Structural and biophysical characterisation of the extra-cellular domains from the mammalian peptide transporters, PepT1 and PepT2

John H. Beale

Department of Biochemistry

Lady Margaret Hall



A thesis submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy* at the University of Oxford

March 4, 2015

Abstract

PepT1 and PepT2 are integral membrane proteins which couple the uptake of di- and tri-peptides to the proton electro-chemical gradient. PepT1 is predominantly expressed in the small intestine and is the main route through which dietary protein is absorbed. PepT2 shares 46% sequence identity with PepT1 and is expressed in the kidneys, lung and central nervous system. Many commonly prescribed drugs, such as penicillin are peptide mimetics, and PepT1 and PepT2 play a direct role in their transport and pharmacokinetic properties. Recent X-ray crystal structures and functional studies of the bacterial peptide transporters have provided significant insight into the likely mechanism by which such drugs are recognised by PepT1 and PepT2. The bacterial peptide transporters share approximately 30% sequence identity within their trans-membrane domains to the mammalian PepT1 and PepT2 transporters. However, a key structural difference exists; an additional 20 kDa extra-cellular loop is inserted between trans-membrane helices 9 and 10 in the animal peptide transporters, and this loop is absent in the bacterial homologues. It was not known, prior to this thesis, if this extra-cellular loop was structured and/or integral to the transport cycle, or whether it served an additional function to assist or regulate peptide transport.

To investigate the role of this domain, the crystal structures of the *Mus musculus* PepT1 and *Rattus norvegicus* PepT2 'loops' were determined to 2.10 and 2.06 Å resolution respectively. The structures indicated that the loop region in both PepT1 and PepT2 forms a bi-lobal, all β-sheet, self-contained extra-cellular domain (ECD). Despite low sequence similarity, the ECDs from PepT1 and PepT2 share a common architecture; two transthyretin-like folds connected by a flexible linker. Sequence and structural analyses have indicated that the lobe interface of *Mm*PepT1^{ECD} is maintained by two highly conserved salt bridges, whereas in *Rn*PepT2^{ECD} only one salt bridge is observed. Small-angle X-ray diffraction experiments indicated that the extra-cellular domains form a compact

structural arrangement, although the lobe conformation of PepT2^{ECD} was more dynamic than PepT1.

Using the X-ray crystal structures of the *M. musculus* and *R. norvegicus* ECDs, and the trans-membrane domain of PepT$_{So}$ from *Shewanella oneidensis*, the first structure-based homology models of *H. sapiens* PepT1 and PepT2 were constructed. The hybrid models indicated that the ECD sits on top of the transporters. Two-electrode voltage clamp studies then revealed that the ECDs do not play a part in the transport mechanism of the transporter, although PepT1^{ECD} may play a role in transporter stability.

Surface plasmon resonance binding assays were performed between the ECDs and the intestinal proteases trypsin and α-chymotrypsin. An interaction between both ECDs and α-chymotrypsin was observed, although this interaction could not be saturated using this assay. Trypsin binding however, could be saturated for both *Mm*PepT1^{ECD} and *Rn*PepT2^{ECD} giving K_ds of 90 ± 20 and 170 ± 30 μM. Physiologically the interaction would give trypsin a predisposition for the peptide transporter ECD; locating the protease in the vicinity of the transporter and aiding the presentation of substrate. The interaction between trypsin and PepT1^{ECD} was explored further with a mutational analysis of potential binding residues. The results could not locate the binding site definitively, however, the work did highlight the probable binding face which is perhaps conserved between PepT1 and PepT2 ECD.

Acknowledgements

The time spent on this project has been made incredibly enjoyable by all of the people that I have been fortunate enough to work with and learn from. From the Newstead lab; Jo Parker, Nicolae Solcan, Sonya Hanson and Katharina Jungnickel, have continuously made me smile, and enjoy laboratory work. I can't imagine another work environment which I would rather immerse myself in and that, to a large part, is down to them. I would also like to thank Anne Barrett and Firdaus Samsudin for the time and effort they put into this project. Their perseverance to get results contributed significantly to my work, and has advanced our understanding of transporter function.

