Journal of Experimental Microbiology and Immunology (JEMI) Vol. 20: 100 - 105

Synthesis, Cloning, and Sequencing of a Codon Optimized Variant of Proteinase Inhibitor II Designed for Expression in *Escherichia coli*

Emily Fogarty, Arshia Alimohammadi, Jacqueline Siu, Alan Stachowiak
Department of Microbiology and Immunology, University of British Columbia

Proteinase Inhibitor II (PI2) is a potato tuber peptidase with many potential applications including pest control, radiation protection, and hunger suppression. Recombinant production using *Escherichia coli* represents a cost-effective approach to produce PI2. In this study, we designed a variant of the *pi2* gene designed for expression in *E. coli* by removal of an intron sequence and optimization of codon usage. The 492 base pair nucleotide sequence termed synPI2 was synthesized and subcloned it into vector pCR2.1-TOPO-TA. The nucleotide sequence of synPI2 was confirmed. The subcloned *pi2* in plasmid pCR2.1-TOPO-TA-PI2 is flanked by several restriction enzyme cut sites and can be easily used to cloned the gene into protein expression vectors such as pET32a in future studies.

Proteins perform a vast array of functions within living organisms and have been harnessed to serve as pharmaceuticals and molecular tools (1). Proteins produced by eukaryotic cells can be difficult to express in large concentrations (1). In addition, the use of eukaryotic cells for protein production may be complex, expensive, and/or unethical, particularly in multicellular organisms. To circumvent these issues, eukaryotic DNA is often transformed into prokaryotic cells (most commonly *Escherichia coli*) and expressed as recombinant protein. However, to express eukaryotic genetic material in prokaryotic cells, GC content and codon frequency need to be optimized since these genetic parameters can be significantly different between prokaryotes and eukaryotes (2). The frequency of the codon usage usually reflects the abundance of cognate tRNA, therefore, an infrequently used codon may not be translated efficiently, resulting in reduced protein expression levels (2).

Proteinase Inhibitor II (PI2) is a 21 kDa, dimeric, cysteine-rich, heat stable, endo-acting peptidase that inhibits chymotrypsin and trypsin (3, 4). It is found in potato tubers and could potentially be used as a pest controlling agent by inhibiting amino acid absorption of herbivores and pathogens by inhibiting their digestive enzymes (5). PI2 could also be used for radiation and UV damage protection by inhibiting proteases involved in the propagation of cellular damage (6, 7). Medically, it could be used as a hunger suppressant as it has been shown to raise cholecystokinin levels, leading to satiety and decreased caloric intake (8).

Previous work has attempted to generate PI2 linked to thioredoxin (TrxA) in pET32a plasmid constructs. These attempts, however, had a PI2 sequence containing an intron that was inserted into the plasmid, and did not result in production of PI2 protein (9, 10). Neither of these studies accounted for the presence of this intron, nor did they codon optimize for expression in *E. coli.* In one of these attempts PI2 was absent in the putative clones (10). Our study revisited the initial phases of the PI2 project. We described the analysis and design of an intron-free, codon optimized *pi2* sequence with appropriate GC content for expression in *E. coli*.

MATERIALS AND METHODS

Analyzing previous PI2 sequence. NCBI Standard Nucleotide BLAST was used to compare the PI2 sequences reported by Keil *et al.* and by Przeworski *et al.* (10, 11). GenScript's Rare Codon Analysis Tool was used to assess GC content and codon usage for *E. coli*. This tool can be found on the Genescript webpage (http://www.genscript.com/cgi-bin/tools/rare_codon_analysis).

Designing PI2 gBlock gene fragment. The synPI2 sequence was codon optimized for *E. coli* expression using OPTIMIZER (12). To facilitate cloning, NcoI and EagI restriction enzyme cut sites were added to the 5' and 3' ends of the synPI2 gene block, respectively.

Strains and plasmids. pCR2.1-TOPO-TA purchased from Invitrogen was used as a storage vector. If not otherwise stated, DH5α *E. coli* was used for all other transformations and plasmid isolations throughout this study. Cells were grown overnight at 37°C at 180 RPM in Lysogeny Broth (LB) liquid medium (pH 7.0) containing tryptone (1.0%), yeast extract (0.5%) and NaCl (1.0%). Ampicillin or Kanamycin was used in LB plates at 100 µg/ml.

