Expression and purification of the predicted \textit{PEG10} aspartyl protease domain

Caillan Crowe-McAuliffe

A thesis submitted for the degree of
Master of Science

Department of Biochemistry
University of Otago, Dunedin,
New Zealand.

April 2013
Abstract

_Paternally Expressed Gene 10 (PEG10)_ is an imprinted, retrotransposon-derived gene found in mammals. Although many of the retrotransposon domains have become degenerated in _PEG10_, a predicted retroviral-type aspartyl protease (AP) domain has been highly conserved. Retroviral-type APs play a crucial role in the replication of some retroviruses such as the Human Immunodeficiency Virus (HIV) and are therefore important drug targets. Consequently, extensive biochemical and structural data are available for this class of proteins, although the vast majority of this has been gathered from only a small number of retroviral enzymes. Preliminary evidence indicates that the PEG10 AP is an active protease, although proteolysis by this enzyme has yet to be observed _in vitro_ (Clark _et al._, 2007). This study aimed to express, purify, and characterise the predicted PEG10 AP.

A number of _PEG10_ AP clones, each with different termini and across more than one recombinant expression system, were expressed to produce the PEG10 AP domain in _E. coli_. The majority of expressed proteins were largely insoluble and unsuitable for further characterisation. One clone, however, produced soluble PEG10 AP in sufficient quantities for purification and further analysis. Several lines of evidence indicated that the purified protein was dimeric in solution, consistent with the quaternary structure of other retroviral-type APs.

The results presented in this thesis support the hypothesis that the PEG10 AP is active and has retained characteristics from the ancestral retrotransposon enzyme. The expression and purification protocol that has been developed can now be used to generate PEG10 AP for detailed biophysical and functional characterisation. Widely used protease inhibitors used for treatment of infection with HIV have pleiotropic effects in patients and are known to inhibit a diverse range of proteases. The development of an _in vitro_ activity assay and testing whether the PEG10 AP is a serendipitous target of these inhibiting compounds will not only enhance our understanding of retroviral-type APs but potentially contribute to their use as therapeutic agents.
Acknowledgements

I would firstly like to acknowledge my two supervisors, Dr Elizabeth Poole and Prof. Warren Tate. The support you provided throughout the project was outstanding, and the freedom you gave me early on was much appreciated. Liz, your practical and level-headed approach was invaluable, especially during the more volatile periods of the project. Warren, I have always been thankful for your openness and generosity of spirit. Cheers to both of you!

Several other staff members and fellow students also deserve thanks. Ryan Hill gave abundant, valuable advice for ligation-independent cloning. Tina, Katie, and Yosuke endured a constant barrage of questions and badgering with aplomb and humour, and—along with other members of the Tate group—deserve special mention for maintaining such an enjoyable environment to work in. Prof. Catherine Day gave much-appreciated advice (both technical and otherwise), and, in collaboration with Rhesa Budidarmo, kindly gifted reagents. Feedback and expertise provided by members of the Krause, Cutfield, and Patrick groups was likewise appreciated. I would finally like to thank the Blair trust, Marsden fund, and University of Otago for providing funding for this project.
# Contents

## 1 Introduction

1.1 Aspartyl proteases ........................................... 1
   1.1.1 Broad-range inhibitors of aspartyl proteases ............ 2
   1.1.2 Aspartyl proteases in retroelements ................... 3
   1.1.3 Substrate specificity of retroviral-type aspartyl proteases ... 10
   1.1.4 Inhibitors of retroviral-type aspartyl proteases ........ 12
   1.1.5 Retroviral-type aspartyl proteases in eukaryotic genomes ... 13

1.2 Paternally Expressed Gene 10 ................................ 14
   1.2.1 LTR-retrotransposons, retroviruses, and PEG10 ......... 15
   1.2.2 Expression and function of PEG10 ....................... 16
   1.2.3 The predicted aspartyl protease domain of PEG10 ........ 19

1.3 Aims .......................................................... 20

## 2 Methods

2.1 Bioinformatic analysis of the PEG10 aspartyl protease domain .... 21
2.2 Analysis of mammalian PEG10 sequences ........................ 22
2.3 Sequence alignment ........................................... 22
2.4 DNA manipulation ............................................. 22
   2.4.1 Preparation of competent E. coli cells .................. 22
   2.4.2 Transformation of competent E. coli cells ............... 23
   2.4.3 Plasmid preparations ..................................... 23
   2.4.4 Restriction endonuclease digests ....................... 24
   2.4.5 Cloning .................................................. 24
   2.4.6 Site-directed mutagenesis ................................ 26
   2.4.7 Plasmid sequencing ...................................... 27
   2.4.8 Agarose gel electrophoresis ............................. 27
2.5 Analysis, expression, and purification of proteins .............. 28
   2.5.1 Protein quantitation ..................................... 28
2.5.2 Colloidal Coomassie staining of polyacrylamide gels . . . . . . 28
2.5.3 SDS-PAGE .......................................................... 28
2.5.4 Western immunoblot ................................................ 29
2.5.5 Mass spectrometry .................................................. 31
2.6 Expression and purification of the PEG10-aspartyl protease fusion proteins in E. coli Rosetta 2 (DE3) ........................................ 31
  2.6.1 Small-scale expression of PEG10-aspartyl protease fusion proteins .......................................................... 31
  2.6.2 Medium-scale expression of MBP–PEG10-AP clone 4 fusion proteins .......................................................... 32
  2.6.3 Small-scale amylose purification of MBP–PEG10 AP fusion proteins .......................................................... 33
  2.6.4 Medium-scale amylose-resin purification of MBP–PEG10 AP clone 4 fusion protein ........................................... 33
  2.6.5 Fast-performance liquid chromatography ........................ 34
  2.6.6 Reciprocal amylose- and cobalt-resin purifications ......... 34
  2.6.7 Medium-scale cobalt-resin purification of the his–PEG10-AP clone 4 protein ..................................................... 35
2.7 Proteolysis assays ........................................................ 35
  2.7.1 Autoproteolysis assays ............................................. 35
  2.7.2 Post-lysis autoproteolysis assays ................................ 36
  2.7.3 Proteolysis of GST–Peg10-ORF1 ............................... 36
2.8 Glutaraldehyde cross-linking ......................................... 37
2.9 Primers used in this study .............................................. 37

Results ........................................................................... 40

3 Bioinformatic analysis and cloning of the predicted PEG10 aspartyl protease 40
  3.1 Bioinformatic analysis to define the predicted PEG10 aspartyl protease domain ......................................................... 40
    3.1.1 Analysis of mammalian PEG10 sequence is insufficient to determine domain boundaries ........................................... 40
    3.1.2 Ty3/Gypsy LTR-retrotransposon relatives of PEG10 contain a phylogenetic signature of retroviral-type aspartyl proteases but do not inform domain boundaries ........................................ 41
3.1.3 Homology modelling fails to generate a plausible structure of the PEG10 aspartyl protease but reveals unusual proline enrichment .................................................. 44
3.1.4 Comparative analysis of the PEG10 aspartyl protease with structurally characterised, homologous retroviral-type APs ........ 47
3.1.5 Choosing several predicted termini for the predicted PEG10 aspartyl protease domain .................................................. 47

3.2 Cloning of possible PEG10 aspartyl protease sequences ........ 52
3.2.1 Cloning of PEG10 aspartyl protease fused to a large partner . 52
3.2.2 Creation of a pMAL-TEV plasmid ................................. 52
3.2.3 Creating a small library of PEG10 aspartyl protease clones ... 55
3.2.4 Amplification of inserts by PCR ..................................... 56
3.2.5 Site-directed mutagenesis of the predicted active site ......... 59

3.3 Conclusions ................................................................. 60

4 Expression and purification of the predicted PEG10 aspartyl protease 61
4.1 Small-scale expressions and purifications of 30 predicted PEG10 aspartyl protease clones ............................... 61
4.1.1 The major cleavage event during expression is active-site dependent .................................................. 64
4.1.2 A map of cleavage events occurring during expression .... 64
4.2 Purification of the MBP–PEG10 aspartyl protease fusion proteins. .... 65
4.2.1 Small-scale purification of the MBP–PEG10-AP fusion proteins 65
4.2.2 Autoproteolysis does not occur post-purification .......... 66
4.2.3 Does autoproteolysis occur after expression? .......... 69
4.2.4 Conclusions .............................................................. 74
4.3 Removal of the MBP from the predicted PEG10 aspartyl protease .... 75
4.3.1 Cleavage of the MBP–PEG10 aspartyl protease fusion proteins with TEV protease .................................................. 75
4.3.2 Confirmation of the identity of the PEG10 aspartyl protease by Western immunoblotting and mass spectrometry .... 75
4.3.3 Chromatography to further purify the PEG10 aspartyl protease domain .................................................. 78
4.4 Does the PEG10 aspartyl protease interact with MBP? .......... 81
4.4.1 Summary ................................................................. 85
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II. Alignment of mammalian PEG10 sequence surrounding the predicted as-</td>
<td>133</td>
</tr>
<tr>
<td>partyl protease domain</td>
<td></td>
</tr>
<tr>
<td>III. Prediction of globular domains throughout PEG10</td>
<td>134</td>
</tr>
<tr>
<td>IV. Analysis of rare codons in the predicted PEG10 aspartyl protease domain</td>
<td>135</td>
</tr>
<tr>
<td>V. OD_{600} values</td>
<td>136</td>
</tr>
<tr>
<td>VI. Solubility of PEG10 AP subsequent to separation from MBP by the TEV protease</td>
<td>137</td>
</tr>
<tr>
<td>VII. Mass spectrometric sequencing of MBP–PEG10-AP clone 4 products</td>
<td>138</td>
</tr>
<tr>
<td>VIII. Mass spectrometric sequencing of the his–PEG10-AP clone 4 and his-</td>
<td>139</td>
</tr>
<tr>
<td>TEV–PEG10-AP clone 10 products</td>
<td></td>
</tr>
<tr>
<td>IX. Size markers used for size-exclusion chromatography</td>
<td>140</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Abbreviated proposed catalytic mechanism of the HIV-1 aspartyl protease .......................... 3
1.2 The Gag-Pol fusion protein and diagram of HIV-1 with annotated protease cleavage positions .................. 5
1.3 Structural features of the HIV-1 aspartyl protease .................. 6
1.4 Conservation of Pol domains among diverse *Ty3/Gypsy* LTR-retrotransposons .......................... 7
1.5 Termini from homologous clan AA aspartyl protease domains have diverse morphologies .................. 9
1.6 Sequence logo of HIV-1 AP substrates based on observed cleavage events 11
1.7 Structure of the HIV protease inhibitor saquinavir .................. 13
1.8 Comparison of the *PEG10* domain structure with parasitic retroelements .................. 15
1.9 Mutagenesis experiments imply that the *PEG10* aspartyl protease is active .......................... 19

3.1 Alignment of the predicted human PEG10 AP domain with closely related LTR-retrotransposons .......................... 43
3.2 Proline and pre-proline residues of PEG10 AP homology models have unfavourable backbone conformations .................. 45
3.3 LTR-retrotransposon relatives of *PEG10* are enriched in proline residues .................. 46
3.4 Optimal alignment of PEG10 AP with three characterised homologues .......................... 48
3.5 Insertions in PEG10 relative to other retroviral-type APs are preferentially located in loops .......................... 49
3.6 Termini of constructs explored in this study .................. 51
3.7 Replacing the factor Xa cleavage site with a TEV-protease cleavage site .................. 53
3.8 Creation of a linear pMAL-TEV amplicon .......................... 54
3.9 Diagnostic restriction digest of pMAL-TEV plasmid preparations .......................... 55
3.10 The modified CPEC method used to clone predicted PEG10 AP am-
plicons into pMAL-TEV .............................................. 57
3.11 PCR amplification of inserts corresponding to predicted PEG10 AP
domains ................................................................. 58
3.12 Linearising pMAL-TEV with Bam HI ............................ 58
3.13 PCR screen for clones 29 and 30 ............................... 59
3.14 Sample sequence from PEG10 AP clone 11 in pMAL-TEV .... 59
3.15 Active-site mutants of the PEG10 AP ............................. 60

4.1 SDS-PAGE analysis of MBP–PEG10-AP expression in E. coli .... 62
4.2 Anti-MBP Western immunoblot of post-induction E. coli lysates .. 63
4.3 MBP–PEG10-AP proteins with a variant active-site have a different
banding pattern compared to wild-type products .................... 64
4.4 Deduced cleavage events from immunoblotting data ............... 65
4.5 Pilot small-scale expression and purification of an MBP–PEG10 fusion
protein ................................................................. 66
4.6 Autoproteolysis assays of semi-purified fractions in high salt and at
low pH ..................................................................... 67
4.7 Autoproteolysis assay of semi-purified fractions in PBS ........... 68
4.8 A Coomassie-stained gel showing a medium-scale expression and pu-
rification of MBP–PEG10-AP clone 4 ............................. 69
4.9 Mixing purified MBP–PEG10-AP clone 4 protein with E. coli lysate
does not restore autoproteolytic activity ............................... 70
4.10 Minor autoproteolysis occurs post-lysis ............................ 72
4.11 Autoproteolysis proceeds over induction .......................... 73
4.12 Effects of variant active-site triads on autoproteolysis during expression 74
4.13 Cleavage of the MBP–PEG10-AP fusion proteins with the TEV protease 76
4.14 SDS-PAGE and Western immunoblot analysis of partially purified and
TEV-protease-cleaved MBP–PEG10-AP clone 4 protein .......... 77
4.15 Tandem MS/MS sequencing spectra of PEG10-AP clone 4 ........ 79
4.16 The PEG10-AP clone 4 protein co-purifies with MBP during size-
exclusion chromatography ............................................. 80
4.17 The PEG10-AP clone 4 protein co-purifies with MBP during ion-exchange
chromatography ....................................................... 81
4.18 Insertion of a C-terminal his-tag into PEG10-AP clone 4 ........ 82
4.19 Experimental strategy for testing an association between PEG10-AP and MBP ................................................................. 83
4.20 Tandem pulldowns imply interaction between PEG10-AP and MBP .............................................................. 84
5.1 Creation of his-tagged constructs ................................................................. 87
5.2 Expression and solubility trials of his-TEV– constructs ................................................................. 90
5.3 Expression and solubility trials of his- constructs ................................................................. 91
5.4 Purification of the his–PEG10-AP clone 4 product with cobalt resin ................................................................. 93
5.5 His–PEG10-AP clone 4 does not cleave GST–Peg10-ORF1 in vitro ................................................................. 94
5.6 The his–PEG10-AP clone 4 protein migrates as a dimer in size-exclusion chromatography ................................................................. 95
5.7 Glutaraldehyde cross-linking and non-reducing SDS-PAGE suggest that the his–PEG10-AP clone 4 product is a dimer ................................................................. 96
6.1 The site of −1 frameshifting in PEG10 mRNA ................................................................. 103
6.2 The hypothesised inhibition of PEG10 AP by HIV PIs is consistent with existing literature ................................................................. 111
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>PCR primers used to generate inserts for cloning</td>
<td>38</td>
</tr>
<tr>
<td>2.2</td>
<td>PCR Primers used for site-directed mutagenesis</td>
<td>39</td>
</tr>
<tr>
<td>2.3</td>
<td>Sequencing primers</td>
<td>39</td>
</tr>
<tr>
<td>2.4</td>
<td><em>E. coli</em> strains used in this study</td>
<td>39</td>
</tr>
<tr>
<td>3.1</td>
<td>Predicted protein sequences most similar to <em>PEG10</em></td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Termini matrix</td>
<td>50</td>
</tr>
</tbody>
</table>
List of abbreviations

φ Either of leucine, isoleucine, or valine
AIDS Acquired immunodeficiency syndrome
ALK1 Activin receptor-like kinase 1
AP Aspartyl protease
BLAST Basic local alignment search tool
BLAT BLAST-like alignment tool
C-terminus Carboxyl-terminus
Coomassie Coomassie Brilliant Blue staining solution
CPEC Circular polymerase-extension cloning
Cracking buffer Strongly denaturing and reducing SDS-PAGE loading buffer
Ddi1 DNA-damage inducible [homologue] 1
dNTP Deoxyribonucleotide triphosphate
d.p.c. Days post-coitum
dpi Dots per inch
DTT Dithiothreitol
EIAV Equine infectious anemia virus
eIF4G Eukaryotic [translation] initiation factor 4 G
EDTA Ethylenediaminetetraacetic acid
EMBOSS European molecular biology open software suite
EPA Environmental protection authority
ERMA Environmental risk management authority
FPLC Fast-performance liquid chromatography
his Hexahistidine tag
his-TEV Hexahistidine tag with TEV-protease cleavage site
HIV Human immunodeficiency virus
IPTG Isopropyl β-D-1 thiogalactoside
LB broth Luria–Bertani broth
LTQ Linear trap quadrupole
LTR Long terminal repeat
MALDI Matrix-assisted laser desorption/ionisation
MBP Maltose-binding protein
MLV Murine leukemia virus
MMP Metallomatrix protease
MUSCLE Multiple sequence comparison by log-expectation
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminus</td>
<td>Amino-terminus</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NIX-1</td>
<td>Neuronal interacting factor X 1</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate-buffered saline</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDB ID</td>
<td>Protein data bank identifier</td>
</tr>
<tr>
<td>PEG10</td>
<td>Paternally expressed gene 10</td>
</tr>
<tr>
<td>PI</td>
<td>Protease inhibitor</td>
</tr>
<tr>
<td>PMSF</td>
<td>Phenylmethanesulfonylfluoride</td>
</tr>
<tr>
<td>rpm</td>
<td>Revolutions per minute</td>
</tr>
<tr>
<td>RTL1</td>
<td>Retrotransposon-like 1</td>
</tr>
<tr>
<td>SASPase</td>
<td>Skin-associated aspartic protease</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural classification of proteins</td>
</tr>
<tr>
<td>SDM</td>
<td>Site-directed mutagenesis</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium-dodecyl sulphate</td>
</tr>
<tr>
<td>SDS-PAG</td>
<td>Sodium-dodecyl sulphate-polyacrylamide gel</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium-dodecyl sulphate-polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SEC</td>
<td>Size-exclusion chromatography</td>
</tr>
<tr>
<td>SIAH</td>
<td>Seven in absentia homologue</td>
</tr>
<tr>
<td>TBST</td>
<td>Tris-buffered saline + Tween20</td>
</tr>
<tr>
<td>TEV</td>
<td>Tobacco etch virus</td>
</tr>
<tr>
<td>TEMED</td>
<td>Tetramethylethylenediamine</td>
</tr>
<tr>
<td>TIFF</td>
<td>Tagged image file format</td>
</tr>
<tr>
<td>TGF</td>
<td>Transforming growth factor</td>
</tr>
<tr>
<td>TOF</td>
<td>Time of flight</td>
</tr>
<tr>
<td>Tris</td>
<td>Tris(hydroxymethyl)aminomethane</td>
</tr>
<tr>
<td>Tween 20</td>
<td>Polysorbate 20</td>
</tr>
<tr>
<td>US FDA</td>
<td>United States Food and Drug Administration</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>XMRV</td>
<td>Xenotropic murine leukemia virus-related virus</td>
</tr>
</tbody>
</table>
A note on nomenclature and abbreviations

Conventions for rendering the names of genes and proteins vary between field and organism. Throughout this document, gene names are rendered in italics, and the corresponding protein product—which often has the same name—is rendered in upright text. These terms are often, but not always redundant: an alignment of PEG10 sequence would consist of nucleotides, while an alignment of PEG10 sequence would consist of amino acid residues, for example. Upright characters have been used when naming clones. The names of human genes and proteins are written entirely in uppercase (e.g. PEG10), while for mouse only the first letter is capitalised (e.g. Peg10). Convention has been given precedence over consistency for some abbreviations, so that “mouse Peg10 expression begins at 9.5 d.p.c”. Where no succinct, suitable alternatives were available, abbreviated colloquialisms have been used. For example, strongly denaturing and reducing SDS-PAGE loading buffer is referred to as ‘cracking buffer’.
Chapter 1

Introduction

Paternally Expressed Gene 10 (PEG10) is a highly unusual imprinted gene found in the genome of all therian mammals examined thus far. There is notable sequence homology between long-terminal repeat (LTR)-retrotransposons—genomic parasites found throughout eukaryotes—and PEG10. This implies that a process of ‘molecular domestication’ has occurred, in which an LTR-retrotransposon became partially degenerated and was co-opted by the host as a cellular gene. In the case of PEG10, this occurred in an early mammal (Butler et al., 2001; Ono et al., 2001). PEG10 has lost some of the domains required to replicate as an independent element and instead has gained an important endogenous function, namely involvement in placental development (Ono et al., 2006). An association between aberrant PEG10 expression and some cancers has also been reported (Tsou et al., 2003; Li et al., 2006).

Relatively little work has been done to explore the molecular functions of PEG10. Functional analysis is complicated by the presence of a second open reading frame (ORF) that is synthesised as part of the PEG10 protein in only a fraction of translational events (Shigemoto et al., 2001; Lux et al., 2005). The second open reading frame includes a predicted aspartyl protease (AP), that might cleave the PEG10 polypeptide at several positions in vivo. This project focuses on the part of the PEG10 protein that contains the AP domain. APs will first be reviewed as an important and intensively studied class of enzymes, and our present understanding of PEG10 expression and function will then be discussed.

1.1 Aspartyl proteases

APs are a common family of endopeptidases, found in all three kingdoms of life and involved in diverse biological processes. The best-characterised class of APs, clan AA,
are soluble enzymes that can be divided into two broad groups: eukaryotic proteins such as pepsin, and retroviral-type APs, the most studied example of which is from HIV (Rawlings et al., 2012). Enzymes from both groups contain a DTG (Asp-Thr-Gly) or DSG (Asp-Ser-Gly) active-site triad. Eukaryotic APs have two domains which interact to form two halves of a bilobed enzyme; by contrast, retroviral-type APs are single-domain proteins, weakly homologous to the N-terminal lobe of the eukaryotic enzymes, and require homodimerisation for stability and activity (Toh et al., 1985; Rawlings and Barrett, 2004). Despite such differences, the structure of the active site cleft is highly conserved between eukaryotic and retroviral-type APs (Wlodawer and Vondrasek, 1998).

Clan AA APs contain two catalytic aspartic acid residues in close proximity at the interface of two β-barrel domains. The carboxyl moieties of the two catalytic aspartic acid residues, one protonated and the other anionic, activate a water molecule which is required to cleave the scissile peptide bond (Figure 1.1). The exact mechanism of catalysis remains in dispute, however (Brik and Wong, 2003).

Most eukaryotic, bilobed clan AA proteases have a low pH optimum. For enzymes such as pepsin, which is active in gastric fluid (pH ∼2), or cathepsin D, a lysosomal protein (pH ∼5), the active-site triad is followed by a serine or threonine residue, which by hydrogen bonding lowers the pKₐ of the functional carboxylic acid group required for activation of the catalytic water. An exception, renin, notably has an alanine three amino acids downstream of the catalytic aspartic acid of the C-terminal lobe and is functional at physiological pH. An alanine residue in the equivalent position is found in most retroviral APs, and substitution of this residue with serine results in diminished enzymatic activity at higher pH (Ido et al., 1991).

### 1.1.1 Broad-range inhibitors of aspartyl proteases

A broad-range inhibitor of APs, termed pepstatin, was isolated from *Ascomycetes* ssp. by Umezawa and Aoyagi (1970). Independently, Fukumara et al. (1971) purified a very similar compound, differing from pepstatin only by N-acylation, from *Streptomyces naniwaensis*. Pepstatin and derivatives are largely hydrophobic peptides containing an unusual β-amino acid, statine, and are capable of inhibiting soluble bilobed APs.

---

¹For the purposes of this document, ‘retroviral-type AP’ refers to all clan AA proteases excluding members of family A1, as catalogued in the MEROPS database as of January 2013 [http://merops.sanger.ac.uk/] This grouping includes all known and predicted dimeric aspartyl proteases. Clan AA, family A1 enzymes are referred to as ‘eukaryotic APs’.

2
monomeric retroviral-type APs, as well as distantly related membrane APs in some cases (Zhang et al., 2001). Nonetheless, several well-studied clan AA APs, such as human renin, are largely insensitive to pepstatin inhibition (Fisher and Hollenberg, 2005; Bernard et al., 2005). Structural analyses revealed that the hydroxyl group of the central statine in pepstatin interacts directly with the catalytic aspartic acids of the AP, thereby excluding a water molecule usually found in the catalytic cleft and mimicking a tetrahedral transition state (Fitzgerald, 1990; Fujinaga and Cherniaia, 2008; see also the central panel of Figure 1.1). Although the pepstatins are not used clinically, the concept of a transition-state mimic embedded within a peptide or peptide-like molecule has been important for subsequent drug discovery.

### 1.1.2 Aspartyl proteases in retroelements

Retroelements are genomic parasites, including retroviruses and retrotransposons, that replicate via an RNA intermediate (Hansen and Heslop-Harrison, 2004). All known LTR-retrotransposons and retroviruses contain an AP domain, usually in the Polyprotein (Pol) gene that is synthesised as an extension of Group-specific antigen (Gag). In most such elements, Pol expression is dependent upon a ribosomal recoding event that occurs only in a fraction of translational passages. Thus, both Gag and Gag-Pol translation products are synthesised from a single mRNA. Gag and Gag-Pol then form spherical nascent virus or virus-like particles in the cytoplasm, which can either
bud from the membrane (in the case of retroviruses) or mature and translocate to the nucleus (in the case of LTR-retrotransposons). The retroelement AP domain, when associated as a homodimer within the nascent virus or virus-like particle, cleaves at various positions along the precursor Gag-Pol polypeptide to release mature structural proteins (e.g. matrix and capsid) or enzymes that are required for replication of the element (e.g. reverse transcriptase and integrase; Figure 1.2). In addition to cleaving at domain junctions within precursor proteins, the HIV AP also cleaves and inactivates itself in a process called autoproteolysis (Rosé et al., 1993).

In HIV-1, the most well-studied example of retroviral-type AP processing, cleavage occurs along the Gag-Pol protein in an ordered manner around the time that the virus buds from the plasma membrane (Erickson-Viitanen et al., 1989; Pettit et al., 2005). Morphological changes in the nascent virus, such as capsid condensation (observed by electron microscopy) track with ordered cleavage events (Wiegers et al., 1998). Ordered or semi-ordered processing of Gag-Pol in LTR-retrotransposons has also been reported (Merkulov et al., 2001).

An active AP is essential for replication of retroviruses and LTR-retrotransposons, evident from both the clinical effectiveness of protease inhibitors (PIs) as an anti-HIV therapy, and from in vitro studies in which the catalytic aspartic acid residue of the AP domain was substituted with loss of infectivity (Kohl et al., 1988; Wensing et al., 2010; Kirchner and Sandmeyer, 1993). Interest in extending the use of PIs to a wider range of infectious retroviruses has promoted research into other retroviral-type APs (Li et al., 2000, 2011).

1.1.2.1 Sequence and structure of retroviral-type aspartyl proteases

Retroviral APs have been the subject of intense study, largely because of drug development efforts, leading to a wealth of available structural data for this class of enzymes. For example, more than 400 structures of the HIV AP have been deposited in the protein data bank, and there are structural models available for the AP domains of seven retroviruses. As indexed by SCOP and Astral [http://scop.berkeley.edu/] and the MEROPS database [http://merops.sanger.ac.uk/] as of February 2013.

