Protein Expression and Folding Optimization for High-Throughput Proteomics

Kate Drahos

Henry Rutgers Scholars

A thesis in fulfillment of the requirements of

the Henry Rutgers Scholars Program

Written under the direction of

Dr. Gaetano T. Montelione

Department of Molecular Biology

and Biochemistry

Rutgers College

April 2004

Table of Contents

Abbreviations

NESG, Northeast Structural Genomics Consortium; MBP, Maltose Binding Protein; SUMO, Small Ubiquitin-related Modifier; TEV, Tobacco Etch Virus Protease; ORF, Open Reading Frame; HTP, High-Throughput; PDB, Protein Data Bank; 6X-His, Hexahistidine tag; SDS-PAGE, Sodium dodecyl sulfate-polyacrylamide gel electrophoresis; T, Total; S, Soluble; HSQC, Heteronuclear Single Quantum Coherence; NE, Not Expressed; E/NS, Expressed/Not Soluble; E/S, Expressed/Soluble

Introduction

The Human Genome Sequencing Project and its offshoots, including the sequencing of several eukaryotic model organisms has led to significant technological advances in DNA sequencing and a wealth of DNA sequence data. Currently, a vast number of genomes have been sequenced both from prokaryotic and eukaryotic origins. For example, 132 bacterial and 16 archeal genomes have been completely sequenced including species ranging from human pathogens such as *Salmonella typhimurium* (McClelland *et al*., 2001) to those from thermal vents such as *Pyrococcus horikoshii* (Gonzalez *et al.,* 1998). The number of eukaryotic organisms that have been sequenced is significantly less owing to the larger size and complexity of their genomes. However, many of the eukaryotic model organisms have been completely sequenced, the first of which was *Saccharomyces cerevisiae* (Goffeau *et al.,* 1996), others include the invertebrates *Caenorhabditis elegans* (No authors listed, 1998) and *Drosophila melanogaster* (Adams *et al.,* 2000), the plant *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), and of course *Homo sapiens* (Lander *et al.,* 2001). This wealth of sequence data has provided researchers with a list of all the protein-coding sequences in each of these organisms. For example, yeast was shown to possess roughly 6000 protein coding open reading frames (ORFs) (Goffeau *et al.,* 1996), while the human sequence indicates somewhere between 30,000 and 40,000 ORFs (Lander *et al.,* 2001). However, based on sequence homology, only about 50% of the yeast genes can be assigned to a known function. This number drops even lower for humans, such that roughly 40% of the ORFs can be functionally annotated (Lander *et al.,* 2001). In order to more deeply understand the biology of life, the next challenge for the biological community is to identify the functions of these ORFs. Currently, there are

many large-scale projects with the aim of elucidating the function of these proteins. For example, genome-wide yeast two-hybrid screens have been developed to identify protein-protein interactions. Other functional genomic assays such as microarray data show the expression profile for each ORF, giving clues to their functions (Winzeler *et al.*, 1999).

A well-known tenet of biology is that protein structure determines function. This has led to the field of structural genomics (Montelione & Anderson, 1999). One of the main goals of this field is to provide clues to the function of unknown proteins based on their three-dimensional (3D) structures (Montelione & Anderson, 1999). Based on the hypotheses generated from the structural information, functional studies can then be carried out in a more directed and efficient manner.

The Northeast Structural Genomics Consortium

The Northeast Structural Genomics Consortium (NESG) is one of nine National Institute of Health (NIH)-funded pilot projects; their primary goal is to develop the necessary technologies for high-throughput (HTP) protein production and 3D structure determination. The primary protein targets of the consortium are proteins from the eukaryotic model organisms *Arabidopsis thaliana, Saccharomyces cerevisiae, Caenorhabditis elegans*, *Drosophila melanogaster,* and humans. The NESG aims to use their established protein target list to develop automated HTP methods for expression plasmid construction, analytical-scale expression screening, and preparative-scale protein production. Using this vast number of protein targets, the consortium will develop and

optimize new HTP crystallization methods, improved solution state NMR data collection methods, and automated methods of structure determination from the raw data.

The project focuses on relatively small proteins, less than 340 amino acids in length, whose structures cannot be accurately predicted using homology modeling techniques. As the number of structures determined increases, a combination of sequence and structural data can be used to uncover evolutionary relationships among proteins. This can be extended to formulate hypotheses about biochemical and cellular functions of proteins, as well as allowing researchers to accurately model their proteins of interest based on structures submitted to the Protein Data Bank (PDB). The available 3D structures can be applied to the coverage of "protein fold space", which will map out new relationships among proteins. By obtaining structural examples of all types of protein folds, it can be determined which folds are evolutionary combinations of older, more simplified folds, and how different folds of similar "evolutionary ages" may or may not be related. Using protein structures from these model organisms, the NESG will be able to provide the scientific community with a vast amount of information about the relationship between a protein's structure and its function, as well as the relationships between many proteins' structures and functions.

High-Throughput Methodologies

To make the vision of projects such as the NESG a reality, it will be necessary to develop HTP methods of protein production and structure determination by both NMR and X-ray crystallography. Of critical importance to the success of these methods is the preparation

of protein samples suitable for 3D structure determination. This is markedly different from previous genome-wide studies, which focus on the preparation of nucleic acid samples (Lander, 1999; Lander *et al.,* 2001, Winzeler *et al.*, 1999), which have very similar biophysical properties enabling identical, well-defined purification and preparation techniques to be employed in an HTP setting. Even genome-scale studies focusing on proteins, such as Yeast Two-Hybrid Screens (Uetz *et al.,* 2000; Ito *et al.,* 2001), require relatively small amounts of protein expressed in a eukaryotic system. Furthermore, it is not necessary to purify these proteins to obtain experimental data. In stark contrast to this, structural genomics efforts require the production of large amounts of soluble, highly purified protein. Unlike the DNA samples prepared for sequencing studies, proteins have extremely diverse biophysical properties, including hydrophobic patches, transmembrane regions, and varied ligand binding sites, making the purification and preparation of suitable samples exceedingly more difficult, especially when considering the need for HTP methodologies.

In addition to the issues raised by the innate properties of proteins, the NESG is focused on both NMR and X-ray crystallographic methods of structure elucidation. This further complicates the need for obtaining suitable protein samples because each of these experimental approaches requires samples that are prepared in significantly different manners. While NMR requires isotope enrichment, usually ^{15}N and ^{13}C, X-ray crystallography requires selenomethionine incorporation. However, the well-developed technologies for these enrichment procedures have mostly been limited to prokaryotic expression systems. This creates yet another obstacle of expressing eukaryotic proteins

in a prokaryotic host. It is clear that one of the major contributions structural genomics will have on science is the addition of technologies that enhance our capabilities in the area of protein chemistry and our ability to provide suitable protein samples in order to carry out diverse biological studies.

The Insolubility Barrier

The common methods of protein overexpression using *E. coli* systems are often unsuccessful, for a variety of reasons. Many proteins may not be expressed at all, or they may fold incorrectly leading to proteolysis. In other cases, improper folding may lead to insolubility and inclusion body formation. It is estimated that one-third to one half of prokaryotic proteins cannot be expressed in a soluble form using *E. coli* expression systems and this estimate most likely increases for eukaryotic proteins (Edwards *et al.*, 2000). Protein insolubility may be caused by a variety of factors including the need for general molecular chaperones, cofactors, or protein partners for proper folding. Innate properties of proteins, such as hydrophobic regions or incompatibility with folding mechanisms of the host cell, may lead to insolubility (Edwards *et al.*, 2000).

Without soluble protein samples for study, structural genomics efforts grind to a halt. If efforts focus only on the small proportion of proteins easily expressed in *E. coli* systems, an unrepresentative set of structures will be generated which do not meet the needs, of projects such as the NESG. Complications due to protein expression and solubility issues also limit biochemical studies that may provide functional information, as an incorrectly folded protein will not perform its biochemical function. The need for soluble proteins

adopting their native structure is critical for 3D structure studies as well as biochemical studies that investigate specific protein function.

Emerging Technologies for Expression and Solubility Enhancement

Insolubility has thus been identified as a major barrier in the fields of structural genomics and protein chemistry. This is not, however, a situation without options, and many strategies are being developed to circumvent this problem. The different strategies under investigation can be classified into two major categories: (1) changes to the expression vector or construct and (2) changes to the host system. Strategies that involve changes to the expression vector include the use of fusion protein expression vectors (Kapust & Waugh, 1999; Hammarström *et al.,* 2002; www.lifesensors.com), the use of cold shock expression vectors (Qing *et al.*, submitted), and the rearrangement of affinity tags from N to C-terminus (Acton *et al.,* in press). Host systems other than the typical *E. coli* system include cell-free expression systems utilizing wheat embryos (Sawasaki *et al.,* 2002), yeast expression systems using *Pichia pastoris* (Cereghino & Cregg, 1999), and insect and human expression systems (Edwards *et al.,* 2000). A distinct third strategy involves the solubilization of inclusion bodies and the refolding of recombinant proteins (Tsumoto *et al.,* 2003).

Cold shock expression vectors, termed pCOLD, are currently being developed as complementary expression vectors to the widely used T7 systems. pCOLD vectors utilize key features of the *cspA* gene, which codes for the *E. coli* cold-shock protein CspA. Specifically, the vectors contain the *cspA* promoter, 5’ UTR, and 3’ transcription

termination sequence, enabling protein expression to be cold-shock inducible, in addition to the usual method of induction using isopropyl-β-d-thiogalactoside (IPTG). Thus far the pCOLD vectors have proved to be complementary to expression in pET vectors (Qing *et al.,* submitted) using our standard T7 Multiplex Expression System (Acton *et al.,* in press).

The rearrangement of small affinity tags is another common strategy employed to obtain soluble recombinant proteins. This laboratory, for example, utilizes modified pET14 and pET15 vectors (Novagen), which encode N-terminal hexahistidine tags (6X-His tag), and the pET21 (Novagen) vector, which encodes a C-terminal 6X-His tag, is also employed (Acton *et al.,* in press). If the N-terminus of the protein of interest is buried within the protein core, the affinity tag may interfere with proper folding, resulting in unfolded molecules. The protein may be rescued however, if the ORF is subcloned into a vector encoding a C-terminal affinity tag, provided that the C-terminus is not buried within the protein structure. Two examples include the human proteins targets HR945 and HR1869. Expressed with an N-terminal 6X-His tag, HR945 was expressed but insoluble. When subcloned into pET21 and expressed with a C-terminal 6X-His tag, HR945 became soluble. HR1869 was not expressed from pET15, but when subcloned into pET21, the protein was expressed and soluble. This data shows that in some cases, the position of the affinity tag can have profound effects on the expression and solubility of proteins.

A greater proportion of eukaryotic proteins are observed to be insoluble when expressed in *E. coli* systems, and it is worthwhile to explore alternative expression systems that do

not rely on prokaryotic hosts. Yaeta Endo's group in Japan has developed a cell-free expression system based on the translational machinery present in wheat embryos (Sawasaki *et al.*, 2002). This system can reach expression levels close to those of *in vivo* systems without the issue of potentially toxic proteins. This cell-free system appears to maintain stability for periods up to two weeks and it has proved amenable to HTP protein production needs. Within two days the system can produce hundreds of micrograms of protein for up to 50 unique genes in parallel.