Plasmid isolation. Plasmid DNA was isolated using the Invitrogen PureLink HQ Mini Plasmid DNA Purification Kit (Cat# K2100-01). DNA concentration and purity was measured via absorbance at 260 nm and 280 nm using a ThermoScientific NanoDrop 2000c Spectrophotometer.

Preparing electrocompetent DH5α *E. coli.* An isolated colony was grown overnight in 1 ml of LB broth and added to 99 ml of LB broth for a culture volume of 100 ml. Culture was shaken at 180 RPM at 37°C until an OD_{600} of 0.8 - 0.9 was achieved. All centrifugation steps were performed at 4000 x g for 10 minutes. Suspended cultures were pre-cooled and centrifuged. Pellets were resuspended in 100 ml of sterile ice-cold dH_2O and centrifuged. Pellets were again resuspended in 50 ml of sterile ice-cold dH_2O and centrifuged. Pellets were resuspended in 20 ml of sterile ice-cold 10% glycerol and centrifuged. Finally, pellets were resuspended in 0.5 ml of ice-cold sterile 10% glycerol. 40 µl of resuspension was aliquoted into pre-cooled tubes and flash-frozen in liquid nitrogen. Competent cells were stored at -80°C.

Electroporation. Competent DH5α *E. coli* cells aliquots were thawed on ice and transferred into plastic cuvettes with 10 ng of isolated DNA. Using a BioRad MicroPulser (2.5 V, 200 Ohm, 500 uF), cells were transformed with DNA and added to 1 mL LB broth and incubated at 37°C at 350 RPM for 60 minutes. After

incubation, cells were spread onto antibiotic selective LB plates and incubated overnight at 37°C. Colonies were enumerated the following day.

PCR amplification of synPI2 gene block. The following oligonucleotide primers were used to PCR amplify the synPI2 gene block: forward primer synPI2-F (5'-CATCCATGGCTATGGACGTT-3') and reverse primer synPI2-R (5'-CTTGCCGGCCGCATTATTAC-3'). All reactions were carried out in volumes of 25 µl. Each reaction contained 2.5 µl of 10X polymerase buffer, 2.5 µl of 50 mM $MgCl_2$, 0.5 µl of 10 mM dNTPs, 0.5 µl of 10 uM synPI2-F and synPI2-R, 0.2 µl of Platinum Taq DNA polymerase (Invitrogen, Cat# 10966018), 1 µl of 0.1 - 1 ng template DNA and PCR grade dH_2O. PCR cycle conditions used were: an initial denaturation at 94°C for 2 minutes followed by 35 cycles of denaturation 94°C for 30 seconds, annealing at 55°C for 30 seconds and extension at 72°C for 30 seconds, with a final extension at 72°C for 5 minutes. For colony PCR, colonies were gently scraped off plates with a pipet tip, added to the appropriate reaction tube, and lysed by heating for 8 minutes at 94°C to release template DNA. All PCR products were resolved on 1.5% agarose gels supplemented with SYBR at 120V for 60 minutes and visualized by exposure to UV light. Log-2 (0.1-10kb) DNA markers from NEB (Cat# 3200S) were used to determine band sizes.

Isolation of PI2 PCR product. Amplified synPI2, which presented as a distinct band around 500 bp, was excised from the 1.5% agarose gel and purified using an Invitrogen PureLink Quick Gel Extraction Kit (Cat# K2100-12) according to the manufacturer's instructions. The product of the gel extraction was used as the template for subsequent PCR reactions and insertion into various vectors.

Insertion of PI2 into pCR2.1-TOPO vector. In order to store synPI2 for future use, a pCR2.1-TOPO Cloning Kit for Sequencing with One Shot TOP10 Chemically Competent kit (Invitrogen, Cat# K4575-01) was used. Following the manufacturer's instructions, the PCR amplified synPI2 product was inserted into pCR2.1-TOPO and transformed into TOP10 competent cells using heat shock. Transformed cells were plated on LB plates supplemented with kanamycin and X-gal for plasmid selection and blue/white colony screening, respectively.