Each monomer of a retroviral-type AP is approximately 100–130 amino acids and forms a (6,10) β-barrel with a hydrophobic core (Figure 1.3 and Murzin et al., 1995). Secondary structure is dominated by loops and β-sheet, with only a short α-helical segment present near the C-terminus in most retroviral-type APs. The active site cleft,
Figure 1.2: The Gag-Pol fusion protein and diagram of HIV-1 with annotated protease cleavage positions. a The fusion Gag-Pol polypeptide. Dotted lines indicate sites of cleavage. MA (matrix), CA (capsid) and NC (nucleocapsid) are structural proteins; AP (aspartyl protease), RT (reverse transcriptase), RNase (RNase H), and IN (integrase) are enzymes. Based on map provided by the Los Alamos HIV sequence database [http://www.hiv.lanl.gov/]. b Diagram of the mature HIV-1 particle. Matrix, capsid, and nucleocapsid are shown in blue, green, and purple, respectively. Orange circles represent enzymes derived from the Gag-Pol precursor. Grey circles represent other cellular or HIV-derived proteins such as cyclophilin A, Vif, and Nef. Derived from a diagram provided by L. Henderson and L. Arthur (National Cancer Institute, USA) [https://www.aidsreagent.org/].
Figure 1.3: **Structural features of the HIV-1 aspartyl protease.** Individual subunits are coloured orange or green, and the inhibitor saquinavir is shown in dark blue.  
I The active site. Catalytic aspartic acid residues are shown as sticks.  
II Dynamic ‘flap’ structures, shown in the closed state, hold the inhibitor in the active site.  
III A conserved φ-φ-Gly (where φ is Leu, Ile, or Val) motif.  
IV The interleaved β-sheet platform. ‘N’ and ‘C’ denote N- and C- termini, respectively. Image derived from PDB entry 3OXC.
Figure 1.4: **Conservation of Pol domains among diverse Ty3/Gypsy LTR-retrotransposons.** Blue shading indicates at least 80% identical residues in a position; grey shading indicates aligned amino acids. Alignment begins one residue prior to the active-site catalytic triad of the aspartyl protease. Sequences, which are approximately 750 amino acids each, were obtained from the GyDB [http://gydb.org], and alignment, beginning one residue before the active site, was performed with MUSCLE. Plot generated with TeXshade.

running along part of the dimer interface, is stabilised by a network of intermolecular hydrogen bonds, termed the ‘fireman’s grip’, between side-chain and backbone atoms of the DTG or DSG active-site triad (Figure 1.3 and Dunn *et al.*, 2002). The active site is covered by ‘flaps’, β-hairpins that hold the substrate in place and contribute to substrate binding (Figure 1.3 II). In simulations by molecular dynamics of the HIV-1 AP apoenzyme, supported by NMR studies, the flaps spontaneously switch between closed, open, and semi-open conformations, thereby controlling access of the substrate to the active site (Hornak *et al.*, 2006). Along with the active site, a conserved φ-φ-Gly motif (where φ is Leu, Ile, or Val) forms one of the few motifs conserved in the primary, as well as secondary and tertiary, structure of retroviral-type APs (Figure 1.3 III and Llorens *et al.*, 2008). A β-sheet platform forms an important part of the dimer interface (Figure 1.3 IV, discussed subsequently).

There is a large degree of primary sequence variation among retroviral-type APs, despite significant conservation of tertiary structure (Sirkis *et al.*, 2006). Indeed, aside from the active-site triad (DTG or DSG) and the conserved φ-φ-Gly motif, sequence analyses across retroviral-type APs yield little information. Conserved motifs within clades, for example in the flaps, are significantly more useful, however (Llorens *et al.*, 2008). Compared to the other Pol domains in LTR-retrotransposons, AP is the most poorly conserved at the level of primary sequence (Figure 1.4).
1.1.2.2 The termini of aspartyl proteases are functionally important despite sequence and structural divergence

Remarkably, in the HIV-1 AP the majority of the stabilisation energy is contributed by the dimer interface. In particular, the four-stranded $\beta$-sheet composed of residues at the extreme N- and C-termini of the enzyme are 'hotspots' that contribute nearly 75% of the stabilisation energy for the dimeric state (Todd et al., 1998 and Figure 1.3 IV). Structural comparisons of diverse clan AA APs, spanning bilobed eukaryotic and retroviral-type enzymes, have shown that an intermolecular $\beta$-sheet motif in this position is ubiquitous in this class of enzymes. Nonetheless, the number of strands and the positions of residues contributing to the $\beta$-sheet platform are divergent, even between retroviral-type APs (Figure 1.5 and Li et al., 2011). In some cases, such as the AP domains of xenotropic MLV-related virus (XMRV) and the DNA-damage inducible 1 (Ddi1) protein, only C-terminal residues contribute to this platform. The $\beta$-sheet platform of clan AA APs is thus, paradoxically, both a functionally significant and poorly conserved structure.

Only one structure of a monomeric retroviral-type AP has been reported, from the Mason-Pfizer monkey virus (Gilski et al., 2011). Both the N- and C-termini in this structure are disordered. In vivo, cysteines at the Mason-Pfizer monkey virus N- and C-termini contribute to an intramolecular disulphide bridge, stabilising the folded conformation (Zábranská et al., 2007). (The N-terminal cysteine was substituted for alanine in the solved monomeric structure.) These observations further argue in favour of a critical role for the termini in the dimerisation of retroviral-type APs.

1.1.2.3 Redox regulation of retroviral aspartyl proteases in mammalian cells

Retroviral APs become active only at certain points of the retroviral lifecycle. Indeed, premature activation of the HIV-1 AP results in disrupted particle formation, reduced virus infectivity, and cellular toxicity (Kräusslich, 1991). The HIV AP is regulated by a ‘redox switch’ in vivo: a conserved thiol-containing residue at the C-terminus (C95 in HIV-1 and M95 in HIV-2) can be glutathionylated, which greatly reduces proteolytic activity (Davis et al., 1996). The cellular enzyme glutaredoxin-1 is incorporated into HIV-1 virions and reverses this glutathionylation, thereby activating the protease once a contained nascent virion has been formed (Davis et al., 1997). Methionine sulphoxide reductase may play an analogous role in HIV-1 (Davis et al., 2000). Recently, terminal cysteine oxidation (e.g. by glutathionylation) was shown to inhibit, in particular, the initial cleavage made by the HIV-1 AP, although the addition of glutathione to a differ-
Figure 1.5: **Termini from homologous clan AA aspartyl protease domains have diverse morphologies.** Each subunit is coloured either green or orange. Surface-exposed terminal β-sheet structures were visualised from outside, rather than within, the structure of each enzyme (as in Figure 1.3). a HIV-1 AP (PDB identifier KB62) b S. cerevisiae Ddi1 AP domain (2I1A) c XMRV AP (NR6) d Human pepsin 3A (1QRP). Figure inspired by Li et al. (2011).
different cysteine residue caused instead an increase in $k_{cat}$ (Davis et al., 1996; Daniels et al., 2010). HIV is not the only retrovirus to utilise such a ‘redox switch’: Mason-Pfizer monkey virus, a betaretrovirus, and human T-cell leukemia virus-1, a deltaretrovirus, both require a reducing environment for AP activation (Parker and Hunter, 2001; Davis et al., 2003). Analysis of published structures combined with sequence alignments indicated that most retroviral-type APs have a sulphur-containing residue at or near the (predicted) dimer interface, with the notable exception of the gammaretroviruses (e.g. XMRV; Davis et al. 2003). Extension of such a regulatory mechanism to LTR-retrotransposons has not been explored, although Sirkis et al. (2006), reporting the structure of a cellular retroviral-like AP domain, noted that two cysteines were within disulphide-forming range on the $\beta$-sheet platform at the dimer interface.

### 1.1.3 Substrate specificity of retroviral-type aspartyl proteases

The amino acid positions of protease substrates are labelled, from N- to C- terminus, PX–P1 and P1’–PX’, in which the scissile peptide bond is located between the P1 and P1’ positions and X is the number of residues distant to the cleavage site. The sequence specificity of retroviral-type APs cannot be reduced to a consensus motif. Rather, a conserved secondary structure—an extended $\beta$ conformation—as well as a network of enzyme–substrate hydrogen bonds and P1–P3 contacts forming toroidal electron density, are common among native substrates of the HIV-1 AP (Prabu-Jeyabalan et al., 2002). Nonetheless, at the primary sequence level cleavage sites are usually divided into two classes: type 1 sites with a preference for Phe or Tyr at P1 and, unusually, Pro at P1’, or the largely hydrophobic type 2 sites (Tözsér, 2010). Despite such a classification, prediction of cleavage sites remains an active area of research, requiring sophisticated computation methods (Niu et al., 2009 and references therein). Several lines of evidence support a substantial role for context in the determination of retroviral AP substrate specificity, at least in HIV-1:

1. Figure 1.6 is a sequence logo automatically generated from 1049 observed cleavage events. Little information can be obtained from primary sequence alone, to the point that in most positions the logo is impossible to interpret. If primary sequence was the major determinant of substrate specificity, it would be expected that certain residues would be observed at a high frequency in some positions, represented by taller letters in Figure 1.6.

2. Mutagenesis studies, in which cleavage sites were placed in non-native contexts,
demonstrated large changes in cleavage efficiency and kinetics for some cleaved sequences in HIV-1 (Tritch and Cheng, 1991; Lee et al., 2012).

3. Molecular dynamics simulations predict that many randomly selected portions of the HIV-1 Gag-Pol precursor have a greater affinity for the HIV-1 AP domains than known cleavage sites. One interpretation of these results is that tertiary (in addition to primary) structure is a major determinant of substrate specificity of the HIV-1 AP (Perez et al., 2010).

Such a reliance on context surrounding the substrate is consistent with the ordered sequence of cleavage events that has been described for retroviruses (Section 1.1.2). In this model, retroviral APs act, in their native context, as part of a ‘molecular machine’ encompassing the Gag and Gag-Pol proteins that have formed into a nascent viral particle. Proteolysis events are inferred to be part of a choreographed cascade that produces a particular morphological shift. Access to some cleavage sites is likely occluded.

In addition to native cleavage sites within viral proteins, there have been numerous reports of cellular substrates of the HIV-1 AP. For example, Strack et al. (1996) showed that recombinant HIV-1 AP cuts the anti-apoptotic protein bcl-2. This was rationalised as a strategy for the virus to activate its own transcription, as bcl-2 is upstream of NF-κB, which when cleaved in turn activates HIV-1 transcription. Whether this effect occurs in actual viral infections, however, was not assessed. Ventoso et al. (2001) showed that HIV-1 AP cleaves the translation initiation factor eIF4GI in three separate positions. Such cleavage is ablated by a specific inhibitor of HIV-1 AP, and can be observed in vitro, both in cultured cells expressing the HIV AP domain alone as well as in HIV-infected cells. Cleavage of eIF4GI specifically decreased cap-dependent initiation of protein synthesis but enhanced translation from HIV genomic RNA, which occurs via an internal ribosome entry site (Ventoso et al., 2001).
1.1.4 Inhibitors of retroviral-type aspartyl proteases

Soon after the discovery of HIV, it was noted that pepstatin A prevented proteolytic processing of Gag and Gag-Pol proteins, indicating that HIV contained an aspartyl-type protease (Seelmeier et al., 1988). Pepstatin A, however, has pharmacologically unfavourable attributes and too broad a substrate specificity for clinical use (Eder et al., 2007). With the urgent need to address the HIV epidemic and the paucity of anti-retroviral drugs, intense drug discovery efforts were focused on the HIV AP in the late 1980s and early 1990s. Rational drug design approaches incorporated a non-hydrolysable transition-state analogue of the scissile peptide bond and the unusual ability of retroviral-type APs to cleave between Phe and Pro. Early work focused on symmetrical inhibitors, which matched the symmetry of the HIV-1 AP observed in structural studies. This approach was abandoned, however, because of problems with oral bioavailability, and later it was shown that symmetrical inhibitors can in fact bind asymmetrically to HIV-1 AP (Wlodawer and Vondrasek, 1998; Dreyer et al., 1993). Eventually, pharmacologically favourable compounds that inhibited the HIV AP specifically, with less than 50% inhibition against the cellular bilobed APs, were developed (Roberts et al., 1990). Saquinavir, the first clinically used anti-HIV PI, is a peptidomimetic small molecule drug that contains a hydroxyethylamine moiety replacing the scissile peptide bond between Phe- and Pro- like moieties (Figure 1.7 and Wlodawer and Vondrasek, 1998). It was approved by the FDA for clinical use in 1995. Subsequently, nine inhibitors of this class have been approved for use in the clinic, most of which are based on similar principles to saquinavir (Wensing et al., 2010).

Used in combination with other anti-retroviral drugs, HIV PIs have played an important part in containing the worldwide AIDS epidemic in developed countries. Despite such successes, PIs are known to cause severe side-effects in some patients, including lipodystrophy, insulin resistance, and various cutaneous conditions (Carr et al., 1998; García-Silva et al., 2000). When PIs were first being used for the treatment of HIV, it was noted that Kaposi’s sarcoma, an angiogenic tumour common in AIDS patients, would rapidly regress upon the commencement of anti-retroviral therapy. In some cases, this occurred more quickly than T cell counts recovered, indicating that the regression was not mediated solely by immune reconstitution (Lebbé et al., 1998). Additionally, oral candidiasis—an opportunistic infection by the genus of fungus Candida that affects immunocompromised patients—can be treated by some HIV PIs alone. This antifungal activity is dependent on the inhibition of secreted APs that act as virulence factors for the invasive Candida (reviewed by Tacconelli et al., 2004).
In addition to side-effects, drug resistance is a significant problem for the treatment of HIV. Resistance to all PIs can develop over the course of a patient’s infection, as error-prone reverse transcription during viral replication leads to a diverse population of HIV within each individual (Wensing et al., 2010). Development of such resistance has stimulated further research into the development of new PIs, some of which have non-overlapping resistance profiles with established drugs (Poveda et al., 2007).

1.1.5 Retroviral-type aspartyl proteases in eukaryotic genomes

Independent of AP domains derived directly from retroelements, several retroviral-like AP domains are found within cellular genes in eukaryotes.

*DNA-damage inducible 1 (DDI1)*, present in all eukaryotic genomes yet examined, encodes a central retroviral-like AP domain flanked by ubiquitin-like and ubiquitin-binding domains (Krylov and Koonin, 2001). Ddi1 (among a range of other cellular functions) transports ubiquitylated substrates to the proteasome for degradation (Elsasser and Finley, 2005). The retroviral-type AP domain mediates the homodimerisation of the multidomain Ddi1 protein (Gabriely et al., 2008). The crystal structure of the Ddi1 retroviral-type AP domain revealed marked structural similarity to the active site of the HIV-1 protease, together with a similar overall fold (Sirkis et al., 2006).

Although no proteolytic substrate of DDI1 has yet been identified, mutation of the putative catalytic aspartic acid residue results in a unique phenotype in yeast (Gabriely et al., 2008). In a transgenic *Saccharomyces cerevisiae* system, expression of Ddi1 was negatively correlated with protein secretion (White et al., 2011a). For yeast expressing only a *Leishmania* Ddi1, protein secretion was inhibited by the HIV-PI nelfinavir, resulting in the same phenotype as an active-site variant of *Leishmania* Ddi1 (White et al., 2011b). Collectively, these data imply strongly that Ddi1 contains an active protease, and that the Ddi1 of at least some eukaryotes is inhibited by HIV PIs. Such an effect was proposed to partially account for the apparent anti-parasite activity of HIV PIs (reviewed by Pozio and Morales (2005)). A paralogue of *DDI1, DDI2*, is also present in most eukaryotic genomes.
Skin-Associated Aspartic Protease (SASPase, also known as ASPRV1), identified as a mammalian protein that contains a retroviral-type AP domain, is expressed specifically in the epidermis (Bernard et al., 2005). SASPase is autocatalytically active and can be inhibited by the HIV PI indinavir, a possible mechanism for the undesirable cutaneous side-effects of PIs (Bernard et al., 2005; García-Silva et al., 2000). Neuronal interacting factor X 1 (NIX-1, also known as NIRP2) is a mammalian protein, expressed only in the central nervous system, that binds to retinoic acid and thyroid hormone receptors (Greiner et al., 2000). From sequence similarity, NIX-1 is likely to be an active retroviral-type AP domain, although little work has been done on this gene (Llorens et al., 2008). Two predicted paralogues of NIX-1 are additionally present in mammalian genomes.

In addition to endogenous retroviruses, LTR-retrotransposons form a significant portion of what was formerly known as ‘junk DNA’ in vertebrate and mammalian genomes. For example, AP-containing autonomous retroelements account for approximately 8% of the human genome, although the vast majority of these sequences are degenerated (Cordaux and Batzer, 2009). The MEROPS database of proteases predicts twenty-two instances of retroviral-type APs from endogenous retroviral genes or pseudogenes in the human genome, and this number is probably artificially low due to the poor sequence conservation within this class of enzymes (see Section 1.1.1.1). Because the active-site triad is an important sequence motif of retroviral-type APs, it is likely that proteins that retain the gross structure of retroviral-type APs but are not catalytically active are also present in mammalian genomes, but are not detected by searches reliant upon primary sequence conservation alone.

At least twelve instances of ‘molecular domestication’ of LTR-retrotransposons—a process in which parasitic genomic elements are co-opted by the host for endogenous function—have occurred in mammals (Ono et al., 2011). The protein-coding genes resulting from at least two such events, PEG10 and Retrotransposon-like 1 (RTL1, also known as PEG11), have conserved Pol-like regions with sequence motifs characteristic of retroviral-type APs.

### 1.2 Paternally Expressed Gene 10

Human PEG10 is located adjacent to the Sarcoglycan-ε gene on chromosome 7q21 (Ono et al., 2001; Suzuki et al., 2007). Expression of PEG10 mRNA occurs only from

---

Figure 1.8: **Comparison of the PEG10 domain structure with parasitic retroelements.**

a PEG10  
b A typical chromovirus, a type of LTR-retrotransposon closely related to PEG10  
c a simple retrovirus. Figure not to scale. Asterisks indicate –1 frameshift elements.  
*Gag*, group-specific antigen; *Pol*, Polyprotein; *Env*, envelope;  
MHR, major homology region, ZF, zinc finger; AP, aspartyl protease; RT, reverse transcriptase;  
RH, RNase H; IN, integrase; CHR, chromodomain.

the paternally inherited allele; transcription from the maternal allele is prevented by  
DNA methylation, a phenomenon termed genomic imprinting. PEG10 mRNA contains two open reading frames (ORFs) separated by a recoding element, so that one  
mRNA essentially codes for two proteins (Shigemoto et al., 2001). PEG10 ORF1 contains  
Gag-like sequence elements. The second ORF of PEG10, which contains sequence motifs of an AP domain that are conserved across therian mammals, is expressed in human tissues via a ribosomal –1 frameshifting event—a mechanism commonly used by LTR-retroelements (Figure 1.8). This is the only described example of a cellular gene using –1 frameshifting in the animal kingdom.

**1.2.1 LTR-retrotransposons, retroviruses, and PEG10**

LTR-retrotransposons and retroviruses are a subgroup of autonomous retroelements,  
genomic parasites that replicate via an RNA intermediate (Craig et al., 2001). Retroviruses are derivatives of LTR-retrotransposons, specifically the Ty3/Gypsy clade, which  
mostly contain a –1 frameshift element separating the Gag and Pol ORFs (Llorens et al., 2011). A stop signal in frame with the Gag ORF is located soon after the –1 frameshift element. If frameshifting does not occur, the ribosome encounters the proximal stop signal and translation ceases: in this scenario, only the first ORF, encoding
$Gag$, is synthesised. When $-1$ frameshifting occurs the ribosome slips back one nucleotide on the mRNA before resuming translation in the new reading frame. The stop signal, now out of frame, is bypassed and a long Gag-Pol fusion protein is translated.

Frameshifting is typically a minor event during ribosomal passages through the mRNA—in the order of 5% for most retroviruses—and as a result Gag proteins are more abundant than Gag-Pol fusion products. By contrast, human PEG10 mRNA promotes high-efficiency frameshifting, the highest reported for any gene. Typically more than 70% of translational passages resulting in the translation of the PEG10 Pol-like ORF2, although this value appears to be dynamic throughout placental development (Clark et al., 2007).

PEG10 is highly homologous to Sushi-ichi, a Ty3/Gypsy chromovirus-like element found in the pufferfish Takafugu rubripes (Butler et al., 2001). Although LTR-retrotransposons are usually inherited vertically, the broad phylogenetic tree of the Ty3/Gypsy group does not recapitulate eukaryotic phylogeny, implying that infrequent horizontal transmission occurs (Llorens et al., 2009). Initial insertion of such an element that eventually has become PEG10 probably occurred after monotremes split from therian mammals (Suzuki et al., 2007).

The PEG10 ORF1 has retained significant similarity to retroelement Gag domains, and features a conserved major homology region (involved in Gag oligomerisation in HIV) and CCHC zinc finger (involved in RNA binding in retroelements and possibly DNA binding in PEG10; Clark et al., 2007; Provitera et al., 2001; Steplewski et al., 1998). By contrast, the PEG10 ORF2 has lost most of the enzymatic domains of Pol required for retrotransposition other than the AP domain. The reverse transcriptase domain of PEG10, although lacking critical residues required for enzymatic activity, is nonetheless somewhat conserved and may still form an ordered fold (unpublished observations). The significance of this with regards to the predicted AP domain is unclear.

1.2.2 Expression and function of PEG10

The murine PEG10 homologue, Peg10, is required for normal placental development and function (Ono et al., 2006). Peg10-knockout mice, in which expression from the paternal allele is abolished, die in utero by 10.5 d.p.c. (days post coitum) due to a malformed placenta. In particular, the labyrinth layer is poorly developed and the central spongiotrophoblast cells are completely absent. If the Peg10-knockout embryos are given a non-transgenic placental graft (i.e. a wild-type placenta), the mice survive
through birth, are fertile, and have no obvious phenotypic abnormalities in adulthood (Ono et al., 2006). Thus, although Peg10 is essential for the development of the placenta in mice, it is dispensable for other cellular functions.

In the mouse placenta, Peg10 expression begins at 9.5 d.p.c. (of a ~20-day pregnancy; Ono et al. 2006). Clark et al. (2007) showed that Peg10 expression was absent at 8.5 d.p.c., began at 9.5 d.p.c., and then remained constant throughout subsequent stages of pregnancy. In the human placenta, PEG10 mRNA is present as early as week seven of gestation, although becomes much more pronounced around week eleven (Smallwood et al., 2003). In both cases, the increase in expression approximately coincides with the end of the hypoxic period of placental development.

Despite the definitive work done in mouse models, a report describing several human patients with large, hemizygous deletions encompassing PEG10 and several surrounding genes has been published. Two of these patients have a deletion of the paternal allele, and suffer from a condition reflecting loss of expression of a different paternally expressed gene within the PEG10 imprinted region. The authors thereby concluded that deficiency of PEG10 is not obligately lethal in humans, although no account of patient placental morphology was included in the study and expression of PEG10 was not assessed (Asmus et al., 2007). Further analysis, preferentially involving placental tissue, is required to determine the role of PEG10 in human development. Despite such apparent dispensability, expression of PEG10 in the human placenta is perturbed in pregnancies with preeclampsia and intrauterine growth restriction, implying that PEG10 plays a role in human placental development and hinting at an involvement in pregnancy-associated disease processes (Diplas et al., 2009; Chen et al., 2011).

Expression of PEG10 enhances resistance to apoptosis, a phenomenon especially noted in primary liver cancer (hepatocellular carcinoma) cell lines, and correlates with progression of cancer (Okabe et al., 2003; Ip et al., 2007; Tsuji et al., 2010). Consistent with this molecular function, Peg10 is also involved in liver regeneration after partial resection (Tsou et al., 2003). Cancers of the endometrium, breast, prostate, gallbladder and B-cells have also been reported to express PEG10, despite the absence of PEG10 in these tissues normally (van der Horst et al., 2012; Li et al., 2006; Liu et al., 2011; Kainz et al., 2007). In support of a link between PEG10 and cancer, the oncogene MYC, as well as members of the E2F family of transcription factors and signalling through oestrogen receptor β, influence the transcriptional control of PEG10 (Li et al., 2006; Wang et al., 2008; Zhao et al., 2009).

Recently, four high-throughput studies using diverse cell lines have reported that
PEG10 ORF2 is ubiquitylated in two positions (Kim et al., 2011; Lee et al., 2011a; Shi et al., 2011; Wagner et al., 2011), although one paper following a similar methodology did not detect ubiquitylation of PEG10 (Danielsen et al., 2011). Regulation of PEG10 at the post-translational, rather than or in addition to the post-transcriptional, level may explain the apparent discrepancy between expression of PEG10 mRNA and PEG10 protein (Lux et al., 2005; Clark et al., 2007; Su et al., 2004).

The subcellular localisation of PEG10 has been variously reported as cytoplasmic, perinuclear, and both cytoplasmic and nuclear (Tsou et al. (2003); Lux et al. (2005); Okabe et al. (2003); respectively). Notably, Tsou et al. (2003) and Lux et al. (2005), in reporting cytoplasmic localisation of PEG10, both used epitopes at the N-terminus of ORF1 for immunofluorescence. By contrast, Okabe et al. (2003) used a polyclonal anti-body raised against the whole of ORF1. One investigator noted that portions of mouse Peg10 ORF2 appeared to localise to either the nucleus or cytoplasm depending on the presence of a binding partner (Długaszewska, 2005).

Lux et al. (2005) identified Activin-like kinase 1, a cell-surface receptor of the transforming growth factor (TGF)-β1 family of signalling molecules, as a binding partner for PEG10, although the functional significance of this interaction was not explored experimentally. The seven in absentia homologue 1 and 2 (SIAH1/2) gene products were also shown to interact with PEG10 in cultured cells, and PEG10 protects SNU423 cells (a hepatocellular carcinoma cell line) from SIAH1-mediated apoptosis (Okabe et al., 2003). SIAH1/2 are pro-apoptotic E3 ubiquitin ligases involved in the p53 pathway, and it has been proposed that SIAH2 modulates the cellular response to hypoxia, a condition common between the developing placenta and some cancers that express PEG10 (Matsuzawa and Reed, 2001; Nakayama et al., 2004). In searching for binding partners, these studies only examined PEG10-ORF1. An association between Hoxd13 and Peg10-ORF2, assessed by yeast two-hybrid, immunoprecipitation, and co-localisation experiments, has been noted in mouse limb buds (Długaszewska, 2005).

The mature PEG10 mRNA is 6.6 kb, two thirds of which is untranslated and 3′ to the open reading frames. This long, relatively conserved 3′ UTR may regulate the translation of PEG10 (Saurat, 2009). Lux et al. (2010) proposed that translation of PEG10 may begin at an upstream, non-canonical start codon, but did not present definitive evidence for an N-terminal extension of the PEG10 protein. A modified ribosome profiling method coupled with RNA sequencing supported the dominance of an upstream, in-frame start codon, along with the presence of small translated upstream ORFs (Ingolia et al., 2011, supplementary material).

The conserved CCHC-type zinc-finger motif in PEG10 ORF1 has been shown, by
electrophoretic mobility shift assays, to bind part of the myelin basic protein promoter DNA. No evidence that this occurs \textit{in vivo} was presented, however (Steplewski \textit{et al.}, 1998). The disputed subcellular localisation of PEG10, discussed above, in aggregate argues against PEG10 binding DNA in the cell. High-throughput studies have implicated PEG10 in binding to polyadenylated RNA, a function consistent with the ancestral genomic-RNA–binding function of the PEG10 zinc finger motif (Castello \textit{et al.}, 2012, supplementary material).

### 1.2.3 The predicted aspartyl protease domain of \textit{PEG10}

In transfection studies, Lux \textit{et al.} (2005) noted that small fragments of PEG10 were detectable in cultured mammalian cells, dependent of the inclusion of the AP-containing \textit{PEG10} ORF2. Corroborating these results, Clark \textit{et al.} (2007) used transfected cultured cells combined with site-directed mutagenesis to substitute the predicted catalytic aspartic acid of the \textit{PEG10} AP domain. Low-molecular-weight PEG10 fragments were detected in lysates of cells expressing the native PEG10 protein, but absent in cells expressing the AP-variant PEG10 (Figure 1.9). Collectively, these data strongly imply that \textit{PEG10} contains an active AP domain which cleaves at several points along the PEG10 polypeptide, although proteolysis \textit{in vitro} has not yet been observed. The identity of the cleavage fragments, or the effect of proteolysis on PEG10 function, are yet to be examined.
1.3 Aims

APs are an abundant class of enzymes in retroelements and eukaryotes that fulfill important diverse functions. Substantial work on retroviral APs has lead to the development of clinically effective PIs for the treatment of HIV infections, and generated a wealth of structural and biochemical data for a limited subgroup of these enzymes. For retroviral-type APs found in LTR-retrotransposons or endogenous cellular genes like PEG10, however, relatively little is known.

The aims of this study were to investigate the putative PEG10 AP domain by

1. Determining the boundaries of the AP domain within the full-length PEG10 protein.

2. Cloning, expressing, and purifying the putative PEG10 AP domain.

3. Characterising the expressed PEG10 AP domain, attempting the development of an in vitro activity assay, and obtaining an inhibition profile with special emphasis on inhibitors of the HIV-1 AP currently used in humans.
Chapter 2

Methods

Work with genetically modified *E. coli* was conducted with approval from the New Zealand ERMA (now EPA) under the codes GMD0000067 and GMD00181. Work with *E. coli* was performed in ‘physical containment 1’ conditions.

See Appendix I for a complete list of materials and solutions used in this study.

2.1 Bioinformatic analysis of the PEG10 aspartyl protease domain

PSI-BLAST was used to search for homologues of PEG10 in non-mammalian species. The predicted PEG10 AP domain, boundaries corresponding to clone 6 in this study, was submitted to PSI-BLAST on the NCBI server [blast.ncbi.nlm.nih.gov] using default settings. The non-redundant protein sequences database, excluding mammals (taxonomy ID 40674) was searched using these parameters. Results from synthetic constructs were ignored during analysis. These results were confirmed by use of profile hidden markov models.