While *in vitro* expression appears to be highly efficient, there is no substitute for *in vivo* protein expression. Eukaryotic systems are available and the yeast *Pichia pastoris* is a viable option for recombinant protein expression (Cregg *et al.*, 2000). One of the most attractive features of this eukaryote is its ability to perform some of the post-translational modifications common to eukaryotic proteins. *E. coli* is unable to catalyze these modifications, and many proteins may be lost due to this inability. The *P. pastoris* system can express both intra- and extracellular proteins at high-levels, and it shows promise as a large-scale protein production system (Cereghino *et al.*, 2002). Other eukaryotic expression systems include the use of insect (Jarvis 2003) and mammalian cell culture (Andersen & Krummen 2002). Although these systems may prove to be more adept at performing necessary post-translational modifications, such as glycosylation, to higher eukaryotic proteins, they are more costly than *E. coli* systems and do not fit the needs of HTP methodologies at the current time.

The third strategy differs from those previously discussed in that it is not based on changing the expression vector or host system, but rather starting with the expressed proteins as insoluble inclusion bodies. The inclusion bodies are solubilized by addition of detergents, 8 M urea, or 6 M guanidine-HCl. After solubilization, the protein is refolded by decreasing the concentration of denaturant present in the buffer, by methods such as dialysis or gel filtration. However, this method is quite time-consuming and it does not guarantee that proteins will fold into their native conformation. Therefore, it is desired that soluble, folded protein can be obtained without the addition of these refolding steps.

The aim of this study is to determine whether expression of recombinant proteins as fusion proteins in *E. coli* will increase their expression and, in particular, their solubility levels, as well as encourage proper folding of the proteins into their native state. We are exploring two different fusion protein expression systems. The first is the Gateway-Maltose Binding Protein (MBP) system, developed by our collaborator Dr. David Waugh at the National Cancer Institute, Frederick, MD (Kapust & Waugh, 1999). The second is the SUMO-fusion system, developed by Tauseef Butt at LifeSensors, Inc., Malvern, PA (please see www.lifesensors.com).

Maltose Binding Protein

MBP is a 40-kDa protein belonging to the periplasmic binding protein superfamily (Evenäs *et al.*, 2001). MBP is a dimer in its biologically active state, in which it binds various maltodextrins for transport into Gram-negative bacteria. In complex with the

proteins MalF, MalG, and MalK, these sugars are transported into the bacterial cell (Spurlino *et al.,* 1991). The structure of MBP is ellipsoid shaped and composed of two globular domains with a deep groove in the middle. The core of each domain is composed of a five-strand β-sheet surrounded by two α-helices on one side and three α-helices on the other (Spurlino *et al.*, 1991). The groove is the site of sugar binding and when bound, maltose is almost completely solvent shielded.

Gateway-MBP Fusion System

The MBP fusion system (New England Biolabs) was first designed to allow for one-step affinity purification of proteins of interest. This was achieved by fusing an ORF in frame with an N-terminal MBP gene. When expressed, the fusion protein could then be purified utilizing the affinity of MBP for its substrate amylose. Briefly, the fusion protein is expressed in *E. coli* and the soluble cell lysate passed over a column containing amylose linked to a matrix of agarose beads. The fusion protein is bound to the column based on the MBP-amylose interactions and purified away from the contaminating bacterial proteins that pass through the column. The protein of interest is then separated from the MBP using a protease specific for an amino acid sequence directly downstream of the MBP protein. As researchers incorporated this technology, several reports surfaced indicating that fusion with MBP can fortuitously increase the solubility of the recombinant protein (Kapust & Waugh, 1999, Pryor & Leiting 1997).

Our collaborator, David Waugh, has undertaken studies involving the solubility enhancement activity of MBP and conclusively shown that MBP can dramatically

increase the solubility of many proteins (Kapust & Waugh, 1999). This research showed that a variety of proteins could be rescued from inclusion body formation through fusion with MBP. The increased solubility was retained after proteolytic cleavage to separate the passenger proteins from MBP, suggesting that the solubilizing effect is more than a large soluble protein (MBP) keeping its fusion partner in solution. These observations suggest a chaperone-like ability attributable to MBP.

The purification of these fusion proteins, however, is not always a simple one step procedure. MBP-amylose chromatography is not always effective and the product is often not of sufficient purity after a single purification step (Routzahn & Waugh, 2002). Therefore, the Gateway-MBP system (Figure 1A and 1B) was developed by David Waugh, which incorporates other affinity tags, particularly the biotin acceptor peptide (BAP), and most importantly the 6X-His tag, both of which have been shown to be effective in the purification of MBP fusion proteins. These other affinity tags are small additions to the protein that enable purification regardless of amino acid composition, much like the larger MBP. Again, purification is achieved by exploiting properties of these tags, instead of having to utilize properties of each individual protein, which are dissimilar and not easily predictable. This allows for high-throughput parallelizable purification techniques (a requirement for structural genomics). This combination of fusion protein with supplementary affinity purification tag can result in the production of protein samples suitable for structural studies that would otherwise not be available if the common T7 systems were used.

The Gateway-MBP system allows for purification of the MBP fusion protein and subsequent cleavage of MBP from the protein of interest using the Tobacco Etch Virus (TEV) protease. The TEV protease recognizes the amino acid sequence Glu-X-X-Tyr-X-Gln-Ser, cleaving between Gln and Ser (Dougherty *et al.*, 1989). This cleavage step is necessary to carry out structural studies using NMR because of the large size of MBP (40 kDa). Although X-ray crystallography can be used to determine structures of high molecular weight proteins, cleavage is desirable as the long, flexible linker of the fusion protein may inhibit crystal formation or cause heterogeneity of crystals. Crystal structures of uncleaved fusion proteins have been determined by replacing the long, flexible linker with a short, rigid linker (see review by Smyth *et al.*, 2003). This cleavage step is also necessary to determine whether the protein of interest is soluble and properly folded, or whether the large size of MBP is merely dragging the passenger protein into solution. It is interesting to note that TEV protease is insoluble when expressed in *E. coli* and must be expressed as an MBP-fusion to achieve it native conformation and catalytic activity (Kapust & Waugh, 1999).

MBP as a Molecular Chaperone

As demonstrated by Kapust and Waugh (1999), not just any highly soluble protein will serve to increase solubility of its partner in the context of a fusion protein. Three highly soluble, frequently employed fusion partners, MBP, glutathione-S-transferase (GST), and thioredoxin, were chosen to examine their solubilizing activities on a range of highly insoluble proteins. It was found that only MBP displayed increased solubilizing activities. Fox *et al.* (2003) have subsequently shown that homologs of MBP in a variety of

microorganisms can serve as potent solubility enhancers when fused with proteins of limited solubility, indicating that this may be a general feature of this class of proteins. This raises the question: how is MBP functioning as a possible general molecular chaperone in the context of a fusion protein?

A model has been proposed to explain these observations (Figure 2) (Kapust & Waugh, 1999). After translation of the fusion protein, an initial Folding Intermediate forms, consisting of the properly folded "chaperone protein" and an unfolded passenger protein. If the passenger protein is able to fold into its native state, the product is a soluble fusion protein (Native Structure). If passenger proteins are not completely folded, they may self-associate, forming insoluble fusion protein aggregates (Insoluble Aggregates). It is hypothesized that physical interactions can occur between the surface of MBP and the passenger protein to form a Sequestered Intermediate. This interaction may help prevent the self-association of the passenger proteins that lead to Insoluble Aggregates.

It has also been suggested that some unfolded proteins can bind to potential hydrophobic patches on the surface of MBP, and hydrophobic residues have been identified that interact with other membrane transport components (Richarme & Caldas, 1997). These residues may serve as a non-specific hydrophobic binding site for unfolded proteins as well as the native protein components of the bacterial transport machinery. A mutational analysis of MBP identified surface residues important for the solubilizing effects that did not perturb the structure of the protein, providing further evidence of this chaperone-like activity (Fox *et al.*, 2001). Although single amino acid substitutions in the deep groove

of MBP did not disrupt its solubilizing activity, three substitutions on the surface of the protein, in close proximity to the groove, profoundly decreased this solubilizing activity. These mutations, W232E, Y242E, and I317E, form a small surface patch and demonstrate that these solvent exposed side chains are critical for the chaperone-like effect of MBP. Upon further investigation, it was shown that these mutations decrease the global stability of MBP in its unfused form (Fox *et al.,* 2001). These observations are consistent with the general molecular chaperone theory, in that rapid formation of native MBP is necessary to promote solubility of the passenger protein (Fox *et al.,* 2001) and a disruption of the protein interaction site of MBP would decrease the formation of the sequestered intermediate, which may be vital to the prevention of passenger protein aggregation.

Small Ubiquitin-Related Modifier

Ubiquitin and ubiquitin-like proteins (Ubl) are small (~100 amino acid) proteins that modulate protein function by way of post-translational covalent bonding to lysine residues of the target proteins (Bernier-Villamor *et al.*, 2002). Small ubiquitin-related modifier (SUMO) is a member of this protein family, found in eukaryotes ranging from yeast to humans. Yeast SUMO, also known as Smt3, is an 11 kDa protein that exhibits 18.4% sequence identity with ubiquitin (44.7% sequence similarity) (Saitoh *et al.*, 1997). While yeast contain one SUMO protein, three members of the SUMO family have been discovered in mammals (SUMO-1, SUMO-2, and SUMO-3) (Verger *et al.,* 2003). Polysumoylation by SUMO-1 has not been observed, while this has been seen for SUMO-2 and SUMO-3, indicating different functions among these proteins.

While ubiquitination has been shown to target proteins to the 26S proteosome for degradation (Hershko & Cienchanover, 1998), sumoylation appears to modify protein function in several ways, by modulating protein-protein interactions, altering cellular localization, or protecting proteins from degradation. A number of proteins have been identified that rely on sumoylation for proper function. Mammalian RanGAP1 protein, required for nucleocytoplasmic transport, is sumoylated on Lys526; this covalent modification is required for proper association of RanGAP1 with protein components of the nuclear pore complex (NPC) (Bernier-Villamor *et al.,* 2002, Saitoh *et al.,* 1997). These studies suggest a general role for SUMO in the regulation of nuclear transport. $I_\kappa B_\alpha$, an inhibitor of the transcription factor $NF_\kappa B$, is targeted for degradation by ubiquitination of its Lys21 residue, but it has been demonstrated that $I_\kappa B_\alpha$ is also sumoylated at the same residue to inhibit its degradation and the subsequent activation of $NF_\kappa B$ (Desterro *et al.,* 1998). In contrast, SUMO has also been shown to covalently bond to Lys298 of the heat shock transcription factor HSF1 to stimulate DNA-binding activity (Hong *et al.,* 2001), resulting in increased transcription.

A single enzyme, Ulp1, catalyzes two essential functions in the SUMO pathway. Ulp1 processes the C-terminus of the SUMO ortholog in yeast, Smt3, to its mature form, removing three residues to yield $–GG\text{-}COO^-$ from $–GGATY\text{-}COO^-$. This same enzyme also catalyzes the removal of SUMO from the lysine residue to which it is sometimes covalently bound (Li & Hochstrasser, 1999). A second enzyme found in yeast, Ulp2, is not essential but mutants lacking this protein exhibit a range of abnormal phenotypes (Li

& Hochstrasser, 2002). The simpler Ulp1/2 system contrasts with the ubiquitin deconjugating system, which is comprised of at least 17 members in yeast (Mossessova & Lima, 2000).

SUMO System

Expression of recombinant proteins as ubiquitin-fusion proteins in both *E. coli*, and *S. cerevisiae* has been shown to dramatically increase passenger protein expression and solubility levels (Butt *et al.*, 1989, Ecker *et al.*, 1989). As SUMO bears 44.7% sequence similarity to ubiquitin, fusion to SUMO (Smt3) has also been identified as a method for increasing recombinant protein expression and solubility. Lifesensors, Inc. (Malvern, PA), has designed a protein expression and purification platform based on Ubls and their related enzymes (please see www.lifesensors.com) (US Patent Application Numbers 20040018591, 20030153045). A similar SUMO fusion protein expression system has been designed in the lab of Christopher Lima (Weill Cornell Medical College, New York, NY) for expression and purification of insoluble or poorly expressed proteins (US Patent Application Number 20030086918).