I would also like to thank members of the OPPF-UK, in particular, Ray Owens, Louise Bird and Anil Verma for their help and experience. Paul Emsley, Ed Lowe and Pietro Roversi for giving up their time to help me gain a better understanding of crystallography and David Scott, David Staunton and Jo Claridge for helping me with various biophysical techniques.

Primarily, I wish to thank my supervisor, Simon Newstead, for giving me this opportunity and continued support throughout. I could not have asked anything more from this project in terms of both the experience that I have gained and my personal and professional development. The directions that the project took, were due to guidance and a number of great ideas from both Simon and his partner Jo. They have made working as a DPhil student very interesting and a lot of fun.

More significantly, if Simon had not given me this opportunity, I would never have met Emma. This is the first time I have been with someone who has made me feel like the member of a team, and I have no doubt that the next steps of our lives together will continue to be as entertaining and as awesome as the last two years.

Contents

1 Introduction **1**

1.1 Nutrient transport 1

1.2 Mammalian di- and tri-peptide transport 2

1.2.1 Dietary absorption of protein by the small intestine 3

1.2.2 Peptide absorption in the kidney 6

1.2.3 Peptide transport is driven by H^+ electro-chemical gradient . 6

1.3 Characterisation of the PepT1 and PepT2 genes 7

1.3.1 PepT1 and PepT2 tissue localisation 8

1.3.2 Role in drug transport 10

1.3.3 Evolutionary conservation of the eukaryotic and bacterial peptide transporters 12

1.4 Transporter classification and mechanism 15

1.4.1 Secondary active transport 15

1.4.2 Major facilitator superfamily 17

1.4.3 The alternating access mechanism of solute transport 18

1.4.4 Alternating access model in bacterial peptide transport . . . 20

1.4.5 Mechanistic model of bacterial peptide transport 27

1.5 Conservation of peptide transport model from bacteria into mammalian PepT1 and PepT2 29

1.5.1 Chimeric studies on mammalian PepT1 and PepT2 29

1.6 Characterisation of the extra-cellular loop 30

1.6.1 SLC protein extra-cellular loops 30

1.6.2 Structure prediction analysis 32

1.7 Aims 35

2 Materials and methods **36**

2.1 Summary 36

2.2 Materials 36

2.2.1 Instruments 36

2.2.2 Bacteria strains, media, chemicals and kits 38

2.3 Alignments and figures 39

2.4 Gels and other quantification techniques 39

2.4.1 SDS-PAGE gels 39

2.4.2 Trichloroacetic acid precipitation 41

2.5 Bicinchoninic acid assay 41

2.6 Cloning and test expression of ECD constructs . . . 42
2.6.1 OPPF-UK high-throughput cloning and test-expression . . . 42
2.6.2 Cloning and test expression of pLou3 and pEHisTev ECD constructs . . . 47
2.6.3 Large scale expression and cell lysate preparation. . . . 51
2.6.4 ECD-3CX construct cloning . . . 52
2.6.5 ECD-3CX aspartate mutant cloning . . . 53
2.6.6 *Rn*PepT2^{ECD} D505A and D600A mutant cloning . . . 54
2.7 Crystallisation . . . 55
2.7.1 96-well sitting drop experiments . . . 55
2.7.2 24-well hanging drop experiments . . . 55
2.7.3 Reductive methylation . . . 56
2.7.4 SeMet derivatisation . . . 56
2.7.5 Sparse matrix micro-seeding protocol . . . 56
2.8 Biophysics . . . 57
2.8.1 Fluorescence scanning calorimetry . . . 57
2.8.2 Heteronuclear single quantum coherence NMR . . . 58
2.8.3 S75 10/300 globular protein calibration . . . 59
2.8.4 Analytical ultracentrifugation . . . 59
2.8.5 Circular dichroism protocol . . . 60
2.8.6 Two electrode voltage clamp *X. laevis* oocyte assays . . . 62
2.8.7 Surface plasmon resonance protocol . . . 63