The HMMER3 server [http://hmmer.janelia.org] was used to search for PEG10 homologues using a hidden Markov model (HMM) framework (Finn *et al.*, 2011). Default parameters and the ‘nr’ (non-redundant) database were used. PEG10 isoform 3 (NP_001165908.1) was used as a query. The top sequence hit for each non-mammalian species was noted, and searched against known LTR-retroelements using the BLAST function built into GyDB [http://gydb.org/index.php/Blast]. The blastp program against the ‘cores’ database was used for these searches. Similarity to documented retroelements was assessed by eye and noted. Unless otherwise stated, all sequence identifiers refer to nucleotide, protein, or predicted protein se-
quences housed by the NCBI.

2.2 Analysis of mammalian PEG10 sequences

The BLAT function of the UCSC genome browser was used to find mammalian homologues of PEG10 which were not annotated in the NCBI database. An online in silico translation tool [web.expasy.org/translate/] was used to obtain protein sequence from the gathered DNA for each species. In the case of elephant and cat, frameshift mutations were found near the frameshift signal i.e. sequences were truncated. Because the amino acid sequence was conserved in each frame, nucleotides were artificially inserted or deleted to bring these sequences into agreement with the consensus sequence.

2.3 Sequence alignment

Sequences were aligned using the MUSCLE sequence alignment software v3.8.31 unless stated otherwise (Edgar, 2004). Default options were used in all cases. The EMBOSS tool ‘seqret’ was used to trim sequences post-alignment, and the TeXshade package was used for typesetting (Beitz, 2000).

2.4 DNA manipulation

2.4.1 Preparation of competent E. coli cells

To prepare competent E. coli Rosetta 2 (DE3) cells, a single colony from an agar plate was seeded into 5 mL LB broth, 30 μg.mL\(^{-1}\) chloramphenicol, and grown overnight at 37°C, 200 rpm with orbital shaking. Of this starter culture, 1 mL was seeded into 100 mL LB broth, 30 μg.mL\(^{-1}\) chloramphenicol, and grown at 37°C, 200 rpm with shaking until the OD\(_{600}\) reached 0.4–0.6. The culture was then chilled on ice for 15 min, and centrifuged for 10 min at 6 000 \(\times\) g, 4°C. The cell pellet was resuspended in 20 mL sterile, chilled 100 mM CaCl\(_2\), held on ice for 15 min, then centrifuged as before. The supernatant was discarded, and the cell pellet resuspended in 5 mL 100 mM CaCl\(_2\), 2.5 mL 60% (v/v) glycerol. Samples were divided into multiple aliquots, snap-frozen on dry ice, and stored at –80°C.

To prepare competent E. coli DH5α cells, a single colony from an agar plate was seeded into 10 mL LB broth and grown overnight at 37°C, 200 rpm in an orbital shaker.
The next day the starter culture was added to 1 L LB broth and grown at 37°C with shaking until the OD$_{600}$ reached 0.4–0.6. The culture was then chilled on ice for 30 min and centrifuged for 10 min at 6 000 × g, 4°C. Cell pellets were then resuspended for washing in 1 L, 500 mL, 500 mL, and 20 mL sterile, chilled ddH$_2$O in four steps, with the supernatant discarded after pelleting the cells in each case. Cells were finally resuspended in 3 mL sterile 10% (v/v) glycerol. Samples were aliquoted, snap-frozen on dry ice, and stored at –80°C.

2.4.2 Transformation of competent *E. coli* cells

Plasmid DNA (stored at –20°C) was thawed on ice or at room temperature. Competent *E. coli* cells (either DH5α or Rosetta 2 (DE3)), stored at –80°C, were thawed on ice. DNA to be transformed was mixed with 100 µL of competent cells in a 1.5 mL microcentrifuge tube on ice, and allowed to associate for 10 min. Reactions were then heat-shocked in a 42°C water bath for 45 s and promptly returned to the ice for at least 1 min. Reaction volume was then increased to 1 mL with 2 × YT and incubated at 37°C for 45 min at 200 rpm in an orbital shaker. After 45 min, 100 µL was plated onto a pre-warmed agar plate with either 100 µg.mL$^{-1}$ ampicillin (DH5α) or 100 µg.mL$^{-1}$ ampicillin and 30 µg.mL$^{-1}$ chloramphenicol (Rosetta 2 (DE3)). For some transformations into Rosetta 2 (DE3) *E. coli*, no enrichment in 2 × YT was used, and all 100 µL of cells was plated after the heat-shock step. Plates were inverted and left to grow at 37°C overnight (16–18 h for DH5α or 18–20 h for Rosetta 2 (DE3)). See Table 2.4 for details of the *E. coli* strains used in this study.

2.4.3 Plasmid preparations

Single colonies from agar plates were seeded into 5 mL LB, 100 µg.mL$^{-1}$ ampicillin in a 15 mL conical centrifuge tube, and grown overnight (16–18 h) at 37°C, 200 rpm in an orbital shaker. Cultures were pelleted at 6000 × g for 5 min at 4°C. From this step, the instructions from the QIAprep spin miniprep kit (Qiagen GmbH, Germany) were followed. The optional buffer PB wash was not performed. A single elution with 50 µL of ddH$_2$O, incubated for approximately 10 min before centrifugation, was harvested. Final yield was quantified, and purity estimated, with a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, USA).
2.4.4  Restriction endonuclease digests

Buffers and restriction enzymes were sourced from Roche Applied Science, Germany. For the diagnostic digest shown in Figure 3.9, each reaction (20 µL) contained 300 ng of DNA, 1× buffer H, 0.5 µL of each restriction enzyme, and ddH2O. Reactions were incubated at 37°C for 2 h, and the enzymes heat-inactivated at 65°C for 20 min. All 20 µL was loaded onto the agarose gel for analysis.

To prepare linear vector for cloning by restriction digestion (e.g. in Figures 3.12 and 5.1c), each reaction (100 µL) contained 1 µg of DNA, 1× buffer B, 3 µL of Bam HI, and ddH2O. Reactions were incubated at 37°C for 3 h, and the enzyme heat-inactivated at 65°C for 20 min. The linear DNA was stored at –20°C.

2.4.5  Cloning

See Figure 3.10 for a schematic overview of the circular polymerase extension cloning (CPEC) technique used in this study. A complete list of primers used for amplification and cloning can be found in Table 2.1.

2.4.5.1  PCR amplification of inserts for cloning

Polymerase chain reactions (PCRs) used the Phusion II HotStart DNA polymerase and supplied buffer (Thermo Fisher Scientific, USA, under the brand name ‘Finnzyme’). In this system, the manufacturer recommends using 98°C for denaturation, and adding 3°C to the calculated Tm of primers because of the high salt concentration in the supplied buffer. Primers were designed and checked for specificity using the Serial Cloner 2.0 software (Serial Basics, France). The NetPrimer applet (Premier Biosoft, USA) was used to estimate the Tm and the potential for self- or cross-dimerisation of each primer. Where possible, primers were designed to avoid any potential dimers or secondary structures with a predicted free energy of less than –10 kcal/mol.

Each reaction (20 µL) contained 200 µM dNTPs, 500 nM of each primer, 10 ng template DNA, 0.4 U Phusion HotStart II DNA polymerase, 1× Phusion HF buffer, and ddH2O in a 0.2 mL thin-walled microcentrifuge tube. Full-length human PEG10 ORFs in the pcDNA 3.1(+) vector (Clark et al., 2007) or clones produced in this study were used as a template for these reactions. A C-1000 Thermal Cycler, (Bio-Rad Laboratories, Inc., USA) with a heated lid at 105°C was used for all thermal cycling protocols:
The 60°C annealing temperature was changed to 56°C for amplification of his– and his-TEV– inserts. On completion, reactions were stored at –20°C.

### 2.4.5.2 Cloning of inserts into target plasmids

After agarose gel analysis to assess purity and yield of the desired insert, 1 µL of the PCR mixture per 100 ng of cut vector was used for each cloning reaction. For cloning into the his-TEV vector, amplicons were purified using a QIAquick PCR purification kit following the manufacturer’s instructions (Qiagen GmbH, Germany) and then used in a insert:vector 5:1 molar ratio with the linearised vector. Each reaction (20 µL) additionally contained 200 µM dNTPs, 0.4 U Phusion HotStart II DNA polymerase, and 1 × Phusion HF buffer, and ddH₂O. Reactions were cycled as follows:

- **98°C 30 s** Initial denaturation
- **98°C 10 s**
- **60°C 30 s** × 30
- **72°C 15 s**
- **72°C 3 min** Final extension
- **4°C** Until further processing

A final denaturation (98°C for 30 s) with gentle annealing (85°C, 75°C, 55°C, and 37°C, all for 30s) was carried out and the products were digested with 10 U *Dpn I* (New England Biolabs, USA) for 2 h at 37°C. All 20 µL of the reaction product was then transformed into *E. coli* DH5α.

### 2.4.5.3 PCR-screening for inserts

PCR screens were conducted using plasmid-appropriate sequencing primers. Each 10 µL reaction contained 1 × PCR buffer with Mg²⁺, 200 µM dNTPs, 200 nM of each primer, and 0.5 U Expand Taq polymerase (Roche Applied Science, Germany). Colonies from transformation plates were picked with a pipette tip, transferred to a
master plate, and the tip then swirled in a PCR reaction mixture. A colony-free section of a colony-containing plate was ‘picked’ as a negative control, one reaction was spiked with empty vector for another negative control, and where available a plasmid with insert was spiked into another reaction as a positive control. The following thermocycling protocol was used:

\[
\begin{align*}
95^\circ C & \quad 5 \text{ min} & \text{Initial denaturation} \\
95^\circ C & \quad 30 \text{ s} \\
55^\circ C & \quad 30 \text{ s} & \times 5 \\
68^\circ C & \quad 1 \text{ min} \\
68^\circ C & \quad 5 \text{ min} & \text{Final extension} \\
4^\circ C & \quad \text{Until further processing}
\end{align*} \]

All 10 µL of each reaction, mixed with 1 µL of DNA loading buffer, was analysed by agarose gel electrophoresis. PCR-positive colonies were then picked from the master plate that had been still incubating at 37°C for the duration of the PCR screen and analysis, and these were used in cultures for plasmid preparation (Section 2.4.3).

### 2.4.6 Site-directed mutagenesis

Table 2.2 lists all primers used for site-directed mutagenesis (SDM) in this study. The T\text{m} of the ‘rightward’ (R) primers is the annealing temperature used for initial amplification rounds. The T\text{m} of the ‘leftward’ (L) primers is the annealing temperature used for the final round of amplification.

#### 2.4.6.1 Inserting a TEV-protease cleavage site into the pMAL-c2 vector

See Figure 3.7 for a schematic of the process used to replace the existing Factor Xa cleavage site with TEV-protease cleavage site in the pMAL-c2 vector. Primers pMAL-TEV F and pMAL-TEV R were used to amplify pMAL-c2 in a long-template PCR. Each reaction (50 µL) contained 350 µM dNTPs, 500 nM of each primer, 11.2 ng pMAL-c2, 0.4 U Phusion HotStart II DNA polymerase, and 1× HF buffer, and ddH₂O. A modified touchdown PCR protocol, which enhances the specificity of amplification, was used for this reaction (Korbie and Mattick, 2008):
98°C  30 s  Initial denaturation
98°C  10 s
61°C → 57°C  30 s  \(\times 35\)
72°C  1.5 min
72°C  3 min  Final extension
4°C  Until further processing

The middle annealing step was decreased by 1°C for each cycle for the first five cycles, then was held at 57°C for the remainder of the thermocycling.

2.4.6.2 Other site-directed mutagenesis

Insertions of his-tags, substitutions around the predicted PEG10 AP active site, and insertion of a TEV-protease cleavage site into the pQE-80L vector were performed essentially as described by Liu and Naismith (2008). For some reactions, a two-step protocol was used (i.e. an annealing temperature of 72°C). All 50 µL was transformed into E. coli DH5α, and 8–12 ng of starting vector was the template for these reactions.

2.4.7 Plasmid sequencing

Plasmids were sequenced by the University of Otago genetic analysis service [http://gas.otago.ac.nz/], using an ABI 3730xl DNA Analyser. Target DNA (~150 ng), and 3.2 pmol of sequencing primer were combined in a 5 µL reaction. See Table 2.3 for sequencing primers. Both forward and reverse sequencing reactions were used to confirm the integrity of each clone. Where there was interference from ‘dye bleeds’ near the beginning of the sequence profile, ambiguous base calls were edited by hand using the 4Peaks software v1.7 (obtainable at [http://nucleobytes.com/]). Sequences were checked against the human genome using the ‘blastx’ functionality of NCBI BLAST [http://blast.ncbi.nlm.nih.gov/]. Complete sequencing data is included in a CD at the back of this document.

2.4.8 Agarose gel electrophoresis

Agarose was added to sodium–borate buffer to a final concentration of 0.8% (w/v) unless otherwise stated, heated in a microwave, and the suspension swirled to dissolve. The solution was left to cool to approximately 55°C, 0.01% (w/v) ethidium bromide added, and it was either poured into a sealed gel tank or stored at 60°C for later use.
The Bio-Rad Mini-sub cell GT system and a Bio-Rad Power-PAC 300 were used for electrophoresis. Separated products were detected using a Bio-Rad GelDoc system. For molecular weight markers, either λ genomic DNA or the DNA molecular weight marker XIV (both from Roche Applied Science, Germany) were used. λ genomic DNA (100 µL) was digested with 4 µL of Hind III and Eco RI (or, if specified, 4 µL of Hind III only) for 2 h at 37°C in 120 µL total volume, and the enzymes heat-inactivated at 60°C for 20 min. Marker lanes were loaded with 1 µL of each marker, corresponding to 210 ng and 150 ng of the λ and XIV markers, respectively.

2.5 Analysis, expression, and purification of proteins

2.5.1 Protein quantitation

A Qubit 2.0 fluorometer (Invitrogen Corporation, USA) was used to measure protein concentration. This instrument was used according to the manufacturer’s recommendations.

2.5.2 Colloidal Coomassie staining of polyacrylamide gels

Colloidal formulations of the popular Coomassie brilliant blue dyes offer enhanced sensitivity for protein detection. A recipe derived from Candiano et al. (2004) was used throughout this study. To make 500 mL of colloidal Coomassie solution, 50 g of ammonium sulphate was added to 100 mL of 50% (v/v) orthophosphoric acid. This was agitated until most of the ammonium sulphate had dissolved and the solution was saturated. To this, 600 mg of Coomassie Blue G-250 was added, and the volume was increased to 400 mL with ddH₂O. The solution was made to 500 mL with absolute ethanol. This solution was stored at room temperature in a bottle protected from light, and shaken prior to use.

2.5.3 SDS-PAGE

The Bio-Rad Mini-Protean 3 gel electrophoresis system was used for PAGE. Unless specified otherwise, gel thickness was 0.75 mm and 16% (w/v) polyacrylamide was used in all separating gels. Stacking gels were 4% polyacrylamide. Polymerisation was induced by the addition of 0.75 µL.mL⁻¹ TEMED, followed by 10 µL.mL⁻¹ of a 10% ammonium persulphate (w/v) stock. For stacking gels, 1 µL.mL⁻¹ TEMED was used instead. The Bio-Rad broad-range molecular weight markers, used according
to the manufacturer’s recommendations, were loaded in either the central lane or the flanking lanes in all gels. Gels were run at 190V (constant voltage) for 1.3 h unless otherwise stated. After electrophoresis, gels were washed for 10 min–1 h in 20% (v/v) ethanol, 8% (v/v) orthophosphoric acid to remove SDS, and then left overnight in colloidal Coomassie stain solution with gentle agitation. Colloidal Coomassie can be re-used, and following removal the gel was destained with multiple washes of dH₂O, typically over the course of a day.

Gels were scanned using an ImageScanner III scanner and the associated Labscan 6.0 software (GE Healthcare, UK), at 300 dpi, using a green filter. Gels were saved in the TIFF format. Adobe Photoshop CS4 and Adobe Illustrator CS4 were used for assembling figures.

Throughout this document, indicated sample volumes used for SDS-PAGE are the volume of the original fraction within the loaded cocktail. For example, samples from purifications were, prior to loading, mixed 1:1 with strongly denaturing and reducing SDS-PAGE sample loading buffer (cracking buffer). Where ‘2 µL loaded’ is indicated, this means 4 µL of the 1:1 sample mixture was the actual load volume on the gel.

### 2.5.4 Western immunoblot

Equipment for Western transfer was obtained from Bio-Rad Laboratories Inc., USA, and was compatible with the Mini-Protean 3 system used for SDS-PAGE. Proteins were transferred to a Whatman PROTRAN nitrocellulose membrane with a 0.45 µM pore size (GE Healthcare, UK). After electrophoresis, gels were assembled in a cassette according to the manufacturer’s instructions with two pieces of Whatman No. 1 paper adjacent to either side of the membrane and gel, and subjected to 100 V for 1 h (constant voltage). An ice pack and small flea were included in the electrophoresis tank, which itself was packed in ice and set atop a magnetic stirrer for agitation.

After transfer, membranes were stained for approximately 5 min in Ponceau red with gentle agitation. The stain was then saved for re-use and the membrane washed with dH₂O until protein bands were visible. Marker bands were then marked with a 6B pencil and the membrane was further washed with dH₂O until the Ponceau staining of bands was faint. Membranes were blocked for 1 h or overnight in TBST + 5% (w/v) skim milk powder (Pams Food, New Zealand) at room temperature with gentle agitation.

For incubation with the primary antibody, membranes were transferred either to a 50 mL conical plastic centrifuge tube and rotated with 4 mL of TBST + 5% (w/v)
skim milk powder, or left in a 8 cm × 11.5 cm plastic tray and placed on a rocking table with 10 mL of TBST + 5% (w/v) skim milk powder. The primary antibody was added at the appropriate dilution (1:1250 or 1:1500 for the anti-MBP, 1:2000 for anti-PEG10–ORF2, and 1:2000 for anti-his) and tubes were left rotating for 90 min at room temperature. Membranes were then placed into rectangular plastic containers and washed three times with TBST + 5% (w/v) skim milk powder for approximately 10 min each. The secondary antibody (anti-rabbit-IgG–HRP for anti-MBP and anti-PEG10–ORF2 primary antibodies, and anti-mouse-IgG–HRP for the anti-his primary) was added at a 1:5000 dilution from a 50% (v/v) glycerol stock stored at –20°C (i.e. 1:10 000 dilution from the manufacturer’s supplied concentration) in a total volume of 12.5 mL TBST + 5% (w/v) skim milk powder. After incubation at room temperature with gentle agitation for 90 min, membranes were washed a further three times with TBST + 5% (w/v) skim milk powder for approximately 10 min each. Prior to development, membranes were then rinsed with TBST only.

Two chemiluminescent reagents were used throughout this study. The first was the Amersham ECL Plus Western blotting detection reagents (GE Healthcare, USA), which was used according to the manufacturer’s instructions. This commercial solution was used for experiments described in Chapter 4. The second system was derived from Haan and Behrmann (2007), and involved mixing stocks to generate a final solution of 100 mM Tris-HCl, pH 8.8, 1.25 mM luminol, 2 mM 4-iodophenylboronic acid, and 5.3 mM H$_2$O$_2$. A total of 2 mL of each solution was used for a typical membrane. A LAS-300 FujiFilm Imager (Alphatech Systems Ltd., New Zealand) was used in development. Unless otherwise stated, the ‘standard’ sensitivity setting was used to preserve the resolution of obtained images. Images were converted to 16-bit greyscale TIFFs and Adobe Photoshop CS4 and Adobe Illustrator CS4 were used for assembling figures.

Anti-MBP primary antibody (Santa Cruz Biotechnology Inc., USA) was stored at 4°C; anti-his (Sigma-Aldrich Corporation, USA) was stored at –20°C. The PEG10-ORF2 primary antibody, raised in rabbit, was stored at –80°C in small aliquots and freeze–thaw cycles avoided (Clark et al., 2007). Secondary antibody–HRP conjugates were anti-rabbit–HRP (from Thermo Fisher Scientific, USA, marketed under the Pierce brand name, stored in 50% (v/v) glycerol at –20°C, used with MBP and PEG10-ORF2 primary antibodies) and anti-mouse–HRP (Sigma-Aldrich Corporation, USA, stored at 4°C).
2.5.5 Mass spectrometry

Mass spectrometry was carried out by the Centre for Protein Research at the University of Otago. Both an ABI 4800 MALDI tandem time-of-flight mass spectrometer and an LTQ-Orbitrap hybrid mass spectrometer were used for sequencing proteins in excised bands after trypsin digestion. The Mascot search engine (Matrix Science Inc, USA), with an expect score cut-off of 0.05 for matching spectra to predicted fragments, was used for bioinformatic analysis.

2.6 Expression and purification of the PEG10-aspartyl protease fusion proteins in E. coli Rosetta 2 (DE3)

For all expressions, (1 mL) pre- and post-induction samples for SDS-PAGE analysis were removed from the culture, the OD$_{600}$ measured with an LKB Biochrom Ultrospec II spectrophotometer, and 0.5 mL of the 1 mL sample was pelleted for 1 min at 17 000 $\times$ g at room temperature. The supernatant was discarded and the E. coli pellet resuspended in 200 $\mu$L $\times$ OD$_{600}$ of cracking buffer.

At each step for all purifications, 50 $\mu$L samples for SDS-PAGE analysis were mixed 1:1 with cracking buffer, briefly vortexed, and heated at 95$^\circ$C for 5 min.

2.6.1 Small-scale expression of PEG10-aspartyl protease fusion proteins

Plasmid preparations (20–100 ng) were transformed into E. coli Rosetta 2 (DE3) as described in section 2.4.2. Plates were grown for 18–20 h at 37$^\circ$C and then stored at 4$^\circ$C for no more than a week. Single colonies were picked and seeded into 1 mL LB broth, 100 $\mu$g.mL$^{-1}$ ampicillin, 30 $\mu$g.mL$^{-1}$ chloramphenicol and grown for 16–20 h at 37$^\circ$C, 200 rpm in an orbital shaker. The next day, 0.5 mL of the starter culture was seeded into 9.5 mL LB broth, 0.2% (w/v) glucose, 100 $\mu$g.mL$^{-1}$ ampicillin, 30 $\mu$g.mL$^{-1}$ chloramphenicol in a conical 15 mL plastic centrifuge tube. The culture tubes were attached horizontally (with masking tape) to the shaking plate of the incubator to aid in aeration. After 2 h of growth at 37$^\circ$C, 200 rpm, the OD$_{600}$ was measured. If the OD$_{600}$ was between 0.50 and 0.75 (mid log phase), cultures were induced with 0.3 mM IPTG and allowed to grow as before for an additional 2 h. If a culture had a very low OD$_{600}$ after 2 h of growth, it was discarded and the culture was grown again, seeded from a fresh starter culture. After 2 h of post-induction growth, another OD$_{600}$
reading and 1 mL sample was taken as before. The final volume of the cultures was therefore 8 mL; in some cases an additional 3 mL was discarded to reduce the volume to 5 mL. Cultures were cooled on ice and pelleted at 6000 \( \times g \) at 4°C for 10 min. The supernatant was decanted, and the pellets were either processed immediately or stored at -20°C.

A commercial passive lysis buffer, BugBuster (Merck KGaA, Germany), was used to lyse cell pellets essentially as described by the manufacturer, in 1 \( \times \) PBS. Lysozyme (50 \( \mu \)g.mL\(^{-1}\)), DNase I (25 \( \mu \)g.mL\(^{-1}\)), and RNase A (25 \( \mu \)g.mL\(^{-1}\)) were added to aid in lysis and digest DNA and RNA, and PMSF (5 mM) was added to inhibit serine proteases of \( E. \) coli. Cell pellets were resuspended with a pipette in 100 \( \mu \)L of the passive lysis buffer per 2 mL of culture, transferred to a 1.5 mL microcentrifuge tube, and agitated gently at room temperature for 20 min. Agitation was performed at 150 rpm or on a rocking table with occasional inversion by hand to aid in mixing. Lysis solutions were then pelleted at 17 000\( \times g \) for 20 min at 4°C to remove any intact cells or large protein aggregates.

A 20 \( \mu \)L sample of soluble fraction was mixed 1:1 with cracking buffer, heated at 95°C for approximately 5 min, vortexed briefly, and stored at -20°C. The pellet remaining after cell lysis was resuspended in 500 \( \mu \)L cracking buffer and treated as above but with more vigorous vortexing, so that the relative concentrations proteins in the pellet fraction and soluble lysate could be compared.

### 2.6.2 Medium-scale expression of MBP–PEG10-AP clone 4 fusion proteins

Starter cultures were seeded and grown as above, except the total volume was 5 mL rather than 1 mL. Starter cultures were pelleted and resuspended in fresh LB broth to remove secreted \( \beta \)-lactamase, and 5 mL was seeded into 95 mL LB broth in glass bevelled shake flasks. Glucose (0.2% w/v), ampicillin (100 \( \mu \)g.mL\(^{-1}\)), and chloramphenicol (30 \( \mu \)g.mL\(^{-1}\)) were added to the 100 mL culture. This was grown for 1.5–2 h, until the \( \text{OD}_{600} \) was 0.4–0.6. At this point, cultures were chilled on ice, induced with 0.1 mM IPTG, and cultures were grown at 18°C overnight at 200 rpm in an orbital shaker. After approximately 18 h, another \( \text{OD}_{600} \) measurement was taken and cultures were pelleted at 6 000 \( \times g \) for 10 min at 4°C in a 250 mL centrifuge bottle. Pellets were either processed immediately or stored at -80°C. Samples for SDS-PAGE analysis were collected as described in Section 2.6.1.

Pellets were thawed at 4°C where appropriate, resuspended in 5 mL Tris-based ly-
sis buffer, and transferred to a pre-chilled 50 mL plastic centrifuge tube. Cell lysis by sonication was performed using a Sonics Vibra-Cell VCX 500 (Sonics and Materials, Inc., USA) set at 30% amplitude. The sonicate was held on ice at all times, and care was taken to ensure it did not warm significantly. One-minute periods of pulsed sonication (2 s on, 2 s off) were used. Eight rounds of sonication, with gaps of at least one minute in between, were typically employed. To test the efficacy of sonication, the sonicate was diluted 1/40 in Tris-based lysis buffer and the OD\textsubscript{600} measured. Sonication was halted as the OD\textsubscript{600} approached 10% of the pre-sonication value. Post-sonication lysates were centrifuged at 17 000 ×g for 20 min at 4°C, and soluble lysates were either processed immediately or snap-frozen on dry ice and stored at −80°C. Samples for SDS-PAGE analysis were taken as in Section 2.6.1.

2.6.3 Small-scale amylose purification of MBP–PEG10 AP fusion proteins

All steps for the small-scale purification of MBP–PEG10 AP fusion proteins were carried out at room temperature and all fractions were collected by decanting unless otherwise specified. To avoid freeze-thaw cycles, soluble lysate was immediately applied to 400 µL of amylose resin slurry (New England Biolabs Inc., USA; resin slurry is 50% (v/v) of amylose resin) in a 1.5 mL microcentrifuge tube. The resin had been washed with a large excess of amylose binding buffer (at least 10× slurry volume) prior to sample application. The total volume of binding solution was increased to 750 µL with amylose binding buffer to facilitate mixing, and binding reactions were left on a rotating wheel at 4°C overnight. Samples were then gently pelleted at 700 ×g for 1 min and the resin bed allowed to settle by gravity. The unbound supernatant (flow-through fraction) was then collected. The resin was then washed five times with 1 mL amylose wash buffer, with inversion by hand until the resin was completely resuspended, and allowed to settle each time by gravity. Two 1 mL fractions were collected using either amylose elution buffer (Figure 4.6) or PBS + 20 mM maltose (Figure 4.7). The first elution was performed for 1 h, the second for 30 min.

2.6.4 Medium-scale amylose-resin purification of MBP–PEG10 AP clone 4 fusion protein

Purifications were performed in 15 mL plastic centrifuge tubes, either at 4°C or on ice for all steps. Samples for SDS-PAGE analysis (50 µL) were collected at all steps and
mixed 1:1 with cracking buffer. For all steps, resin was agitated on a rotating wheel, pelleted at 700 × g for 1 min at 4°C and allowed to settle by gravity flow. Fractions were collected by decanting (rather than pipetting). For equilibration, 2 mL of amylose resin slurry (i.e. 1 mL amylose resin bed volume; New England Bioscience, USA) was washed 5× with ~12 mL amylose binding buffer. Approximately 5 mL of soluble post-sonication fraction was applied to the resin and left to bind on a rotating wheel overnight. The next day, the bound resin was washed with 12 mL amylose wash buffer (5× for 5–10 min). Two 4 mL elutions were performed for 2 h with amylose elution buffer. Subsequently, the resin was resuspended in 4 mL amylose binding buffer and a sample of the resuspended resin was taken and mixed 1:1 with cracking buffer for SDS-PAGE analysis. Elution fractions were processed immediately or snap-frozen on dry ice and stored at –80°C.