A fusion protein system has been developed based on SUMO and its deconjugating enzyme Ulp1, for expression of recombinant proteins in *E. coli*, yeast, or insect cells (Figure 1C and 1D). ORFs are cloned into the pSUMO expression vector using endonuclease/ligase dependent cloning. A 6X-His tag is present at the N-terminus of the SUMO ORF to allow for Ni-affinity purification. Ulp1 cleavage, conducted *in vitro* following purification, is dependent upon recognition of the entire SUMO sequence. As

SUMO retains the 6X-His tag after cleavage by Ulp1, the protein of interest can be obtained through a second round of Ni-affinity chromatography. Expression of proteins as SUMO fusions has been shown to increase expression and solubility levels close to those observed with ubiquitin-fusion proteins.

Goals

We are investigating the use of two different fusion protein expression systems to determine their effects on expression and solubility levels of recombinant proteins in *E. coli*. A set of 50 eukaryotic targets has been selected for cloning into both the Gateway-MBP and SUMO expression vectors. We have expressed subsets of our target list in each of these vectors on an analytical scale for expression and solubility screening. We are experimenting with these fusion protein systems to evaluate their worth as alternative expression vectors for targets that are insoluble or not expressed in our standard T7 Multiplex Expression system. Furthermore, we are evaluating the amenability of these systems to the high-throughput screening methods necessary for our structural genomics effort.

To analyze the MBP-fusion system, a target set of *C. elegans* proteins were expressed as C-terminal MBP fusion proteins in *E. coli*. We have examined the differences in solubility levels between *in vivo* and *in vitro* cleavage of MBP using the TEV protease (Park *et al*., 1994), and compared this data with the expression and solubility levels of these proteins in our T7 Multiplex Expression system. A subset of our eukaryotic target list was cloned into the pSUMO vector. Expression and solubility screening was carried

out for the uncleaved SUMO-fusion proteins. These expression and solubility levels were compared with those of our T7 system as a preliminary analysis of the SUMO system.

Materials & Methods

MBP *in vivo* cleavage analysis

Transformations

MBP-fusion constructs were prepared in the lab of David Waugh (National Cancer Institute, Frederick, MD). 1-2 µl of plasmid DNA for each target was mixed with 50 µl BL21 pRK603 pKC1 *E. coli* competent cells (Kapust & Waugh, 1999) (Figure 3) on ice for 30 minutes, heat shocked at 42°C for 60 seconds and returned to ice for 5 minutes. 100 µl SOC (20 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 0.5g/L NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 20 mM glucose pH 7.0) was added to each transformation and cells were allowed to grow without antibiotic selection for 1 hour at 37°C. The entire transformation volume was plated on LB/Ampicillin/Kanamycin/Chloramphenicol + 0.5% glucose and incubated overnight at 37°C.

Protein Expression and SDS-PAGE Analysis

Single colonies were picked from transformations and grown for 6-8 hours in 0.5 ml LB/Ampicillin/Kanamycin/Chloramphenicol + 0.5% glucose at 37°C. After this time 0.5 ml MJ9 minimal media (2.5 g/l $(NH_4)SO_4$, 9.0 g/l KH_2PO_4, 6.0 g/l K_2HPO_4, 0.5 g/l sodium citrate, with added supplements) (Jansson *et al.*, 1996) precultures with appropriate antibiotics were inoculated with 50 µl of the LB precultures and grown overnight at 37°C. 2 ml MJ9 minimal media cultures were inoculated with 200-250 µl of the overnight MJ9 precultures to achieve an initial $OD_{600}\approx0.15$ units and incubated at 37°C until $OD_{600}\approx0.6$ units. At this time, protein expression was induced with IPTG at a final concentration of 1 mM and cultures were further incubated at 17°C to for growth

overnight. Anhydrous tetracycline was not added to induce expression of TEV protease for the reason that we observed cleavage of the fusion proteins without induction. This suggests that leakage of the Tet promoter was sufficient for producing enough active TEV to completely separate MBP from the passenger protein.

The overnight expressions were separated into 1 ml portions in 96-well S-blocks (QIAGEN) and the cells were harvested by centrifugation at 3000x*g* at 4°C for 15 minutes. The supernatant was removed and one set of pellets was stored at -20°C. The second set was resuspended in 100 µl Native Lysis Buffer per pellet (50 mM NaH_2PO_4, 300 mM NaCl pH 8.0, 10 mM β-mercaptoethanol added at time of use) on ice and transferred to 96-well PCR plate for sonication. Cells were sonicated for 60-80 minutes using the following program: 5 minute pulse, 1 minute off, output level 7, and the solution was resuspended every 20 minutes. 15 µl total cell lysate (T) samples were removed after sonication for SDS-PAGE analysis. Cell lysates were then centrifuged for 10 minutes at 3000x*g* to pellet the insoluble portion of the cell lysate. 15 µl soluble cell lysate (S) samples were then taken for SDS-PAGE analysis.

Reducing sample buffer was prepared by mixing 5:1 NuPage Sample buffer (Invitrogen # NP0007) and 1 M DTT. 6 µl of this mixture was added to each T and S sample and samples were heated at 70°C for 10 minutes. Samples were then either run on precast NuPage gels (Invitrogen # NP0322BOX) or stored at -20°C to be run on a later date. SDS-PAGE gels were rinsed three times with deionized water and stained overnight with

GelCode Blue (Pierce # 24592). Expression and solubility levels were evaluated by presence or absence of the correct size band in both total and soluble cell lysate samples.

MBP *in vitro* **cleavage analysis**

Transformations

Transformations were performed as described above, but with slight modifications. Fusion protein constructs were transformed into BL21 (DE3) pMgk (Figure 3) *E. coli* competent cells (Stratagene). All plates and liquid cultures required the use of only ampicillin and kanamycin as selective agents. Kanamycin selects for cells containing the pMgk plasmid and ampicillin selects for the MBP-fusion plasmid.

Protein Expression and Purification

Protein expression was carried out in the same manner as described above except as noted. Two sets of 5 ml expression cultures were grown instead of 2 ml cultures. These cells were harvested after overnight induction with IPTG. One set of pellets was resuspended in 0.5 ml Native Lysis Buffer and sonicated as described above. T and S SDS-PAGE samples were taken and prepared as described above. The uncleaved fusion proteins were then Ni-NTA (QIAGEN) purified in automated 96-well format using the QIAGEN Biorobot 8000. Purified (P) samples were taken for SDS-PAGE analysis. Samples were prepared and gels were run as described above. For fusion proteins whose concentrations were relatively high when analyzed by SDS-PAGE, buffer exchange was performed using NAP-5™ (Amersham Biosciences) desalting columns to exchange

fusion proteins into TEV cleavage buffer (20 mM NaPi, 150 mM NaCl pH 7.3) (Parks *et al.,* 1994).

TEV Cleavage and SDS-PAGE Analysis

24 units of TEV protease (US Biological) were added to each fusion protein and the cleavage was allowed to proceed overnight at 30°C. 40 µl samples (C) were taken for SDS-PAGE analysis. The cleaved protein was then concentrated at 4°C to 1/10 of its original volume and 15 µl samples (CC) were taken for SDS-PAGE analysis. SDS-PAGE samples were prepared as described above and gels were analyzed for proportion of cleaved fusion protein and solubility of these proteins after *in vitro* cleavage.

SUMO

Target DNA Amplification

Selected NESG targets were amplified by PCR from previously prepared plasmid templates. General primers were designed for the targets set using vector sequence from the original clones (Table 1). PCR products were then separated on 2% agarose gel. Bands of the correct size were cut out and gel purified using the QIAGEN BioRobot 8000. Inserts were digested with BsaI/XhoI (or BsmBI/XhoI in the case of an internal BsaI recognition sequence). Digested inserts were purified in 96-well format using the QIAGEN BioRobot 8000, lyophilized to dryness, and desalted using Centri•sep 96 filter plates (Princeton Separations).

Vector Preparation

pSUMO vector was obtained from LifeSensors, Inc. (Malvern, PA). pSUMO was transformed and amplified in XL-10 Gold *E. coli* competent cells (Stratagene) and purified using QIAfilter Maxiprep Kit (QIAGEN). pSUMO was digested for 3 hours, replenishing enzyme after each hour, with either BsaI or BsmBI and separated on a 1% agarose gel. Correct size bands were cut out, gel purified, and digested overnight with XhoI. pSUMO was then treated with calf intestinal alkaline phosphatase (CIP) (New England BioLabs) for 45 minutes at 37°C to remove phosphate groups from DNA overhangs and prevent vector from reannealing without an insert. pSUMO was then purified and concentrated using QIAGEN Minelute spin columns.

Ligations, Transformations, and Construct Validation

Ligations were performed overnight using T4 DNA Ligase (New England BioLabs at 16°C. pSUMO was added at concentrations between 40-100 ng/µl. After 16-hour ligations, ligase was heat inactivated at 65°C for 10 minutes. Ligation digestions were performed at 37°C for 1 hour using either BamHI or HindIII to digest vector that did not contain an insert.

1-5 µl ligation digestions were transformed into X-10 Gold *E. coli* competent cells following the same procedure as described above. The entire transformation volume was plated on LB/Kanamycin and incubated overnight at 37°C. Colonies that formed were resuspended in 50 µl deionized water. Colony PCR was performed using 10 µl colony

resuspension. 15 µl of each PCR reaction was run on 2% agarose gel to check for bands of the correct size.

Minipreps

For positive clones (one per target), 10 µl colony resuspension was added to 3 ml Superbroth/Kanamycin and grown overnight at 37°C. After overnight growth, glycerol stocks were prepared and chloramphenicol was added to amplify plasmid DNA. Cultures were grown another 4-6 hours and 1.5 ml were centrifuged for 1 minute to pellet cells. The pellet was isolated and the plasmid DNA was purified using the QIAGEN Miniprep Spin Column protocol. Plasmid DNA was eluted in 50 µl Buffer EB and stored at -20°C.

Expression Transformations

1.2 µl plasmid miniprep DNA was mixed with 15 µl BL21 Codonplus (DE3)-RIL *E. coli* competent cells (SUMO System Expression SOP) (Stratagene). These competent cells encode rare codon tRNAs for the amino acids arginine, isoleucine, and leucine as well as the gene for chloramphenicol resistance. As pSUMO contains the kanamycin resistance gene, an *E. coli* strain such as BL21 (DE3) pMgk could not be used for expression screening. Transformations were carried out as described above. The entire transformation volume was plated on LB/Kanamycin/Chloramphenicol + 0.5% glucose and incubate overnight at 37°C.

Protein Expression and SDS-PAGE Analysis

Fusion protein expression was carried out in the same manner as described above. SDS-PAGE analysis was performed as described above to determine the expression and solubility levels of uncleaved SUMO-fusion proteins.

Archiving

MBP-fusion protein glycerol stocks

Glycerol stocks of MBP-fusions in *E. coli* cloning strains were previously prepared. MBP-fusion glycerol stocks (in BL21 pRK603 pKC1 and BL21 (DE3) pMgk) were stored in S-block format. To a specified volume of LB preculture (300-400 µl), 80% glycerol was added to a final concentration of ~20%. Glycerol stocks were mixed well and frozen on dry ice for 30 minutes, then stored at -80°C.

SUMO-fusion protein glycerol stocks

Glycerol stocks were prepared in the same manner as described above except as noted. SUMO-fusion glycerol stocks were prepared in individual tube format. Separate stocks exist for both *E. coli* XL-10 Gold and BL21 (DE3) Codonplus-RIL strains.