3 Cloning, expression and purification of the PepT1 and PepT2 ECDs 66
3.1 Summary . . . 66
3.2 Heterologous protein expression . . . 67
3.3 OPPF-UK high throughput cloning and test expression . . . 68
3.3.1 ECD construct design . . . 70
3.3.2 Cloning of ECDs into the pOPIN suite . . . 73
3.3.3 OPPF-UK HT small scale test expression of ECDs in pOPIN vectors . . . 74
3.4 Large scale expression of OPPF-UK constructs . . . 81
3.4.1 *Mm*PepT1^{ECD} pOPINM purification . . . 81
3.4.2 *Mm*PepT2^{ECD} 410 – 609 pOPINS3C purification . . . 84
3.5 Optimised PepT2^{ECD} construct cloning and test expression . . . 86
3.5.1 Cloning of the new PepT2 constructs into the pLou3 and pE-HisTev expression vectors . . . 86
3.5.2 Test expression of the PepT2^{ECD} pLou3 and pEHisTev constructs . . . 87
3.5.3 Purification of *Mm*PepT2^{ECD} and *Rn*PepT2^{ECD} 410 – 601 pLou3 constructs . . . 87
3.6 Discussion . . . 90

4 Crystal structure of *M. musculus* PepT1^{ECD} **92**
4.1 Summary . . . 92
4.2 Crystallisation . . . 93
4.2.1 Sparse matrix Screening . . . 93
4.2.2 Crystal optimisation . . . 94
4.3 Data Collection . . . 99
4.3.1 Bragg Diffraction . . . 99
4.3.2 Radiation damage and cryo-protection . . . 101
4.3.3 Indexing for data collection . . . 102
4.3.4 Data-collection Parameters . . . 104
4.3.5 Detectors . . . 104
4.4 Data reduction . . . 105
4.4.1 Indexing . . . 106
4.4.2 Integration . . . 107
4.4.3 Scaling . . . 107
4.4.4 Merging . . . 107
4.5 Asymmetric unit analysis . . . 112
4.6 Experimental Phasing . . . 114
4.6.1 Phase calculation . . . 114
4.6.2 Isomorphous replacement . . . 115
4.6.3 Single wavelength anomalous dispersion . . . 116
4.6.4 Density modification . . . 119
4.7 Structure building/refinement . . . 120
4.8 *Mm*PepT1^{ECD} ASU Model . . . 124
4.9 Discussion . . . 126

5 Crystal structure of the *R. norvegicus* PepT2^{ECD} **128**
5.1 Summary . . . 128
5.2 *Mm*PepT2^{ECD} 410 – 609 spare matrix crystallisation experiments. . . 129
5.2.1 Fluorescence scanning calorimetry of the *Mm*PepT2^{ECD} 410 – 609 construct . . . 130
5.2.2 Reductive methylation of the *Mm*PepT2^{ECD} 410 – 609 construct 132
5.2.3 *In situ* limited proteolysis of the *Mm*PepT2^{ECD} 410 – 609 construct . . . 132
5.2.4 Using NMR to investigate the structural flexibility in the 410 – 601 *Mm*PepT2^{ECD} construct . . . 135
5.3 *Mm*PepT2^{ECD} 410 – 601 sparse matrix crystallisation experiments. . 135
5.3.1 HSQC experiment of the *Mm*PepT2^{ECD} 410 – 601 construct 136
5.4 *Rn*PepT2^{ECD} 410 – 601 crystallisation experiments . . . 137
5.4.1 Sparse matrix screening . . . 137
5.4.2 Crystal optimisation . . . 138
5.5 Data collection . . . 139
5.6 Data reduction . . . 141

5.6.1 Trigonal crystals . . . 141
5.7 Asymmetric unit analysis . . . 144
5.8 Experimental phasing of the trigonal crystal . . . 145
5.8.1 SeMet SAD phasing and model building . . . 145
5.9 Refinement . . . 147
5.10 Higher resolution *Rn*PepT2^{ECD} 410 – 601 structure . . . 150
5.10.1 Sparse matrix micro-seeding . . . 150
5.10.2 Data collection and reduction . . . 152
5.10.3 Molecular replacement . . . 152
5.10.4 Model building and refinement . . . 153
5.11 The model of *Rn*PepT2^{ECD} . . . 155
5.12 Discussion . . . 155