2.6.5 Fast-performance liquid chromatography

An ÄKTA Purifier (GE Healthcare, UK) at room temperature was used for automated chromatography. All solutions used were filtered through a 0.22 μM nitrocellulose filter (Millipore, USA) by suction and, on the day of use, were degassed for 10 min in a Soniclean sonicating water bath (Transtek Systems Pty Ltd, Australia). Loading loops were 500 μL (for size-exclusion chromatography) or 5 mL (for ion exchange chromatography). All columns were handled and washed as the manufacturers recommended. Flow rates of 0.5 mL.min⁻¹, 1 mL.min⁻¹, and 0.5 mL.min⁻¹ and fraction volumes of 1 mL, 0.5 mL, and 0.5 mL were used during purification for the Superdex 75 10/300 GL, Resource Q, and Superdex 200 10/300 columns, respectively. Fractions of 0.5 mL were collected throughout purification. For ion exchange chromatography, sample was applied and, before the salt gradient was begun, the A₂₈₀ was allowed to drop to baseline and the column was washed with approximately 5 mL of binding buffer. Samples were centrifuged at 17 000 × g at 4°C for 10 min or 20 min to remove aggregated proteins prior to loading onto the column.

2.6.6 Reciprocal amylose- and cobalt-resin purifications

MBP-PEG10 AP clone 4–his was expressed and lysate gathered as described in Section 2.6.2, except that the cell pellet was resuspended in phosphate-based lysis buffer (without imidazole), with 1× Roche Complete EDTA-free protease inhibitor cocktail. Soluble lysate (2.5 mL) was added to 200 μL amylose resin (New England Biolabs, USA) and 100 μL TALON metal affinity resin (Clontech Laboratories Inc., USA) in

34
15 mL conical centrifuge tubes. Resins were handled, and samples taken, as described in Section 2.6.4. The resins had been washed with appropriate binding buffers in large excess (> 10× resin bed volume) prior to sample application. The total volume of binding solution was increased to 5 mL with phosphate binding buffer (the 10 mM imidazole was excluded for the binding and wash steps of the amylose-resin purification). Binding was allowed to proceed overnight with gentle agitation. The following day, 3× 5 mL washes with phosphate wash buffer were performed for 5–10 min each, and two 2 mL elutions were performed with phosphate–imidazole elution buffer for 1 h each.

2.6.7 Medium-scale cobalt-resin purification of the his–PEG10-AP clone 4 protein

All steps were performed at 4°C or on ice. Incubations were performed on a rotating wheel, and resin was harvested by pelleting at 700 × g for 1 min and allowed to settle by gravity. Fractions were collected by decanting. Soluble lysate (5 mL) was applied to 200 µL of cobalt TALON metal affinity resin (Clontech Laboratories Inc., USA) in a 15 mL conical centrifuge tube. The resin had been washed with phosphate wash buffer in large excess (> 10× resin bed volume) beforehand. Phosphate buffer was added so that a final volume of 10 mL and an imidazole concentration of 10 mM was reached for overnight binding. The next day, the resin was washed twice with 10 mL phosphate wash buffer, and once with 10 mL phosphate wash buffer with 50 mM (rather than 10 mM) imidazole. Washes were incubated for 5–10 min each. Four 2 mL elutions of 15 min each in phosphate elution buffer were performed. Subsequently, the resin slurry was resuspended in 2 mL phosphate elution buffer, and a sample taken for SDS-PAGE analysis while the resin was resuspended.

2.7 Proteolysis assays

2.7.1 Autoproteolysis assays

For the experiments shown in Figure 4.6, elution fractions were gently thawed and pelleted at 17 000 × g for 20 min at 4°C, and 600 µL was dialysed into 10 mM tri-sodium citrate, 1 M NaCl, pH 5.5 at 4°C using microcollodion dialysis capsules (Sartorius, Germany). Dialyses were left overnight with gentle agitation, and 100 µL fractions were allowed to equilibrate to room temperature before the autoproteolysis reaction
was considered to have begun. For the experiments shown in Figure 4.7, elution fractions were processed immediately and were already in the desired buffer. For assays containing a reducing agent, 2 mM DTT was added from a 100 mM DTT stock solution. To halt the reactions, 70 µL of the reaction was added to 35 µL of cracking buffer. The samples were heated at 95°C, briefly vortexed, and stored at –20°C.

### 2.7.2 Post-lysis autoproteolysis assays

*E. coli* Rosetta 2 (DE3) transformed with the plasmid encoding MBP–PEG10-AP clone 4 was seeded with 5 mL from an overnight starter culture into 100 mL LB broth with 30 µg.mL⁻¹ chloramphenicol only. After 2 h of growth at 37°C, 200 rpm in an orbital shaker, the OD₆₀₀ was 0.588. Cells were pelleted, frozen, thawed, and sonicated (in 1.3 mL PBS + 0.5 mM PMSF), and the insoluble material pelleted as described in Section 2.6.2. A 1:1 mixture of non-expressing *E. coli* lysate and ‘elution 1’ fraction of MBP–PEG10 AP clone 4 protein (see Figure 4.8) was in a total volume of 100 µL. For one reaction, 10 mM DTT was added from a 1 M stock solution of DTT in ddH₂O. Reactions were left at room temperature for 2 h and stopped with the addition of 100 µL of cracking buffer.

### 2.7.3 Proteolysis of GST–Peg10-ORF1

His–PEG10-AP clone 4 protein (800 µL) and GST–Peg10-ORF1 (700 µL) were dialysed into neutral proteolysis buffer or low-pH proteolysis buffer at 4°C overnight. The protein concentration was measured using a Qubit fluorometer (Section 2.5.1), and samples were diluted to 4 µM for his–PEG10-AP clone 4 (using a predicted molecular mass of 18.87 kDa) and 6 µM for GST–Peg10-ORF1 (using a predicted molecular mass of 65 kDa). Proteolysis reactions (40 µL) contained a final concentration of 1 µM his–PEG10-AP clone 4, 3 µM GST–Peg10-ORF1, and 1× Roche Complete EDTA-free protease inhibitor cocktail. (Roche complete EDTA-free protease inhibitor cocktail does not contain any agent known to inhibit aspartyl proteases and was added to diminish possible background protease activity arising from contaminant proteins in the final reaction.) Reactions were held at room temperature for 2 h, then halted with the addition of 40 µL cracking buffer, vortexed, and heated at 95°C for 5 min.
2.8 Glutaraldehyde cross-linking

Glutaraldehyde was diluted to a concentration of 2.5% (v/v) in proteolysis buffer. This solution was mixed in a 1:20 ratio with his–PEG10-AP clone 4 elution 2 that had also been dialysed into proteolysis buffer (Section 5.2.1), in a final reaction volume of 100 µL. This was held at 37°C for 5 min, and the reaction stopped with the addition of 10 µL of 1 M Tris-HCl, pH 7.2 to quench the glutaraldehyde. Soon afterwards, 110 µL of cracking buffer was added to create a sample for analysis by SDS-PAGE.

2.9 Primers used in this study

Primers were synthesised by Sigma-Aldrich Corporation, USA, on a 0.025 µmol scale, and had been purified by desalting. Dry primer was briefly centrifuged and resuspended in a volume of ddH₂O so that the final concentration was 100 µM. Vigorous vortexing was used to ensure that the primer powder was completely resuspended. Working stocks of 10 µM (for PCR) or 6.4 µM (for sequencing) were made by dilution in ddH₂O. Dashes within primer sequences in the following tables indicate boundaries between sequence complementary to the template and sequence complementary to the vector backbone (for Table 2.1) or between primer–primer overlapping and primer–primer non-overlapping regions as defined by Liu and Naismith (2008) (for Table 2.2).
<table>
<thead>
<tr>
<th>Target plasmid/primer name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pMAL-TEV</strong></td>
<td></td>
</tr>
<tr>
<td>N-terminus 1</td>
<td>GTACTTCCAG GGCACAATC- GGCGGGAAAC TCCCCG</td>
</tr>
<tr>
<td>N-terminus 2</td>
<td>GTACTTCCAG GGCACAATC- CCTTCAGCGA CCGGGC</td>
</tr>
<tr>
<td>N-terminus 3</td>
<td>GTACTTCCAG GGCACAATC- GCCTCATC TCAACCTGCA</td>
</tr>
<tr>
<td>N-terminus 4</td>
<td>GTACTTCCAG GGCACAATC- CAAGTGATGC TCCAGATTCA TCTT</td>
</tr>
<tr>
<td>N-terminus 5</td>
<td>GTACTTCCAG GGCACAATC- CCAGAAATAA TAAGGTCGCC ACA</td>
</tr>
<tr>
<td>C-terminus 1</td>
<td>CTGCAGGTCG ACTCTAGAGG ATCTTA-GCT CAGGCCACGA ACCC</td>
</tr>
<tr>
<td>C-terminus 2</td>
<td>CTGCAGGTCG ACTCTAGAGG ATCTTA-TTA AAAGACGATA GATCGAGTGC TCC</td>
</tr>
<tr>
<td>C-terminus 3</td>
<td>CTGCAGGTCG ACTCTAGAGG ATCTTA-GCG GCAGTATTCA GAATCAAG</td>
</tr>
<tr>
<td>C-terminus 4</td>
<td>CTGCAGGTCG ACTCTAGAGG ATCTTA-CCG GCAGTGGTAG CGG</td>
</tr>
<tr>
<td>C-terminus 5</td>
<td>CTGCAGGTCG ACTCTAGAGG ATCTTA-ATA ATAGAGTGGC GGTTGTTG</td>
</tr>
<tr>
<td><strong>pQE-80L</strong></td>
<td></td>
</tr>
<tr>
<td>N-terminus 1</td>
<td>CACCATCACG ATACCGGATCT -GGCGGGAA ACTCCCCG</td>
</tr>
<tr>
<td>N-terminus 2</td>
<td>CACCATCACG ATACCGGATCT -CTCTTCA GCAGGGGC</td>
</tr>
<tr>
<td>N-terminus 3</td>
<td>CACCATCACG ATACCGGATCT -GCTCATC TCCACACTTG CA</td>
</tr>
<tr>
<td>N-terminus 4</td>
<td>CACCATCACG ATACCGGATCT -CAAGTGAT GCTCCAGATT CATCTT</td>
</tr>
<tr>
<td>N-terminus 5</td>
<td>CACCATCACG ATACCGGATCT -CGAAGAT AATAGGGTCC CCACA</td>
</tr>
<tr>
<td>C-terminus 4</td>
<td>CGGAGCTCGG ATGCGG-ATCT TAGCCGCCAG TGGTAGCGG</td>
</tr>
<tr>
<td><strong>pQE-TEV</strong></td>
<td></td>
</tr>
<tr>
<td>N-terminus 1</td>
<td>CGASAACCTG TACTTCCAG- GA-GCGGGG AAACCTCCCCG</td>
</tr>
<tr>
<td>N-terminus 2</td>
<td>CGASAACCTG TACTTCCAG- GA-CCTTCA GGCAGCGGGC</td>
</tr>
<tr>
<td>N-terminus 3</td>
<td>CGASAACCTG TACTTCCAG- GA-GCTCCA TCTCCACACTTGCA</td>
</tr>
<tr>
<td>N-terminus 4</td>
<td>CGASAACCTG TACTTCCAG- GA-CAAGTG ATGCTCCAGA TTACATCTT</td>
</tr>
<tr>
<td>N-terminus 5</td>
<td>CGASAACCTG TACTTCCAG- GA-CCAGAA ATAAATAAGGT CCCACA</td>
</tr>
<tr>
<td>C-terminus 4</td>
<td>CGGAGCTCGG ATGCGG-ATCT TAGCCGCCAG TGGTAGCGG</td>
</tr>
</tbody>
</table>
## Table 2.2: PCR Primers used for site-directed mutagenesis

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
<th>$T_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>pMAL-TEV R</td>
<td>GCCCTGGAAG TACAGGTTTT C-GAGGTTGT TGTTATTGTT ATTGT</td>
<td>N/A</td>
</tr>
<tr>
<td>pMAL-TEV L</td>
<td>GAAAACCTGT ACTTCCAGGG C-CGGATCCT CTAGAGTCGA CCT</td>
<td>N/A</td>
</tr>
<tr>
<td>DCG R</td>
<td>CCATGATCGA TTGTTGTC- TTCTGCAAC TTTATGATC ACG</td>
<td>72$^\circ$C</td>
</tr>
<tr>
<td>DCG L</td>
<td>GCACCAACAT CGATCATGG- CTGGACGAA CGGGTGTGT C</td>
<td>72$^\circ$C</td>
</tr>
<tr>
<td>DTG R</td>
<td>CCATGATCGA TACTGGTGC- TTCTGCAAC TTTATGATC ACG</td>
<td>72$^\circ$C</td>
</tr>
<tr>
<td>DTG L</td>
<td>GCACCAATAT CGATCATGG- CTGGACGAA CGGGTGTGT C</td>
<td>72$^\circ$C</td>
</tr>
<tr>
<td>NSG R</td>
<td>CCATGATCAA CTCTGGTGC- TTCTGCAAC TTTATGATC ACG</td>
<td>72$^\circ$C</td>
</tr>
<tr>
<td>NSG L</td>
<td>GCACCAGATGT TGATCATGG- CTGGACGAA CGGGTGTGT C</td>
<td>72$^\circ$C</td>
</tr>
<tr>
<td>C-term his R</td>
<td>TACCACTGCC GG-CACCATC ACCATCATCA C-TAAGATCC TCTAGAGTCG ACCTGC</td>
<td>59$^\circ$C</td>
</tr>
<tr>
<td>C-term his L</td>
<td>GTGCCGGCAG TGGTA-GCGG CAGTATTCAG AATCAAGAG</td>
<td>50$^\circ$C</td>
</tr>
<tr>
<td>pQE-TEV R</td>
<td>ATACACCTCA CGGATCC-GA AAACCTGATAC TTCCAGGTT- GCATGCGAGC TCCTGACC</td>
<td>58$^\circ$C</td>
</tr>
<tr>
<td>pQE-TEV L</td>
<td>GGATCGGTGA TGTTGAT-GG TGATGCGATC CTCTCATAGT TA</td>
<td>49$^\circ$C</td>
</tr>
</tbody>
</table>

## Table 2.3: Sequencing primers

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>pMAL-TEV F</td>
<td>TGCGTACTGC GGTGATCAAC</td>
</tr>
<tr>
<td>pMAL-TEV R</td>
<td>CAAGCTGCCA TTCGCCATTC</td>
</tr>
<tr>
<td>pQE- F</td>
<td>ATAAATAGATT CAAATTGAG CG</td>
</tr>
<tr>
<td>PQE- R</td>
<td>AAGCTAGGTT GGATTCTCAC</td>
</tr>
</tbody>
</table>

## Table 2.4: *E. coli* strains used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH5α</td>
<td>F$^-$ φ80lacZΔM15 Δ(lacZYA-argF)U169 endA1 recA1 hsdR17(r$^-$m$^+$k$^+$)/ deoR thi-1 phoA supE44 λ$^-$ gyrA96 relA1</td>
</tr>
<tr>
<td>Rosetta 2 (DE3)</td>
<td>F$^-$ ompT hdsB$_B$(r$_B^{-}$m$_B^{-}$) gal dcm (DE3) pRARE2 (Cam$^R$)</td>
</tr>
</tbody>
</table>
Chapter 3

Bioinformatic analysis and cloning of the predicted PEG10 aspartyl protease

3.1 Bioinformatic analysis to define the predicted PEG10 aspartyl protease domain

Retroviral-type APs typically rely on termini for dimerisation, and this structural property is essential for enzymatic activity as well as stability in some cases. The correct termini for production of an active AP must therefore be included in any useful expressed variant of the PEG10 AP. Additionally, because of the relatively small size of retroviral-type APs, any disordered terminal amino acids in an expressed construct may especially affect solubility and monodispersity of a PEG10 AP construct (Klock et al., 2008). Initially, bioinformatic approaches were used to predict the likely termini of the PEG10 AP.

3.1.1 Analysis of mammalian PEG10 sequence is insufficient to determine domain boundaries

Investigating the cellular retroviral-type AP Ddi1, Sirkis et al. (2006) used a sequence alignment across several diverse eukaryotic species and looked for poorly conserved stretches of amino acids that correspond to largely disordered inter-domain linkages. This strategy was successful and resulted in the production of a protein that was crystallised for structure determination. Within PEG10 the predicted AP domain is flanked, on one side, by a highly conserved −1 frameshift element, and on the other by a potential AP cleavage site and degenerated reverse transcriptase/RNase H domain. The
<table>
<thead>
<tr>
<th>Host species</th>
<th>Retroelement name</th>
<th>Sequence identifier</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Xenopus tropicalis</em></td>
<td>Amn-san</td>
<td>XP_002937446</td>
<td>2.8^{−81}</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>Amn-ni</td>
<td>XP_003199426.1</td>
<td>7.5^{−79}</td>
</tr>
<tr>
<td><em>Dicentrarchus labrax</em></td>
<td>Unknown</td>
<td>CBN81178.1</td>
<td>5.5^{−75}</td>
</tr>
<tr>
<td><em>Anolis carolinensis</em></td>
<td>Amn-ichi</td>
<td>XP_003219856.1</td>
<td>6.5^{−52}</td>
</tr>
<tr>
<td><em>Takifugu rubripes</em></td>
<td>Sushi-ichi</td>
<td>AAG60684.1</td>
<td>7.5^{−43}</td>
</tr>
<tr>
<td><em>Aspergillus flavus</em></td>
<td>Unknown</td>
<td>AAR29046.2</td>
<td>1.2^{−36}</td>
</tr>
<tr>
<td><em>Filobasidiella depauperata</em></td>
<td>Unknown</td>
<td>ACZ80693.1</td>
<td>2.5^{−35}</td>
</tr>
<tr>
<td><em>Botryotinia fuckeliana</em></td>
<td>Unknown</td>
<td>CCD34677.1</td>
<td>2.4^{−32}</td>
</tr>
<tr>
<td><em>Magnaporthe grisea</em></td>
<td>Maggy</td>
<td>PIR: T18348</td>
<td>7.4^{−27}</td>
</tr>
<tr>
<td><em>Vitis vinifera</em></td>
<td>Unknown</td>
<td>CAN62233.1</td>
<td>1.9^{−16}</td>
</tr>
<tr>
<td><em>Schistosoma japonicum</em></td>
<td>Unknown</td>
<td>AAX25963.2</td>
<td>4.2^{−11}</td>
</tr>
</tbody>
</table>

Table 3.1: **Predicted protein sequences most similar to PEG10.** HMMER3 searches using full-length human PEG10 isoform 3 (NP_001165908.1) were searched against a non-redundant protein database. One selected sequence from each of the top-ranked non-mammalian organisms was checked for similarity to documented retroelements.

strategy applied by Sirkis *et al.* (2006) was therefore not appropriate for estimating the domain boundaries of PEG10 AP. PEG10 is restricted to mammals and highly conserved, and upon sequence alignment there is poor contrast between conserved domain boundaries and potential disordered linker regions with relaxed evolutionary constraints (Appendix II).

Because of the difficulties associated with analysing only mammalian PEG10 protein sequences, close retroelement relatives of PEG10 were included in the analysis and also examined to help find the potential domain boundaries of the PEG10 AP.

### 3.1.2 Ty3/Gypsy LTR-retrotransposon relatives of PEG10 contain a phylogenetic signature of retroviral-type aspartyl proteases but do not inform domain boundaries

The sequence analysis tools and databases PSI-BLAST, HMMER3 and the TBLASTN function of the GyDB retroelement database were used to search for homologues of the predicted PEG10 AP (Finn *et al.* 2011; Llorens *et al.* 2011; Section 2.1). Mammalian proteins were excluded from analyses. The PEG10 AP was most similar to predicted protein sequences from *Xenopus tropicalis*, a high-copy-number predicted LTR-
retrotransposon annotated as ‘Amn-san’ by Llorens et al. (2011). This homology was corroborated with analysis of full-length PEG10 protein sequence (NP_001165908.1; see also Table 3.1). There was also significant similarity to related retroelements from the zebrafish Danio rerio, the European seabass Dicentrarchus labrax, and the green anole (a model reptile) Anolis carolinensis. The GyDB, a curated database of retroelements, clusters predicted elements from these five species as ‘V-clade’ Ty3/Gypsy LTR-retrotransposons. No functional studies have been performed on members of this clade of LTR-retrotransposons.

After vertebrate LTR-retrotransposons, predicted protein sequences from fungi were the next-most related sequences to PEG10. Botryotinia fuckeliana, Magnaporthe grisea, Filobasidiella depauperata—a diverse group of fungi spanning both Ascomycota and Basidiomycota—all had predicted elements highly similar to PEG10. Such homology is in agreement with the Ty3/Gypsy LTR-retrotransposon phylogeny outlined by Llorens et al. (2011). When mammalian proteins were included in the PSI-BLAST and HMMER3 searches, RTL1, also known as PEG11, was found to be highly homologous to PEG10, in some cases more similar than the closest Amn-san retrotransposon from X. tropicalis. Interestingly, no sequences with significant sequence similarity to PEG10 were found when searching the chicken Gallus gallus translated genome.

An amino acid sequence alignment containing sequences from V-clade Ty3/Gypsy LTR-retrotransposons, the close mammalian homologue to PEG10, RTL1, and the fungal retrotransposon Dane-1 was compiled to search for functionally important amino acids in the predicted PEG10 AP. All protein sequence from V-clade retrotransposons is predicted from nucleotide sequence and found by homology searches, rather than from direct amino acid sequencing of isolated proteins. Available predicted protein sequence begins only one amino acid prior to the retroviral-type AP active-site triad due to the close proximity of the –1 frameshift element. Thus, analysis of the N-terminal region of the PEG10 AP is precluded using this approach.

As expected, the active site triad and φ-φ-Gly motif were both highly conserved between PEG10 and close homologues (Figure 3.1). An Asp-Gly (DG) motif with absolute conservation was also observed in the predicted flap region. This DG motif forms part of the ‘pattern 3’ noted as being informative within, rather than between, clades of retroviral-type APs and aligns with the well-known diglycine flap motif of the HIV-1 AP (Figure 3.4 and Llorens et al., 2011). Highly conserved C-terminal cysteines were also apparent, and are possibly analogous to the terminal cysteines/methionines found in retroviruses (Davis et al., 2003). Sirkis et al. (2006) noted that cysteines near this region in Ddi1 were close enough in the tertiary structure to form disulphide
Figure 3.1: Alignment of the predicted human PEG10 AP domain with closely related LTR-retrotransposons. Sequences were obtained via GyDB and aligned using MUSCLE. RTL1, H. sapiens, NP_001128360.1; Amn-ni, D. rerio, XP_003199463.1; Sushi-ichi, T. rubripes, AAC33526.2; Amn-ichi, A. carolinensis, XP_003220873.1; Amn-san, X. tropicalis, XP_002932346.1; Dane-1, A. nidulans, AF295689.1. Dark blue to light blue shading indicates high to moderate conservation, and full stops represent gaps in the alignment.
bonds, further hinting at functional importance for these amino acids.

3.1.3 Homology modelling fails to generate a plausible structure of the PEG10 aspartyl protease but reveals unusual proline enrichment

The HHpred and Phyre 2 automated homology modelling pipelines were used to search for structural analogues of human PEG10 AP (Söding et al., 2005; Kelley and Sternberg, 2009). Briefly, these servers search for close, structurally characterised homologues of an input protein sequence, predict the secondary structure of the input sequence, build a 3D model using close homologues as a template, and refine the final model using loop libraries and/or simple molecular dynamics. PEG10 AP was found to have identity to homologues well into the ‘twilight zone’ of homology modelling (in which the reliability of modelling is highly questionable), with a maximum of 18% identity to yeast Ddi1. Accurate homology modelling is difficult with < 20% percentage identity. After Ddi1, the retroviral XMRV and equine infectious anaemia virus (EAIV) APs were the next-most related, structurally characterised, homologues of PEG10 AP.

All models of PEG10 AP generated by HHpred and Phyre 2 yielded Ramachandran plots with amino acids placed in strongly unfavourable regions, indicating that the accuracy of the predicted models was low. (In addition to unfavourable placements on a Ramachandran plot, the top-ranked model produced by Phyre 2 excluded 20 amino acids in the predicted flap region of PEG10 AP and was excluded from further analysis.) In particular, proline or pre-proline residues, which have unusual backbone conformational restraints, had unacceptable $\phi$-$\psi$ bond angles (Figure 3.2). Full-length human PEG10 isoform 3 has a molar proline percentage of 12.6%, significantly greater than the average of 5.2% in a set of representative proteins (the Dayhoff statistic, as defined by EMBOSS).

To further investigate the apparently unusual position of prolines within the predicted PEG10 AP, alignments of chromovirus Gag-Pol proteins, a clade of LTR-retrotransposons with a wide range of host species and containing the closest relatives of PEG10, were downloaded from GyDB (Llorens et al., 2011). The molar percentage proline of both full-length elements and predicted AP domains was tabulated and compared throughout chromoviruses, and normalised to the average molar percentage of prolines across a diverse set of proteins (Figure 3.3). Marginal enrichment of proline was observed across whole Gag-Pol in vertebrate and some fungal chromoviruses. As
Figure 3.2: Proline and pre-proline residues of PEG10 AP homology models have unfavourable backbone conformations. Ramachandran plots from four top-ranked homology models are shown. Lightly shaded regions indicate disfavoured but ‘allowed’ Ramachandran space, and residues occupying such space are coloured orange (∼2% of residues are expected to reside in this area in high-quality structures). Residues in red occupy forbidden Ramachandran space. The template structure is indicated above each group of plots; models are from Phyre 2 unless otherwise specified. Glycine, proline, and pre-proline residues have different φ-ψ bond angle constraints and are hence presented separately. Plots generated using RAMPAGE (Lovell and Davis, 2003).
Figure 3.3: **LTR-retrotransposon relatives of PEG10 are enriched in proline residues.** Left, phylogenetic tree obtained from GyDB (majority rule consensus tree from 100 bootstraps, Pol domains of Ty3/gypsy LTR-retrotransposons). Centre, Dayhoff statistic for the molar percentage of proline residues in either full-length elements (Gag-Pol) or the predicted AP domain only. See text for further explanation. Right, cladistics of chromoviruses obtained from GyDB. A green box outlines the clade of interest. Graphs were created using R, and the Dayhoff statistics calculated using the ‘pepstat’ EMBOSS tool.

very few chromoviruses have characterised AP domains, only the region between the active site and conserved φ-φ-Gly motif (which is predicted to be part of the AP with high confidence) was analysed for the AP domain in Figure 3.3. Proline enrichment in the predicted AP region was restricted to vertebrate chromoviruses and more pronounced than in the whole of Gag-Pol (Figure 3.3).
3.1.4 Comparative analysis of the PEG10 aspartyl protease with structurally characterised, homologous retroviral-type APs

Despite failing to generate a plausible model of the PEG10 AP fold to inform domain boundaries, automated homology modelling was useful in identifying the structurally characterised proteins most homologous to PEG10 AP. Three such retroviral-type AP domains, from HIV-1, XMRV, and the eukaryotic protein Ddi1, were used for further analysis.

An optimal alignment of the selected homologous domains was generated with Expresso, a sequence alignment algorithm that incorporates structural information for more accurate alignment (Figure 3.4; Armougom et al. 2006). It should be noted that the Ddi1 and XMRV structures have disordered flaps, and so this region may have suffered from lack of structural information relative to adjacent sequence during alignment. As expected, high-confidence alignment was generated around the relatively conserved active site and φ-φ-Gly motifs. An alanine three residues downstream of the predicted catalytic aspartic acid was conserved across all four proteins; an alanine in this position is known to bring the pH optima of APs closer to neutral (Section 1.1.2.2 and Ido et al., 1991). PEG10 AP was found to have insertions at several positions throughout the alignment relative to all other sequences, and alignment was especially poor at both termini. Although the HIV-1 AP has well-characterised termini, Ddi1 and XMRV retroviral-type APs do not: the termini shown in Figure 3.4 for these proteins are merely those used for crystallography, and may not correspond to genuine domain boundaries in vivo. The majority of indels in the resulting alignment were insertions in PEG10 AP.

Regions corresponding to insertions in PEG10 AP relative to the HIV-1, Ddi1, and XMRV APs were mapped to structural models and visualised with PyMol (Figure 3.5). Insertions and deletions at the poorly aligned termini were not used in this analysis. The majority of insertions occurred in the flaps or peripheral loops away from the dimer interface (Figure 3.5).