Construct information and analytical-scale expression and purification data can be found in the SPiNE database, http://mining.nesg.org (Goh *et al.*, 2003; Bertone *et al.*, 2001).

Section 1

Gateway-MBP Fusion Expression & Solubility Screening

Experimental Results

24 *C. elegans* targets (Table 2) were screened for protein expression and solubility levels as MBP fusion proteins. Expression and solubility levels of fusion proteins cleaved both *in vivo* and *in vitro* were compared with levels of proteins expressed in pET14 and pET15 vectors. Using our standard pET vector system, 8 targets were not expressed, 7 were expressed but insoluble, and 8 were expressed and soluble in at least one pET vector system (Table 3). Target WR5 was unavailable for expression analysis. One target, WR18 was very poorly expressed and solubility was difficult to judge.

MBP-fusions with *in vivo* cleavage

When expressed as MBP-fusion proteins and cleaved with TEV protease *in vivo*, 11 targets were not expressed, 4 targets were expressed but insoluble, and 8 targets were expressed and soluble (Table 3). Again, WR18 was poorly expressed and solubility could not be evaluated. Although the percentage of unexpressed proteins rose by 13%, only one soluble protein, WR8, was lost using the Gateway-MBP system. Three additional proteins that were expressed but insoluble using pET vectors were unexpressed as fusion proteins. One protein, WR3, was highly expressed but completely insoluble when expressed in both pET14 and pET15, but as an MBP-fusion, WR3 became highly soluble (Figure 4).

MBP-fusions without cleavage

Our target set was next analyzed for expression and solubility of MBP-fusion proteins uncleaved by TEV protease. Expression and solubility levels were compared to the observations reported above. 6 targets were observed to be unexpressed, while 17 targets were expressed and soluble before cleavage (Table 3). One target, WR26, was found to be expressed, but only slightly soluble. Of the expressed and soluble targets, 8 were expressed and soluble using pET vectors, 6 proteins, WR2, WR3, WR13, WR27, WR49, and WR54 were expressed but insoluble using pET vectors, and 3 proteins, WR6, and WR10, WR18 were unexpressed of very poorly expressed using pET vectors. In total, 8 targets were solubilized when expressed as uncleaved MBP-fusion proteins.

Purification and *in vitro* cleavage

The uncleaved fusion proteins were then Ni-NTA purified in automated 96-well format using the QIAGEN BioRobot 8000. Of the 17 soluble uncleaved MBP-fusion proteins, 15 (88%) were purified (Table 3). Two targets, WR13 and WR54, although expressed and soluble, did not appear to bind to the Ni-NTA resin and were not present upon SDS-PAGE analysis of the purified fusion proteins. This may be due to inability of the 6X-His tag to interact with the Ni-NTA resin, most likely because of steric hindrance. Of the 15 purified protein targets, 8 (53%) were subjected to *in vitro* cleavage by TEV protease. The remaining 7 purified proteins were not significantly concentrated to yield appreciable results (labeled **low** in Table 3). All 8 fusion proteins cleaved *in vitro* were observable by SDS-PAGE and upon 10X concentration did not appear to fall out of solution. Of the 8

cleaved targets, WR3 (Figure 4) and WR6 were solubilized by expression as MBP-fusion proteins with subsequent *in vitro* cleavage by TEV protease.

Discussion

Our analysis of the Gateway-MBP fusion protein expression system has proved that over all, the system may be beneficial to our structural genomics effort. As previously discussed, MBP has been shown to have a molecular chaperone-like solubilizing effect on many proteins to which it is fused. Our analysis has yielded two clear examples of MBP increasing the solubility of its fusion partners, WR3 and WR6. In addition, MBP has not proven to have a significant negative effect on proteins that are soluble when expressed in out pET system. Only one protein soluble in our pET system, WR8 was found to be insoluble when cleaved *in vivo*. WR8 appeared to remain soluble when cleaved from MBP *in vitro*, suggesting that our *in vivo* cleavage result may be a false negative. It would be beneficial to run expression screenings a second time for this target to verify our result.

An analysis of the overall effects of MBP-fusion protein expression is presented in Figure 5. By examining this figure, we can determine how MBP-fusions affect the different expression categories observed with pET vector systems. Of the 34% (8) of unexpressed proteins in our pET systems, 87% (7) were still unexpressed as MBP fusions cleaved *in vivo* and 13% (1) were expressed but not soluble as fusions. Of the 29% (7) of insoluble proteins expressed in the pET vectors, 43% (3) were not expressed as MBP fusions cleaved *in vivo*, 43% (3) were expressed but not soluble, and 14% (1) were expressed and

soluble. Of the 34% (8) of proteins that were expressed and soluble in pET vectors, 87% (7) were expressed and soluble as MBP fusions cleaved *in vivo,* while 13% (1) were not expressed.

Our observations changed somewhat when the MBP-fusions were expressed in BL21 (DE3) pMgk, which does not allow *in vivo* cleavage of MBP from the target protein. Under these circumstances, we found that 62% (5) of proteins not expressed in pET vectors remained unexpressed, while 38% (3) of proteins were expressed and soluble. Increases in solubility were also seen for proteins that were expressed but not soluble using pET vectors, such that 75% (6) of proteins were expressed and soluble, while 12.5% (1) were unexpressed and 12.5% (1) were expressed but not soluble. Finally, we observed that 100% (8) of expressed and soluble proteins were also expressed and soluble as uncleaved MBP-fusions. As stated above, 88% (15) of the 17 soluble uncleaved MBP-fusions were Ni-NTA purified. However, due to low yields of purified protein, *in vitro* TEV cleavage data was only collected for 8 targets. 100% of these targets appeared to be soluble after cleavage from MBP and subsequent centrifugal concentration.

It may be true that proteins soluble as MBP-fusion have similar overall qualities. As a starting point, we examined the predicted pI of each target protein to find possible correlations with the ability of MBP to increase the solubility of our targets. This type of analysis may provide clues to the surface interactions that may be occurring between MBP and its passenger protein, thereby allowing an analysis of other possible candidates

for expression as MBP fusions. This search, however, did not produce any significant correlations. Proteins found to be expressed and soluble after either *in vivo* or *in vitro* cleavage with TEV protease had pI values ranging from 4.39 (WR11) to 9.79 (WR39). A similar range was observed for unexpressed and insoluble proteins, suggesting that pI does not affect the ability of MBP to aid in solubilizing its partner. As more targets are analyzed for expression and solubility levels, similar analyses will be carried out for protein characteristics such as hydrophobicity and hydrophilicity.

Around ? of targets were not expressed from pET vectors. It is possible that some of these ORFs contain the class II T7 terminator sequence (Lyakhov *et al.*, 1998). This 7 bp sequence, 5'-ATCTGTT-3' is found on the non-template strand. When this sequence is present during transcription, T7 RNA polymerase (RNAP) pauses, and transcription termination occurs 6-8 bp further downstream. The natural occurrence of this sequence in the T7 bacteriophage genome seems to be involved in the initiation of DNA packaging. However, its presence in ORFs of recombinant proteins will inhibit their complete transcription when T7 RNAP and the T7 promoter are employed. Our targets have not been screened for the T7 terminator sequence, but our data does not seem to support its presence in our target set. The T7 terminator sequence is recognized by T7 RNAP, while the MBP-fusion protein plasmid utilizes the *tac* promoter and the ORF is transcribed by host RNAP. If any of our targets were not expressed from pET vectors due to the presence of a T7 terminator sequence, this should be overcome when transcribed by bacterial RNAP recognizing the *tac* promoter.

By inspection of Table 3, it can be seen that a larger proportion of targets were expressed and soluble as uncleaved MBP-fusions than when cleaved *in vivo*. One difference between the two *E. coli* strains used for expression is the accessory plasmid encoding rare codon tRNAs (Figure 3). *E. coli* BL21 pRK603 pKC1, which encodes two rare codons for Arg and Ile, was used for *in vivo* cleavage analysis. BL21 (DE3) pMgk, used for *in vitro* cleavage, encodes three rare codons, two for Arg and one for Ile. This additional rare codon tRNA found in BL21 (DE3) pMgk cells may result in increased protein expression relative to BL21 pRK603 pKC1. It would be interesting to test *in vivo* cleavage of MBP-fusions produced with the help of the pMgk plasmid. This hypothesis cannot be tested at this time however, because both pMgk and pRK603 are kanamycin resistant and this would not provide a method of selection for both plasmids.

As 47% of purified MBP-fusions were too dilute for cleavage analysis, it would be beneficial to attempt purification of these targets again. Many of the dilute purified fusions were insoluble in pET systems and it would be worthwhile to discover if a larger proportion of targets can be rescued through *in vitro* cleavage of MBP. Larger cultures (~50 ml) could be grown to provide more fusion protein for purification and cleavage analysis.

In addition to low yield of certain Ni-NTA purified fusions, we have observed incomplete cleavage of MBP from its fusion partner (see Figure 4 for example). This is an undesirable result in that after cleavage, over one half of the soluble protein is lost. In comparison, it appears that close to 100% of fusion protein is cleaved when TEV

protease is expressed *in vivo* (see Figure 4 for example). The *in vivo* and *in vitro* cleavage of MBP utilizes TEV from two different sources. TEV is produced by the pRK603 plasmid for *in vivo* cleavage, while *in vitro* cleavage uses a commercially purchased protease. Differences in cleavage will be evaluated by purification of milligram quantities of TEV produced from pRK603 for use in *in vitro* cleavage of our MBP-fusions. It may be true that TEV is much more active *in vivo* and the source of the enzyme is not an issue. However, it is also possible that TEV protease produced from pRK603 exhibits more catalytic activity than the commercially purchased enzyme.

After the expression and solubility analyses described here, another issue becomes quite important. This is the difference between "solubility" and "concentratability". Thus far we have defined solubility as the presence of the recombinant protein in the cell extract after low-speed centrifugation at 3,000x*g*. The concept of concentratability arises after purification, during preparation of protein samples for crystallization and NMR experiments, where protein concentrations of 0.5-1.0 mM are required. Concentratability refers to the protein remaining soluble after purification and subsequent concentration to experimentally significant levels. Some proportion of proteins appear to be soluble in the cell extract, but upon purification and concentration they form insoluble aggregates. This problem can be encountered with any type of expression system, as the aggregation is not an issue until final sample preparation. There is no simple method to screen for this phenomenon in an HTP setting, as this requires large amounts of protein, subsequent purification, buffer exchange and concentration to at least 0.5 mM. This is an issue of both screening abilities and documentation, because a simple screen is not available to

determine “limits of concentratability”, or the highest concentration at which a protein remains soluble, and a documentation system is not established to differentiate between solubility and concentratability for each protein target. This distinction cannot currently be made and concentratability will continue to be a problem for some proteins determined to be initially soluble in the cell extract.

After further analysis and optimization of *in vitro* cleavage condition, our next step will be to examine the “foldedness” of proteins produced as MBP-fusions, as well as their amenability to structure determination by NMR. This can be determined by heteronuclear single quantum coherence (HSQC) NMR screens. It is sometimes the case that although a protein appears to be soluble, HSQC screening reveals that it is not folded and it does not produce clean spectra. In addition to the solubilizing effects of MBP-fusion proteins, we are interested in whether MBP has a positive influence on folding (supporting the chaperone model) and can improve the HSQC spectra of problematic proteins. Our first step in this analysis would be to produce ^{15}N-MBP-WR11 and ^{15}N-MBP-WR14 (which were identified as having poor HSQC spectra when expressed from pET vectors) on a large scale. HSQC experiments would be run on the purified proteins to examine whether expression as an MBP-fusion produces a clean HSQC, indicating that MBP had some effect on the folding of the passenger proteins. This type of analysis may reveal differences in folding of proteins produced from pET and the two types of MBP expression systems, *in vivo* and *in vitro* cleavage by TEV. Depending on the results of these experiments, further analyses may be carried out on a larger number of expressed and soluble MBP-fusions.