6 Structural and biophysical analysis of the PepT1 and PepT2 ECD lobe orientation 160
6.1 Summary . . . 160
6.2 The structures of *Mm*PepT1 and *Rn*PepT2 ECD . . . 161
6.3 Alignment analyses of the ECDs . . . 164
6.3.1 Sequence and structural database searches . . . 164
6.3.2 Sequence analysis of the ECDs lobes and TTR . . . 167
6.3.3 Sequence analysis of the ECD lobes . . . 167
6.4 Bioinformatics analysis of the ECDs . . . 170
6.4.1 PISA physiological interface predictions . . . 173
6.4.2 PROPKA analysis of the conserved aspartate residues in the ECDs . . . 173
6.5 Linker mutation study . . . 175
6.5.1 *Mm*PepT1$^{ECD\text{-}3CX}$ linker cleavage experiment results . . . 176
6.6 PepT2^{ECD} lobe dynamics . . . 180
6.6.1 Small-angle X-ray scattering . . . 181
6.6.2 SAXS data collection . . . 182
6.6.3 SAXS data processing . . . 182
6.6.4 Analytical ultra-centrifugation . . . 188
6.7 Discussion . . . 192

7 Investigations into the function of the ECDs 195
7.1 Summary . . . 195
7.2 Regulation of peptide transporter function . . . 196
7.3 Hybrid model construction . . . 198
7.3.1 Method . . . 198
7.3.2 Hybrid model validity . . . 204
7.4 Investigating the role of the ECD in the transport cycle . . . 204
7.5 Proteolytic cleavage of dietary protein . . . 208
7.5.1 The ECD as a protease binding protein . . . 209

7.5.2 Trypsin, α-chymotrypsin and elastase 211
7.6 ECD binding experiments 211
7.6.1 α-chymotrypsin and pepsin 212
7.6.2 Trypsin 212
7.6.3 Characterising the *Mm*PepT1^{ECD} trypsin binding site 214
7.7 Discussion 223

8 Final discussion **228**
8.1 Final analysis 228
8.2 Future work 231

A Complete X-ray crystallographic data statistics **234**

B AUC data **237**
B.1 Raw data Lamm equation fits 237
B.2 c(s) distributions 242

List of Tables

1.1 Chimeric study mutant constitution and results. 31
1.2 The members of the SLC families which have large intra-helical extracellular sections. 32

2.1 A list of the genes used for alignments and bioinformatic analysis in this thesis. 40
2.2 Components and volumes used for Tris-Gly SDS-PAGE gels. 41
2.3 The homologous sequences used for the OPPF-UK pOPIN vectors. . . 43
2.4 Priming sequences for the peptide transporter ECDs used in the OPPF-UK cloning. 43
2.5 PCR reaction volumes used in OPPF-UK cloning. 44
2.6 Reaction times and temperatures for PCR amplifications using Phusion Flash or KOD Xtreme^{TM} DNA polymerases. 44
2.7 The composition of the In-fusion reaction mixture. 45
2.8 Primers used for the cloning of the ECDs into the pLou3 and pEHisTev expression vectors. 48
2.9 Volumes and parameters used for KOD Hot Start PCRs. 49
2.10 The composition of a restriction enzyme digest. 49
2.11 A list of the primers used in the lobe conformation mutation studies. 53
2.12 Thermocycler parameters for KOD Hot Start site directed mutagenesis PCR. 54
2.13 A list of the sparse-matrix screens use during the crystallisation. . . . 55
2.14 The composition of PASM-5052 auto-inducing SeMet *E. coli* growth media. 57
2.15 The composition of N-5052 auto-inducing ^{15}N *E. coli* growth media. . 58
2.16 The globular protein MW standard V_es from the Superdex S75 10/300 GF column. 59
2.17 *Hs*PepT1 mutant site directed mutagenesis primers used during the oocyte assays. 62
2.18 *Mm*PepT1^{ECD} mutant site directed mutagenesis primers used to elucidate the trypsin binding site. 64
2.19 Final immobilisation levels of the ECDs and mutants used in the SPR protease assays. 65

