrogels from self-assembling artificial proteins. *Science* 1998**,** 281, (5375), 389-392.

[112] Panitch, A.; Yamaoka, T.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A., Design and biosynthesis of elastin-like artificial extracellular matrix proteins containing periodically spaced fibronectin CS5 domains. *Macromolecules* 1999**,** 32, (5), 1701-1703.

[113] McPherson, D. T.; Morrow, C.; Minehan, D. S.; Wu, J. G.; Hunter, E.; Urry, D. W., Production and Purification of a Recombinant Elastomeric Polypeptide, G-(Vpgvg)19-Vpgv, from Escherichia-Coli. *Biotechnol. Prog.* 1992**,** 8, (4), 347-352.

[114] Urry, D. W., Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers. *J. Phys. Chem. B* 1997**,** 101, (51), 11007-11028.

[115] Urry, D. W., Five axioms for the functional design of peptide-based polymers as molecular machines and materials: Principle for macromolecular assemblies. *Biopolymers* 1998**,** 47, (2), 167-178.

[116] Lee, J.; Macosko, C. W.; Urry, D. W., Phase transition and elasticity of protein-based hydrogels. *Journal of Biomaterials Science-Polymer Edition* 2001**,** 12, (2), 229-242.

[117] Woods, T. C.; Urry, D. W., Controlled release of phosphorothioates by protein-based polymers. *Drug Deliv.* 2006**,** 13, (4), 253-259.

[118] Qi, M.; O'Brien, J. P.; Yang, J. J., A recombinant triblock protein polymer with dispersant and binding properties for digital printing. *Biopolymers* 2008**,** 90, (1), 28-36.

[119] Kumar, M.; Cuevas, W. A. Use of repeat sequence protein polymers in personal care compositions. US 07297678, Nov 20 2007, 2007.

[120] Won, J. I.; Meagher, R. J.; Barron, A. E., Characterization of glutamine deamidation in a long, repetitive protein polymer via bioconjugate capillary electrophoresis (vol 5, pg 619, 2004). *Biomacromolecules* 2004**,** 5, (4), 1624-1624.

[121] Meagher, R. J.; Won, J. I.; McCormick, L. C.; Nedelcu, S.; Bertrand, M. M.; Bertram, J. L.; Drouin, G.; Barron, A. E.; Slater, G. W., End-labeled free-solution electrophoresis of DNA. *Electrophoresis* 2005**,** 26, (2), 331-350.

[122] Won, J. I.; Meagher, R. J.; Barron, A. E., Protein polymer drag-tags for DNA separations by end-labeled free-solution electrophoresis. *Electrophoresis* 2005**,** 26, (11), 2138-2148.

[123] Meagher, R. J.; Won, J. I.; Coyne, J. A.; Lin, J.; Barron, A. E., Sequencing of DNA by Free-Solution Capillary Electrophoresis Using a Genetically Engineered Protein Polymer Drag-Tag. *Anal. Chem.* 2008**,** 80, (8), 2842-2848.

[124] Sambrook, J.; Fritsch, E. F.; Maniatis, T., *Molecular Cloning A Laboratory Manual Second Edition Vols. 1 2 and 3*. 1989; p XXXIX+PAGINATION VARIES(VOL 1), XXXIII+PAGINATION VARIES(VOL 2), XXXII+PAGINATION VARIES(VOL 3).

[125] pET System Manual. In [Online] 11 ed.; Novagen: Madison, WI, 2006. http://www.emdbiosciences.com/docs/docs/PROT/TB055.pdf.

[126] Karplus, P. A., Hydrophobicity regained. *Protein Sci.* 1997**,** 6, (6), 1302-1307.

[127] Grossman, P. D.; Colburn, J. C., *Capillary Electrophoresis: Theory and Practice*. Academic Press: San Diego, 1992.

[128] Garnier, J.; Gibrat, J. F.; Robson, B., GOR method for predicting protein secondary structure from amino acid sequence. In *Computer Methods for Macromolecular Sequence Analysis*, 1996; Vol. 266, pp 540-553.

[129] Kishon, M. Computational Design of Polymeric End-Labels for Microchannel DNA Sequencing by End-Labeled Free-Solution Electrophoresis. Northwestern University, Evanston, IL, 2000.