The Gateway-MBP system has already proved to be beneficial in certain instances, and does not appear to be disadvantageous to proteins known to be soluble in pET expression systems. Thus, Gateway-MBP may serve as an important tool in the expression of certain, "high value" proteins, such as those implicated in disease states. As a larger set of protein targets are analyzed, more proteins may be found that benefit from the chaperone-like effects of MBP. Proteins rescued from insolubility by fusion with MBP may then be analyzed for primary or secondary structure similarities to uncover any conserved features among rescued proteins. A bioinformatic analysis of this type could uncover many additional protein targets amenable to expression as MBP-fusion proteins.

Section 2

SUMO System Cloning, Expression & Solubility Screening

Experimental Results

A core set of 50 eukaryotic protein targets were designated for cloning into the pSUMO fusion protein expression vector. Of this total, 8 (16%) targets have been analyzed for expression and solubility as uncleaved SUMO-fusion proteins to date. One additional protein has been cloned and is ready for expression analysis. 27 (54%) additional targets are currently in the process of cloning into pSUMO (Table 4).

Of the 8 targets analyzed, 3 (37.5%) were not expressed using our standard pET vectors. These were targets AR5, AR22, and HR1697. The remaining 5 targets (62.5%) were expressed but insoluble in our standard expression system (Table 5).

When expressed as SUMO-fusion proteins, the expression and solubility patterns for these proteins were markedly different. 4 of 8 targets (50%) were highly expressed and soluble. Two of these targets, AR22 and HR1697, were previously unexpressed, and the remaining two targets, FR10 and HR1553, were previously expressed but insoluble. Of the remaining proteins, 3 targets (37.5%), HR894, HR1686, and WR26 were expressed but insoluble, and 1 target, AR5, was unexpressed (Figure 6, Table 5). A band is present in both the T and S fractions for AR5, but this appears to be only SUMO according to the molecular weight marker. A possible explanation is that AR5 was unfolded after translation and degraded by host proteases, while the properly folded SUMO was not. As we cannot be sure of this, AR5 is said to be not expressed.

Discussion

The data discussed above may be regarded as preliminary, as the target set for expression and solubility is quite small. However, this set does contain proteins spanning four eukaryotic organisms: AR *Arabidopsis thaliana,* FR *Drosophila melanogaster*, HR *Homo sapiens*, and WR *Caenorhabditis elegans*. Recombinant protein solubility increased from 0% using our pET expression system to 50% using the SUMO system. In addition, the SUMO system does not appear to have any detrimental effects on protein expression and solubility thus far. No protein targets were lost by expression as SUMO-fusion proteins.

As the cloning of many more proteins into pSUMO is currently underway, this data set may soon be expanded to provide a more complete picture of the effects of SUMO-fusion

proteins on expression and solubility levels. If further data follows the same trend as that seen here, SUMO-fusion proteins may provide a reliable method for expression of many eukaryotic proteins in soluble form.

The next logical step will be to examine solubility levels of purified SUMO-fusion protein after cleavage with SUMO deconjugating enzyme Ulp1. This data will provide a clearer picture of the benefits of the SUMO System. We will then be able to determine whether SUMO possesses some molecular chaperone-like qualities in the context of a fusion protein, or whether its high solubility is simply keeping unfolded passenger proteins in solution. If SUMO does in fact possess chaperone-like qualities, it would be interesting to determine whether any hydrophobic surface residues play a role in the solubilization of fusion partners, similar to the studies published by Fox *et al.* (2001).

Alongside the analysis of *in vitro* Ulp1 cleavage, it could be worthwhile to examine *in vivo* cleavage with this enzyme. As this is a eukaryotic protein, an Ulp1 plasmid could be constructed for expression in *E. coli*. Of course, this type of cleavage analysis would only be possible if Ulp1 is soluble and catalytically active when expressed in *E. coli*. Biochemical studies carried out by Li and Hochstrasser (1999) have shown that Ulp1 can be expressed as a glutathione-S-transferase (GST)-fusion protein in *E. coli* and retain its catalytic activities. A comparison of *in vivo* versus *in vitro* Ulp1 cleavage of SUMO-fusion proteins could then be carried out similar to that discussed for MBP-fusion proteins. This may reveal important differences between the two cleavage methods.

Although the current SUMO-fusion data set is small, other issues discussed above in relation to MBP-fusion proteins are also relevant to the SUMO system. In particular, the T7 terminator sequence may present a problem as the SUMO expression vector is constructed using a T7 promoter. We may also encounter low purification yields such as those observed with the purification of MBP-fusion proteins. Low yields could result in the inability to cleavage fusion proteins. If this presents itself as a major barrier, it may be necessary to design an Ulp1 plasmid for *in vivo* cleavage as discussed above. Again, any protein rescued from insolubility by this system, may have a low "limit of concentratability", rendering the protein unsuitable for NMR studies. Some or all of these issues will no doubt be raised as we investigate the SUMO-fusion system more thoroughly.

Figure & Table Legends

Figure 1: (A) A schematic of the Gateway-MBP fusion protein ORF and recombinational attR cloning sites of the destination vector. Endonuclease/ligase dependent cloning is not employed to create MBP-fusion protein constructs. **(B)** A schematic of the expressed MBP-fusion protein, indicating the locations of the TEV protease cleavage site and the 6X-His tag. **(C)** A schematic of the SUMO fusion protein ORF and multiple cloning site (MCS). Expression the fusion protein is dependent upon the T7 promoter. **(D)** A schematic of the expressed SUMO fusion protein, indicating the locations of the Ulp1 cleavage site and the 6X-His tag. In contrast to MBP-fusion proteins, after cleavage by Ulp1, the 6X-His tag will be lost.

Figure 2: Illustration of the model proposed by Kapust and Waugh (1999) in the explanation of how MBP may act as a solubility enhancer in the context of a fusion protein. Spheres represent MBP, attached strings represent unfolded passenger protein, and helices represent properly folded passenger protein. (Figure courtesy of Kapust & Waugh, 1999)

Figure 3: Diagrams of the cell lines used for expression of MBP-fusion proteins. **(A)** *E. coli* BL21 pRK603 pKC1 provided by David Waugh. This cell line is used for *in vivo* cleavage using TEV protease. pRK603 encodes the MBP-TEV fusion protein (Kapust & Waugh 2000). pKC1 encodes two rare codon tRNAs for Arg and Ile for enhanced expression of recombinant proteins (Evdokimov *et al.,* 2002). **(B)** *E. coli* BL21 (DE3) pMgk is used for *in vitro* cleavage using TEV protease. This cell line does not contain

the TEV protease expression plasmid. pMgk encodes three rare codon tRNAs: two Arg tRNAs and one Ile tRNA, for enhanced expression of recombinant proteins. This cell line also takes advantage of the DE3 lysogen for increased control of protein expression.

Figure 4: SDS-PAGE analysis of expression and solubility levels of WR3. **(A)** WR3 is completely insoluble when expressed in either pET14 or pET15 (Total (**T**) and soluble (**S**)). Black arrows indicate the protein band, or lack thereof. **(B)** When expressed as an MBP-fusion and cleaved *in vivo*, WR3 is highly expressed and highly soluble. **(C)** When expressed as an MBP fusion protein and cleaved *in vitro* WR3 remains highly soluble. A band of uncleaved MBP-WR3 can also be seen, indicating that the cleavage is incomplete (Purified and Cleaved (**P**/**C**)).

Figure 5: Pie chart analysis of MBP-fusion data. **(A)** Pie chart gives the percentage of targets that are not expressed, expressed/not soluble, expressed/soluble, or other*, for either expression vector, pET14 or pET15. **(B)** Each chart analyzes the fate of each expression category when expressed as and MBP-fusion and cleaved *in vivo*. **(C)** Each chart analyzes the fate of each expression category when expressed as an MBP-fusion with out cleavage. The arrows point to the fraction of expressed/soluble proteins that were soluble after *in vitro* cleavage with TEV. The fraction of proteins whose concentrations were too low for cleavage analysis is also given.
*The category "other" refers to WR5, which was not available in either pET14 or pET15 for expression analysis.

Figure 6: SDS-PAGE analysis of expression and solubility levels of SUMO-fusion proteins. Red boxes indicate band of interest. **Gel 1:** Lanes 2 & 3 are AR5 Total (T) and Soluble (S), respectively. The boxed band in these lanes appears to be SUMO without a passenger protein, by the molecular weight. Lanes 4 & 5 contain HR894 T and S. Although HR894 appears to be expressed, it does not appear to be appreciably soluble. Lanes 6 & 7 contain WR26 T and S. WR26 is expressed at a lower level than other fusions, but it is not soluble. Lanes 8 & 9 are AR22 T and S. This protein is highly expressed and soluble as a SUMO-fusion. Lanes 10 & 11 are HR1553 T and S. HR1553 is also highly expressed and soluble as a SUMO-fusion. **Gel 2:** Lanes 2 & 3 contain FR10 T and S. FR10 is highly expressed and soluble. Lanes 4 & 5 contain HR1686 T and S. Although highly expressed, it is not soluble as a SUMO fusion. Finally, lanes 6 & 7 contain HR1697 T and S. HR1697 is highly expressed and soluble as a SUMO-fusion protein.

Table 1: This table gives the primer sequences (5'-3') used to PCR amplify our target ORFs for cloning into the pSUMO expression vector. Primers were designed to produce ORFs with and without N-terminal 6X-His tags. Thus far all amplified targets have used primers that include the 6X-His tag.

Table 2: This table lists each target cloned into the Gateway-MBP expression vector. Molecular weights are listed of the target protein and the target protein + MBP (40 kDa). These molecular weights were used for SDS-PAGE analysis. Molecular weights of the target protein were used when analyzing *in vivo* cleavage by TEV. Molecular weights of

the fusion protein were used when analyzing uncleaved MBP fusions, before and after Ni-NTA purification.

Table 3: This table compares expression and solubility data among targets expressed in pET14, pET15, MBP (*in vivo* cleavage), and MBP (*in vitro* cleavage). The legend explains color-coding. Boxes in white were single instances and did not fall into one of the larger categories. This table summarizes data such as the example shown in Figure 4.

Table 4: This table summarizes target features for those targets to be cloned into the pSUMO expression vector. Size is given of the ORF in base pairs (bp) and molecular weight is given for the protein product, both cleaved and uncleaved from SUMO (11 kDa). External restriction endonuclease (Ext. RE) sites are those utilized in cloning targets. Internal restriction endonuclease (Int. RE) sites are those found within the ORF of the target. These restriction endonucleases may not be used for ligation digestions, which are performed to cleave ligated vector that does not contain an insert. If enzymes are used that recognize sites within the target sequence, positive clones will be digested. The last column indicates, by color, the progress of each target toward complete expression and solubility analysis.

Table 5: This table summarizes the SUMO-fusion protein expression and solubility data shown in Figure 5. Color-coding is the same as that shown in Table 3. By examining this table, it can easily be seen that 0% of targets are soluble when expressed in pET, but 50% of targets are expressed and soluble as SUMO-fusion proteins.