3.1 A list of the OPPF-UK vectors used for the *E. coli* intra-cellular expression. 69

3.2 The genes and construct lengths used in the OPPF-UK test expressions. 70
3.3 Results from the OPPF-UK intra-cellular *E. coli* test expression of the ECD pOPIN constructs. 80
3.4 Summary of the ECD pOPIN *E. coli* intra-cellular test-expression results. 81
3.5 Genes and constructs used in the Oxford Biochemistry *E. coli* ECD test expression. 86

4.1 Data collection parameters from the data collection of *Mm*PepT1^{ECD} crystal datasets. 105
4.2 Data processing statistics of the *Mm*PepT1^{ECD} datasets. 111
4.3 A table showing the V_m probability distribution for dataset N1. . . . 112
4.4 The Friedel mate difference intensities and initial HA positions for the Hg-SAD dataset. 118
4.5 The acentric phasing statistics calculated during HA sub-structure refinement of the Hg-SAD dataset. 119
4.6 The `Solomon` contrast and $|E|^2$ values for different solvent contents of the ASU. 120
4.7 Refinement statistics for the *Mm*PepT1^{ECD} model. 123

5.1 A list of the conditions which grew *Mm*PepT2^{ECD} micro-crystals or crystalline like precipitate. 131
5.2 A table showing the conditions from sparse matrix screens which gave *Rn*PepT2^{ECD} 410 – 601 crystallisation hits. 139
5.3 Data collection parameters for the trigonal *Rn*PepT2^{ECD} 410 – 601 datasets. 141
5.4 Data processing statistics of the *Rn*PepT2^{ECD} trigonal datasets. . . . 143
5.5 A table showing the V_m probability distribution for the $P3_221$ Native dataset. 144
5.6 HA positions found by `HySS`. 146
5.7 The Se-SAD dataset acentric phasing statistics. 148
5.8 A list of the SeMet sites built into the density calculated by `AutoSOL`. 149
5.9 Refinement statistics for the *Rn*PepT2^{ECD} model from the trigonal crystal. 151

6.1 The lobe interface `PISA` predictions for the *Mm*PepT1 and *Rn*PepT2 ECD ASUs. 174
6.2 The pK_as and percentage exposures of the conserved aspartate residues in the *Mm*PepT1 and *Rn*PepT2 ECDs. 175
6.3 The SAXS parameters for *Mm*PepT1 and *Rn*PepT2 ECDs equilibrated at pH 5.5, 7.5, 8.0 and 9.5. 185
6.4 The calculated R_gs in `Hydropro` for the *Mm*PepT1 and *Rn*PepT2 ECD models shown in Figure 6.13. 188
6.5 The sedimentation coefficients calculated in `Hydropro` for the PepT1 and PepT2 ECD models shown in Figure 6.13. 192

7.1 Percentage sequence identity between the bacterial MFS structures and *H. sapiens* PepT1 and PepT2. 198
7.2 `Molprobity` analysis of the hybrid models. 204
7.3 The T_m of the WT and mutant ECDs at pH 7.5 as determined by CD thermal melt. 204

A.1 Complete X-ray crystallographic data statistics for the *Mm*PepT1^{ECD} datasets. 235
A.2 Complete X-ray crystallographic data statistics for the *Rn*PepT2^{ECD} datasets. 236

List of Figures

1.1 A simplified diagram of cellular nutrient uptake. 2
1.2 A comparison between GlyGly and the non-hydrolysable peptide glycyl-sarcosine. 4
1.3 Representation of peptide transport in the intestinal and kidney brush boarder epithelia. 7
1.4 Peptide transporter topology. 8
1.5 PepT1 and PepT2 localisation in the kidney nephron. 9
1.6 Two families of β-lactam antibiotics. 11
1.7 The structures of valacyclovir, oseltamivir and ubenimex. 12
1.8 A sequence alignment of a subset of POT family transporters. 14
1.9 Secondary active transporter classifications. 15
1.10 The MFS fold of GlpT and $PepT_{So}$. 17
1.11 The alternating access rocker switch mechanism of MFS symport. . . 19
1.12 Rocker switch conformations of MFS symporters. 21
1.13 The salt bridges which line the substrate channel in the bacterial peptide transporters. 23
1.14 The substrate binding site of the bacterial peptide transporters. . . . 25
1.15 A model of bacterial peptide transport. 28
1.16 Secondary structure predictions of the *H. sapiens* PepT1 and PepT2 extra-cellular loop. 33
1.17 Residue disorder predictions for the extra-cellular loop regions of *H. sapiens* PepT1 and PepT2. 34