[130] Moriyama, E. N.; Powell, J. R., Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. *Nucleic Acids Res.* 1998**,** 26, (13), 3188-3193.

[131] Bulmer, M., The selection-mutation-drift theory of synonymous codon usage. *Genetics* 1991**,** 129, (3), 897-907.

[132] Henderson, D. B.; Davis, R. M.; Ducker, W. A.; Van Cott, K. E., Cloning strategy for producing brush-forming protein-based polymers. *Biomacromolecules* 2005**,** 6, (4), 1912-1920.

[133] Won, J. I.; Barron, A. E., A new cloning method for the preparation of long repetitive polypeptides without a sequence requirement. *Macromolecules* 2002**,** 35, (22), 8281-8287.

[134] Voet, D.; Voet, J. G.; Pratt, C. W., Fundamentals of Biochemistry. In John Wiley & Sons, Inc.: New York, 1999.

[135] Nilsson, J.; Stahl, S.; Lundeberg, J.; Uhlen, M.; Nygren, P. A., Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Expression Purif.* 1997**,** 11, (1), 1-16.

[136] Porath, J.; Carlsson, J.; Olsson, I.; Belfrage, G., Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature* 1975**,** 258, (5536), 598-599.

[137] Terpe, K., Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* 2003**,** 60, (5), 523-533.

[138] TALON Metal Affinity Resins User Manual. In [Online] Clontech Laboratories, Inc.: Mountain View, CA, 2007. http://www.clontech.com/images/pt/PT1320-1.pdf.

[139] Gross, E., The cyanogen bromide reaction. *Methods Enzymol.* 1967**,** 11, 238-255.

[140] *Protein Purification Protocols*. 2 ed.; Humana Press: Totowa, NJ, 2003; p 496.

[141] Price, N. C., *Proteins*. Academic Press, Inc.: San Diego. CA, 1996; p 318.

[142] Occupational Outlook Handbook. http://www.bls.gov/oco/ocos115.htm (5/1/08),

[143] Arnau, J.; Lauritzen, C.; Petersen, G. E.; Pedersen, J., Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. *Protein Expression Purif.* 2006**,** 48, (1), 1-13.

[144] Trabbic-Carlson, K.; Liu, L.; Kim, B.; Chilkoti, A., Expression and purification of recombinant proteins from Escherichia coli: Comparison of an elastin-like polypeptide fusion with an oligohistidine fusion. *Protein Sci.* 2004**,** 13, (12), 3274-3284.

[145] Banki, M. R.; Wood, D. W., Inteins and affinity resin substitutes for protein purification and scale up. *Microbial Cell Factories* 2005**,** 4.

[146] Ge, X.; Yang, D. S. C.; Trabbic-Carlson, K.; Kim, B.; Chilkoti, A.; Filipe, C. D. M., Self-cleavable stimulus responsive tags for protein purification without chromatography. *J. Am. Chem. Soc.* 2005**,** 127, (32), 11228-11229.

[147] Wu, W. Y.; Mee, C.; Califano, F.; Banki, R.; Wood, D. W., Recombinant protein purification by self-cleaving aggregation tag. *Nature Protocols* 2006**,** 1, (5), 2257-2262.

[148] Banki, M. R.; Feng, L. A.; Wood, D. W., Simple bioseparations using self-cleaving elastin-like polypeptide tags. *Nat. Methods* 2005**,** 2, (9), 659-661.

[149] Banki, M. R.; Gerngross, T. U.; Wood, D. W., Novel and economical purification of recombinant proteins: Intein-mediated protein purification using in vivo polyhydroxybutyrate (PHB) matrix association. *Protein Sci.* 2005**,** 14, (6), 1387-1395.

[150] Chow, D. C.; Dreher, M. R.; Trabbic-Carlson, K.; Chilkoti, A., Ultra-high expression of a thermally responsive recombinant fusion protein in E-coli. *Biotechnol. Prog.* 2006**,** 22, (3), 638-646.

[151] Mee, C.; Banki, M. R.; Wood, D. W., Towards the elimination of chromatography in protein purification: Expressing proteins engineered to purify themselves. *Chem. Eng. J.* 2008**,** 135, (1-2), 56-62.