Figure 1

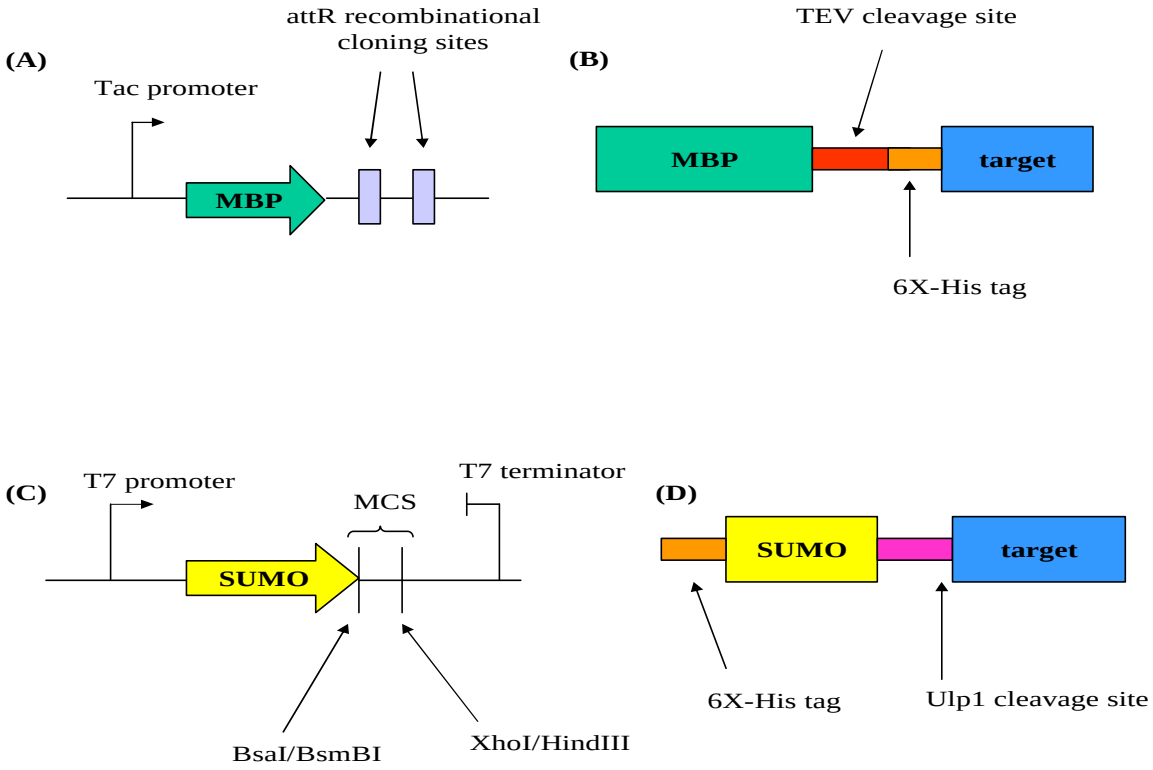
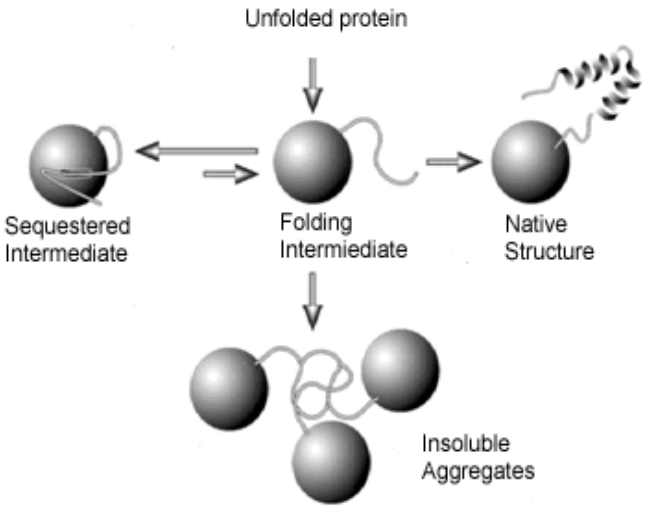

(A)
attR recombinational cloning sites
Tac promoter
MBP
(B)
TEV cleavage site
MBP
target
6X-His tag
(C)
T7 promoter
MCS
T7 terminator
SUMO
BsaI/BsmBI
XhoI/HindIII
(D)
SUMO
target
6X-His tag
Ulp1 cleavage site

Figure 2



(Kapust & Waugh 1999)

Figure 3

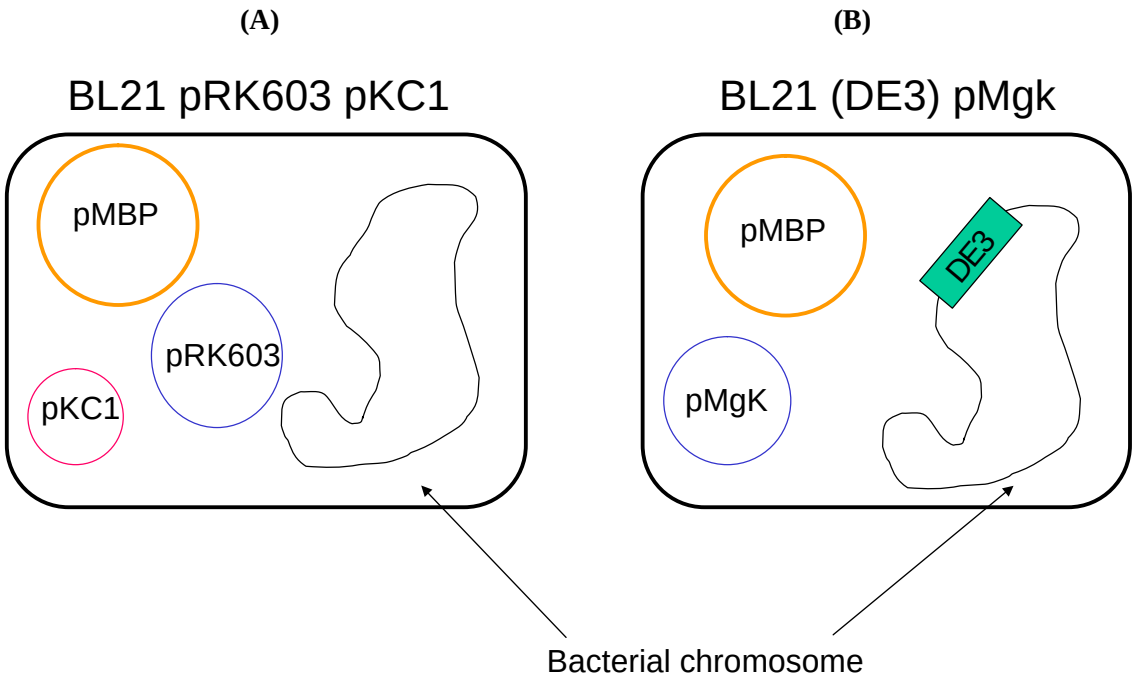

(A)
(B)
BL21 pRK603 pKC1
BL21 (DE3) pMgk
pMBP
pRK603
pKC1
pMBP
DE3
pMgK
Bacterial chromosome

Table 1

Primer Type	5'-3' Sequence
Forward Primer including 6X-His tag	GGTCTCAAGGTATGGGCCATCACCATCACCATC
Forward Primer NO 6X-His tag	CACCATCACCATCACAGCCATGGTCTCAAGGTATG
Reverse Primer	CTTTGTTAGCAGCCGGATCTCGAG
Forward Primer including 6X-His tag internal BSAI site	CGTCTCAAGGTATGGGCCATCACCATCACCATC
Forward Primer NO 6X-His tag internal BSAI site	CACCATCACCATCACAGCCATCGTCTCAAGGTATG
Reverse Primer for internal XhoI site	CTTTGTTAGCAGCCGGATAAGCTT

Table 2

Target	ORF (kDa)	MBP+ORF (kDa)
WR2	10.8	50.8
WR3	11.2	51.2
WR4	11.6	51.6
WR5	10.4	50.4
WR6	8.2	48.2
WR8	18.9	58.9
WR9	20.6	60.6
WR10	14.9	54.9
WR11	9.9	49.9
WR13	12.2	52.2
WR14	11.9	51.9
WR15	17.4	57.4
WR16	15.6	55.6
WR18	18.5	58.5
WR26	24.8	64.8
WR27	16.4	56.4
WR28	18.4	58.4
WR39	23.2	63.2
WR41	9.8	49.8
WR43	21.5	61.5
WR44	14.8	54.8
WR49	23.3	63.3
WR53	17.4	57.4
WR54	9.6	49.6

Figure 4

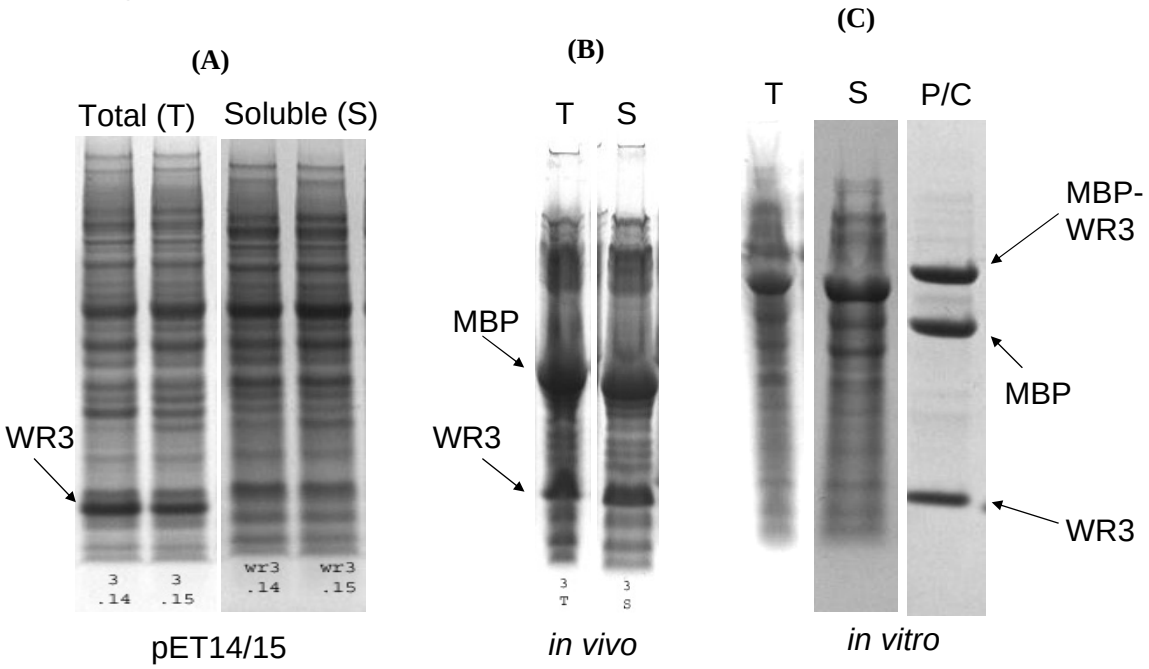

(A)
Total (T)
Soluble (S)
WR3
3 .14
3 .15
wr3 .14
wr3 .15
pET14/15
(B)
T
S
MBP
WR3
3 T
3 S
in vivo
(C)
T
S
P/C
MBP-WR3
MBP
WR3
in vitro

Table 3

Target	pET14	pET15	MBP-fusion in vivo cleavage	MBP-fusion uncleaved	MBP-fusion purified	MBP-fusion in vitro cleavage
WR2	E/NS	NE	NE	E/S	low	
WR3	E/NS	E/NS	E/S	E/S	Y	Y
WR4	E/S	NE	E/S	E/S	Y	Y
WR5			NE	NE		
WR6	NE	NE	E/NS	E/S	Y	Y
WR8	E/S	E/S	NE	E/S	Y	Y
WR9	NE	NE	NE	NE		
WR10		NE	NE	E/S	low	
WR11*	E/S	E/S	E/S	E/S	Y	Y
WR13	E/NS	E/NS	E/NS	E/S	N	
WR14*	E/S	E/S	E/S	E/S	Y	Y
WR15	NE		NE	NE		
WR16	NE	NE	NE	NE		
WR18	PE		PE	E/S	low	
WR26	E/NS	NE	NE	E/LS		
WR27	E/NS	NE	E/NS	E/S	low	
WR28	NE	NE	NE	NE		
WR39	E/S	E/S	E/S	E/S	low	
WR41	E/S	E/S	E/S	E/S	Y	Y
WR43	E/S	E/S	E/S	E/S	low	
WR44	NE	NE	NE	NE		
WR49	E/NS	E/NS	E/NS	E/S	low	
WR53	E/S		E/S	E/S	Y	Y
WR54	E/NS	PE	NE	E/S	N	

not expressed(NE)
expressed/not soluble (E/NS)
expressed/soluble (E/S)
purified
in vitro cleaved
not available