DNA sequencing of pCR2.1-TOPO(+)PI2 vector. Sanger sequencing was used to identify colonies with potential inserts. Isolated colonies were grown overnight to 5 ml and subject to DNA extraction using Invitrogen PureLink HQ Mini Plasmid DNA Purification Kit (Cat# K2100-01). Sample concentration was confirmed to be approximately 200 ng/µl using a ThermoScientific NanoDrop 2000c Spectrophotometer. An M13 forward primer was used for Sanger sequencing performed by the Nucleic Acid Protein Service Unit at University of British Columbia.

RESULTS

Analysis and optimization of *pi2* nucleotide sequence for expression in *E. coli*. Initially, sequences reported by Keil *et al* and Przeworski *et al.* were compared. Kiel *et al.* first analyzed the sequence of the *pi2* gene, and discovered a 117 nucleotide long intron that is spliced out in the eukaryotic cell post-transcriptionally (11). Bacterial cells are not capable of splicing out introns. Przeworski *et al*. attempted to express this gene in *E. coli,* however were unsuccessful due to the presence of the intron and lack of codon optimization for expression of a eukaryotic gene in a prokaryotic host (10, 11). As shown in Figure 1, Przeworski *et al*. reported an open reading frame (orf) of 579 nucleotides (10) while Keil *et al.* reported a 462 nucleotides orf determined by RNA protection experiments (11). By aligning the reported *pi2* sequences, we show that the *pi2* gene sequence reported by Przeworski *et al.* includes a 117 base pair region that is not present in the cDNA sequence of PI2 reported by Keil et al. (Fig 1a, S1) (10, 11). Keil PI2-cDNA sequence consists of two exons in different reading frames spanning nucleotides 1 - 53 and 171 - 468 of the Keil-PI2 nucleotide sequence. We conclude that Przeworski-PI2 DNA sequence includes a 117 base pair intron between bases 53 and 171.

In addition, based on a study reported by Sharp *et al.* describing *E. coli* codon usage, we determined that *pi2* contains 53 codons used less than 50% of the time in *E. coli* (2). This would potentially decrease the expression levels of PI2 if the protein was expressed in *E. coli.* (Fig 1c) (2). We also found that the GC content of *pi2* sequence ranged from 10% - 63% (Fig 1d). GC content affects DNA stability and mRNA secondary structure; thus a higher and more consistent GC content could lead to improved expression levels (13,14). In addition to this, Knight *et al.* propose that GC content helps to drive codon bias to some extent (19). To address these issues, we re-designed the *pi2* gene sequence to accommodate expression in *E. coli*. The optimized DNA sequence (titled synPI2) was chemically synthesized as a gene block fragment (Fig 1c, d) (2).

PCR amplification and cloning of *pi2* into pCR2.1-TOPO-TA. Our initial attempts at directly cloning the synPI2 gene block fragment into a pET32a expression vector were unsuccessful, likely due to EagI having lower enzymatic activity in the double digest. To ensure that the gene block wasn't depleted, we decided to amplify the synPI2 DNA using PCR. Primers were designed to amplify the 500 bp sequence. As shown in Figure 2a, the gene block was amplified as evidenced by distinct bands migrating at 500 bp in an agarose gel. Fainter bands are visible at 1000 bp at all different template DNA concentrations used (Fig 2a). When water was used as a negative control for template bands were not observed in the gel (Fig 2a).

Next, we cloned the synPI2 PCR product into pCR2.1-TOPO using topoisomerase I. Single deoxyadenosine residues added to ends of the *pi2* PCR product by Taq polymerase allow insertion into pCR2.1-TOPO (15). After pCR2.1-TOPO *pi2* ligation and electrotransformation into DH5α *E. coli*, blue/white colony screening on X-Gal and selection on a kanamycin agar plate was used to identify colonies containing plasmids with the *pi2* insertion. Five white colonies (suspected synPI2 positive colonies), and one blue colony were screened by PCR using the *pi2* primers described above. PCR products were resolved by gel electrophoresis (Fig 2b). All five white colonies have a band at 500 bp. The 500 bp band is not seen in PCR performed on the blue colony. The water control also has a faint band at 500 bp, possibly a result of synPI2 template contamination from another sample. Two of the white colonies have 100 bp bands, possibly a result of primer dimerization. The observation of a band at 500 bp in white colonies but not in the blue colony support the conclusion that synPI*2* has been inserted into pCR2.1-TOPO plasmid.