[152] Trabbic-Carlson, K.; Meyer, D. E.; Liu, L.; Piervincenzi, R.; Nath, N.; LaBean, T.; Chilkoti, A., Effect of protein fusion on the transition temperature of an environmentally responsive elastin-like polypeptide: a role for surface hydrophobicity? *Protein Engineering Design & Selection* 2004**,** 17, (1), 57-66.

[153] Manning, G. S., Limiting laws and counterion condensation in poly-electrolyte solutions .7. Electrophoretic mobility and conductance. *J. Phys. Chem.* 1981**,** 85, (11), 1506-1515.

[154] Scotchle, J. W.; Robinson, A. B., Deamidation of glutaminyl residues - dependence on ph, temperature, and ionic-strength. *Anal. Biochem.* 1974**,** 59, (1), 319-322.

[155] Joshi, A. B.; Kirsch, L. E., The relative rates of glutamine and asparagine deamidation in glucagon fragment 22-29 under acidic conditions. *J. Pharm. Sci.* 2002**,** 91, (11), 2332-2345.

[156] Nakamura, Y.; Gojobori, T.; Ikemura, T., Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res.* 1998**,** 26, (1), 334-334.

[157] Gerber, S.; Kirchhof, K.; Kressler, J.; Schmelzer, C. E. H.; Scholz, C.; Hertel, T. C.; Pietzsch, M., Cloning, expression, purification, and characterization of a designer protein with repetitive sequences. *Protein Expression Purif.* 2008**,** 59, (2), 203-214.

[158] Belfort, M., Derbyshire, V., Stoddard, B.L., Wood, D.W., *Homing Endonucleases and Inteins*. Springer: Heidelberg, Germany, 2005; p 377.

[159] Perler, F. B.; Davis, E. O.; Dean, G. E.; Gimble, F. S.; Jack, W. E.; Neff, N.; Noren, C. J.; Thorner, J.; Belfort, M., Protein splicing elements - inteins and exteins - a definition of terms and recommended nomenclature. *Nucleic Acids Res.* 1994**,** 22, (7), 1125-1127.

[160] Chong, S. R.; Mersha, F. B.; Comb, D. G.; Scott, M. E.; Landry, D.; Vence, L. M.; Perler, F. B.; Benner, J.; Kucera, R. B.; Hirvonen, C. A.; Pelletier, J. J.; Paulus, H.; Xu, M. Q., Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. *Gene* 1997**,** 192, (2), 271-281.

[161] Zhang, A. H.; Gonzalez, S. M.; Cantor, E. J.; Chong, S. R., Construction of a mini-intein fusion system to allow both direct monitoring of soluble protein expression and rapid purification of target proteins. *Gene* 2001**,** 275, (2), 241-252.

[162] Xu, M. Q.; Paulus, H.; Chong, S. R., Fusions to self-splicing inteins for protein purification. In *Applications of Chimeric Genes and Hybrid Proteins, Pt A*, 2000; Vol. 326, pp 376-418.

[163] Wood, D. W.; Wu, W.; Belfort, G.; Derbyshire, V.; Belfort, M., A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol.* 1999**,** 17, (9), 889-892.

[164] Ma, J. F.; Cooney, C. L., Application of vortex flow adsorption technology to intein-mediated recovery of recombinant human alpha 1-antitrypsin. *Biotechnol. Prog.* 2004**,** 20, (1), 269-276.

[165] Sharma, S. S.; Chong, S. R.; Harcum, S. W., Simulation of large-scale production of a soluble recombinant protein expressed in Escherichia coli using an intein-mediated purification system. *Appl. Biochem. Biotechnol.* 2005**,** 126, (2), 93-117.

[166] IMPACT-CN Instruction Manual. In [Online] New England BioLabs: Ipswich, MA, 2001.

[167] Chong, S. R.; Montello, G. E.; Zhang, A. H.; Cantor, E. J.; Liao, W.; Xu, M. Q.; Benner, J., Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. *Nucleic Acids Res.* 1998**,** 26, (22), 5109-5115.

[168] Wood, D. W.; Derbyshire, V.; Wu, W.; Chartrain, M.; Belfort, M.; Belfort, G., Optimized single-step affinity purification with a self-cleaving intein applied to human acidic fibroblast growth factor. *Biotechnology Progress* 2000**,** 16, (6), 1055-1063.