3.1.5 Choosing several predicted termini for the predicted PEG10 aspartyl protease domain

Lack of reliable sequence from closely related retroelements, strong sequence conservation in mammals, and possible conserved adjacent motifs all confounded comparative methods of domain prediction in the PEG10 AP. Additionally, terminal amino
Figure 3.4: **Optimal alignment of PEG10 AP with three characterised homologues.** Shading from blue to red indicates poor to good alignment quality. Salient structural elements from HIV-1 are annotated above the alignment. Alignment was generated using the Expresso software. Genbank accessions: HIV-1, AAP45767.1; Ddi1, EDN63120.1. Sequence for the XMRV retroviral-type AP was obtained from the 3NR6 PDB file and (Li et al., 2011).
Figure 3.5: **Insertions in PEG10 relative to other retroviral-type APs are preferentially located in loops.** Orange and green denote different polypeptides. The active site catalytic triad is in blue. Four amino acids surrounding insertions in Figure 3.4 are coloured magenta. a the HIV-1 AP in complex with an inhibitor that is not shown in this representation, (PDB ID KB62) b the XMRV AP apoenzyme (PDB ID 3NR6) c *S. cerevisiae* Ddi1 apoenzyme (PDB ID 211A). Note that the ‘flaps’ (positioned at the top of the structures in this view) are not resolved in the XMRV and Ddi1 structures.
Table 3.2: **Termini matrix.** Numbers indicate the identity of the PEG10 AP construct produced. The four beginning and ending amino acids are shown for each N- and C-terminus, respectively. See also Figure 3.6.

A number of predictions for both N- and C-termini of human PEG10 AP were made based on diverse lines of available evidence. For example, the disorder prediction programme Globplot 2.3 was used to determine N-terminus 3 and C-terminus 4, despite potential problems with this method outlined above (Appendix III). Crucially, this software is homology-independent and therefore not redundant with other analyses conducted in this study (Linding et al., 2003). C-terminus 6 ends at the site of a small insertion in rodent PEG10, and C-terminus 1 was estimated by alignment to the Ty3 LTR-retrotransposon that has characterised AP termini. Retroviral-type APs are known to cleave substrates with a proline in the P1’ position, so preference was given for this amino acid when predicting N-termini. Other termini were chosen by position alone, so that a range of sizes and sequence space would be explored within the generated clones. See Figure 3.6 and Table 3.2 for the position of each termini within PEG10 and labelling of corresponding clones.
Figure 3.6: **Termini of constructs explored in this study.** Superscript numbers indicate amino acid numbering of human PEG10 isoform 3 (NP_001164908.1). The catalytic active site (DSG) and conserved hydrophobic loop motifs (VLG) are underlined in red, and a naturally occurring Factor Xa protease cleavage site (IDGR) in underlined in green. The first or last four amino acids of each selected terminus are underlined in black. See also Table 3.2.
In this system, the shortest construct, numbered 19 (Table 3.2), has a total of 120 amino acids from PEG10. Although this is longer than the HIV AP, which is 99 amino acids, sequence alignment shows several insertions in the PEG10 sequence (Figures 3.4 and 3.5). The longest sequence, clone 6, of 176 amino acids is substantially longer than most retroviral-type APs, and is therefore likely to contain the full-length monomer of PEG10 AP with perhaps extra disordered amino acids at either end.

3.2 Cloning of possible PEG10 aspartyl protease sequences

3.2.1 Cloning of PEG10 aspartyl protease fused to a large partner

An initial construct of the PEG10 AP was shown to be >95% insoluble when cloned in fusion with glutathione-S-transferase. By contrast, a construct of PEG10 AP in fusion with the maltose-binding protein (MBP) could be produced as a soluble protein at a high yield (Crowe-McAuliffe, 2010). On these grounds, a library of 30 constructs with each combination of N- and C-termini were prepared with PEG10 sequences in fusion with the MBP using a rapid high-throughput cloning method. Active-site mutants were also generated to assess the proteolytic activity of the PEG10 AP domain.

3.2.2 Creation of a pMAL-TEV plasmid

The pMAL-c2 vector contains sequence coding for maltose-binding protein (MBP) and a multi-cloning site separated by a Factor Xa protease cleavage site (I[DE]GR). By coincidence, a Factor Xa protease cleavage site is also found within all of the predicted PEG10 AP constructs (Figure 3.6, green underline). Because the MBP is a large protein and may affect the function of any fusion partners, it is desirable to remove it for functional studies of the protein of interest. To this end, a novel cloning technique was used to replace the Factor Xa site with a Tobacco Etch Virus (TEV)-protease cleavage site (ENLYFQ↓G). This technique involves i) the production of a linear pMAL-c2 template, excluding the Factor Xa cleavage site and with complementary ends encoding a TEV-protease cleavage site, ii) digestion with the Dpn I restriction enzyme to remove circular template, iii) circularisation of the linear template, and iv) transformation into E. coli bacteria. See Figure 3.7 for a detailed explanation of this technique.

The initial long-template PCR of pMAL-c2 that created a product of the predicted size (6.6 kb) that would circularise was eventually successful. However, the PCR reaction was difficult to replicate, requiring subtle changes in experimental parameters.
Figure 3.7: Replacing the factor Xa cleavage site with a TEV-protease cleavage site. a The existing plasmid, pMAL-c2. b The method used for replacing the factor Xa site with a TEV-protease site. pMAL-c2 is shown in green; primers are black (regions homologous to pMAL-c2) and orange (TEV site-encoding regions). c The altered plasmid, pMAL-TEV.
Figure 3.8: **Creation of a linear pMAL-TEV amplicon.** Lane 1, a positive control reaction used an alternative forward primer that had previously been shown to amplify but yielded an amplicon unable to circularise. Lane 2, the experimental reaction. Lane 3, a negative control reaction, identical to the experimental reaction but containing ddH₂O instead of template DNA. Each lane contains 2.5 µL of PCR reaction product, except for the leftmost lane, which contains λ DNA marker. This agarose gel was run at 180 V for 50 min.

Due to excessive production of amplicons with heterogeneous lengths (manifested as a smear on a gel), poor amplification, or the presence of non-specific products (Figure 3.8 and data not shown). The upward smear observed in both amplification lanes in Figure 3.8 is common in long-template PCRs but is fortunately not prohibitive to cloning with the reaction product (see, for example, Quan and Tian 2009).

The PCR product shown in lane 2 of Figure 3.8 was circularised and transformed into *E. coli* DH5α as described in Section 2.4.6.1. The agar plate seeded with this reaction yielded 38 colonies. Plasmid preparations from four colonies were digested with either *Eco* RI alone, which cuts pMAL-c2 but not pMAL-TEV, or *Eco* RI and *Eco* RV in combination, which results in a single linear product for pMAL-TEV and two products for pMAL-c2. (The *Eco* RI cleavage motif is part of the pMAL multi-cloning site, and is directly 5′ to the *Bam* HI site in Figure 3.7; the *Eco* RV cleavage site is 1.8 kb away from the *Eco* RI cleavage site in pMAL-c2.) The higher molecular-weight species in the pMAL-TEV/*Eco* RI lanes in Figure 3.9 were deduced to be undigested, supercoiled plasmids. All four isolated plasmid preparations from positive colonies have digestion patterns consistent with the predicted pMAL-TEV, rather than pMAL-c2 (Figure 3.9). Success in cloning coupled with sequencing confirmed that an in-frame TEV-protease cleavage site had been created, although the presence of point mutations throughout the plasmid cannot be excluded.
<table>
<thead>
<tr>
<th>Colony 1</th>
<th>Colony 2</th>
<th>Colony 3</th>
<th>Colony 4</th>
<th>pMAL-c2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcoRI/V</td>
<td>EcoRI/V</td>
<td>EcoRI/V</td>
<td>EcoRI/V</td>
<td>EcoRI/V</td>
</tr>
</tbody>
</table>

**Figure 3.9: Diagnostic restriction digest of pMAL-TEV plasmid preparations.** Four plasmid preparations from pMAL-TEV clones and one of pMAL-c2 were digested with either Eco RI and Eco RV (Eco RI/V), or Eco RI alone. Each lane contains 200 ng of DNA except for the outside marker lanes, for which λ DNA digested with Hind III was loaded. This agarose gel was run at 150 V for 40 min.

### 3.2.3 Creating a small library of PEG10 aspartyl protease clones

Ligation-independent cloning methods are more efficient than conventional restriction-based procedures, and favoured by structural genomics consortia when constructing large libraries of constructs for expression (Gräslund et al., 2008). Due to the difficulties encountered in amplifying the pMAL-c2 vector while creating the pMAL-TEV vector (Figure 3.8), methods which required a high-yield long-template PCR of the vector backbone were avoided.

A derivative of the circular polymerase extension cloning (CPEC) method (Quan and Tian, 2009) was used to insert the 30 predicted PEG10 AP domains (Table 3.2), each with different termini, into the pMAL-TEV vector. In this system, inserts are first amplified with ∼20 nt extensions homologous to a linearised vector. Vector and insert are then combined in a PCR-like reaction without primers. During the annealing phase, the complementary overhang of the insert anneals to the vector, allowing the thermostable DNA polymerase to synthesise around the backbone during the extension step (Figure 3.10). A Dpn I digestion then removes template DNA and any uncut pMAL-TEV that contains 6-methyladenosine within GATC (e.g. that has been prepared in a Dam+ strain of E. coli such as DH5α) that may contaminate the reac-
tions. A reaction held on ice instead of undergoing thermocycling was used a negative control. A significant advantage of this procedure is that no gel extractions, DNA precipitations, ligations, or column-based purifications are required, although use of a high-fidelity DNA polymerase is crucial.

### 3.2.4 Amplification of inserts by PCR

The Phusion II HotStart PCR system (Thermo Scientific) was used to amplify the inserts, ranging in size between 351 and 579 nt, using full-length human *PEG10* in the pcDNA 3.1 (+) vector as a template (see Figures 3.10b and 3.11). All desired inserts were successfully amplified by PCR, albeit with varying efficiencies (Figure 3.11). To create the backbone to receive the inserts (Figure 3.10a), the 6.6 kb pMAL-TEV vector was efficiently linearised with *Bam* HI, as assessed on an agarose gel by the absence of a high-molecular-weight band representing a supercoiled uncut product (Figure 3.12).

In small-scale pilot cloning reactions, in which 20 ng of vector was mixed with 0.2 µL of PCR product in the CPEC reaction, approximately one third of the experimental plates yielded colonies. Ten clones (3, 4, 5, 6, 7, 8, 10, 11, 15, and 25) were obtained using this small-scale protocol alone. Similar reactions with greater concentrations of both insert and vector DNA (100 ng of *Bam* HI-linearised pMAL-TEV and 1 µL of PCR product), as well as an additional denaturation and slow annealing step prior to *Dpn* I digestion, were used to obtain the remaining clones (see Section 2.4.5 for details). These reactions were more efficient, with an average of 27 colonies per clone, and 3 on the negative control plate. Nevertheless, clones 12, 21, 27, 29, and 30 proved more difficult to clone using this direct method. For these sequences, insert amplicons were purified using a spin column, quantified, and used in a 1:1 molar ratio with 100 ng of the linearised pMAL-TEV.

Colonies from experimental plates were screened for the presence of an insert in pMAL-TEV by PCR, using pMAL sequencing primers which flank the insertion region (see Section 2.4.5.3 for methods). This system allows discrimination between pMAL-TEV with no construct (a small, ∼350 bp band), pMAL-TEV with insert (sizes vary, but all are over 600 bp), and the presence of a vector other than pMAL-TEV (the absence of any band, for example if pcDNA 3.1 (+) template was carried over from the initial insert PCR reaction). Typically, one third to one half of the screened colonies contained an insert and negative colonies contained pMAL-TEV without an insert (see Figure 3.13 for a sample PCR screen). Although all negative colonies amplified inserts of a similar size, some small size differences were noted (e.g. lanes 13 and 14 in Figure
Figure 3.10: The modified CPEC method used to clone predicted PEG10 AP amplicons into pMAL-TEV. a pMAL-TEV is linearised with Bam HI. b The predicted PEG10-AP inserts are amplified by PCR. Extensions to the primers (black lines) are homologous to the ends of the linearised pMAL-TEV vector. c Linear pMAL-TEV and insert were mixed at approximately equimolar concentrations in a PCR-like reaction without primers. During the annealing phase, the overlaps at the ends of the insert amplicon anneal to linear pMAL-TEV, and DNA polymerase begins to synthesise (arrows and dotted lines) using these regions as primers. The resulting double-stranded plasmids contain the insert and have nicks at positions indicated by asterisks. Five cycles of denaturation, annealing, and extension are performed per reaction. d Dpn I is used to remove transformation-competent contaminants, and samples were transformed directly into E. coli DH5α.
**Figure 3.11:** PCR amplification of inserts corresponding to predicted PEG10 AP domains. Each lane contains the final PCR reaction (2.5 µL loaded); outer lanes contain 1 µL of 100 bp marker. See Figure 3.6 and Table 3.2 for details of these clones. Primer sequences are listed in Table 2.1. ‘–’, no-template control; ‘+’, positive control. These gels were run at 150 V for 22 min.

**Figure 3.12:** Linearising pMAL-TEV with Bam HI. Each lane contains 150 ng of either pMAL-TEV digested with Bam HI (lane 1) or intact pMAL-TEV (lane 2). The leftmost lane contains λ DNA marker. This agarose gel was run at 180 V for 50 min.
3.13, implying that more than one pathway to an undesired but transformable product exists in this cloning system.

Plasmid preparations were prepared from two colonies of each clone that had been identified as positive by PCR-screening. Plasmids were sequenced using pMAL sequencing primers (Section 2.4.7) to confirm the presence of an in-frame PEG10 sequence after the TEV-AP cleavage site (Figure 3.14), and checked for non-synonymous substitutions using BLAST and BLASTX (see Section 2.1 for details). Although no insert-positive clones were out of frame, several sequences had non-synonymous substitutions and required re-cloning.

### 3.2.5 Site-directed mutagenesis of the predicted active site

Three active-site mutants, DTG, DCG, and NSG, were made using a modified Quick-Change protocol (Section 2.4.6.2 and Liu and Naismith 2008). The NSG mutation
Figure 3.15: **Active-site mutants of the PEG10 AP.** Site-directed mutagenesis was performed to mutate the active site of PEG10 AP (natively DSG) to DTG (the most common site among retroviral-type APs), NSG (an enzymatically inactive mutant), or DCG (unknown activity). Mutated residues are marked with an asterisk. *In silico* translation is in black.

abolishes activity in retroviral-type APs whilst maintaining the natural structure around the active site (Kohl *et al.*, 1988). Most retroviral-type APs studied to date have a DTG active site. Some retrotransposons lack a DTG or DSG active-site motif, and instead harbour a DCG triad; a retroviral-type AP containing a DCG motif is predicted to be active, but no biochemical investigation of such an enzyme has been published (Butler *et al.*, 2001). Despite the absence of colonies from negative controls (reactions held on ice instead of undergoing the thermocycling protocol), some randomly picked colonies from experimental plates did not have the desired mutation. Nonetheless, validated sequence change was obtained in colonies for all desired mutants (Figure 3.15).

### 3.3 Conclusions

Bioinformatic analysis of predicted PEG10 AP sequence revealed conserved active-site and φ-φ-Gly motifs, but did not inform domain boundaries. Likewise, homology modelling did not produce a plausible structure of the PEG10 AP but hinted that prolines within PEG10 (and vertebrate LTR-retrotransposon APs) may disrupt structures found in other retroviral-type APs. Several lines of evidence were used to predict five possible N-termini and six possible C-termini for PEG10 AP, and a ligation-independent cloning technique was used to produce the resulting 30 constructs in fusion with MBP. A TEV-protease cleavage site was included in the linker region to allow the removal of MBP if required. Active-site mutants were created by SDM to facilitate analysis of proteolytic activity by the PEG10 AP.
Chapter 4

Expression and purification of the predicted PEG10 aspartyl protease

Each of the 30 MBP–PEG10-AP fusion proteins was expressed in E. coli, initially in small-scale cultures to assess parameters such as solubility, activity, and stability. The aim of this strategy was to select one or several clones for large-scale expression and purification.

4.1 Small-scale expressions and purifications of 30 predicted PEG10 aspartyl protease clones

The predicted PEG10 AP contains codons that are scarce in E. coli, potentially decreasing efficiency of expression of the protein (Appendix IV). E. coli Rosetta 2 (DE3), a strain optimised to express proteins with rare codons, was therefore used as an expression host for the MBP–PEG10-AP clones. Cultures were grown at 37°C and induced to express protein in mid-to-late log phase with 0.3 mM IPTG for 2 h (see Appendix IV for an example set of OD$_{600}$ values). A detailed expression protocol is described in Section 2.6.1.

The predicted relative molecular mass of the MBP, polyasparagine linker, and TEV-protease cleavage site that precede PEG10-AP is 43 kDa; the PEG10-AP clone products have predicted molecular masses between 11.3 kDa and 19.6 kDa. SDS-PAGE analysis of post-induction E. coli lysates showed robust expression of all clones in the expected zone between the 45 kDa and 66 kDa molecular weight markers. For some constructs, however (for example, clones 23 and 24), more than one new band was apparent in the post-induction lane (Figure 4.1).
Figure 4.1: SDS-PAGE analysis of MBP–PEG10-AP expression in *E. coli*. Each lane contains a sample of whole-cell lysate (5 µL). ‘0’ indicates pre-induction and ‘2’ indicates hours post-induction. The numbers above indicate the identity of the each clone (see Table 3.2). SDS-PAGs were stained with Coomassie.

When post-lysate samples were analysed by Western immunoblotting using an MBP antibody, all 30 clones contained either two or three immunoreactive bands, with apparent molecular weights within the range of 45–66 kDa (Figure 4.2a and b). The smallest immunoreactive species was of variable size across clones, dependent only upon the N-terminus, and had a molecular weight greater than MBP alone in all instances (each row in Table 3.2 shares the same-sized smaller fragment). This is most clearly shown in Figure 4.2c, in which the banding pattern of clones that share the same C-terminus and differ only by N-termini are compared side by side. Collectively, these observations imply that the lowest molecular-weight band contains MBP and an N-terminal portion of PEG10-AP and is the result of a single major cleavage event.

All clones yielding three immunoreactive bands in post-induction whole-lysate immunoblots shared one of the two longest C-termini chosen for this study (Figure 4.2). These clones (5, 6, 11, 12, 17, 18, 23, 24, 29, and 30) correspond to the two rightmost columns of Table 3.2 and contain extended C-termini compared to other clones. This implies that a second, minor, cleavage event is occurring between C-termini 4 and 5.
Figure 4.2: **Anti-MBP Western immunoblot of post-induction E. coli lysates.**

a, Matched Coomassie-stained SDS-PAGE (left) and Western immunoblot (right) of post-induction whole-cell lysates from MBP–PEG10-AP clones 1–7. b, Western immunoblots for all clones. No smaller immunoreactive bands were visible in any experiments. c Western immunoblot of selected lysates that have different N-termini but share the same C-terminus. Numbers above lanes indicate the identity of the MBP–PEG10-AP clone (see Table 3.2). ‘−’ lanes contain lysates from *E. coli* expressing MBP only. A 1:1500 dilution of anti-MBP and 10 min development was used for these experiments.
4.1.1 The major cleavage event during expression is active-site dependent.

Site-directed mutagenesis was used to replace the predicted catalytic aspartate residue with asparagine (Figures 3.15 and 4.3). D→N is a conservative substitution that results in inactivity in other retroviral-type APs without substantially altering conformation (Kohl et al., 1988). Four mutant clones, corresponding to all combinations of the longest and shortest termini (1, 6, 19, and 24), were expressed in E. coli Rosetta 2 (DE3) as above, and post-induction whole-cell lysates were analysed by Western immunoblotting. The major cleavage event, common to all clones, did not occur in samples expressing D→N variants. By contrast, the second, minor, cleavage event, which occurred only in clones with long C-termini, was independent of the replacement of the active-site residue (Figure 4.3).

Figure 4.3: MBP–PEG10-AP proteins with a variant active-site have a different banding pattern compared to wild-type products. Lysates and samples were treated as in Figure 4.2. ‘N’, NSG (variant) active-site; ‘D’, DSG (natural) active site.

4.1.2 A map of cleavage events occurring during expression

A crude map of the cleavage events occurring during MBP–PEG10-AP expressions was constructed from data obtained from the Western immunoblots of post-induction samples of each clone and the site-directed mutagenesis experiments (Figure 4.4). The major cleavage event, observed in all clones, yields a protein consisting of full-length MBP and a partial PEG10 AP sequence. Evidence for the complementary C-terminal fragment, consisting of partial PEG10-AP, was not pursued rigorously as the predicted size of this fragment meant it would not easily be detected by Western immunoblot. Additionally, the epitope profile of the PEG10-ORF2 antibody used in this study has not been characterised, and so the antibody may not detect the predicted small C-terminal fragment. The major cleavage event was dependent upon the presence of a predicted catalytic aspartic acid residue, and therefore may indicate it was caused by autoproteolysis i.e. self-cleavage by the PEG10-AP, analogous to a process undergone by the HIV AP (Rosé et al., 1993). By contrast, the second, minor cleavage event that
### Figure 4.4: Deduced cleavage events from immunoblotting data.

The major cleavage event occurs within the PEG10-AP (red arrow), as the size of the major cleavage fragment is dependent on the N-terminus. An intact active site is required for this event. A minor cleavage event (blue arrow), observed only for clones with extended C-termini, also occurs in active-site variants.

 occurred in the extended C-termini, was independent of the active-site and therefore most probably mediated by endogenous proteases of *E. coli*.

### 4.2 Purification of the MBP–PEG10 aspartyl protease fusion proteins.

Despite the observed cleavage events, the full-length product from all clones was still evident by Coomassie staining after SDS-PAGE (Figure 4.1). In order to further assess autoproteolysis, each fusion protein was expressed as described above and purified over amylose resin, a method dependent upon the interaction of MBP with amylose.

#### 4.2.1 Small-scale purification of the MBP–PEG10-AP fusion proteins

After two hours of induction, the *E. coli* Rosetta 2 (DE3) cultures were pelleted and re-suspended in a commercial passive lysis buffer. After lysis, reactions were centrifuged to remove any intact bacteria or large protein aggregates, and affinity purification of the MBP–PEG10-AP fusion proteins was performed using amylose beads (see Section 2.6.3 for detailed methods). Figure 4.5 shows an example of the results from such a purification, the MBP–PEG10-AP clone 19 protein with the NSG variant active site.
Figure 4.5: **Pilot small-scale expression and purification of an MBP–PEG10 fusion protein.** This Coomassie-stained SDS-PAG depicts the purification of a clone 19 NSG active-site variant. ‘S’, soluble lysate; ‘I’, post-lysis insoluble pellet; ‘U’, unbound fraction (0.6 µL loaded). For wash and elution lanes, 2.7 µL was loaded (proportionate loading).

Cell lysis and binding of the fusion protein to the amylose resin were both efficient, and although some MBP–PEG10-AP fusion protein was present in the insoluble fraction, a band of the same apparent molecular weight was the most prominent protein in the soluble post-lysis fraction (Figure 4.5 lanes ‘S’ and ‘I’). A fraction of the MBP–fusion protein did not bind the resin (Figure 4.5 lane ‘U’)) but most eluted from the amylose beads with maltose-containing buffer (Figure 4.5 elutions 1 and 2). The eluted protein was relatively pure in the elution fractions, although small amounts of low-molecular-weight contaminants were nonetheless evident (Figure 4.5). The ‘elution 1’ fraction was judged to be sufficiently pure for autoproteolysis assays.

### 4.2.2 Autoproteolysis does not occur post-purification

Autoproteolysis assays were performed as described in Section 2.7.1. Briefly, fractions purified from amylose resin were dialysed into the desired buffer at 4°C and incubated at room temperature for 2.5 h. Samples were analysed by SDS-PAGE with Coomassie staining. Initially, high-salt, low-pH conditions similar to those used by Zhang *et al.* (1991) for the HIV-1 AP were used. No autoproteolysis was observed under these conditions (Figure 4.6a). Due to the presence of a conserved cysteine residue at the C-terminus of the predicted PEG10 AP (Figure 3.1), and because of the precedence for redox regulation of retroviral APs (Section 1.1.2.3), autoproteolysis assays were repeated but with the addition of 2 mM DTT as a reducing agent. Again, no autoproteolysis was observed in the purified samples (Figure 4.6b). A further matched set of reduced and non-reduced autolysis assays were performed in PBS, to more closely mimic physiological conditions in which the PEG10 AP may be active *in vivo*. No cleavage was observed for any clone product under these conditions (Figure 4.7).
Figure 4.6: *Autoproteolysis assays of semi-purified fractions in high salt and at low pH.* **a** Purified samples were incubated in 10 mM tri-sodium citrate, pH 5.5, 1 M NaCl for 2.5 h at room temperature. For **b**, 2 mM DTT was added. Each lane of these Coomassie-stained SDS-PAGs contains 10 µL of sample. ‘0’ and ‘2’ indicates no or 2.5 h of incubation, respectively, and the top number indicates the identity of the clone.
Figure 4.7: Autoproteolysis assay of semi-purified fractions in PBS. Samples were eluted from amylose resin with PBS + maltose buffer and were allowed to incubate at room temperature for 2.5 h. For b, 2 mM DTT was added. These fractions were analysed by SDS-PAGE and Coomassie staining as in Figure 4.6. The major protein species in each lane, MBP–PEG10-AP and fragments, were all between the 45 kDa and 66 kDa protein markers (centre lanes).
4.2.3 Does autoproteolysis occur after expression?

To more closely analyse the apparent autoproteolysis that occurs during expression but not after purification, MBP–PEG10-AP clone 4 was selected for closer scrutiny. This construct is the largest that will express with an intact C-terminus in *E. coli*, and has the longest N-terminal region of clones used in this study (See Figures 3.6 and 4.4). A 100 mL culture of MBP–PEG10-AP clone 4 was induced at 18°C and expression allowed to occur overnight (Section 2.6.2). The cells were lysed and the protein purified in a procedure similar to that outlined in Section 4.2.1 but adapted to a larger scale (Section 2.6.4 and Figure 4.8). Cell lysis was less efficient in this experiment than for the small-scale cultures (lane ‘I’, Figure 4.8), although elution fractions were significantly more concentrated than for purifications as it was practical to use a smaller relative elution volume (0.6 μL loaded in Figure 4.8 versus 2.7 μL in Figure 4.5).

It was hypothesised that autoproteolytic activity was observed during, but not after, expression of the MBP–PEG10-AP fusion proteins perhaps due to the presence of a co-factor in, or the environment of, the *E. coli* protoplasm. The ‘elution 1’ fraction (50 μL) from an amylose purification of MBP–PEG10-AP clone 4 (Figure 4.8) was thereby mixed with non-expressing *E. coli* Rosetta 2 (DE3) soluble lysate and allowed to incubate for 2.5 h at room temperature (see Section 2.7.1 for detailed methods). Due to the lower concentration of MBP–PEG10-AP in these mixtures compared to expressing lysates, non-specific bands were evident in a Western immunoblot using an anti-MBP primary antibody (Figure 4.9a and the left-most lane of Figure 4.9b). Nonetheless, bands corresponding to the expected MBP-containing species were clearly visible at the expected sizes in the appropriate lanes, although the membrane required high sensitivity of exposure to clearly detect them. No reconstitution of autoproteolysis activity was apparent after 2.5 h, regardless of the presence or absence of 10 mM DTT (Figure 4.9b). No cleavage of *E. coli* proteins was apparent by Coomassie staining after
Figure 4.9: **Mixing purified MBP–PEG10-AP clone 4 protein with *E. coli* lysate does not restore autoproteolytic activity.** Reactions containing either *E. coli* lysate only or *E. coli* lysate mixed in a 1:1 ratio with purified MBP–PEG10-AP clone 4 product were either not incubated (‘0’) or incubated for 2.5 h (‘2.5’), with or without the addition of DTT, at room temperature. Lanes contain 5 µL of the final reaction. a, Coomassie-stained SDS-PAG b, Western immunoblot probed with a 1:1500 dilution of an anti-MBP primary antibody, exposed for 15 min. The contrast was digitally scaled in order to better detect faint bands.

SDS-PAGE (Figure 4.9a).

To further explore the autoproteolysis during expression, degradation of the MBP–PEG10-AP clone 4 product was examined during *E. coli* cell lysis. One 50 mL culture expressing MBP–PEG10-AP clone 4 was grown as in Section 2.6.1. After 2 h of induction, four 8 mL aliquots were pelleted and lysed in passive lysis buffer as per Section 2.6.3, except with a range of proteases inhibitors included or excluded from the lysis reaction. The tested conditions were i), no protease inhibitors, ii) 2 µM pepstatin A (to inhibit APs), iii) 5 mM PMSF (to inhibit serine proteases), or iv) a commercial cocktail of protease inhibitors (to inhibit serine and cysteine proteases).¹ Samples were taken after 20 min (the lysis period used in other experiments) or 2 h of lysis at room temperature (see Section 2.7.2 for detailed methods).