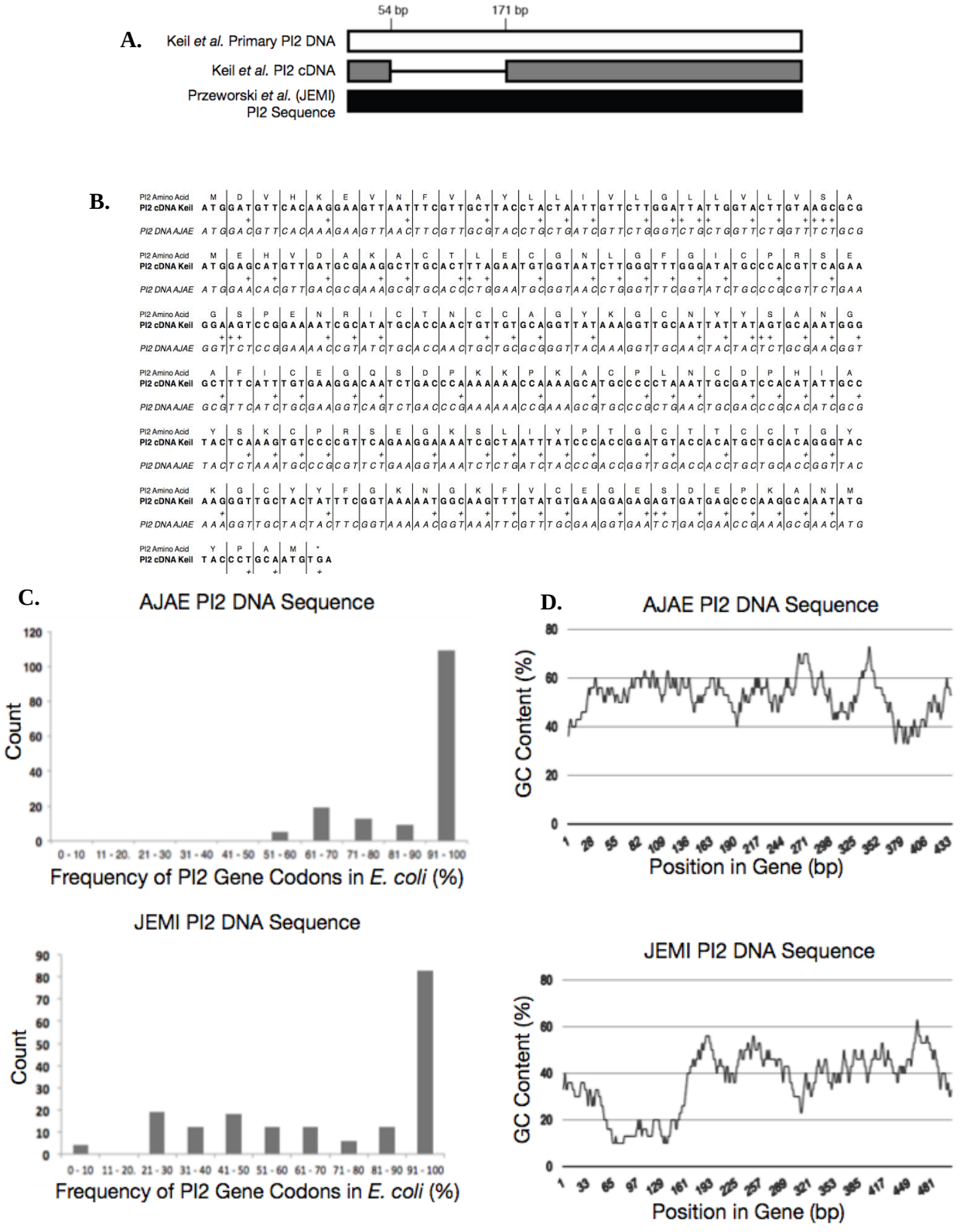

FIG. 1. Analyzing previous PI2 sequence and designing new PI2 gBlock. (A) Representation of BLAST results comparing Keil *et al.*'s primary DNA and cDNA sequence with Przeworski *et al.*'s JEMI PI2 sequence, showing the presence of an intron, represented as a line, between 54 bp and 171 bp (B) Comparing Keil *et al.*'s cDNA sequence with AJAE-PI2's *E. coli* codon optimized PI2 DNA sequence. The associated single letter amino acid codes are indicated above, vertical lines mark the open reading frame, and (+) indicates which nucleotide base are changed. (C) Histogram of how frequently PI2 codons are used (%) relative to *E. coli* codon usage patterns. (D) GC content % over JEMI-PI2 or AJAE-PI2 DNA sequence.

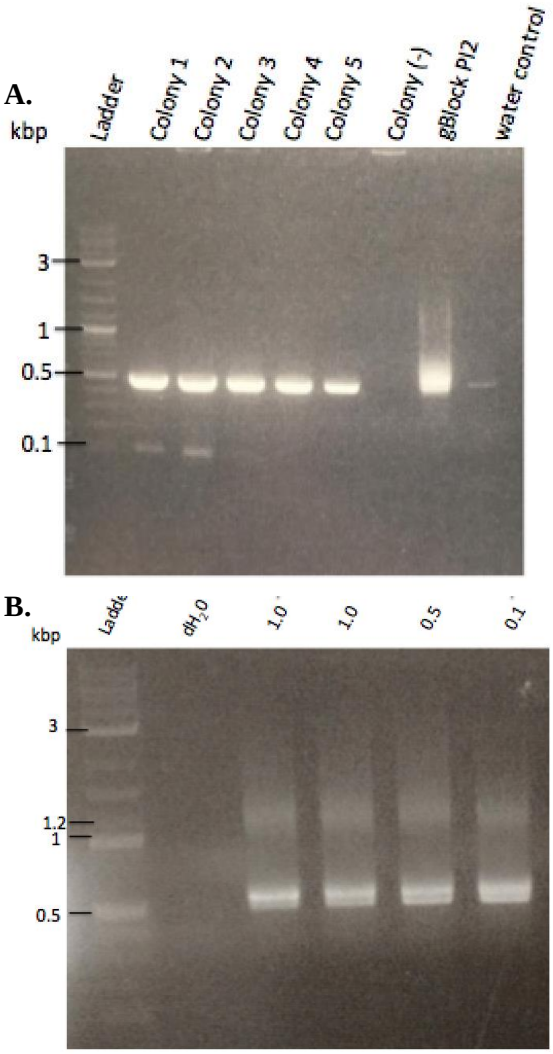

FIG. 2. PI2 PCR products visualized via gel electrophoresis in 1.5% agarose. (A) PCR amplification of PI2 gDNA. (B) Colony PCR of pCR2.1-TOPO(+)PI2 containing colonies to screen for PI2 inserts. 2-log ladder from NEB.

Sequencing results show *pi2* gene inserted into pCR2.1-TOPO. DNA sequencing was performed in order to confirm that the synPI2 gene fragment had been cloned into pCR2.1-TOPO DNA. Analysis of the results sequence data indicated that the synPI2 fragment had been inserted into pCR2.1-TOPO with the restriction enzyme site NcoI located at the 5' of the inserted *pi2* gene and EcoRI located at both ends of the inserted gene. Of the five colonies screened and sequenced, only plasmid pCR2.1-TOPO-1-1eta_M13R had no mutations. The other four sequenced plasmids each had the same silent point mutation at the 81st nucleotide where the codon changed from CCA to CCG. The sequencing results for plasmid pCR2.1-TOPO-1-1eta_M13R are shown in Figure S5.