[169] Gangopadhyay, J. P.; Jiang, S.-q.; van Berkel, P.; Paulus, H., In vitro splicing of erythropoietin by the Mycobacterium tuberculosis RecA intein without substituting amino acids at the splice junctions. *Biochimica et Biophysica Acta (BBA) - General Subjects* 2003**,** 1619, (2), 193-200.

[170] Meagher, R. J. DNA Sequencing, Genotyping, and Analysis by End-Labeled Free-Solution Electrophoresis (ELFSE). Northwestern University, Evanston, IL, 2005.

[171] Hermanson, G. T., *Bioconjugate Techniques*. Academic Press: San Diego, CA, 1996; p 785.

[172] Kolb, H. C.; Finn, M. G.; Sharpless, K. B., Click chemistry: Diverse chemical function from a few good reactions. *Angewandte Chemie-International Edition* 2001**,** 40, (11), 2004-+.

[173] Kolb, H. C.; Sharpless, K. B., The growing impact of click chemistry on drug discovery. *Drug Discov. Today* 2003**,** 8, (24), 1128-1137.

[174] Meagher, R. J.; McCormick, L. C.; Haynes, R. D.; Won, J. I.; Lin, J. S.; Slater, G. W.; Barron, A. E., Free-solution electrophoresis of DNA modified with drag-tags at both ends. *Electrophoresis* 2006**,** 27, (9), 1702-1712.

[175] Kaiser, R.; Metzka, L., Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds. *Anal. Biochem.* 1999**,** 266, (1), 1-8.

[176] Goodlett, D. R.; Armstrong, F. B.; Creech, R. J.; Vanbreemen, R. B., Formylated peptides from cyanogen-bromide digests identified by fast atom bombardment mass-spectrometry. *Anal. Biochem.* 1990**,** 186, (1), 116-120.

[177] Duewell, H. S.; Honek, J. F., CNBr/formic acid reactions of methionine- and trifluoromethionine-containing lambda lysozyme: Probing chemical and positional reactivity and formylation side reactions by mass spectrometry. *J. Protein Chem.* 1998**,** 17, (4), 337-350.

[178] Buchholz, B. A.; Zahn, J. M.; Kenward, M.; Slater, G. W.; Barron, A. E., Flow-induced chain scission as a physical route to narrowly distributed, high molar mass polymers. *Polymer* 2004**,** 45, (4), 1223-1234.

[179] Yang, L.; Liu, Z.-R., Bacterially expressed recombinant p68 RNA helicase is phosphorylated on serine, threonine, and tyrosine residues. *Protein Expression and Purification* 2004**,** 35, (2), 327-333.

[180] Winkler, S.; Wilson, D.; Kaplan, D. L., Controlling beta-sheet assembly in genetically engineered silk by enzymatic phosphorylation/dephosphorylation. *Biochemistry (Mosc).* 2000**,** 39, (41), 12739-12746.

[181] Fahnestock, S. R.; Irwin, S. L., Synthetic spider dragline silk proteins and their production in Escherichia coli. *Appl. Microbiol. Biotechnol.* 1997**,** 47, (1), 23-32.

[182] Robinson, M.; Lilley, R.; Little, S.; Emtage, J. S.; Yarranton, G.; Stephens, P.; Millican, A.; Eaton, M.; Humphreys, G., Codon usage can affect efficency of translation of genes in Escherichia-coli. *Nucleic Acids Res.* 1984**,** 12, (17), 6663-6671.

[183] Rosenberg, A. H.; Goldman, E.; Dunn, J. J.; Studier, F. W.; Zubay, G., Effects of consecutive AGG codons on translation in Escherichia-coliI, demonstrated with a versatile codon test system. *J. Bacteriol.* 1993**,** 175, (3), 716-722.

[184] Kuliopulos, A.; Walsh, C. T., Production, purification, and cleavage of tandem repeats of recombinant peptides. *J. Am. Chem. Soc.* 1994**,** 116, (11), 4599-4607.

[185] Armstrong, M. D., The relationship between homoserine and its lactone. *J. Am. Chem. Soc.* 1949**,** 71, (10), 3399-3402.

[186] Murphy, C. M.; Fenselau, C., Recognition of the carboxy-terminal peptide in cyanogen-bromide digests of proteins. *Anal. Chem.* 1995**,** 67, (9), 1644-1645.

[187] Lee, T. D.; Shively, J. E., Enzymatic and chemical digestion of proteins for mass-spectrometry. *Methods Enzymol.* 1990**,** 193, 361-374.