Few differences were observed between samples lysed in the presence or absence of different protease inhibitors (Figure 4.10). Notably, the major cleavage event observed during expression—a reduction of the high-molecular weight band to a smaller fragment that migrates above the 45 kDa marker—did not apparently occur during

¹When EDTA, an inhibitor of metalloproteases, was included in lysis reactions, large pellets were observed after centrifugation. Samples from this experiment were not analysed by SDS-PAGE, as it was concluded that EDTA was affecting the lysis efficiency.
cell lysis (Figure 4.10b). Minor events, involving the further degradation of the major cleavage fragment, were evident under all conditions, however. After 20 min of the lysis protocol, doublet bands had replaced the major cleavage fragment in all conditions, implying that this fragment is unstable. By 2 h of lysis, the lower doublet band, with an apparent molecular weight of approximately 45 kDa, had become the dominant observed cleavage fragment. The observed banding pattern across time was similar in all conditions, except for a faint band (marked with an asterisk in Figure 4.10b) that is most intense in the pepstatin A-containing lysis reaction after 2 h. This observation is best explained if this fragment is further cleaved by an aspartyl protease, and may be evidence of further proteolytic activity by MBP–PEG10-AP clone 4. This specific proteolysis event is difficult to detect and therefore not suitable to assay for activity of MBP–PEG10-AP, however. A slight drop in the intensity of the band representing full-length MBP–PEG10-AP clone 4 protein was observed between 20 and 120 min for all incubations except for that containing the commercial protease inhibitor cocktail (Figure 4.10b). Such slight degradation may therefore be mediated by cysteine proteases from *E. coli*.

In order to assess how autoproteolysis proceeds over time during expression, a small-scale MBP–PEG10-AP clone 4 culture was grown and induced as described in Section 2.6.3, and samples were taken at 30 min intervals over the course of a 3 h induction. In a parallel experiment, an identical culture was treated with a high concentration of tetracycline to inhibit translation at 1 h post-induction. Analyses of fractions by SDS-PAGE and Western immunoblot revealed that autoproteolysis, indicated by the presence of a minor immunoreactive band, begins soon after induction (Figure 4.11). The ratio of the full-length product and major cleavage fragment appears to stay constant by simple examination of the immunoblot. No additional smaller fragments appeared by 3 h post-induction (Figure 4.11a). The addition of 20 μg.mL⁻¹ tetracycline at 1 h post-induction halted cell growth as measured by OD₆₀₀ and effectively halted translation (Figure 4.11b and data not shown). Remarkably, the major cleavage fragment was then completely degraded over the subsequent 2 h, while the full-length translation product remained stable, with only a slight drop in intensity over time (Figure 4.11b). Importantly, no large-scale proteolytic programme in *E. coli* was triggered by the addition of tetracycline, as assessed by SDS-PAGE (Figure 4.11b, left).

Several active-site variants of MBP–PEG10-AP clone 4 were expressed as in Sec-

---

2There is a wide dynamic range across the Western immunoblot in Figure 4.11a. As a result, the intensity of each band is unlikely to be linearly proportional to the amount of protein in each band, which precluded densitometric analysis.
Figure 4.10: **Minor autoproteolysis occurs post-lysis.** MBP–PEG10-AP clone 4 was expressed and bacteria lysed as described. Immediately before cell lysis in passive lysis buffer, a sample of culture was pelleted and resuspended in cracking buffer (lanes ‘B’; 3.5 µL loaded). Cell pellets were lysed either without any protease inhibitors (‘None’), with 2 µg.mL⁻¹ pepstatin A (‘Pep. A’), 5 mM PMSF, or a commercial protease inhibitor cocktail (‘Cocktail’). Lanes labelled ‘20’ or ‘120’ indicate samples taken 20 min and 120 min, respectively, after cells were resuspended in lysis buffer (1 µL loaded). **a** A Coomassie-stained gel. **b** Western immunoblots of the above gel, probed with a 1:1 000 dilution of an anti-MBP primary antibody. Exposed for 5 min (left) or 15 min (right). Asterisks indicate the faint band that appears only in the samples lysed in the presence of pepstatin A after 120 min.
Figure 4.11: **Autoproteolysis proceeds over induction.** Cultures expressing MBP-PEG10-AP clone 4 were induced and samples (5 µL loaded) taken as described. **a.** A matched Coomassie stained gel and Western immunoblot of post-induction samples. **b.** As in a, except tetracycline was added 1 h after induction (indicated with an asterisk). A 1:1500 dilution of anti-MBP primary antibody was used for these Western immunoblots. Membranes were exposed for 15 min and 20 min for **a** and **b** respectively.
Figure 4.12: **Effects of variant active-site triads on autoproteolysis during expression.** DCG, DTG, and NSG active-site variants of MBP–PEG10-AP clone 4 (see Figure 3.15) as well as the wild-type variant (DSG) were expressed for 2 h. To best observe minor bands, 10 µL of post-induction whole-cell lysate was loaded in each lane. This Western immunoblot was probed with a 1:1250 dilution of an anti-MBP primary antibody and exposed for 20 min.

4.2.4 Conclusions

Two cleavage events were observed by Western immunoblotting during expression of MBP–PEG10-AP in *E. coli*. A minor cleavage event occurred only in clones with extended C-termini and was likely mediated by endogenous proteases of *E. coli*. Other explanations, however, such as the falloff of the ribosome from the mRNA, cannot be excluded to account for this event. The second, major cleavage event was observed for all 30 MBP–PEG10-AP clones during expression, but no proteolysis was observed post-purification in a variety of conditions (Figures 4.2, 4.6 and 4.7). This major cleavage event requires a DSG or DTG active-site triad typical of active retroviral-type APs and occurs within the PEG10 portion of the MBP–PEG10-AP fusion protein, although the exact site was not determined. Taken together, these observations imply that the major cleavage fragment is a product of autoproteolysis—a phenomenon known to occur in some other retroviral-type APs. The major cleavage fragment was unstable both in *E. coli* during expression (Figure 4.11) and during cell lysis (Figure 4.10).
4.3 Removal of the MBP from the predicted PEG10 aspartyl protease

The MBP is a large, solubilising fusion partner that may, in some cases, interfere with the activity of the fused protein of interest. Prior to further functional studies of the predicted PEG10 AP, removal of the MBP fusion partner was thought to be desirable.

4.3.1 Cleavage of the MBP–PEG10 aspartyl protease fusion proteins with TEV protease

Relatively pure first elution fractions from amylose purifications (see ‘0 h’ fractions from Figures 4.6 and 4.7) were mixed with 5 µg TEV protease (AcTEV protease, Invitrogen Corporation, USA) and left at 4°C overnight. Fractions were centrifuged to remove any insoluble aggregates that might have formed and the soluble fraction of each clone was analysed by Coomassie staining after SDS-PAGE. The resulting bands were at the expected molecular weights of both MBP and that of the particular PEG10-AP protein (Figure 4.13). The intensity of each intact MBP–PEG10-AP fusion protein can be used to assess the qualitative efficiency of cleavage by the TEV protease. Despite the MBP–PEG10-AP products sharing the same TEV-protease cleavage sequence (ENLYFQ↓G), and an additional three amino acids (RIG) prior to any unique sequence, there was a wide range of TEV-protease cleavage efficiencies among them. Proteins containing N-terminus 4 (18–24) had no appreciable cleavage under these conditions, while clones with N-termini 1 or 2 (1–12) had nearly complete cleavage. Constructs with other N-termini had intermediate cleavage efficiencies (Figure 4.13).

4.3.2 Confirmation of the identity of the PEG10 aspartyl protease by Western immunoblotting and mass spectrometry

MBP–PEG10-AP clone 4 was chosen for further analysis as it contains the longest intact C-terminus and there was near-complete cleavage by the TEV protease under gentle conditions. Following expression and purification (Figure 4.8), the first elution fraction of MBP–PEG10-AP clone 4 protein was dialysed into PBS and incubated overnight at 4°C with 20 µg TEV protease, 2 mM DTT. Salient fractions from this

---

3No visible pellets were evident after centrifugation, and ‘insoluble’ fractions in all cases resembled the soluble counterpart when analysed by SDS-PAGE, implying carry-over contamination (Appendix VI).
Figure 4.13: **Cleavage of the MBP–PEG10-AP fusion proteins with the TEV protease.** Numbers above each lane indicate the identity of the clone. Three zones in each lane, corresponding to the predicted sizes of the MBP–PEG10 fusion proteins, MBP, and the PEG10 proteins are indicated on the left. Each lane of these Coomassie-stained SDS-PAGs contains 6.7µL of each TEV-cleavage reaction. The ‘−’ lane contains MBP incubated with TEV protease.
Figure 4.14: SDS-PAGE and Western immunoblot analysis of partially purified and TEV-protease-cleaved MBP–PEG10-AP clone 4 protein. Matched Coomassie-stained SDS-PAGE (left) with anti-MBP and anti-PEG10–ORF2 immunoblots. ‘S’, the soluble lysate after cell disruption by sonication; ‘E’, the elution fraction from amylose purification; ‘C’, elution fraction incubated with TEV protease overnight. One, two and three asterisks indicate the MBP–PEG10-AP, MBP, and PEG10-AP clone 4 proteins, respectively.

process were analysed by SDS-PAGE and Western immunoblotting. This analysis revealed the absence of smaller immunoreactive bands that may correspond to cleavage fragments of the fusion protein post-lysis, although Western analysis implied a lower TEV-protease cleavage efficiency than inferred from Figure 4.13 (Figure 4.14).

In order to confirm the identity of the PEG10-AP clone 4, mass spectrometry was used in addition to Western immunoblotting (see Section 2.5.5 for detailed methods). MALDI TOF/TOF sequencing confirmed the identity of the PEG10-AP clone 4 product, with 92% coverage. A minor band with an apparent molecular mass less than the PEG10-AP clone 4 product was evident by Coomassie staining after SDS-PAGE (‘unknown protein’ in Figure 4.15a). This was also observed, albeit faintly, in the Coomassie-stained SDS-PAGE in Figure 4.14. This minor band was not immunoreactive with an anti-PEG10–ORF2 antibody (Figure 4.14, right panel). A more sensitive LTQ-Orbitrap analysis was required to identify this unknown band as also corresponding to the PEG10-AP clone 4 product. Although sequence coverage was comparatively poor at 61%, an N-terminal peptide beginning three amino acids into the predicted PEG10-AP clone 4 product was within this sequence coverage, indicating that this
smaller protein is likely a C-terminal truncation—and may be the PEG10-AP portion of the major cleavage fragment, released from MBP by the TEV protease. The extreme N- and C-termini were not detected for either the PEG10-AP clone 4 product or the truncated band as they are within very small fragments generated by proteolysis. Sample spectra for a peptide spanning the active-site triad are shown in Figure 4.15c and d. Mascot ions scores were 110 and 87, respectively, for these spectra, indicating that the matches were very unlikely to arise by random chance. See Appendix VII for Mascot peptide matching results for both bands.

### 4.3.3 Chromatography to further purify the PEG10 aspartyl protease domain

After cleavage of the fusion partner with TEV protease the resulting fraction was applied to a Superdex 75 10/300 GL size-exclusion chromatography (SEC) column (Section 2.6.5; GE Healthcare, UK). SEC was chosen as an initial chromatography method because it can serve both as a purification step and analytical tool to estimate the monodispersity and oligomerisation state of a protein. Significant co-elution was observed between PEG10-AP clone 4 and MBP (Figure 4.16). PEG10-AP clone 4 has a predicted molecular weight of 17.4 kDa, in contrast to the much larger MBP (42.5 kDa). Dimerisation of PEG10-AP clone 4 may therefore explain the co-elution with MBP over a size-exclusion column. However, two overlapping peaks—one corresponding to MBP and the other to PEG10-AP clone 4—around fractions 4 and 5 in Figure 4.16 might have been expected in this scenario, instead of the observed single peak. This chromatography technique was not pursued further to separate MBP and PEG10-AP clone 4 after TEV-protease cleavage.

Alternatively, a Resource Q (GE Healthcare, UK) anion-exchange column was investigated as a technique to purify PEG10-AP clone 4 after initial amylose purification and TEV-protease cleavage (Figure 4.17). All 4 mL of the E1/TEV fraction in low-salt phosphate buffer was loaded onto the column in this experiment, and elution was performed with a 0–1 M NaCl gradient (Section 2.6.5). Most of the MBP was separated from the PEG10-AP clone 4 product. As with SEC, however, the PEG10-AP clone 4 product co-eluted with a minor fraction of MBP (fractions 4–10). Uncleaved MBP–PEG10-AP clone 4 fusion protein was an additional contaminant. An unusually ‘long tail’, containing all three major protein species, was observed by analysis of fractions by SDS-PAGE with Coomassie staining (Figure 4.17). Such a tail may imply multiple conformations and/or interactions between MBP and the PEG10-AP clone 4 product.
Figure 4.15: **Tandem MS/MS sequencing spectra of PEG10-AP clone 4.**  

**a.** A Coomassie-stained SDS-PAG of PEG10-AP clone 4 elution 1 fraction after cleavage with TEV protease (4 µL loaded).  
**b.** The peptide selected for sequencing, which spans the active-site catalytic triad.  
**c.** MALDI TOF/TOF sequencing of the indicated peptide from the PEG10-AP major band.  
**d.** LTQ-Orbitrap sequencing of the indicated peptide from the PEG10-AP minor band. For **c** and **d**, positions of the dominant y ions, which correspond to breakages of the peptide bond and a charged C-terminus, are indicated.
Figure 4.16: The PEG10-AP clone 4 protein co-purifies with MBP during size exclusion chromatography. MBP–PEG10-AP clone 4 fusion protein that had been eluted from an amylose column was incubated with TEV protease and applied to a Superdex 10/300 GL column as described. Top, the chromatogram of fractions collected. $A_{280}$ is shown in blue; $A_{215}$, shown in pink, is displayed in arbitrary units for comparison of relative peak sizes only. Bottom left, a gel containing protein separated by SDS-PAGE and stained with Coomassie of fractions 1–7 (4.5 µL loaded) with 2 µL of original fraction loaded in the left-most lane after the size marker.
Figure 4.17: The PEG10-AP clone 4 protein co-purifies with MBP during ion-exchange chromatography. Relatively pure MBP–PEG10-AP clone 4 fusion protein that had been eluted from an amylose column was incubated with TEV protease and applied to a Resource Q anion-exchange column as described. Blue, $A_{280}$, brown, conductivity induced by application of a 0–1 M NaCl gradient. $A_{215}$, shown in pink, is displayed in arbitrary units for comparison of peak sizes relative to $A_{280}$ only. Below is a Coomassie-stained gel of fractions 1–10 (7 µL) with 2 µL of original fraction loaded in the left-most lane after the size marker.

Fractions containing all three major protein species (5–10) spanned four peaks in the absorbance profile, two of which overlapped (centred around fraction 5) and one of which was very broad (centred around fraction 8).

4.4 Does the PEG10 aspartyl protease interact with MBP?

It was hypothesised from the results of Section 4.3 that the PEG10-AP clone 4 product and MBP might be associating in a non-covalent manner, as this would explain their co-elution in the two chromatography systems. In order to test this hypothesis and/or facilitate the separation of MBP and the PEG10 AP, site-directed mutagenesis, confirmed by sequencing, was used to insert a hexa-histidine tag onto the C-terminus of MBP-PEG10-AP clone 4 (Figure 4.18 and Section 2.4.6.2). Figure 4.19 outlines the
strategy, somewhat similar to the commonly used pull-down method, and shows initial purification of material used to test the hypothesised interaction. If there were an interaction between MBP and the PEG10-AP clone 4–his product, it would be expected that PEG10-AP clone 4–his would appear to bind amylose resin loaded with MBP, and MBP would appear to bind cobalt resin loaded with PEG10-AP clone 4–his.

The MBP–PEG10-AP clone 4–his protein was expressed in a 100 mL culture and cells were lysed as described previously (Section 2.6.2). Half of the lysate was purified over amylose resin utilising the MBP fusion partner, and the other half over cobalt resin utilising the C-terminal his-tag. The amylose- and cobalt-purified elution fractions were incubated with TEV protease and then reapplied reciprocally to the complementary cobalt and amylose resins, respectively (Figure 4.19). (The tight association between MBP and maltose prevents later binding of MBP to amylose resin subsequent to an elution with maltose (New England Biolabs, pMAL- vector manual)). Expression of MBP–PEG10-AP clone 4–his was at a high level, and cell lysis efficient. Initial purification over the cobalt resin yielded a relatively pure, concentrated elution fraction. Likewise, the elution 1 fraction from the amylose resin was pure, as shown by Comassie staining after SDS-PAGE, although elution was less efficient than for the cobalt-resin purification (Figure 4.19 gels).

Subsequent to digestion with TEV protease, amylose- and cobalt-purified fractions were applied to the complementary resins (see Section 2.6.6 for detailed methods). Resins were washed, and a sample of each was suspended in cracking buffer to assess binding. MBP was not found in cobalt resin fraction loaded with PEG10-AP clone 4–his, as would be expected in the case of an interaction between the two proteins (Figure 4.20a, middle panel; the position of the expected band is indicated with a red dagger). Conversely, PEG10-AP clone 4–his was associated with the amylose resin fraction loaded with MBP, albeit weakly (Figure 4.20b, right panel, band marked with a red asterisk). Thus, Figure 4.20 offers evidence both for and against an interaction between MBP and PEG10-AP clone 4–his.
Figure 4.19: **Experimental strategy for testing an association between PEG10-AP and MBP.** MBP-PEG10–AP–his (clone 4) was expressed and lysed as described. The lysate was split into two equal fractions and purified over either amylose (left) or cobalt resin (right). Elution fractions were cleaved with TEV protease to remove the covalent link between MBP and PEG10-AP–his, then applied to the complementary resin. ‘0’ and ‘18’ reflect pre-induction (0 h) and post-induction (18 h) samples (5 µL loaded), ‘S’ post-lysis soluble fraction (2.5 µL loaded); ‘I’, post-lysis insoluble fraction (2.5 µL loaded); ‘U’, unbound (5 µL loaded); ‘W’, washes (5 µL loaded); ‘E’, elutions (2 µL loaded). Proteins were separated by SDS-PAGE and gels stained with Coomassie. Purification gels were loaded proportionately.
Figure 4.20: **Tandem pulldowns imply interaction between PEG10-AP and MBP.**

All lanes contain 1.5 µL of each fraction. **a**, Amylose–cobalt resin pull-downs. **b**, Cobalt–amylose resin pulldowns. The anti-ORF2 Western immunoblots were both developed for 1 min; anti-MBP Western immunoblots were developed for 5 min (a) and 3 min (b). ‘L’, TEV-protease cleaved E1 fractions loaded onto the resin; ‘U’, unbound; ‘W’, washes; ‘R’, resin. An asterisk in red shows the PEG10-AP–his apparently bound to the amylose resin. A dagger in red indicates the position of the expected band if MBP were apparently binding to the cobalt resin.
Despite the accumulated evidence in favour of an interaction between MBP and PEG10-AP clone 4, the amylose resin → cobalt resin purification scheme was successful in separating the two proteins. Elution fractions, however, would contain contaminating TEV protease as well as uncut fusion proteins. A purification protocol would therefore require at least one further step. Additionally, the resulting pure PEG10-AP domain would contain a his tag at the C-terminus—a site key to this study as it is important for dimerisation in all homologous retroviral-type APs for which structural data are available (Section 1.1.2.2). It was concluded that further work with MBP–PEG10-AP clone 4–his would not be pursued.

4.4.1 Summary

A wide range of TEV-protease cleavage efficiencies, dependent on the N-terminus, was observed across the MBP–PEG10-AP clone products. PEG10-AP clone 4 was selected for further analysis based on both its TEV-protease cleavage efficiency and because it was the longest clone that can express an intact C-terminus in E. coli. Western immunoblot and mass spectrometry confirmed the identity and sequence of PEG10-AP clone 4. Separation of MBP and the PEG10-AP clone 4 protein proved difficult subsequent to TEV-protease cleavage, however, with substantial co-elution observed in two chromatography systems. This lead to the hypothesis that MBP and the PEG10-AP clone 4 protein were associating non-covalently. A modified pull-down approach provided some evidence in favour of an association, but was not conclusive.

4.5 Conclusions

All 30 chosen termini expressed well and underwent limited autoproteolysis during expression. Clones with long C-termini underwent an additional cleavage event mediated by endogenous proteases of E. coli. No autoproteolytic activity was detectable subsequent to purification, but site-directed mutagenesis confirmed that the observed activity during expression required sequence hallmarks of retroviral-type APs. PEG10-AP clone 4 was selected for detailed examination but appeared to associate with the MBP subsequent to cleavage with the TEV-protease. Evidence for an association between the two proteins remained equivocal, despite further analysis. In either scenario, however, the resulting purification strategy was judged unsatisfactory for producing pure PEG10-AP protein for further work.
Chapter 5

Expression and purification of his-tagged PEG10 aspartyl protease

5.1 Creation of his-tagged constructs

A subset of predicted PEG10-AP sequences was selected for further cloning and expression with an N-terminal hexahistidine tag. As clones with C-termini 5 or 6 had C-terminal cleavage in *E. coli*, clones with C-terminus 4, the longest that can be expressed intact in *E. coli* were selected for expression with an N-terminal hexahistidine tag. These clones correspond to numbers 4, 10, 16, 22, and 28 in Table 3.2. An additional set of constructs was generated for these clones that had an N-terminal his-tag followed by a TEV-protease cleavage site (ENLYFQ↓G). The inclusion of a TEV-protease cleavage site enables removal of the his-tag after the use of a convenient purification scheme utilising a his-tagged TEV protease, but the TEV sequence can in some cases decrease the solubility of expressed proteins (Kurz *et al.*, 2006).

The modified ligation-independent cloning technique (Section 2.4.5) was used to insert sequences corresponding to clones 4, 10, 16, 22, and 28 into the pQE-80L vector, which contains an N-terminal his-tag. Site-directed mutagenesis was used to insert a TEV-protease cleavage site (ENLYFQ↓G) directly after the his-tag of pQE-80L (see Section 2.4.6.2 for details). The resulting vector was called ‘pQE-TEV’. The modified ligation-independent cloning technique was then also used to insert the selected clones into this vector (Figure 5.1). All clones were confirmed to be mutation-free by sequencing.
Figure 5.1: **Creation of his-tagged constructs.**  

- **a** N-termini for his- (top) and his-TEV-tagged sequences. An additional G was included at the C-termini of the shown sequences for PEG10-AP clone 4.
- **b** Amplicons (2 µL) used for insertion into the pQE-80L vector to create the his-tagged constructs. Gel run for 25 min at 130 V.
- **c** pQE-80L linearised with *Bam* HI (lane 1) or uncut (lane 2), with 120 ng DNA loaded in each lane. Gel (1% agarose) run for 45 min at 200 V.
- **d** Selected sequences of the pQE-TEV vector (top) or pQE-TEV–PEG10-AP clone 10 (bottom), showing the his-TEV site and the start of the PEG10 AP sequence (arrow). The *Bam* HI cleavage site for pQE-TEV, used for cloning, is indicated in red.
5.2 The his–PEG10-AP clone 4 product is the most soluble of the expressed constructs

Small-scale (10 mL) cultures of his- and his-TEV- constructs were grown as described in Section 2.6.1, except they were cooled on ice prior to induction and expression was at 18°C overnight (18 h). Lysis in phosphate-based passive lysis buffer was conducted at 4°C for 45 min rather than 20 min at room temperature. The TEV protease has low cleavage efficiency at higher salt concentrations, so the lysis buffer contained 100 mM NaCl, in contrast to the 300 mM NaCl included otherwise for lysis of *E. coli* expressing his-TEV–PEG10-AP constructs. Induction strength can sometimes affect the solubility of the expressed protein, so low-IPTG (0.05 mM) expression and solubility trials were conducted in parallel with usual conditions (0.3 mM IPTG). Western immunoblotting and Coomassie staining after SDS-PAGE, although not strictly quantitative, were used to assess expression and estimate the solubility of his– and his-TEV–PEG10-AP proteins.

Initially, the his-TEV–PEG10 AP constructs were assessed for expression and solubility. Unexpectedly, a specific antibody did not detect the his tag in the his-TEV–expression products. This has been observed before in the research group with this antibody and is not easily explained (Y. Shimaki, personal communication).\(^1\) Therefore a PEG10-ORF2 antibody, used in the experiments described in Chapter 4, was used to assess expression and solubility for these constructs in whole-cell *E. coli* lysates and post-lysis samples, respectively (Figure 5.2). The immunoblots were developed for 10 min, 5 min, 3 min, and 3 min for the right panels of Figure 5.2a, b, c, and d, respectively. A non-specific band with an apparent molecular weight near the size of the his-TEV–PEG10-AP clone products was observed in both pre- and post-induction *E. coli* lysates as well as in the soluble post-lysis fraction in the immunoblots. (A faint band in this position can also be seen in the post-lysis fractions shown in Figure 4.14.) His-TEV–PEG10-AP clone products 4 and 10 have apparent molecular weights greater than the non-specific band, 16 and 22 have apparent molecular weights less than, and 28 has an apparent molecular weight similar to the non-specific band. His-TEV–PEG10-AP clone products 4 and 10 were strongly immunoreactive against the

\(^1\)In addition to the DNA sequencing used to assess the N-terminus of all his-TEV–PEG10-AP constructs, mass spectrometric analysis of one selected product, corresponding to his-TEV–PEG10-AP clone 10, contained the expected N-terminus and the his tag epitope (Appendix VIIIa). This excludes the possibility that errors introduced during cloning resulted in the lack of immunogenicity for the his-TEV–PEG10-AP proteins.
PEG10-ORF2 antibody, although Coomassie staining implies that this may be related to the epitope profile of the polyclonal antibody rather than the relative expression of these proteins (clearly demonstrated in a comparison of Figure 5.2c left and right panels). The his-TEV–PEG10-AP clone 22 product, the smallest of the selected clones, had notably lower apparent expression than the other clones. Only the his-TEV–PEG10-AP clone 4 product, with an extended N-terminus, produced detectable soluble protein (red arrows, Figure 5.2), although the majority of the product was nonetheless insoluble under inductions with both low and medium IPTG inductions (Figure 5.2b and d).

When the selected clones were expressed with an N-terminal his-tag without the TEV-protease cleavage site, the primary his antibody was then effective in assessing expression and solubility without the non-specific band (Figure 5.3). The immunoblots were developed for 2 min, 30 s, 20 s, and 10 s for the right panels of Figure 5.3a, b, c, and d, respectively. Products induced with low IPTG could not be detected by Coomassie staining after SDS-PAGE (Figure 5.3a and b, left panels). As with the his-TEV–PEG10-AP expression trials, a wide dynamic range of intensity was observed across each Western immunoblot: the his–PEG10-AP clone 4 product was especially immunoreactive, while the his–PEG10-AP clone 22 product in some cases could barely be detected. Again, this pattern was not replicated in the Coomassie-stained profiles (clearly demonstrated in Figure 5.3d). Nonetheless, the intensity of the his–PEG10-AP clone 22 band was sufficient to determine that this clone is poorly soluble subsequent to cell lysis.

As had been seen for the his-TEV–PEG10-AP constructs, only the his–PEG10-AP clone 4 product, containing the longest N-terminus, was appreciably soluble post-lysis (Figure 5.3, red arrows). This was an unexpected finding, as the N-terminus for this clone begins at the site of the –1 frameshift element unprecedented for any known retroviral-type APs to date. The next-longest construct, the his–PEG10-AP clone 10 product, was barely soluble post-lysis.

Little difference in solubility was observed between low- and medium-strength IPTG inductions (Figures 5.2b and d and 5.3b and d). Although the majority of the his–PEG10-AP clone 4 product remained insoluble, a greater proportion was soluble compared with the his-TEV–PEG10-AP clone 4 product (Figures 5.3b and d and 5.2b and d). From these results, the his–PEG10-AP clone 4 (induced with 0.3 mM IPTG) was selected for expression and protein purification on a larger scale.