DISCUSSION

For PI2 expression in *E. coli*, the proper primary amino acid sequence must be translated. For the majority of eukaryotes, the mRNA transcribed in the nucleus must be modified prior to its translation in the cytoplasm; modification may include intron removal and mRNA splicing (16). However, most prokaryotes, including *E. coli*, do not have introns or intron removal mechanisms. We determined that the PI2 nucleotide sequence reported by Pzrowrsoki *et al*. matched the eukaryotic potato *pi2* gene, which contains a 117 base pair intron (Fig 1a, S1) (10, 11). The presence of an intron is problematic because *E. coli* is unable to splice the intron out, and the wrong mRNA sequence is used for translation. Not only are additional amino acids translated, but the second exon is out of frame with the first intron.

In addition to an intron, the nucleotide sequence of the native pi2 is not optimal for expression in *E. coli* due to a number of codons that are rarely used in *E. coli*. Although parts of the genetic code are degenerate, some species have different preferences for certain amino acid codons (2). As shown in Figure 1c, 30% of the native *pi*2 codons are preferentially used less than 50% of the time in *E. coli*. The literature suggests that it is important to account for codon bias when expressing recombinant proteins because if the distribution of codons is significantly different than the typical *E. coli* codon distribution a reduction in quality or quantity of the heterologous protein could occur (17). Hu *et. al.* show that even though different codons code for the same amino acid, there can be a difference in protein solubility and functionality (18). This is potentially as a result of mRNA secondary structure leading to increased translation initiation, however the authors stress that this topic needs to be investigated further. (18). Thus, while it is important to optimize the codons for *E. coli* codon usage patterns, it is crucial to take into account these changes to ensure that the protein is functioning optimally for downstream processes. In DNA, the guanine and cytosine nucleotide base pairs form a stronger bond with three hydrogen bonds while the adenine and thymine base pairs have a weaker bond with only bound by two hydrogen bonds. Overall, the GC content affects the stability of the DNA and the secondary structure of the mRNA. For *E. coli*, it appears that for most native genes the GC content is between 15% and 70% (19). The model proposed by Knight *et.al.* suggests that some codons are used more frequently with increasing GC content (19). In other words, GC content helps to drive codon bias (19). As shown in Figure 1d, the GC content of *pi2* ranged from 10% to 63%. Because of the presence of an intron, inadequate *E. coli* codon usage, and highly varied GC content, we decided to design and order a new *pi2* sequence from Genescript. The designed synPI2 is based on the published cDNA sequence, and we further optimized the codons for *E. coli* translation and ensured the GC content is within 35%-70% as shown in Figure 1b, c, d.