PE poor expression levels
E/LS expressed with low solubility
* soluble with poor HSQC in pET vectors

Figure 5 (A)

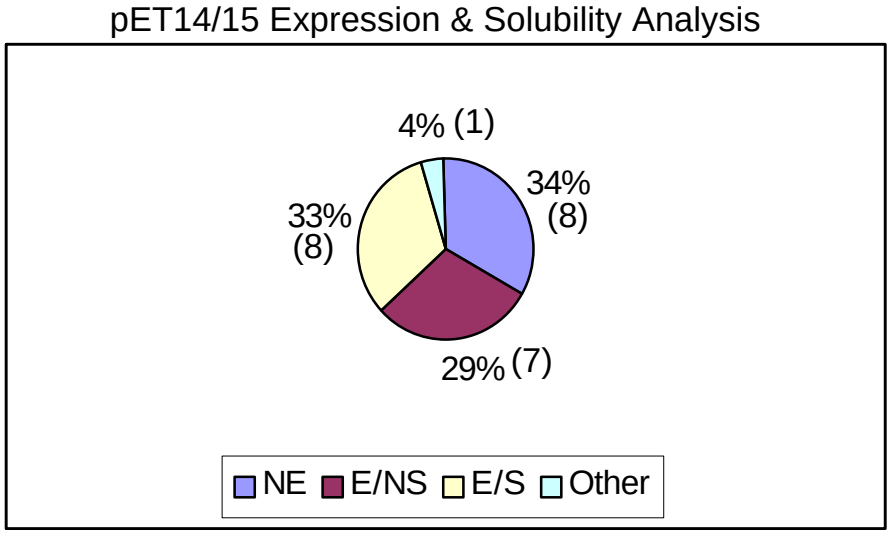

pET14/15 Expression & Solubility Analysis
4% (1)
34% (8)
33% (8)
29% (7)
NE
E/NS
E/S
Other

Figure 5 (B)

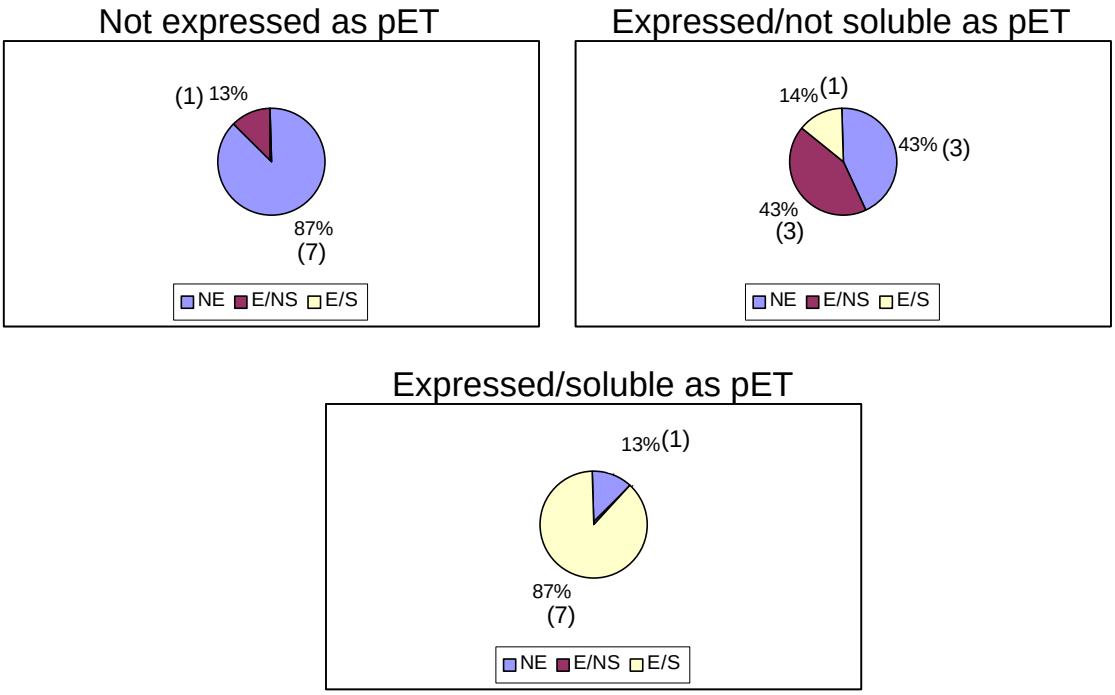

Not expressed as pET
(1) 13%
87%
(7)
NE E/NS E/S
Expressed/not soluble as pET
14% (1)
43% (3)
43%
(3)
NE E/NS E/S
Expressed/soluble as pET
13% (1)
87%
(7)
NE E/NS E/S

Figure 5 (C)

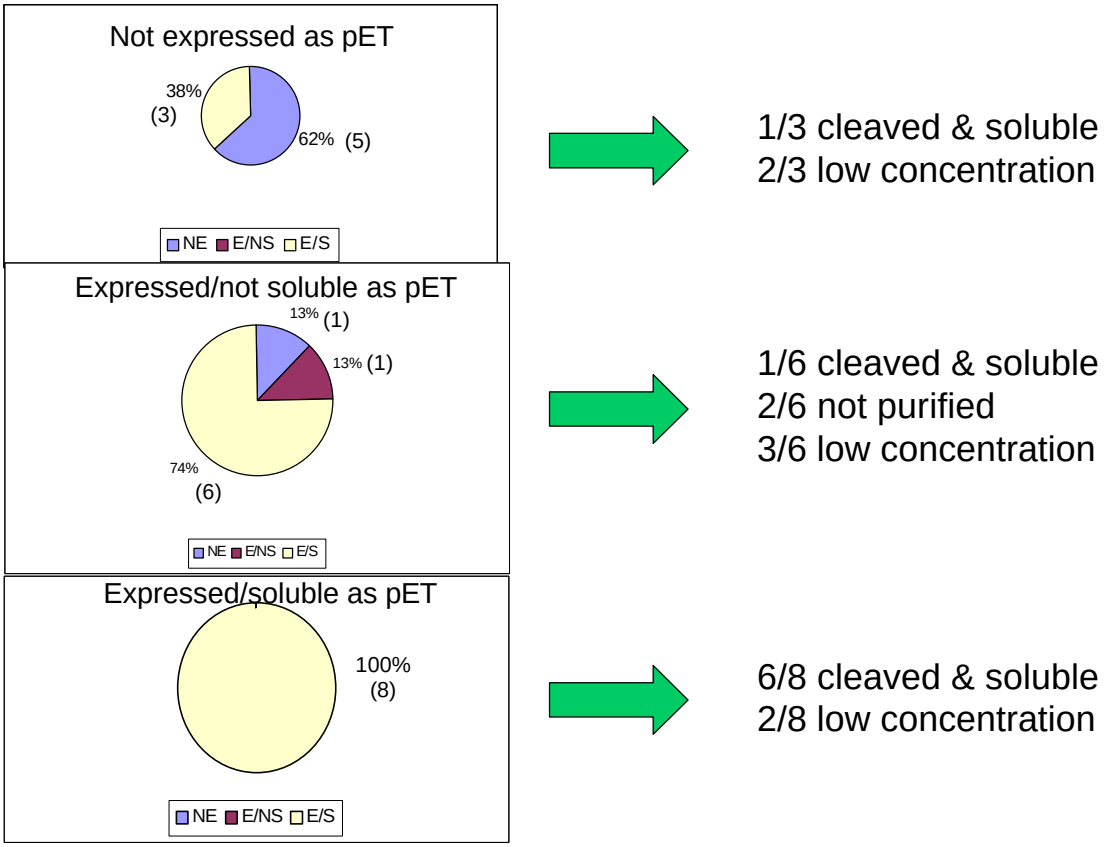

Not expressed as pET
38%
(3)
62% (5)
NE E/NS E/S
Expressed/not soluble as pET
13% (1)
13% (1)
74%
(6)
NE E/NS E/S
Expressed/soluble as pET
100%
(8)
NE E/NS E/S
1/3 cleaved & soluble
2/3 low concentration
1/6 cleaved & soluble
2/6 not purified
3/6 low concentration
6/8 cleaved & soluble
2/8 low concentration

Table 4

Target	Size (bp)	MW (kDa)	Fusion MW	Ext. RE Sites	Int. RE Sites	Progress
AR5	621	23.6	34.6	BsaI/XhoI	SacI, NcoI	
AR15	453	17.4	28.4	BsaI/HindIII	SacI, XhoI	
AR16	663	25.1	36.1	BsaI/XhoI	SalI	
AR21	558	21.3	32.3	BsaI/XhoI	EcoRI	
AR22	423	16.3	27.3	BsaI/XhoI	none	
FR4	381	14.1	25.1	BsaI/XhoI	none	
FR10	462	17.7	28.7	BsaI/XhoI	none	
FR59	534	21	32	BsaI/XhoI	none	
FR79	231	8.9	19.9	BsaI/XhoI	none	
HR91	219	8.2	19.2	BsaI/XhoI	none	
HR547	714	27	38	BsmBI/XhoI	BamHI, BsaI	
HR812	327	12.8	23.8	BsaI/XhoI	none	
HR894	810	30.5	41.5	BsaI/XhoI	KpnI	
HR919	627	23.8	34.8	BsmBI/XhoI	BsaI	
HR945	459	17.6	28.6	BsaI/XhoI	none	
HR969	420	16.2	27.2	BsaI/XhoI	NcoI, HindIII	
HR1553	516	19.7	30.7	BsaI/XhoI	NcoI	
HR1576	198	8.1	19.1	BsaI/XhoI	EcoRI	
HR1686	399	15.4	26.4	BsaI/XhoI	NcoI, HindIII	
HR1697	168	7	18	BsaI/XhoI	none	
HR1719	576	21.9	32.9	BsaI/XhoI	BamHI	
HR1722	477	18.3	29.3	BsmBI/XhoI	BsaI	
HR1738	528	20.2	31.2	BsaI/XhoI	HindIII, EcoRI	
HR1854	570	21.2	32.2	BsaI/XhoI	none	
HR1869	393	14.8	25.8	BsaI/XhoI	none	