89
Figure 5.2: Expression and solubility trials of his-TEV- constructs. ‘0’ and ‘18’ indicate pre- and post-induction culture samples and ‘S’ and ‘I’ indicate soluble and insoluble post-lysis fractions, respectively. Left, Coomassie-stained SDS-PAGs; right, Western immunoblots with the PEG10-ORF2 antibody. a and b were induced with 0.05 mM IPTG, and c and d were induced with 0.3 mM IPTG. Asterisks and arrows indicate the expressed products. Red arrows indicate soluble product.
Figure 5.3: **Expression and solubility trials of his- constructs.** ‘0’ and ‘18’ indicate pre- and post-induction samples of *E. coli* culture, respectively; ‘S’ and ‘I’ indicate soluble and insoluble post-lysis fractions, respectively. Left, Coomassie-stained SDS-PAGs; right, Western immunoblots with the his antibody. **a** and **b** were induced with 0.05 mM IPTG, and **c** and **d** were induced with 0.3 mM IPTG. Asterisks and arrows indicate the expressed products.
5.2.1 Expression and purification of his–PEG10-AP clone 4

A medium-scale (100 mL) culture of bacteria expressing his–PEG10-AP clone 4 was grown, the recombinant protein expressed, cells lysed by sonication, and the his–PEG10-AP clone 4 protein purified over cobalt resin (see Section 2.6.7). Protein expression was high, although cell lysis was relatively poor compared to the passive lysis used for small-scale cultures (indicated by the presence of many E. coli proteins in the insoluble lane of Figure 5.4, left gel). Apparently complete binding of the his–PEG10–AP clone 4 product to cobalt resin was observed overnight. Elution of the his-tagged protein, however, required a very high imidazole concentration (250 mM). The protein concentration in eluted fractions was relatively low, as assessed by SDS-PAGE (Figure 5.4, right gel). A significant amount of his–PEG10-AP clone 4 protein apparently remained bound to the resin even after four elutions with high imidazole, perhaps reflecting unfolded, aggregated protein that pelleted with the resin throughout the ‘batch’ purification procedure (Figure 5.4 right gel, rightmost lane). Eluted his–PEG10-AP clone 4 protein, which has a predicted relative molecular mass of 18.6 kDa, migrated just below the 21 kDa molecular weight marker.

The imidazole was removed from elution fractions 1 and 2 by dialysis, and the protein concentration measured using a commercial fluorometric assay (Section 2.5.1). Post-dialysis, elutions 1 and 2 had a measured concentration of 208 µg.mL⁻¹ and 99.6 µg.mL⁻¹, respectively. Although these concentrations are relatively low for a protein eluted from an affinity resin, a large-scale culture (e.g. 1 L) could be used to produce the his–PEG10-AP clone 4 product in milligram quantities, especially if cell lysis is optimised.

5.2.2 His–PEG10-AP clone 4 does not cleave a heterologous anticipated substrate

A likely substrate for PEG10 AP, recombinant GST–Peg10-ORF1 (from mouse), was available to test proteolytic activity (see Section 2.7.3 for details). It was expected that this would be a substrate of the PEG10 AP from work done in cell culture implying that there is a cleavage within the Peg10-ORF1 (see Section 1.2.3 and Lux et al.,

---

2Unlike other gels containing purification fractions in this document, ‘equivalent loading’—in which the relative abundance of a protein can be assessed across lanes—was not used for Figure 5.4 due to the low concentration of elution fractions. Elution fractions are 5× more concentrated than wash and soluble fractions shown in Figure 5.4.

3Kindly gifted by Y. Shimaki.
2005). The ‘elution 2’ fraction from the above purification, while having a lower concentration than the first elution fraction, had fewer contaminant \( E. \text{coli} \) proteins that might contribute to background proteolysis, and so was selected for testing. No cleavage of the substrate by the purified his–PEG10-AP clone 4 product was observed at either pH 5.5 or pH 7.0 (Figure 5.5). This substrate is from the mouse Peg10, so the human PEG10 AP may lack activity against it. Alternatively, reaction conditions and enzyme/substrate concentrations may not have been optimal, or the his–PEG10-AP clone 4 product may simply not be active. The concentration of the his–PEG10-AP clone 4 product in the proteolysis reaction was too low to detect in Figure 5.5.

5.2.3 The his–PEG10-AP clone 4 product is a dimer in solution

The lack of observed proteolytic activity could be due to a failure of the his–PEG10-AP clone 4 product to dimerise, since this is necessary for activity of all known retroviral-type APs. A sample of the ‘elution 1’ fraction of his–PEG10-AP clone 4 was therefore applied to a Superdex 200 10/300 GL size-exclusion column (GE Healthcare, UK) to assess the oligomerisation state and test the appropriateness of using SEC as a purification step. Consistent with the reasonably high purity of the loaded fraction as assessed by SDS-PAGE, only one major peak, centred at approximately 15.75 mL
Figure 5.5: **His–PEG10-AP clone 4 does not cleave GST–Peg10-ORF1 in vitro.** A Coomassie-stained SDS-PAGE containing his–PEG10-AP clone 4 protein, dialysed into either pH 5.5 or pH 7.0 buffer, mixed with GST–Peg10-ORF1 (65 kDa), and allowed to incubate at room temperature for 0 h or 2 h. Each lane contains 2.5 µL of the proteolysis reaction.

Post-injection, was observed. The major peak was almost symmetrical, with only a slight rightward shoulder (Figure 5.6a). Fractions collected from this peak contained his–PEG10-AP clone 4 protein (Figure 5.6b), confirmed by mass spectrometry (Appendix VIIIb).

Three molecular weight markers, BSA (66.5 kDa), DNase I (30.1 kDa), and lysozyme (14.3 kDa) were used to calibrate the Superdex 200 10/300 GL column (elution profiles for these proteins are shown in Appendix IX). As a monomer, the his–PEG10-AP clone 4 product (predicted molecular mass of 18.6 kDa) would be expected to elute from the column at a volume between that of DNase I and lysozyme. By contrast, a dimeric his–PEG10-AP clone 4 product would be expected to elute at a volume between that of BSA and DNase I. The position of the major peak corresponding to the his–PEG10-AP clone 4 product unambiguously falls between the volumes for BSA and DNase I elutions, implying that the PEG10 AP was dimeric (Figure 5.6a). No appreciable peak between the DNase I and lysozyme elutions was observed, indicating that a monomer–dimer equilibrium strongly favours the dimeric form. The second fraction collected contained pure protein as assessed by Coomassie staining after SDS-PAGE, although a higher concentration of protein is required for better analysis of purity (Figure 5.6c).

To further assess the dimerisation state of the his–PEG10-AP clone 4 protein, a sample of the ‘elution 2’ fraction from the cobalt-resin purification (Figure 5.4) was subjected to a cross-linking process with glutaraldehyde (see Section 2.8 for methods). Proteins that are interacting have potential to be covalently linked by this method. A sample of the ‘elution 2’ fraction from cobalt-resin purification was also taken up in cracking buffer that lacked the reducing agent β-mercaptoethanol (i.e. a non-reduced sample). These samples were analysed by SDS-PAGE with Coomassie stain-
Figure 5.6: The his–PEG10-AP clone 4 protein migrates as a dimer in size-exclusion chromatography. The ‘elution 1’ fraction (500 µL) from cobalt-resin purification was loaded onto a Superdex 200 10/300 GL column. a Chromatogram of the separation. $A_{280}$ is shown in blue, $A_{215}$, in pink, is shown in arbitrary units for comparison of relative peak heights only. Arrows indicate the centre of peaks from monomeric BSA (66.5 kDa), bovine pancreatic DNase I (30.1 kDa), and hen egg white lysozyme (14.3 kDa; see Appendix IX for chromatograms). ‘0 mL’ is calibrated to the point that sample was injected into the column. b Fractions of 500 µL were collected, and 7.5 µL fractionated by SDS-PAGE before Coomassie staining. The leftmost lane after the marker contains 2 µL of elution 1 from the cobalt column. c The third elution fraction (25 µL) analysed by SDS-PAGE using a 1.5 mm-thick gel and stained with Coomassie.
Glutaraldehyde cross-linking and non-reducing SDS-PAGE suggest that the his–PEG10-AP clone 4 product is a dimer. The ‘elution 2’ fraction from Figure 5.4 (6 µL loaded) was subjected to cross-linking with glutaraldehyde (‘G’; 7.95 µL loaded) as described. The rightmost, non-reduced (‘– R’), lane contains a sample identical to lane E2, except no reducing agent was included in the cracking buffer when the sample was prepared. a Coomassie-stained SDS-PAGE. b Western immunoblot with PEG10-ORF2 primary antibody, exposed for 2 min. Asterisks indicate the expected size of a dimer of the his–PEG10-AP clone 4 product.
5.3 Conclusions

Clones containing each N-terminus and C-termini 4 were created with both his– and his-TEV– N-terminal tags. Strikingly, only the PEG10-AP clone 4 products, with the longest N-termini explored in this study, were appreciably soluble in either system. Solubility was far greater for the his– than the his-TEV– PEG10-AP clone 4 product, and on these grounds his–PEG10-AP clone 4 was selected for larger-scale expression and purification. Elution of the his-tagged protein from a cobalt resin was challenging but ultimately effective, and the purification scheme is predicted to be able to produce low-milligram quantities of protein when scaled up. Purified his–PEG10-AP clone 4 protein did not cleave a likely but heterologous substrate. Size-exclusion chromatography and glutaraldehyde cross-linking, however, indicate that the his–PEG10-AP clone 4 protein exists as a dimer in solution, which is necessary for activity and consistent with the structure of other retroviral-type APs.
Chapter 6

Discussion

PEG10 is a conserved mammalian gene with important normal functions in placental development and abnormal functions in some cancers. Unusually, a second ORF in PEG10 is expressed as an extension in a portion of ribosomal translational passages that undergo –1 frameshifting, and the second ORF contains a conserved retroviral-type AP domain that likely cleaves at one or several positions throughout PEG10 (Clark et al., 2007). The majority of functional work on PEG10 has focused exclusively on the first ORF (Tsou et al., 2003; Okabe et al., 2003; Lux et al., 2005). This study has aimed to express, purify, and analyse the PEG10 AP encoded in the second ORF both to augment our understanding of PEG10 biology and serve as an unusual example of a well-studied class of enzymes, the retroviral-type APs.

6.1 Sequence analysis indicates that the boundaries of the predicted PEG10 aspartyl protease domain are unlike any so far described

Several attributes innate to the PEG10 gene caused difficulty in determining the domain boundaries of the predicted AP domain by sequence analysis. PEG10’s restriction to mammals, coupled with high conservation at the protein level, diminished the utility of multiple sequence alignments. The highly conserved –1 frameshift element near the predicted N-terminus of the AP domain further confounded domain analysis based on conservation.

Previously, the closest relative to PEG10 had been described as the Sushi-ichi LTR-retrotransposon from Takifugu rubripes (Butler et al., 2001; Youngson et al., 2005). Search for homologues of PEG10 using profile hidden Markov models revealed sev-
eral predicted LTR-retrotransposons from vertebrates that were also predicted to be homologues of PEG10 with high confidence (Table 3.1). A tree-based analysis of V-clade Ty3/Gypsy LTR-retrotransposons with the addition of PEG10 and RTL1 may help elucidate the lineage of these co-opted retroelements in mammals. Notably, the mammalian parasites Schistosoma japonicum and Arthroderma gypseum, also host sequences similar to PEG10 and hence are potential vectors for horizontal transmission of Ty3/Gypsy LTR-retroelements as proposed by Poulter and Butler (1998). As little is known about these elements, however—there are few functional and no structural studies of AP domains from Ty3/Gypsy LTR-retrotransposons—little information about the domain boundaries of the predicted PEG10 AP was gathered from these analyses.

Sensitive structural alignment of PEG10 with structurally defined homologous domains showed that PEG10 contained sequence motifs characteristic of retroviral-type APs, bolstering the hypothesis that PEG10 contains an active AP domain (Figure 3.4). Such analyses did not generate reliable alignment of the termini, however, consistent with the diverse range of terminal structures found throughout retroviral-type APs (Section 1.1.2.2). Interestingly, the closest homologue of the predicted PEG10 AP with a defined structure was predicted to be from Ddi1, a cellular gene found throughout eukaryotes (Section 3.1.3). The predicted PEG10 AP domain therefore diverged from its closest structurally characterised homologue sometime around the emergence of Eukaryota (Krylov and Koonin, 2001). This underscores the evolutionary distance between PEG10 and other retroviral-type APs that have served to inform the predictions made, and techniques used, throughout this study. Such evolutionary distance also indicates that PEG10 AP may be useful as an early branch enzyme in analysing retroviral-type APs, a class of proteins intensively studied in the context of drug development.

6.1.1 The predicted APs of PEG10 and vertebrate LTR-retrotransposons are enriched in proline

No plausible homology model could be generated for the predicted PEG10 AP using sensitive automated homology modelling pipelines (Section 3.1.3). In some instances, this appeared to result from the high molar percentage of proline residues within the predicted PEG10 AP (Figure 3.2). Although proline is typically thought of as a destabilising amino acid, analysis of predicted AP domains from LTR-retrotransposons revealed that vertebrate sequences were enriched for proline compared to the closest (fungal) homologues (Figure 3.3). These elements have conserved AP domains that
are very likely required for replication of the element, and proline enrichment therefore has not caused a loss of proteolytic activity in these domains. The AP domains of vertebrate LTR-retrotransposons and PEG10 may represent a somewhat novel fold among retroviral-type APs.

Despite several interesting observations emerging from sequence analysis of the predicted PEG10 AP, no confident predictions of domain boundaries could be made. Several low-confidence domain boundaries were therefore selected based on sequence alignments against other retroviral-type APs, indels within mammalian PEG10 sequences and domain predictions. Predicted termini were cloned combinatorially with each other for functional analysis.

6.2 Expressing the predicted PEG10 AP as a soluble protein in E. coli

For many retroviral APs, expression into inclusion bodies, followed by solubilisation in a strongly denaturing buffer and gentle refolding, has been the standard protocol for generating material for crystallisation and functional studies (Gulnik et al., 1995; Gilski et al., 2011; Hartl et al., 2011). Restriction of the retroviral APs to inclusion bodies during expression avoids toxicity resulting from promiscuous proteolytic activity against E. coli proteins (Baum et al., 1990). Previous work, however, noted that a clone of the predicted PEG10 AP is largely soluble when expressed in fusion with MBP, and so soluble expression was pursued in this study (Crowe-McAuliffe, 2010). The lack of apparent toxicity of PEG10 AP against E. coli observed in this study could result from the production of an inactive protease. This explanation is unlikely, however, given the autoproteolysis of MBP–PEG10-AP observed during expression. Instead, the absence of toxicity during expression in E. coli may instead represent significant narrowing of the substrate specificity of the PEG10 AP compared to its retroviral counterparts. Such a narrowing of substrate specificity in the PEG10 AP would be consistent with the role of LTR-retrotransposons as endogenous parasites under little selective pressure to disrupt cellular function, whereas retroviruses use proteolysis of host proteins to trigger cell death and viral spread (Section 1.1.3). A convergent narrowing of substrate specificity in both the PEG10 AP and Ddi1, both cellular retroviral-type APs, may explain the relatively high sequence identity between these two proteins.

The enrichment of proline in the PEG10 AP also disfavours a refolding method.
Proline exists in a random mixture of cis-trans isomers in unfolded proteins, and cis-trans isomerisation is a slow, rate-limiting step in protein folding that occurs on the timescale of seconds in the absence of enzymatic enhancement (Wedemeyer et al., 2002). When available, soluble production in *E. coli* is usually favoured over refolding from inclusion bodies due to potential low yields of refolded protein and many different refolding methods to explore (Gräslund et al., 2008). The his-tagged constructs examined in this study can be expressed in an insoluble form at 37°C if refolding of the PEG10 AP becomes desirable in the future.

### 6.2.1 MBP–PEG10-AP fusion proteins undergo autoproteolysis during expression

The observation of apparent autoproteolysis in all clones during expression was an important initial observation, but proteolytic activity by the PEG10 AP could not be demonstrated *in vitro* after purification. Remarkably, Fehér et al. (2006) observed a similar phenomenon when expressing the murine leukemia virus (MLV) retroviral AP domain in fusion with MBP. The MLV AP was apparently active and cleaved the MBP fusion partner during expression, dependent on an intact catalytic aspartic acid residue. As in the present study, this cleavage event was not observed after purification when purified MLV AP was mixed with purified MBP. Fehér et al. (2006) hypothesised that the MLV AP was activating a cascade of *E. coli* proteases to cause the cleavage event but did not present any data to support this hypothesis. Activation of a proteolytic cascade within *E. coli* is unlikely to account for the unusual autoproteolysis observed in this study, because incubating partially purified MBP–PEG10-AP clone 4 protein with *E. coli* lysate did not recover the autoproteolytic activity (Figure 4.9).

The reasons for the discrepancy between proteolytic activity of MBP–PEG10-AP fusion proteins *in vivo* during and *in vitro* after expression in *E. coli* remain unclear. Richarme (1982, 1983) reported that native periplasmic MBP is largely dimeric, and becomes monomeric only in the presence of maltose. However, I was unable to replicate these findings in the case of recombinantly expressed MBP in the protoplasm. Because retroviral-type APs require homodimerisation for function, even weak induced dimerisation by MBP during expression may explain the difference between apparent autoproteolytic activity during expression and the inactive state after purification. Alternatively, a very high local concentration of protein during protein synthesis may induce cleavage of substrates that are unfavourable at concentrations achievable post-purification. This latter model would explain both the results presented in Chapter 4.
and those of Fehér et al. (2006). Alternatively, the PEG10 AP may be unstable and only transiently folded in an active conformation after translation. In this model, the unfolded PEG10 AP is ‘dragged’ into solution by the highly soluble MBP.

6.2.2 Do MBP and the PEG10 AP interact?

Co-elution of MBP and PEG10-AP clone 4 protein over two chromatography columns implied a non-covalent, non-specific interaction between the two proteins. An N-terminal MBP fusion partner can act as a ‘passive chaperone’ to help fold proteins of interest. The passive chaperone activity of MBP requires a surface-exposed hydrophobic patch, which may serve as a binding site for folding intermediates of the fusion protein of interest (Fox et al., 2008). In the event that the fusion protein was unstable, non-specific interactions between the MBP hydrophobic patch and unfolded intermediates may continue post-expression. Limited evidence for such an association is presented in Figure 4.20.

Regardless of the reason for an apparent association between MBP and the PEG10 AP, the resulting purification scheme was laborious and inefficient. Further analysis of the apparent interaction between MBP and the PEG10-AP clone 4 protein was not considered relevant to the aims of this project and so was not explored further.

6.2.3 Expression and purification of his-tagged PEG10 AP

Utilising information about C-termini that can be expressed intact in E. coli, five clones were selected for expression with either only a his-tag, or with a his-TEV tag. Inclusion of the TEV-protease cleavage motif resulted in decreased solubility compared to the his-tag only, which precluded the use of a convenient purification system, although this is consistent with published data (Kurz et al., 2006). Unexpectedly, only one of the five constructs, his–PEG10-AP clone 4, yielded soluble protein in this system (Figure 5.3). PEG10 AP clone 4 contains N-terminus 1, which begins at the –1 frameshift site. This terminus was the longest examined in this study, and was selected as a positive control which was thought likely to contain the PEG10 AP domain, but unlikely to accurately represent the N-terminal domain boundary of the predicted PEG10 AP (Section 3.1.5 and Figure 3.6). By contrast, the next-longest clone, his–PEG10-AP clone 10, yielded product that was nearly totally insoluble. Intriguingly, the PEG10-AP clones 4 and 10 products differ only by nine mixed hydrophobic and hydrophilic amino acids, GLKPGPAVE, at the N-terminus.
The **PEG10** –1 frameshift element spans approximately 42 nucleotides and consists of a heptanucleotide slippery sequence, a short linker, and a pseudoknot (a type of structured RNA element; Manktelow *et al.*, 2005). This sequence, which is highly conserved at the RNA level, encodes most of the N-termini 1 and 2 used in this study. This implies that the protein sequence at the N-terminus of PEG10 AP clones 4 and 10 has been evolutionarily constrained by RNA structure as well as acting as the C-terminus of ORF1. Most retroviruses contain AP domains that begin downstream of the RNA recoding element, perhaps to avoid such constraints. Expression of proteins with N-terminus 1 was not apparently affected by the presence of the pseudoknot, which could have acted as a barrier to translation by the ribosome (e.g. Figure 5.3a and c).

The PEG10 amino acid sequence used in this study corresponds to the sequence deposited in the NCBI protein database. Notably, however, the identity of the lysine in the second position of clone 4 (K396 in NP_001165908.1) is unclear and is dependent on the mechanism of –1 frameshifting during translation. If –1 frameshifting occurs after peptidyl transfer, when N is added to the C-terminus of the nascent PEG10 polypeptide, then the sequence of N-terminus 1 would begin GNLPGPAVE, in contrast to the GKLPGPAVE used in this study. If frameshifting occurs prior to peptide bond formation, then the asparaginyl-tRNA with a GUU anticodon in the ribosomal A site would be paired with the unfavourable AAA codon, and therefore may be subject to ribosomal proofreading mechanisms and replaced with a kinetically favourable lysyl-tRNA (Figure 6.1). Many models of –1 frameshifting have been proposed in the literature, at least one of which—the E/P site post-translocational model—is consistent with the presence of N, rather than K, at the second position of the PEG10 AP N-terminus 1 used in this study (Horsfield *et al.*, 1995; Graves, 2005; Brierley *et al.*, 2010). Given the significance of the N-terminal sequence in solubilising the his–PEG10-AP clone 4 protein, resolving this sequence ambiguity, perhaps using targeted mass spectrometry, is important prior to subsequent work.
6.2.4 Does his–PEG10-AP undergo autoproteolysis?

Despite the evident self-cleavage for the MBP–PEG10-AP constructs, no evidence for autoproteolysis was observed during the expression and purification of his-TEV- and his-tagged PEG10 AP constructs (Figures 5.3 and 5.4). Any fragments resulting from autoproteolysis may have been too small and/or unstable in *E. coli* to be detected by Coomassie-staining or Western immunoblot after SDS-PAGE. No evidence for autoproteolysis of the his–PEG10-AP clone 4 product was observed during purification, although this was performed at 4°C and at neutral pH. Results from Fehér *et al.* (2006), discussed above, caution against assuming that autoproteolysis will occur in the his-tagged PEG10 AP proteins simply because of results from MBP-fusion proteins during expression. Nonetheless, experimental exploration of post-purification autoproteolysis, perhaps utilising urea-SDS-PAGE suited to the analysis of small proteins and peptides, may be a tractable and informative path for future work.

It is possible that soluble his–PEG10-AP rapidly undergoes autoproteolysis, explaining the apparent lack of soluble induced protein for most his–PEG10-AP clones. In this model, the extended N-terminus of the his–PEG10-AP clone 4 product would serve to inhibit proteolytic activity and explain why protein expressed from this clone appears intact after cell lysis. Coomassie-stained SDS-PAGE analysis of post-expression lysates from his–PEG10-AP clones with an inactive NSG active site triad, however, showed no appreciable difference in apparent expression, arguing against such a hypothesis.

The HIV-1 AP has been widely studied using convenient autoproteolysis-resistant variants (Mildner *et al.*, 1994). Identification of cleavage sites and generation of equivalent stable variants for the PEG10 AP may be a useful tool for future work.

6.3 Does the PEG10 AP contain a disulphide bond?

A non-reduced sample of PEG10 AP ran largely in monomeric and apparent dimeric forms on SDS-PAGE (Figure 5.7). Minor higher-order oligomers were also observed, although these were relatively scarce and were not evident as major species in SEC (Figure 5.6). The apparent dimer ran at approximately the rate expected of a linear polypeptide corresponding to the relative molecular mass of a dimer. This observation is consistent with the location of the two cysteine residues within the his–PEG10-AP clone 4 product, which are near the C-terminus. The predicted role of the C-terminus as a dimerisation interface is also bolstered by the observation of an intermolecular
disulphide bond in Figure 5.7. Mass spectrometric analysis may be useful in identifying which cysteine is primarily involved in forming the disulphide bond. The most N-terminal cysteine (C546 in NP_001165908.1) is conserved between PEG10 and related LTR-retrotransposons and therefore is a candidate for functional importance, noted in Figure 3.1.

Canonically it would be considered unlikely that the PEG10 AP forms a disulphide bond within the reducing mammalian cytoplasm. There is evidence that disulphide bonds can form specifically and dynamically in some cytoplasmic proteins in response to oxidative stress, however, and redox-sensitive glutathionylation is a well-characterised post-translational modification (Cumming et al., 2004; Giustarini and Rossi, 2004). Oxidative stress occurs during the normal development of the placenta—an organ with a high level of PEG10 expression—at the end of the first trimester, when trophoblast invasion causes perfusion of oxygenated blood after 8–9 weeks of relative ischemia (Myatt and Cui, 2004). This reperfusion period roughly coincides with increased PEG10 expression (Section 1.2.2).

There is also precedent for cysteine modification, an oxidative process, as a regulatory factor in retroviral-type APs in addition to the glutathionylation of retroviral APs that regulates proteolytic activity (Section 1.1.2.3). Yeast DDI1—which, like PEG10, contains a retroviral-type AP domain with C-terminal cysteines that can form an intermolecular disulphide bond when expressed in E. coli—is transcriptionally upregulated in response to oxidative stress (Sirkis et al., 2006). The Maggy and Pyggy fungal Ty3/Gypsy LTR-retrotransposons contain a predicted AP domain with a conserved DCG active site triad (Butler et al., 2001). More concretely, Zábranská et al. (2007) found that an intramolecular disulphide bridge, between N- and C-terminal cysteines, stabilised the monomeric form of the Mason-Pfizer monkey virus AP monomer. The closest N-terminal cysteine in the PEG10 protease is part of a CCHC-type zinc finger that is translated prior to the frameshift element, however, and hence is unlikely to contribute to the N-terminus of the AP domain.

In summary, results presented in Figure 5.7 combined with existing literature are insufficient to conclude whether or not the PEG10 AP contains a disulphide bond under native conditions. Precedence for widespread redox regulation of retroviral-type APs combined with the high conservation of a C-terminal cysteine residue in the predicted PEG10 AP, however, hint at some importance of this structure. Strains of E. coli that have been modified to have a relatively oxidising protoplasm conducive to the formation of disulphide bonds are commercially available. Expression of his–PEG10-AP clone 4 in such a strain, followed by analysis of the soluble/insoluble recombinant
protein ratio and non-reducing SDS-PAGE, may provide information about whether disulphide formation enhances folding of the predicted PEG10 AP.

6.4 Evidence in favour of a dimeric his–PEG10-AP

In addition to the C-terminal intermolecular disulphide bond, both SEC and glutaraldehyde cross-linking implied that the his–PEG10-AP clone 4 product is a dimer in solution (Section 5.2.3). Importantly, glutaraldehyde cross-linking was performed at a low protein concentration (∼0.2 mg.mL⁻¹) to impede covalent linkage of non-interacting proteins. No evidence of higher-order oligomers was observed after glutaraldehyde cross-linking, which additionally implies that the coupling was not a nonspecific process (Figure 5.7). Alone, none of these lines of evidence conclusively demonstrate that the his–PEG10-AP clone 4 product is a dimer. By parsimony, however, it is unlikely that all three supporting results are artefactual. Nonetheless, further validation of the oligomeric state, for example by multi-angle light scattering-coupled SEC, is a potentially valuable future experiment.

Notably, the apparent dimer observed after glutaraldehyde cross-linking had the same apparent relative molecular mass of the dimer formed by terminal disulphide cross-linking (Figure 5.7, discussed above). If glutaraldehyde cross-linking were occurring near the centre of the primary sequence of a protein, it would be expected that the resulting branched polypeptide would have a perturbed electrophoretic mobility compared to the non-reduced dimer, which is a pseudo-linear terminally-linked protein.

Only two lysines, substrates for glutaraldehyde cross-linking, are found in the his–PEG10-AP clone 4 protein. One is near the N-terminus, immediately after the his-tag. The other is situated within the predicted flap region of the enzyme, and is conserved in both the Ddi1 and HIV-1 AP (Figure 3.4). Arginine–lysine conjugates have recently been described after glutaraldehyde cross-linking, and it is notable that the his–PEG10-AP clone 4 product is arginine-rich, with ten arginine residues throughout the protein, including two at the C-terminus (Salem et al., 2010). A tentative conclusion that glutaraldehyde is cross-linking near the termini of the his–PEG10-AP clone 4 product can thus be drawn.

Most retroviral-type APs require dimerisation for stability in solution. The dimerisation state of the PEG10 AP may therefore be used as a proxy for testing the folded state of the protein, for it is unlikely that a soluble but mis-folded protein would happen to specifically dimerise yet not form higher-order oligomers. Confirmation of the
dimerisation state of the PEG10 AP could thus complement commonly used techniques such as circular dichroism to assess the folded state of a purified protein.

6.5 Searching for activity of the predicted PEG10 aspartyl protease domain

An unmet aim of this study was to establish an activity assay for the PEG10 AP. An activity assay would facilitate biochemical examination, allow the establishment of an inhibition profile, and confirm the folded state of the protein for further biophysical studies. Retroviral-type APs usually act within the constrained environment of a nascent viral- or viral-like particle and cleave substrates in an ordered manner (Section 1.1.2). Nonetheless, in vitro activity assays have been widely used for this class of enzymes.