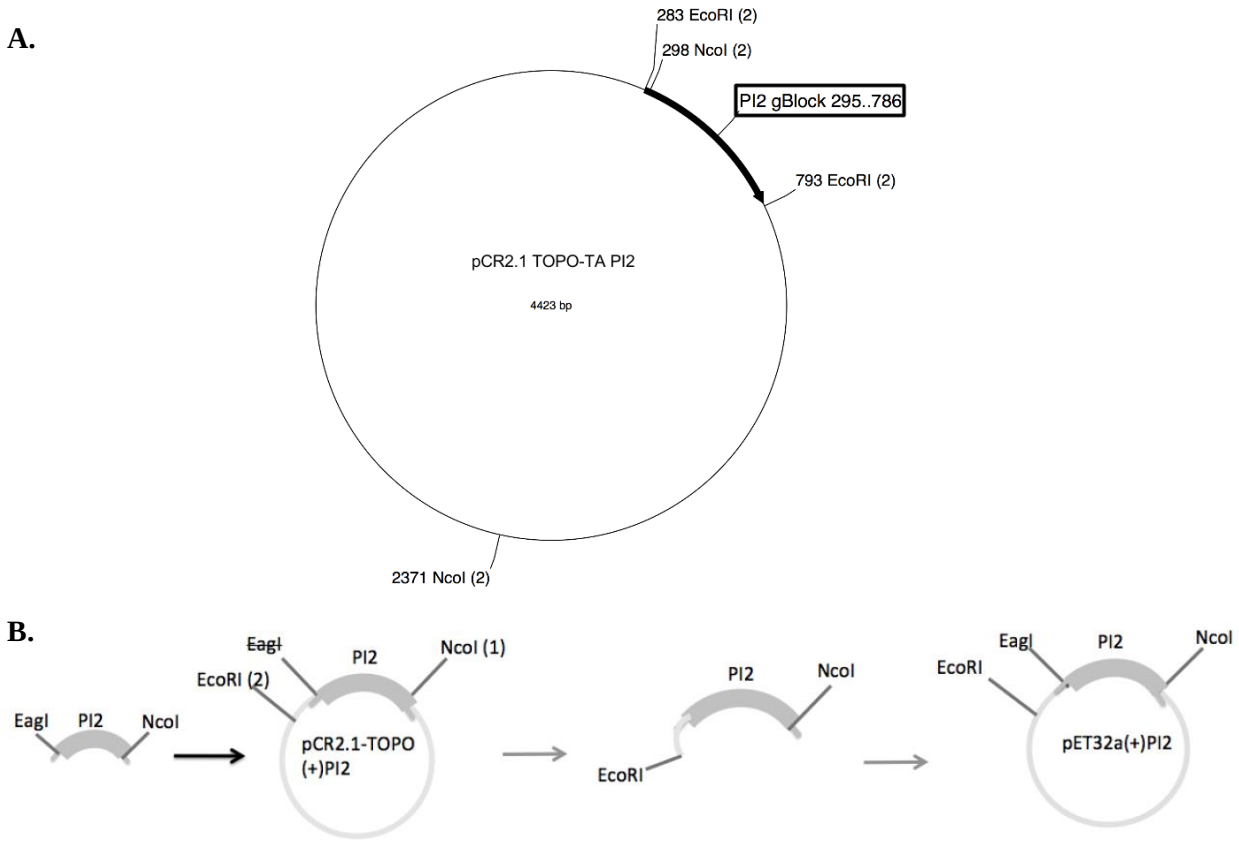

FIG. 3. Cloning representation of PI2 into pCR2.1-TOPO and future insertion into pET32a. (A) Plasmid map of pCR2.1-TOPO(+)PI2 designed in APE program illustrating enzyme cut sites and PI2 gBlock insertion. (B) Pictorial representation of PI2 cloning. Solid black arrow represents successful insertion of PI2 into pCR2.1-TOPO. Grey arrows indicate future digests and ligations to insert PI2 into pET32a. Plasmid map designed in APE program.

In order to preserve the synPI2 gene block DNA for future use it was amplified using PCR and inserted into pCR2.1-TOPO. Figure 3b illustrates the cloning strategy for synPI2. In the future, synPI2 DNA can be excised using restriction enzymes or amplified using PCR from pCR2.1-TOPO and ligated into another plasmid for protein expression (Fig 3b).

FUTURE DIRECTIONS

The insertion of synPI2 into pCR2.1-TOPO allows a future group to subclone this sequence into an expression vector (Fig 3b). Although the restriction enzyme cut sites designed in the synPI2 were EagI and NcoI, using another enzyme in the place of EagI is recommended, as EagI has lower activity in a double digest. An EcoRI site could be added into the synPI2 using a primer designed with this site. Once synPI2 is isolated from pCR2.1-TOPO, it can be ligated directly into an expression vector such as pET32a. PI2 has 16 Cys residues that form 8 disulphide bonds. Because of the complex folding, this protein often misfolds resulting in the formation of inclusion bodies. pET32a is a favourable choice because it includes thioredoxin, a small protein that is known to decrease the formation of inclusion bodies (6). *E. coli* strain Origami would be an interesting expression host to test since it has been engineered to have an oxidative cytoplasm which allows is to facilitate proper disulphide bond formation for proteins with many disulphide bonds (20). The protein expression levels of PI2 can be evaluated by running whole cell lysates of soluble and insoluble fractions on SDS-PAGE. If expression cannot be determined by whole cell lysates, an affinity purification may reduce background and effectively concentrate the PI2 protein.