2.1 The 1.5% w/v Agarose, 1 × TAE gels from the cloning of the MmPepT1$^{\text{ECD-3CX}}$ constructs. 54
2.2 The calibration curve for the Superdex S75 10/300 GF column based on the V_e of the globular protein standards shown in Table 2.16. . . . 59
2.3 The $s^0_{20,w}$ calculation plots for MmPepT1^{ECD}, RnPepT2^{ECD}, D505A and D600A. 61

3.1 A vector map of the pTriEx^{TM}-1.1 Hygro vector. 68
3.2 A sequence alignment of *H. sapiens* and *M. musculus*, PepT1 and PepT2 against $PepT_{So}$. 71
3.3 The N- and C-terminal locations of the ECDs from PepT1 and PepT2. 72
3.4 PCR insert conformations of the ECDs in the pOPIN vectors. 73

3.5 12 reducing NuPAGE 10% Bis-Tris Midi gels from the OPPF-UK HT *E. coli* intra-cellular test expression of the ECD pOPIN constructs. . 75
3.6 The *Mm*PepT1^{ECD} pOPINM purification. 83
3.7 The *Mm*PepT2^{ECD} pOPINS3C purification. 85
3.8 A 1.5% w/v Agarose, 1 × TAE text cleavage gel of the PepT2^{ECD} constructs in the pLou3 and pEHisTev vectors. 87
3.9 An anti-His western blot from a PepT2^{ECD} pLou3 and pEHisTev test expression using TB media. 88
3.10 A *Rn*PepT2^{ECD} 410 – 601 pLou3 purification. 89

4.1 *Mm*PepT1^{ECD} GF trace and gel. 94
4.2 Sitting drop vapour diffusion crystallography. 95
4.3 Crystal hit of *Mm*PepT1^{ECD}. 95
4.4 Crystal hit optimisation. 96
4.5 The two crystal forms of the optimised *Mm*PepT1^{ECD} crystals. 98
4.6 Bragg diffraction. 100
4.7 An image of the diffraction from a *Mm*PepT1^{ECD} crystal. 103
4.8 A screen shot created by `iMosflm` using two images from a *Mm*PepT1^{ECD} crystal taken 90 ° a part. 104
4.9 Temperature factor and space-group assignment for the *Mm*PepT1^{ECD} crystal. 108
4.10 The $\chi = 180°$ SRF section of the *Mm*PepT1^{ECD} dataset N1. 113
4.11 An Argand diagram showing the relative amplitudes and phase components of $\mathbf{F}_{PA}$, $\mathbf{F}_{P}$ and $\mathbf{F}_{A}$. 116
4.12 The peak wavelength dependent factors f' and f'' calculated by `Chooch` from a fluorescence scan collected at Diamond Light Source I04 over the Hg (II) absorptive edge on a *Mm*PepT1^{ECD} crystal soaked in Hg (II) Acetate. 117
4.13 An electron density map generated from the Hg-SAD $|\mathbf{F}_{hkl}|$ and `SHARP` refined phases. 121
4.14 The asymmetric unit of the *Mm*PepT1^{ECD} crystal containing two ECD monomers. 122
4.15 A Ramachandran plot of the *Mm*PepT1^{ECD} ASU calculated using `Rampage`. 124
4.16 Characteristics of the *Mm*PepT1^{ECD}. 125

5.1 *Mm*PepT2^{ECD} 410 – 609 GF trace and gel. 129
5.2 *Mm*PepT2^{ECD} 410 – 609 FSC results. 132
5.3 A comparison between the *Mm*PepT2 410 – 609 WT and methylated protein GF traces. 133
5.4 Results and analysis of the *Mm*PepT2^{ECD} 410 – 609 limited proteolysis experiment. 134