[188] Rydzanicz, R.; Zhao, X. S.; Johnson, P. E., Assembly PCR oligo maker: a tool for designing oligodeoxynucleotides for constructing long DNA molecules for RNA production. *Nucl. Acids Res.* 2005**,** 33, (suppl_2), W521-525.

[189] Mukherjee, S.; Shukla, A.; Guptasarma, P., Single-step purification of a protein-folding catalyst, the SlyD peptidyl prolyl isomerase (PPI), from cytoplasmic extracts of Escherichia coli. *Biotechnol. Appl. Biochem.* 2003**,** 37, 183-186.

[190] Parsy, C. B.; Chapman, C. J.; Barnes, A. C.; Robertson, J. F.; Murray, A., Two-step method to isolate target recombinant protein from co-purified bacterial contaminant SlyD after immobilised metal affinity chromatography. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 2007**,** 853, (1-2), 314-319.

[191] McMurry, J. L., Macnab, R.M., Talon resin does not bind *E. coli* SlyD, a common contaminant in Ni-NTA IMAC. In [Online] Clontech Laboratories, Inc.: Mountain View, CA, 2004. http://www.clontech.com/images/brochures/AN6Z2212_TALON_US.pdf.

[192] Heilshorn, S. C.; Liu, J. C.; Tirrell, D. A., Cell-binding domain context affects cell behavior on engineered proteins. *Biomacromolecules* 2005**,** 6, (1), 318-323.

[193] Mattson, G.; Conklin, E.; Desai, S.; Nielander, G.; Savage, M. D.; Morgensen, S., A practical approach to cross-linking. *Mol. Biol. Rep.* 1993**,** 17, (3), 167-183.

[194] Cuatrecasas, P.; Parikh, I., Adsorbents for affinity chromatography. Use of N-hydroxysuccinimide esters of agarose. *Biochemistry (Mosc).* 1972**,** 11, (12), 2291-2299.

[195] Henaut, A., Danchin, A., *Analysis and predictions from Escherichia coli sequences, or E. coli in silico*. 2 ed.; American Society for Microbiology Press: Washington DC, 1996.

[196] Andre, P.; Long, D.; Ajdari, A., Polyelectrolyte/post collisions during electrophoresis: Influence of hydrodynamic interactions. *European Physical Journal B* 1998**,** 4, (3), 307-312.

[197] McCormick, L. C.; Slater, G. W., The molecular end effect and its critical impact on the behavior of charged-uncharged polymer conjugates during free-solution electrophoresis. *Electrophoresis* 2005**,** 26, (9), 1659-1667.

[198] Meyer, D. E.; Chilkoti, A., Quantification of the effects of chain length and concentration on the thermal behavior of elastin-like polypeptides. *Biomacromolecules* 2004**,** 5, (3), 846-851.

[199] Katayama, S., Chemical condition responsible for thermoswelling of thermoshrinking type of volume phase-transition in gels - effect of relative amounts of hydrophobic to hydrophilic groups in the side-chain. *J. Phys. Chem.* 1992**,** 96, (13), 5209-5210.

[200] Johnson, W. C., Protein secondary structure and circular-dichroism - a practical guide. *Proteins-Structure Function and Genetics* 1990**,** 7, (3), 205-214.

[201] Unneberg, P.; Merelo, J. J.; Chacon, P.; Moran, F., SOMCD: Method for evaluating protein secondary structure from UV circular dichroism spectra. *Proteins-Structure Function and Genetics* 2001**,** 42, (4), 460-470.

[202] Shortle, D., The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* 1996**,** 10, (1), 27-34.

[203] Smith, L. J.; Fiebig, K. M.; Schwalbe, H.; Dobson, C. M., The concept of a random coil - Residual structure in peptides and denatured proteins. *Fold. Des.* 1996**,** 1, (5), R95-R106.

[204] Grossman, P. D.; Soane, D. S., Experimental and theoretical-studies of DNA separations by capillary electrophoresis in entangled polymer-solutions. *Biopolymers* 1991**,** 31, (10), 1221-1228.

[205] Barron, A. E.; Blanch, H. W.; Soane, D. S., A transient entanglement coupling mechanism for DNA separation by capillary electrophoresis in ultradilute polymer-solutions. *Electrophoresis* 1994**,** 15, (5), 597-615.