ACKNOWLEDGEMENTS

This study was facilitated by the University of British Columbia Microbiology and Immunology department. We would like to thank Dr. David Oliver and Chris Deeg for their knowledge and support throughout the course of this study. We would also like to thank the students previously involved with this project for providing us with the foundation of our study.

REFERENCES

1. **Griffiths AJF, Gelbart WM, Miller JH, et al.** Modern Genetic Analysis. New York: W. H. Freeman; 1999. Expressing Eukaryotic Genes in Bacteria.
2. **Sharp P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright**. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and*

Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. **16**:8207–11.
3. **Bryant J, TR Green, T Gurusaddaiah, and CA Ryan**. 1976. Proteinase inhibitor II from potatoes: isolation and characterization of its protomer components. Biochemistry.
4. **Li X.-Q. Q., T. Zhang, and D. Donnelly**. 2011. Selective loss of cysteine residues and disulphide bonds in a potato proteinase inhibitor II family. PLoS ONE **6**:e18615.
5. **Bergey D. R., G. A. Howe, and C. A. Ryan**. 1996. Polypeptide signaling for plant defensive genes exhibits analogies to defense signaling in animals. Proc. Natl. Acad. Sci. U.S.A. **93**:12053–8.
6. **Billings P. C., A. R. Morrow, C. A. Ryan, and A. R. Kennedy**. 1989. Inhibition of radiation-induced transformation of C3H/10T1/2 cells by carboxypeptidase inhibitor 1 and inhibitor II from potatoes. Carcinogenesis **10**:687–91.
7. **Conconi A., M. J. Smerdon, G. A. Howe, and C. A. Ryan**. 1996. The octadecanoid signalling pathway in plants mediates a response to ultraviolet radiation. Nature **383**:826–9.
8. **Hill A. J., S. R. Peikin, C. A. Ryan, and J. E. Blundell**. 1990. Oral administration of proteinase inhibitor II from potatoes reduces energy intake in man. Physiol. Behav. **48**:241–6.
9. **Geum L., Huber R., Leung N, Lowe M.** 2015 Construction of Recombinant Expression Vectors to Study the Effect of Thioredoxin on Heterologous Protein Solubility. J. Exp. Microbiol. Immunol. **19;**1-6.
10. **Przeworski C., Pham D., Wang I., Murillo J.** 2015. Attempted Construction of Recombinant Vectors Designed to Study the Solubility of Overexpressed Proteinase Inhibitor 2 when Co-expressed with Thioredoxin. J. Exp. Microbiol. Immunol. **19**:1-6.
11. **Keil M., J. Sanchez-Serrano, J. Schell, and L. Willmitzer.** 1986. Primary structure of a proteinase inhibitor II gene from potato (*Solanum tuberosum*). Nucleic Acids Res. **14**:5641–50.
12. **Puigbò P., E. Guzmán, A. Romeu, and S. Garcia-Vallvé**. 2007. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res.**35**:W126–31.
13. **SantaLucia, J, Allawi, HT, Seneviratne, PA.** 1996. Improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry (N. Y.). 35:3555-3562.
14. **Rao, YS, Chai, XW, Wang, ZF, Nie, QH, Zhang, XQ.** 2013. Impact of GC content on gene expression pattern in chicken. Genetics Selection Evolution. 45:1.
15. **ThermoFisher Scientific. 2015. The technology behind TOPO® cloning.** Accessed 10 Nov 2015 (http://www.thermofisher.com/us/en/home/life-science/cloning/topo/topo-resources/the-technology-behind-topo-cloning.html)
16. **Darnell J. E.** 2013. Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. RNA **19**:443–60.
17. **Kane J. F.** 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr. Opin. Biotechnol. **6**:494–500.
18. **Hu S., M. Wang, G. Cai, and M. He**. 2013. Genetic code-guided protein synthesis and folding in *Escherichia coli*. J. Biol. Chem. **288**:30855–61.
19. **Knight R. D., S. J. Freeland, and L. F. Landweber**. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. **2**: 1-6.
20. **Gross, E, Kastner, DB, Kaiser, CA, Fass, D.** 2004. Structure of Ero1p, source of disulfide bonds for oxidative protein folding in the cell. Cell. 117:601-610.