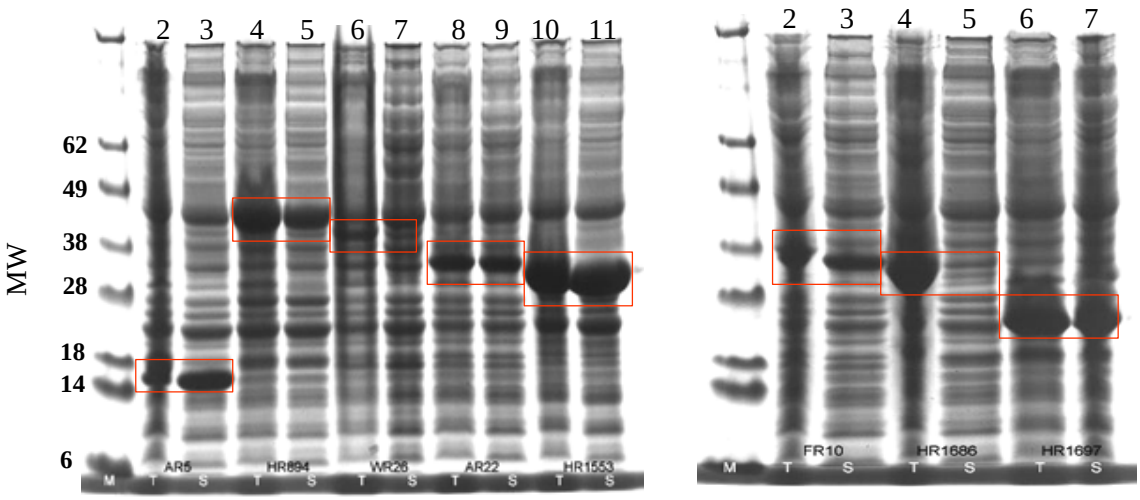
	cloning in progress
	cloned/amplified
	expression analyzed

Table 4 (continued)

Target	Size (bp)	MW (kDa)	Fusion MW	Ext. RE Sites	Int. RE Sites	Progress
HR1889	420	16.4	27.4	BsaI/XhoI	none	
HR1913	447	16.8	27.8	BsaI/XhoI	NcoI	
HR1953	630	23.2	34.2	BsmBI/HindIII	SacI, XhoI, BsaI	
WR12	195	8	19	BsaI/XhoI	none	
WR13	312	12.2	23.2	BsaI/XhoI	none	
WR19	531	20.3	31.3	BsaI/XhoI	EcoRI, HindIII	
WR20	534	20.4	31.4	BsaI/HindIII	NcoI, XhoI	
WR21	240	9.6	20.6	BsaI/HindIII	SacI, XhoI	
WR23	267	10.6	21.6	BsaI/XhoI	BamHI	
WR24	228	9.2	20.2	BsaI/XhoI	EcoRI	
WR26	648	24.6	35.6	BsaI/XhoI	HindIII, SacI	
WR27	429	16.5	27.5	BsaI/XhoI	SacI	
WR28	507	19.4	30.4	BsaI/XhoI	SacI, EcoRI	
WR35	519	19.8	30.8	BsaI/XhoI	HindIII, SacI, SalI, EcoRI	
WR38	570	21.7	32.7	BsaI/XhoI	SacI, SalI	
WR40	246	9.8	20.8	BsaI/XhoI	SacI, EcoRI	
WR41	285	11.3	22.3	BsaI/XhoI	EcoRI	
WR42	300	11.8	22.8	BsaI/XhoI	none	
WR43	576	21.9	32.9	BsaI/XhoI	SalI	
WR44	399	15.4	26.4	BsaI/XhoI	HindIII	
WR49	621	23.6	34.6	BsaI/HindIII	XhoI	
WR50	444	17.1	28.1	BsaI/HindIII	EcoRI, XhoI	
WR51	540	20.6	31.6	BsaI/XhoI	SalI, EcoRI, HindIII	
WR53	462	17.7	28.7	BsaI/XhoI	SalI, EcoRI	
WR54	264	10.5	21.5	BsaI/XhoI	none	

	cloning in progress
	cloned/amplified
	expression analyzed

Figure 6



Gel 1 Gel 2

Table 5

Target	pET	SUMO uncleaved
AR5	NE	NE
AR22	NE	E/S
FR10	E/NS	E/S
HR894	E/NS	E/NS
HR1553	E/NS	E/S
HR1686	E/NS	E/NS
HR1697	NE	E/S
WR26	E/NS	E/NS

	not expressed (NE)
	expressed/not soluble (E/NS)
	expressed/soluble (E/S)

Acknowledgements

I would like to thank Dr. Gaetano Montelione for allowing me to conduct this research and contributing his time and resources to this project. I would also like to thank Tom Acton for endless discussion, encouragement, and assistance. Thank you to Steve Anderson for taking time to sit on my defense committee and providing helpful comments and interesting stories. Ritu Shastry and Chi Kent Ho have taken so much of their time to show me many molecular biology techniques, for which I am very grateful. Thank you to Li-Chung Ma, YiWen Chiang, and Rong Xiao for helpful discussions and tips. Finally, thanks to all members of the Montelione lab for all your support and all the laughs that we have had.

References

The Arabidopsis Genome Initiative. (2000) *Nature* **408**, 796-815.

The C. elegans Sequencing Consortium (no authors listed). (1998) *Science* **282**, 2012-2018.

Acton, TB, Xiao, R, Clement, T, Gunsalus, K, Everett, JK, Shastry, R, Ma, L-C, Chiang, Y, Wu, M, Ho, CK, Palacios, D, and Montelione, GT. (2003) *Journal of Structural and Functional Genomics* (in press).

Adams, MD, *et al.* (2000) *Science* **287**, 2185-2195

Andersen, DC and Krummen, L. (2002) *Curr. Opin. Biotechnol.* **13,** 117-123.

Bernier-Villamor, V, Sampson, DA, Matunis, MJ, and Lima, CD. (2002) *Cell* **108,** 345-356.

Bertone, P, Kluger, Y, Lan, N, Zheng, D, Christendat, D, Yee, A, Edwards, AM, Arrowsmith, CH, Montelione, GT, and Gerstein, M. (2001) *Nucleic Acids Res.* **29,** 2884-2898.

Butt, TR, Jonnalagadda, S, Monia, BP, Sternberg, EJ, Marsh, JA, Stadel, JM, Ecker, DJ, and Crooke, ST. (1989) *Proc. Natl. Acad. Sci.* **86,** 2540-2544.

Cereghino, GPL, Cereghino, JL, Ilgen, C, and Cregg, JM. (2002) *Curr. Opin. Biotechnol*. **13,** 329-332.

Cereghino, GP and Cregg, JM. (1999) *Curr. Opin. Biotechnol.* **10,** 422-427

Cregg, JM, Cereghino, JL, Shi, J, and Higgins, DR. (2000) *Mol. Biotechnol.* **16,** 23-52.

Desterro, JM, Rodriguez, MS, and Hay, RT. (1998) *Mol. Cell* **2,** 233-239.

Dougherty, WG, Cary, SM, and Parks, TD. (1989) *Virology* **171,** 356-364.

Ecker, DJ, Stadel, JM, Butt, TR, Marsh, JA, Monia, BP, Powers, DA, Gorman, JA, Clark, PE, Warren, F, Shatzman, A, and Crooke, ST. (1989) *J. Biol. Chem.* **264,** 7715-7719.

Edwards, AM, Arrowsmith, CH, Christendat, D, Dharamsi, A, Friesen, JD, Greenblatt, JF, and Vedadi, M. (2000) *Nature Struct. Biol.* **7** (suppl), 970-972.

Evenas, J, Tugarinov, V, Skrynnikon, NR, Goto, NK, Muhandiram, R, and Kay LE. (2001) *J. Mol. Bio.* **309,** 961-974.

Evdokimov, AG, Tropea, JE, Routzhan, KM, and Waugh, DS. (2002) *Acta Cryst.* D**58,** 398-406.

Fox, JD, Kapust, RB, and Waugh, DS. (2001) *Prot. Sci.* **10**, 622-630.

Fox, JD, Routzahn, KM, Bucher, MH, and Waugh, DS. (2003) *FEBS Letters* **537**, 53-57.

Goffeau, A, Barrell, BG, Bussey, H, Davis, RW, Dujon, B, Feldmann, H, Galibert, F, Hoheisel, JD, Jacq, C, Johnston, M, Loius, EJ, Mewes, HW, Murakami, Y, Philippsen, P, Tettelin, H, and Oliver, SG. (1996) *Science* **274**, 546, 563-567.

Gonzalez, JM, Masuchi, Y, Robb, FT, Ammerman, JW, Maeder, DL, Yanagibayashi, M, Tamaoka, J, and Kato, C. (1998) *Extremophiles* **2**, 123-130.

Goh, CS, Lan, N, Echols, N, Douglas, SM, Milburn, D, Bertone, P, Xiao, R, Ma, LC, Zheng, D, Wunderlich, Z, Acton, T, Montelione, GT, and Gerstein, M. (2003) *Nucleic Acids Res.* **31,** 2833-2838.

Hammarstrom, M, Hellgren, N, van Den Berg, S, Berglund, H, and Hard, T. (2002) *Prot. Sci.* **11,** 313-321.

Hershko, A and Ciechanover, A. (1998) *Annu. Rev. Biochem.* **67,** 425-479.

Hong Y, Rogers R, Matunis MJ, Mayhew CN, Goodson ML, Park-Sarge OK, Sarge KD, and Goodson M. (2001) *J. Biol. Chem.* **276,** 40263-40267.

Ito, T, *et al.* (2001) *Proc. Natl Acad. Sci. USA* **98**, 4569-4574.

Jansson, M, Li, YC, Jendeberg, L, Anderson, S, Montelione, GT, and Nilsson, B. (1996) *J. Biomol. NMR* **7**, 131-141.

Jarvis, DL. (2003) *Virology* **310,** 1-7.

Kapust, RB and Waugh, DS. (1999) *Prot. Sci.* **8**, 1668-1674.

Kapust, RB and Waugh, DS. (2000) *Protein Expr. Purif.* **19,** 312-318.

Lander, ES. (1999) *Nat. Genet.* **21** Suppl, 3-4.

Lander, ES. *et al.* (2001) *Nature* **409**, 860-921.

Li, S-J and Hochstrasser, M. (1999) *Nature* **398,** 246-251.

Li, S-J and Hochstrasser, M. (2000) *Mol. Cell Biol.* **20,** 2367-2377.

Lyakhov, DL, He, B, Zhang, X, Studier, FW, Dunn, JJ, and McAllister, WT. (1998) *J. Mol. Bio.* **280,** 201-213.

McClelland, M. *et al.* (2001) *Nature* **413**, 852-856.

Montelione, GT and Anderson, S. (1999) *Nature Struct. Biol.* **6**, 11-12.

Mossessova, E and Lima, CD. (2000) *Mol. Cell* **5,** 865-876.

Pryor, KD and Leiting, B. (1997) *Protein Exp. Purif.* **10,** 309-319.

Qing, G, Ma, L-C, Khorchid, A, Swapna, GVT, Mal, TK, Takayama, MM, Xia, B, Phadtare, S, Ke, H, Acton, T, Montelione, GT, Ikura, M, and Inouye, M. (2004) submitted.

Richarme, G and Caldas, TD. (1997) *J. Biol. Chem.* **272,** 15607-15612.

Routzahn, KM and Waugh, DS. (2002) *Journal of Structural and Functional Genomics* **2**, 83-92.

Saitoh, H, Pu, RT, and Dasso, M. (1997) *TIBS* **22,** 374-376.

Sawasaki, T, Ogasawara, T, Morishita, R, and Endo, Y. (2002) *Proc. Natl. Acad. Sci.* **99,** 14652-14657.

Spurlino, JC, Lu, G-Y, and Quiocho, FA. (1991) *J. Biol. Chem.* **266,** 5202-5219.

Smyth, DR, Mrozkiewicz, MK, McGrath, WJ, Listwan, P and Kobe, B. (2003) *Prot. Sci.* **12,** 1313-1322.

Tsumoto, K, Ejima, D, Kumagai, I, and Arakawa, T. (2003) *Protein Exp. Purif.* **28,** 1-8.

Uetz, P, *et al.* (2000) *Nature* **403**, 623-627.

Verger, A, Perdomo, J, and Crossley, M. (2003) *EMBO Reports* **4,** 137-142.

Winzeler, EA, *et al.* (1999) *Methods Enzymol.* **306**, 3-18.

www.lifesensors.com

www.nesg.org