[206] Broseta, D.; Leibler, L.; Lapp, A.; Strazielle, C., Universal properties of semidilute polymer-solutions - a comparison between experiments and theory. *Europhys. Lett.* 1986**,** 2, (9), 733-737.

[207] Barron, A. E.; Soane, D. S.; Blanch, H. W., Capillary electrophoresis of DNA in uncross-linked polymer-solutions. *J. Chromatogr.* 1993**,** 652, (1), 3-16.

Appendix A

PZ-1 to 8 and BB-1 Gene Sequences

PZ-1	ATA TAG AAT TCC TCT TCA GGT AGT GGC CAA GGA GAA AGT GGT AGT GGA CAG GGC GAG TCA GGA AGC GGC CAA GGT GAA AGC GGT AGA AGA GGA ATT CAT ATA
PZ-2	ATA TAG AAT TCC TCT TCA GGT GCG GGC CAA GGA GAA GCA GGT GCT GGA CAG GGC GAG GCA GGA GCT GGC CAA GGT GAA GCG GGT AGA AGA GGA ATT CAT ATA
PZ-3	ATA TAG AAT TCC TCT TCA GGT GTA GGC CAA GGA GAA GTG GGT GTT GGA CAG GGC GAG GTG GGA GTA GGC CAA GGT GAA GTT GGT AGA AGA GGA ATT CAT ATA
PZ-4	ATA TAG AAT TCC TCT TCA GGT CTG GGC CAA GGA GAA TTA GGT CTG GGA CAG GGC GAG TTG GGA CTT GGC CAA GGT GAA TTA GGT AGA AGA GGA ATT CAT ATA
PZ-5	ATA TAG AAT TCC TCT TCA GGT GCG GGC CAA GGA AAC GCA GGT GCT GGA CAG GGC AAT GCA GGA GCT GGC CAA GGT AAC GCG GGT AGA AGA GGA ATT CAT ATA
PZ-6	ATA TAG AAT TCC TCT TCA GGT GCG GGC CAA GGA AGT GCA GGT GCT GGA CAG GGC AGC GCA GGA GCT GGC CAA GGT TCC GCG GGT AGA AGA GGA ATT CAT ATA
PZ-7	ATA TAG AAT TCC TCT TCA GGT GCG GGC TCG GGA AGT GCA GGT GCT GGA TCA GGC AGC GCA GGA GCT GGC AGC GGT TCC GCG GGT AGA AGA GGA ATT CAT ATA
PZ-8	ATA TAG AAT TCC TCT TCA GGT GCG GGC ACC GGA AGT GCA GGT GCT GGA ACG GGC AGC GCA GGA GCT GGC ACC GGT TCC GCG GGT AGA AGA GGA ATT CAT ATA
BB-1	G CTA GCC ATA TGC TCT TCA GGT AAA GGC AGC GCG CAG GCC GGC AAG GGT TCT GCG CAA GCA GGC AAA GGT AGC GCC CAG GCG GGT TGA AGA GGG ATC CAC TAG T

Appendix B

Protein Polymer Production Cost Analysis

Protein Yield		7.50	mg for 4L		
Category	**Item**	**Price ($ per unit)**	**Unit**	**Amt Used**	**Cost**
Growth media and antibiotics	kanamycin	10.19	g	0.12	$1.26
Growth media and antibiotics	tetracycline	0.83	g	0.05	$0.04
Growth media and antibiotics	200 absolute ethanol	0.05	ml	2.00	$0.11
Growth media and antibiotics	LB broth Miller	0.08	g	2.50	$0.19
Growth media and antibiotics	TB media	0.11	g	190.40	$21.67
Growth media and antibiotics	glycerol	1.05	ml	16.00	$16.75
Growth media and antibiotics	agar	0.37	g	0.30	$0.11
Inducer (IPTG)	IPTG	12.50	g	0.95	$11.92
Buffer	urea	0.02	g	25.95	$0.51
Buffer	NaCl	0.10	g	0.95	$0.09
Buffer	imidazole	0.32	g	0.07	$0.02
Buffer	sodium phosphate monobasic	0.09	g	0.15	$0.01
Buffer	sodium phosphate dibasic	0.30	g	0.44	$0.13
Talon resin	Talon resin	6.97	ml	2.00	$13.94
Protease cleavage	Factor Xa cleavage capture kit	0.50	U	75.00	$37.13
Protease cleavage	buffer for second IMAC cleanup step				$0.77
Protease cleavage	repeat dialysis				$3.51
Miscellaneous	nitrile gloves	0.46	pair	13.00	$6.04
Miscellaneous	weigh boats	0.17	ea	8.00	$1.33
Miscellaneous	pipette tips	0.01	ea	18.00	$0.11
Miscellaneous	petri dish	0.30	ea	1.00	$0.30
Miscellaneous	BLR(DE3) competent cells	0.20	ul	20.00	$4.00
Miscellaneous	microcentrifuge tubes	0.10	ea	1.00	$0.10
Miscellaneous	25ml sterile serological pipet	0.91	ea	5.00	$4.53