Reviewing the substrate specificity of retroviral AP domains, Tözsér (2010) noted that “...no generally accepted and widely used standard assay conditions have been established in the retroviral protease field, and the [protease] is highly sensitive to the pH, ionic strength and type of the substrate (e.g., protein versus peptide, or the presence of ionizable side chains such as Glu)...” Such inconsistency has prevented quantitative comparison of data between studies, and in some cases unusual buffering conditions have lead to artefactual results. For example, the AP domain from the foamy virus, an early-branching retrovirus, was found to be monomeric and active only in vitro in 2–3 M NaCl (Hartl et al., 2008). It was subsequently discovered that a C-terminal extension of the foamy virus AP domain contained RNA-binding motifs, and that binding of the foamy virus genomic RNA was required to trigger AP dimerisation and activity under more moderate conditions (Lee et al., 2011b; Hartl et al., 2011).

Selecting conditions for activity assays of the PEG10 AP is therefore not trivial, and in this study only a small number of parameters has been investigated.

In cultured mammalian cells, Clark et al. (2007) detected both full-length PEG10 and cleavage fragments by Western immunoblot. When cells were transfected with a variant PEG10 that was predicted to abolish proteolytic activity, only the full-length forms of PEG10 were evident. These observations strongly indicate that the PEG10 polypeptide itself is a substrate for the PEG10 AP domain, perhaps in an intramolecular reaction in vivo. Recombinant, purified fragments of PEG10 are therefore appealing test substrates when searching for activity of the PEG10 AP domain.

Several potential sites of cleavage within PEG10 can be deduced from existing
data. Lux et al. (2005), using an N-terminal epitope, observed an immunoreactive species with an apparent relative molecular mass less than PEG10-ORF1 alone. From the apparent relative molecular masses of the immunoreactive species, it can thereby be concluded that a cleavage event occurs approximately near the centre of PEG10 ORF1 primary sequence. Consistent with this observation, Clark (2007) used mass spectrometry to identify a partial PEG10 ORF1/2 fusion protein that was missing sequence from the N-terminus.

Expression and purification of soluble human PEG10 has previously been problematic, and no human PEG10-ORF1 was available to test as a substrate in this study (Clark, 2005). An alternative anticipated substrate, a soluble fusion protein of GST and mouse Peg10-ORF1 was tested, but no cleavage was observed (Figure 5.5). The heterologous nature of this substrate or reaction conditions (such as buffer components—notably a lack of potassium ions—and enzyme/substrate concentration) are both plausible reasons for the lack of observed activity. Alternatively, the purified his–PEG10-AP clone 4 product simply may not be an active protease. Using known peptide substrates of related enzymes, in this case from the Ty3 LTR-retrotransposon, or searching for substrates from human cell lysates are potential strategies to overcome the problems encountered in Section 5.2.2 (Kirchner and Sandmeyer, 1993). As discussed above, the substrate specificity of the PEG10 AP is likely to be narrower than for retroviral APs, possibly confounding the search for a substrate of the PEG10 AP.

Another cellular retroviral-type AP, Ddi1, has been extensively characterised in the absence of an activity assay. For example, genetic studies determined Ddi1 to be involved in cell-cycle checkpoint control and protein secretion in yeast, both requiring an intact predicted AP active site (Gabriely et al., 2008; White et al., 2011a). Even sensitivity to HIV PIs was established in a yeast system by observing a cell phenotype rather than a specific proteolytic event with Ddi1 and mutants (Section 1.1.5 and White et al., 2011b). Thus, an activity assay, although desirable, is not strictly essential for further biophysical characterisation of the PEG10 AP.

6.6 Possible interactions between PEG10 and cellular retroviral-type aspartyl proteases

The predicted PEG10 AP domain is one among several known instances of predicted retroviral-type APs in mammalian genomes, as well as at least 22 additional predicted domains from retroelements that have been annotated in the MEROPS database of pro-
teases (Section 1.1.5). Retroviral-type APs in retroelements usually only become active within the confined environment of a nascent viral or virus-like particle. Could cellular retroviral-type APs, diffusing freely in the cytoplasm, heterodimerise to form hybrid enzymes, or some serve as inhibitors to others? The homology between retroviral-type APs and a single domain of bilobed eukaryotic APs serves as a proof of principle that heterodimerisation is possible in this class of enzymes. There is also precedent for cross-regulation of proteolytic pathways by a network of genetically related membrane proteases (Zettl et al., 2011). Pulldown assays using human cell lysates, perhaps utilising the cobalt-binding capability of the his–PEG10-AP clone 4 protein, could aid in testing the hypothesis that heterodimerisation may occur between retroviral-type APs in eukaryotes.

6.7 Safety of HIV-protease inhibitors during pregnancy

The safety of HIV PI use during pregnancy is disputed, and all PIs currently on the market have been given a pregnancy category of either B or C\(^1\) by the US FDA. Anti-retroviral therapy during pregnancy is vital for preventing transmission of HIV between mother and child, and is generally well-tolerated. Nonetheless, there is some controversy in the literature regarding an association between the use of PIs during pregnancy and pre-term birth, with many studies arguing both for and against an association. In a recent review, Andany and Loutfy (2013) note that many studies do not differentiate between particular PIs, although cohorts taking lopinavir in particular seemed more likely to report an association between PIs and pre-term birth. Higher rates of other adverse pregnancy outcomes (such as preeclampsia) have been reported for women taking combination therapies of anti-retroviral agents, although again these findings are disputed (Suy et al., 2006; Chen et al., 2012).

If a positive association indeed exists between pregnancy disorders and the use of HIV PIs, no obvious molecular mechanism is present and the retroviral-type APs of PEG10 and RTL-1, both required for placental development in mice, are thereby plausible mediators of any such association (Sekita et al., 2008; Renfree et al., 2009). Using an in vitro activity assay to test clinically used PIs against the predicted PEG10 and RTL-1 APs would be a first step in testing this hypothesis. Immunoblot analyses of

\(^1\)Category B: “Animal reproduction studies have failed to demonstrate a risk to the fetus and there are no adequate and well-controlled studies in pregnant women.” Category C: “Animal reproduction studies have shown an adverse effect on the fetus and there are no adequate and well-controlled studies in humans, but potential benefits may warrant use of the drug in pregnant women despite potential risks.”
placenta samples from women taking PIs throughout pregnancy and matched controls may also be informative.

6.8 HIV-protease inhibitors as anticancer agents

There has recently been interest in exploring the anti-cancer effects of HIV PIs, which have been well noted against AIDS-associated Kaposi’s sarcoma, and treatment of HIV-negative patients with Kaposi’s sarcoma using HIV PIs has yielded some initial successes (Section 1.1.4, Chow et al. 2009; Monini et al. 2009). Known molecular targets of HIV PIs in this context include the proteasome and some metallomatrix proteases (MMPs), and other cellular effects of HIV PIs seem to be largely downstream of these targets (Piccinini et al., 2005; Sgadari et al., 2002; Monini et al., 2004). For example, dephosphorylation of the signalling-hub protein Akt is ultimately triggered by the unfolded protein response, which itself occurs as misfolded proteins accumulate in the endoplasmic reticulum due to the lack of proteasome activity (Gupta et al., 2007). MMPs -2 and -9 are required for extracellular tissue remodelling, angiogenesis, and tumour invasion, both in the context of neoplasias and during normal organismal development, such as placental invasion into the uterine endothelial stroma (Foda and Zucker, 2001; Onogi et al., 2011).

The proteasome and MMPs are genetically unrelated proteases with distinct mechanisms of proteolysis compared to APs. It might appear unusual, then, that HIV PIs, which are poor inhibitors of cellular APs, would bind to and affect these targets by chance alone. It has been shown, however, that proteases universally recognise substrates in extended β-sheet conformation in their active site (Madala et al., 2010). As HIV PIs are peptide mimetics locked into a β-sheet-like conformation with a non-hydrolysable group in place of the peptide bond, interaction with other proteases is perhaps not surprising.

Interestingly, in vitro work in cultured cells has shown that the PEG10 ORF1 is a positive regulator of both MMP mRNA and protein (Xiong et al., 2012). MMPs are positive regulators of TGF-β signalling, and so PEG10 modulation of MMP expression may therefore act synergistically with PEG10 modulation of TGF-β signalling via the ALK1 surface receptor (Vu, 2001; Lux et al., 2005). This work is also consistent with other known functions of PEG10, as MMPs are essential for formation of the labyrinth layer of the placenta (Szabova et al., 2010).

In contrast to the reduction of invasiveness observed in many tumours, the HIV-PI nelfinavir has been shown instead to promote apoptosis and cause cell cycle arrest in
hepatocellular carcinoma-derived cell lines (Sun et al., 2012). No molecular mechanism for this phenomenon was proposed. Notably, Tsou et al. (2003) and Okabe et al. (2003) reported PEG10 overexpression and, specifically, involvement in resistance to apoptosis and cell cycle progression in hepatocellular carcinoma. Two HIV PIs, ritonavir and saquinavir, have also been shown to inhibit proteolysis by MMPs in cervical intraepithelial neoplasias. Importantly, the MMP mRNA levels were also diminished in these cells (Barillari et al., 2012). Although the AP domain of PEG10 plays an as-yet unknown role in PEG10 regulation and biology, inhibition of the PEG10 AP leading to a decrease in MMP expression by disruption of PEG10 function is a plausible model of HIV PIs active in this system (Figure 6.2). In vitro activity assays combined with transfections in a cell culture system could be used to examine the hypothesised interaction between HIV-PIs, the PEG10 AP, and MMPs.

6.9 Future work

Production of a soluble, pure his–PEG10-AP clone 4 protein has been the major experimental output of this project. Biophysical characterisation of the his–PEG10-AP clone 4 product to identify the dimerisation state and predict secondary structure is an immediate next step. Although the dimerisation observed in this study implies that the purified his–PEG10-AP clone 4 product is properly folded in solution, circular dichroism could be used to assess the folded state and provide information about secondary structure, which is expected to be highly enriched in \( \beta \)-sheet character. Concentrating the pure his–PEG10-AP clone 4 protein is necessary for any subsequent structural analyses, and if homodimerisation is a requirement for proper folding, stability across a range of concentrations would be expected to have some first-order character.

Figure 6.2: The hypothesised inhibition of PEG10 AP by HIV PIs is consistent with existing literature. A model derived from the results of Xiong et al. (2012) and Barillari et al. (2012). The dotted lined indicates the hypothesised inhibition of the PEG10 AP by HIV PIs.
Identifying a substrate and developing an activity assay would confirm that the PEG10 AP is properly folded for further biophysical studies and facilitate the identification of cleavage sites along the full-length PEG10 polypeptide. Confident identification and characterisation of PEG10 cleavage fragments may resolve conflicting reports of PEG10 function and activity (e.g. subcellular localisation) and provide insights into retroviral-type APs as a class of enzymes. Probing the inhibition profile of the PEG10 AP, beginning with a broad-range inhibitor of APs, could be used to test the hypothesis that the PEG10 AP is a target of HIV PIs that are widely used in clinical practice.

6.10 Conclusion

Although PEG10 has been implicated in important processes such as placental development and oncogenesis, there is a poor understanding of this protein at a molecular level, and so far little work has been conducted into the conserved predicted AP domain of PEG10. Site-directed mutagenesis experiments presented in this study and elsewhere strongly imply that this domain is active, but no direct observation of proteolytic activity has yet been described. The promiscuous inhibition of diverse proteases by HIV PIs, combined with the pleiotropic effects of these drugs on processes with known PEG10 involvement, make the development of an activity assay for the PEG10 AP an especially attractive prospect. The results presented in this study—namely the purification and initial characterisation of the isolated PEG10 AP domain—will facilitate the characterisation of an unusual member of a well-studied class of enzymes, the retroviral-type APs, and aid in elucidating the function of PEG10 in human biology.
References


Clark, M. (2005) Expression and purification of the two proteins encoded by Paternally Expressed Gene 10 for polyclonal antibody production. BSc (Hons) thesis, University of Otago, New Zealand


Crowe-McAuliffe, C. (2010) *Does the human gene PEG10 have a functional protease domain?* BSc (Hons) thesis, University of Otago, New Zealand


Merkulov, G., Lawler, J., Eby, Y., and Boeke, J. (2001) Ty1 proteolytic cleavage sites are required for transposition: all sites are not created equal. *J. Virol.* **75**, 638–644


tammar wallaby SIRH12, derived from a marsupial-specific retrotransposition event. 
*DNA Res.* **18**, 211–219


123


Appendices

I. Materials and solutions

Materials

Unless otherwise stated, all chemicals used were ‘analytical’ or ‘reagent’ grade.

*Acros Organics, USA*
- Ponceau S

*Kock-Light, UK*
- Bromophenol Blue

*Ajax Finechem Pty Ltd, Australia*
- Orthophosphoric acid (85% v/v)

*Merck KGaA, Germany*
- Imidazole
- Peptone from casein
- Yeast extract granules

*AppliChem GmbH, Germany*
- Glycine
- SDS
- TEMED
- Tris

*Roche Diagnostics GmbH, Germany*
- DTT

*Scharlab SL, Spain*
- Acetic acid (glacial)
- Ammonium sulphate
- EDTA
- \( \text{H}_2\text{O}_2 \)
- HCl
- \( \text{KH}_2\text{PO}_4 \)

*Bio-Rad Laboratories Inc., USA*
- 30% Acrylamide/Bis solution (37.5:1)
- Broad-range molecular weight markers

*Invitrogen Corporation, USA*
- Ultrapure agarose
- NaCl
- NaOH
- NaH\(_2\)PO\(_4\)
- Urea
Solutions

All solutions were prepared using ddH$_2$O prepared in a Milli-Q reference apparatus (Thermo Fisher Scientific, USA) unless otherwise noted.

Gel electrophoresis and Western blotting

**Sodium-borate (SB) buffer, 20x stock**
- 0.8% w/v NaOH
- pH adjusted to 8.0 with boric acid (>50g)

**PAGE running buffer (inner)**
- 25 mM Tris
- 190 mM Glycine
- 0.2% (w/v) SDS

**DNA loading buffer**
- 7 M Urea
- 125 nM EDTA
- 25% w/v Sucrose
- 0.5% w/v Bromophenol blue

**Cracking buffer**
- 62.5 mM Tris
- 1 M Urea
- 25% (v/v) Glycerol
- 1% (w/v) SDS
- 2% (v/v) β-mercaptoethanol pH adjusted to 8.0 with HCl

**Ponceau S stain**
- 0.1% (w/v) Ponceau S
- 5% (w/v) Acetic acid

**PAGE buffer (outer)**
- 25 mM Tris
- 190 mM Glycine

**Separating gel buffer**
- 750 mM Tris
- Adjusted with HCl to pH 8.8
Stacking gel buffer

- 125 mM Tris
- 0.1% (w/v) SDS
- Adjusted with HCl to pH 6.8

Western-transfer buffer

- 25 mM Tris

TBST

- 40 mM Tris
- 150 mM NaCl
- 0.5% (v/v) Tween 20
- Adjusted with HCl to pH 7.6

Cell lysis and chromatography

Unless otherwise stated, phosphate buffers were made from mixing 1 M stocks of Na$_2$HPO$_4$ and NaH$_2$PO$_4$ in ratios indicated by Sambrook et al. (1989) for the desired pH. The pH was subsequently checked with a pH meter at the appropriate temperature, and the pH was adjusted with HCl or NaOH as necessary.

**Tris-based lysis buffer**

- 20 mM Tris
- 200 mM NaCl
- 50 µg.mL$^{-1}$ Lysozyme
- 25 µg.mL$^{-1}$ DNase I
- 25 µg.mL$^{-1}$ RNase A
- Adjusted with HCl to pH 7.2

**Phosphate-based lysis buffer**

- 50 mM Phosphate, pH 7.0
- 300 mM NaCl
- 10 mM Imidazole
- 50 µg.mL$^{-1}$ Lysozyme
- 25 µg.mL$^{-1}$ DNase I
- 25 µg.mL$^{-1}$ RNase A
- Adjusted with HCl to pH 7.2

**Amylose resin binding buffer**

- 20 mM Tris
- Adjusted with HCl to pH 7.2

**Amylose resin wash buffer**

- 20 mM Tris
- 1 M NaCl
- Adjusted with HCl to pH 7.2

**Amylose resin elution buffer**

- 20 mM Tris
- 10 mM Maltose
- Adjusted with HCl to pH 7.2

**Low-salt phosphate buffer**

- 8 mM Na$_2$HPO$_4$
- 1 mM KH$_2$PO$_4$
- 2 mM DTT
- Adjusted with HCl to pH 7.3

**High-salt phosphate buffer**

- 8 mM Na$_2$HPO$_4$
- 1 mM KH$_2$PO$_4$
- 1 M NaCl
- 2 mM DTT
Adjusted with HCl to pH 7.3

<table>
<thead>
<tr>
<th><strong>Phosphate wash/binding buffer</strong></th>
<th><strong>Phosphate–imidazole elution buffer</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mM Phosphate, pH 7.0</td>
<td>50 mM Phosphate, pH 7.0</td>
</tr>
<tr>
<td>300 mM NaCl</td>
<td>300 mM NaCl</td>
</tr>
<tr>
<td>10 mM Imidazole</td>
<td>250 mM Imidazole</td>
</tr>
</tbody>
</table>

**Other solutions**

**Luria–Bertani Broth**
- 10 g.L\(^{-1}\) Casein hydrolysate
- 5 g.L\(^{-1}\) Yeast extract
- 10 g.L\(^{-1}\) NaCl

**Agar plates**
- 10 g.L\(^{-1}\) Casein hydrolysate
- 5 g.L\(^{-1}\) Yeast extract
- 10 g.L\(^{-1}\) NaCl
- 1.5–2% Agar

Plates were autoclaved and allowed to cool to ~55°C before pouring.

**Neutral proteolysis buffer**
- 50 mM phosphate, pH 7.0
- 300 mM NaCl

**Low-pH proteolysis buffer**
- 50 mM Citric acid
- 300 mM NaCl
- Adjusted with NaOH to pH 5.5
II. Alignment of mammalian PEG10 sequence surrounding the predicted aspartyl protease domain

Alignment was generated using MUSCLE and trimmed by hand. The N-terminus begins at the –1 frameshifting site and the C-terminus ends at an insertion in rodent Peg10 (corresponding to amino acids 395–560 in human PEG10 [NP_001165908.1]).
III. Prediction of globular domains throughout PEG10

GlotPlot v2.3 software [http://globplot.embl.de/] was used to predict potential globular domains within the full-length PEG10 polypeptide. This software uses empirically derived propensities for each amino acid to form either a random coil or ordered secondary structure in solution. Each amino acid in the input sequence is assigned a disorder value based on the innate propensity as well as the values of surrounding amino acids and a C-terminal weighting. The propensities are plotted, with smoothing, across the input sequence. Ordered and disordered regions are then selected using a minimum window length and the sign of the first-order derivative of the smoothed curve at each amino acid position (Linding et al., 2003). The globular domain from residues 344–475 in this diagram represent the boundaries of PEG10 AP clone 16 used in this study.
IV. Analysis of rare codons in the predicted PEG10 aspartyl protease domain

PEG10-AP clone 6, the longest clone examined in this study, was used for this analysis. The *E. coli* rare codon calculator available at http://people.mbi.ucla.edu/sumchan/caltor.html was used.
V. $\text{OD}_{600}$ values

Pre- and post-induction $\text{OD}_{600}$ values from small-scale cultures, relevant to section 2.6.1. ‘–’ refers to pMAL-TEV–only control.

<table>
<thead>
<tr>
<th>Clone</th>
<th>Pre-induction $\text{OD}_{600}$</th>
<th>Post-induction $\text{OD}_{600}$</th>
<th>Clone</th>
<th>Pre-induction $\text{OD}_{600}$</th>
<th>Post-induction $\text{OD}_{600}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.687</td>
<td>1.104</td>
<td>17</td>
<td>0.652</td>
<td>1.119</td>
</tr>
<tr>
<td>2</td>
<td>0.683</td>
<td>1.075</td>
<td>18</td>
<td>0.708</td>
<td>1.148</td>
</tr>
<tr>
<td>3</td>
<td>0.653</td>
<td>1.072</td>
<td>19</td>
<td>0.671</td>
<td>1.157</td>
</tr>
<tr>
<td>4</td>
<td>0.743</td>
<td>1.130</td>
<td>20</td>
<td>0.669</td>
<td>1.157</td>
</tr>
<tr>
<td>5</td>
<td>0.651</td>
<td>1.089</td>
<td>21</td>
<td>0.639</td>
<td>1.137</td>
</tr>
<tr>
<td>6</td>
<td>0.715</td>
<td>1.121</td>
<td>22</td>
<td>0.556</td>
<td>1.136</td>
</tr>
<tr>
<td>7</td>
<td>0.590</td>
<td>1.043</td>
<td>23</td>
<td>0.673</td>
<td>1.134</td>
</tr>
<tr>
<td>8</td>
<td>0.629</td>
<td>1.070</td>
<td>24</td>
<td>0.711</td>
<td>1.139</td>
</tr>
<tr>
<td>9</td>
<td>0.631</td>
<td>1.076</td>
<td>25</td>
<td>0.639</td>
<td>1.103</td>
</tr>
<tr>
<td>10</td>
<td>0.692</td>
<td>1.106</td>
<td>26</td>
<td>0.683</td>
<td>1.156</td>
</tr>
<tr>
<td>11</td>
<td>0.692</td>
<td>1.098</td>
<td>27</td>
<td>0.640</td>
<td>1.074</td>
</tr>
<tr>
<td>12</td>
<td>0.756</td>
<td>1.146</td>
<td>28</td>
<td>0.630</td>
<td>1.123</td>
</tr>
<tr>
<td>13</td>
<td>0.654</td>
<td>1.104</td>
<td>29</td>
<td>0.676</td>
<td>1.105</td>
</tr>
<tr>
<td>14</td>
<td>0.664</td>
<td>1.107</td>
<td>30</td>
<td>0.672</td>
<td>1.063</td>
</tr>
<tr>
<td>15</td>
<td>0.647</td>
<td>1.100</td>
<td>–</td>
<td>0.638</td>
<td>1.139</td>
</tr>
<tr>
<td>16</td>
<td>0.623</td>
<td>1.125</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VI. Solubility of PEG10 AP subsequent to separation from MBP by the TEV protease

'S', soluble (10 μL); 'I', insoluble (all loaded). Loading is such that the insoluble lanes are proportionately 10× more concentrated than soluble fractions. Because pellets were invisible, no wash steps were carried out and much of the material in the insoluble lanes is likely to be carry-over contamination. The PEG10 AP cannot be seen for clones 19–24 due to poor TEV-protease cleavage efficiency (Figure 4.13). The ‘–’ lane contains MBP incubated with TEV protease as a negative control. The presence of MBP—a particularly soluble protein—in ‘I’ lanes indicates carryover contamination.
VII. Mass spectrometric sequencing of MBP–PEG10-AP clone 4 products

**a**

<table>
<thead>
<tr>
<th>Expect</th>
<th>Rank</th>
<th>Unique</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.9e-006</td>
<td>1</td>
<td>U</td>
<td>R.SIVFDSEYCR.Y + Carbamidomethyl (C)</td>
</tr>
<tr>
<td>7.2e-012</td>
<td>1</td>
<td>U</td>
<td>R.WLSTHPNITWSTR.S</td>
</tr>
<tr>
<td>2.1e-011</td>
<td>1</td>
<td>U</td>
<td>K.LPGPAVEGPSATGPEIIR.S</td>
</tr>
<tr>
<td>9e-012</td>
<td>1</td>
<td>U</td>
<td>R.IGGKLPGPAVEGPSATGPEIIR.S</td>
</tr>
<tr>
<td>9.9e-005</td>
<td>1</td>
<td>U</td>
<td>R.EVLSFDVTQSPFFPPVGLVR.W + Glu-&gt;pyro-Glu (N-term E)</td>
</tr>
<tr>
<td>0.00019</td>
<td>1</td>
<td>U</td>
<td>R.EVLSFDVTQSPFFPPVGLVR.W + Glu-&gt;pyro-Glu (N-term E)</td>
</tr>
<tr>
<td>2.3e-016</td>
<td>1</td>
<td>U</td>
<td>R.EVLSFDVTQSPFFPPVGLVR.W</td>
</tr>
<tr>
<td>5.3e-016</td>
<td>1</td>
<td>U</td>
<td>R.SPQDDASSPHLQVMLQIHLPGR.H</td>
</tr>
<tr>
<td>2.1e-006</td>
<td>1</td>
<td>U</td>
<td>R.SPQDDASSPHLQVMLQIHLPGR.H + Oxidation (M)</td>
</tr>
<tr>
<td>1.2e-016</td>
<td>1</td>
<td>U</td>
<td>R.AMIDSGASGNFIDHEYVAQNGIPLR.I</td>
</tr>
<tr>
<td>3.4e-011</td>
<td>1</td>
<td>U</td>
<td>R.AMIDSGASGNFIDHEYVAQNGIPLR.I + Oxidation (M)</td>
</tr>
<tr>
<td>0.008</td>
<td>1</td>
<td>U</td>
<td>R.IKDWPILEAIDGRPIASGVPVHETHDLIVDLGDHR.E</td>
</tr>
</tbody>
</table>

**b**

<table>
<thead>
<tr>
<th>Expect</th>
<th>Rank</th>
<th>Unique</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3e-005</td>
<td>1</td>
<td>U</td>
<td>E.THDLIVDLGDHR.E</td>
</tr>
<tr>
<td>0.0017</td>
<td>1</td>
<td>U</td>
<td>P.GPAVEGPSATGPEIIR.S</td>
</tr>
<tr>
<td>0.0004</td>
<td>1</td>
<td>U</td>
<td>R.IKDWPILEAIDGR.P</td>
</tr>
<tr>
<td>0.0037</td>
<td>1</td>
<td>U</td>
<td>R.IKDWPILEAIDGR.P</td>
</tr>
<tr>
<td>3.7e-006</td>
<td>1</td>
<td>U</td>
<td>K.LPGPAVEGPSATGPEIIR.S</td>
</tr>
<tr>
<td>3.7e-005</td>
<td>1</td>
<td>U</td>
<td>R.EVLSFDVTQSPFFPPVVL</td>
</tr>
<tr>
<td>7e-005</td>
<td>1</td>
<td>U</td>
<td>R.IGGKLPGPAVEGPSATGPEIIR.S</td>
</tr>
<tr>
<td>1.5e-006</td>
<td>1</td>
<td>U</td>
<td>R.IGGKLPGPAVEGPSATGPEIIR.S</td>
</tr>
<tr>
<td>3.2e-007</td>
<td>1</td>
<td>U</td>
<td>R.EVLSFDVTQSPFFPPVGLVR.W</td>
</tr>
<tr>
<td>0.00026</td>
<td>1</td>
<td>U</td>
<td>R.SPQDDASSPHLQVMLQIHLPGR.H</td>
</tr>
<tr>
<td>0.0021</td>
<td>1</td>
<td>U</td>
<td>R.AMIDSGASGNFIDHEYVAQNGIPLR.I</td>
</tr>
<tr>
<td>9.3e-008</td>
<td>1</td>
<td>U</td>
<td>R.AMIDSGASGNFIDHEYVAQNGIPLR.I</td>
</tr>
<tr>
<td>0.00039</td>
<td>1</td>
<td>U</td>
<td>R.AMIDSGASGNFIDHEYVAQNGIPLR.I + Oxidation (M)</td>
</tr>
<tr>
<td>0.0011</td>
<td>1</td>
<td>U</td>
<td>R.AMIDSGASGNFIDHEYVAQNGIPLR.I + Deamidated (NQ); Oxidation (M)</td>
</tr>
</tbody>
</table>

Mascot results relevant to Figure 4.15. Matched peptides are shown in bold red. **a** Sequence of the MBP–PEG10-AP clone 4 product major band, determined by MALDI TOF/TOF. **b** Sequence of the MBP–PEG10-AP clone 4 product minor band, determined using an LTQ-Orbitrap instrument.
VIII. Mass spectrometric sequencing of the his–PEG10-AP clone 4 and his-TEV–PEG10-AP clone 10 products

MASCOT results pages. Matched peptides are shown in bold red. **a** Coomassie-stained SDS-PAG of the his-TEV–PEG10-AP clone 10 product (insoluble fraction, 2 µL loaded) and matching MASCOT results. The N-terminal hexahistidine sequence with appended TEV-protease cleavage site was detected among the peptides. **b** MALDI-TOF/TOF sequencing of the prominent band in fraction 3 of Figure 5.6 matched against the non-redundant human proteome by NCBI BLAST. Sequence coverage is 48% for this protein.
IX. Size markers used for size-exclusion chromatography

Chromatograms of BSA, DNase I, and lysozyme run over a Superdex 200 10/300 GL column. A total of 500 µL of 1 mg.mL$^{-1}$ (BSA and DNase I) or 0.5 mg.mL$^{-1}$ (lysozyme) was applied to each column as described in Section 2.6.5. Only A$_{280}$ traces are shown in this figure. For Figure 5.6, the larger BSA peak, corresponding to a monomeric species, was indicated. Traces are presented as in Figure 5.6.