Miscellaneous	plastic cuvettes	0.29	ea	2.00	$0.59
Miscellaneous	dialysis membrane	3.51	ft	1.00	$3.51
Miscellaneous	Falcon tube 50ml	0.54	ea	6.00	$3.23
Miscellaneous	Falcon tube 15ml	0.39	ea	3.00	$3.00
Miscellaneous	bottletop filter	7.27	ea	1.00	$7.27
Miscellaneous	foil	0.37	ft	1.00	$0.37
Miscellaneous	autoclave tape	0.03	ft	2.00	$0.06
Labor	protein expression and purification	17.17	hours	17.00	$291.89
Labor	protease cleavage IMAC cleanup	17.17	hours	4.50	$77.27

Protein Yield		**100**	**mg for 4L**		
Category	**Item**	**Price ($ per unit)**	**Unit**	**Amt Used**	**Cost**
Growth media and antibiotics	kanamycin	10.19	g	0.12	$1.26
Growth media and antibiotics	tetracycline	0.83	g	0.05	$0.04
Growth media and antibiotics	200 absolute ethanol	0.05	ml	2.00	$0.11
Growth media and antibiotics	LB broth Miller	0.08	g	2.50	$0.19
Growth media and antibiotics	TB media	0.11	g	190.40	$21.67
Growth media and antibiotics	glycerol	1.05	ml	16.00	$16.75
Growth media and antibiotics	agar	0.37	g	0.30	$0.11
Inducer (IPTG)	IPTG	12.50	g	0.95	$11.92
Buffer	urea	0.02	g	259.46	$5.09
Buffer	NaCl	0.10	g	9.47	$0.90
Buffer	imidazole	0.32	g	0.75	$0.24
Buffer	sodium phosphate monobasic	0.09	g	1.45	$0.12
Buffer	sodium phosphate dibasic	0.30	g	4.42	$1.33
Talon resin	Talon resin	6.97	ml	20.00	$139.40
Protease cleavage	Factor Xa cleavage capture kit	0.50	U	1000.00	$495.00

Protease cleavage	buffer for second IMAC cleanup step				$7.67
Protease cleavage	repeat dialysis				$3.51
Miscellaneous	nitrile gloves	0.46	pair	13.00	$6.04
Miscellaneous	weigh boats	0.17	ea	8.00	$1.33
Miscellaneous	pipette tips	0.01	ea	18.00	$0.11
Miscellaneous	petri dish	0.30	ea	1.00	$0.30
Miscellaneous	BLR(DE3) competent cells	0.20	ul	20.00	$4.00
Miscellaneous	microcentrifuge tubes	0.10	ea	1.00	$0.10
Miscellaneous	25ml sterile serological pipet	0.91	ea	5.00	$4.53
Miscellaneous	plastic cuvettes	0.29	ea	2.00	$0.59
Miscellaneous	dialysis membrane	3.51	ft	1.00	$3.51
Miscellaneous	Falcon tube 50ml	0.54	ea	6.00	$3.23
Miscellaneous	Falcon tube 15ml	0.39	ea	3.00	$3.00
Miscellaneous	bottletop filter	7.27	ea	1.00	$7.27
Miscellaneous	foil	0.37	ft	1.00	$0.37
Miscellaneous	autoclave tape	0.03	ft	2.00	$0.06
Labor	protein expression and purification	17.17	hours	20.00	$343.40
Labor	protease cleavage IMAC cleanup	17.17	hours	7.50	$128.78