5.5 An HSQC spectrum of the *Mm*PepT2^{ECD} 410 – 609 construct collected at pH 5.5. 136
5.6 An HSQC spectrum of the *Mm*PepT2^{ECD} 410 – 601 construct collected at pH 5.5. 137
5.7 *Rn*PepT2^{ECD} 410 – 601 crystals. 138
5.8 Diffraction pattern and beamline processing of a *Rn*PepT2^{ECD} 410 – 601 crystal. 140
5.9 An $I/\sigma(I)$ against c axis plot of the P3 *Rn*PepT2^{ECD} 410–601 intensities. 142
5.10 The SRF of a *Rn*PepT2^{ECD} trigonal crystal. 144
5.11 The peak wavelength dependent factors f' and f'' calculated by `Chooch` from a fluorescence scan collected at Diamond Light Source I04 over the Se absorptive edge on a SeMet derivatised *Rn*PepT2^{ECD} crystal. . 145
5.12 Examples of the Se density in the *Rn*PepT2^{ECD} $P3_221$ crystal. 148
5.13 An electron density map of the trigonal *Rn*PepT2^{ECD} ASU generated from the Se-SAD $|\mathbf{F}_{hkl}|$ and `SHARP` refined φ_{hkl}. 149
5.14 The *Rn*PepT2^{ECD} trigonal crystal structure. 150
5.15 Sparse matrix seed phase diagram. 151
5.16 The dual histidine co-ordination of the Cs atom at the 2-fold crystallographic axis. 154
5.17 A Ramachandran plot of the $P4_12_12$ 410 – 601 ASU calculated using `Rampage`. 154
5.18 Analysis of the *Rn*PepT2^{ECD} $P3_221$ crystal crystal structure. 156
5.19 A Cα alignment of the $P4_12_12$ ASU with chain A from the $P3_221$ ASU. 156
5.20 An analysis of why the *Mm*PepT2^{ECD} construct failed to crystallise. . 158

6.1 The structure and topology of *Mm*PepT1 and *Rn*PepT2 ECD. 162
6.2 An alignment of *H. sapiens*, *M. musculus* and *R. norvegicus*, PepT1 and PepT2 ECD sequences. 163
6.3 A Cα alignment of the monomeric *Rn*PepT2^{ECD} onto the ASU of *Mm*PepT1^{ECD}. 164
6.4 A sequence and structural comparison of TTR and the ECD lobes. . . 166
6.5 An alignment of vertebrate PepT1 and PepT2, lobe 1 and lobe 2 sequences. 168
6.6 Structural analysis of the lobe motifs. 169
6.7 The two potential lobe conformations based on the *Mm*PepT1^{ECD} ASU. 170
6.8 The S75 10/300 GF traces of *Mm*PepT1 and *Rn*PepT2 ECD. 171
6.9 The four conserved residues which predominantly make up the closed interface between the lobes in the *Mm*PepT1 and *Rn*PepT2 crystal structure. 172
6.10 The lobe assignment used in the `PISA` analysis presented in Table 6.1 for *Mm*PepT1 and *Rn*PepT2 $P3_221$. 174
6.11 Final GF traces and reducing gels from the *Mm*PepT1$^{ECD\text{-}3CX}$ and *Mm*PepT1$^{ECD\text{-}3CX}$-D574A purifications. 177

6.12 The results of the *Mm*PepT1$^{\text{ECD-3CX}}$ linker cleavage experiment. . . . 179
6.13 A diagram showing the hypothesised change in the PepT2^{ECD} lobe arrangement with an increase or decrease of H^+. 180
6.14 A diagram of SAXS data collection. 181
6.15 The stacked scattered intensities from the *Mm*PepT1 and *Rn*PepT2 ECD pH screens. 183
6.16 The normalised P(r) distribution functions and dimensionless Kratky plots for the *Mm*PepT1 and *Rn*PepT2 ECD SAXS pH curves. 187
6.17 *Rn*PepT2^{ECD} D505A and D600A mutant purifications. 190
6.18 The $s^0_{20,w}$ plotted against pH for *Mm*PepT1^{ECD}, *Rn*PepT2^{ECD} WT, D505A and D600A. 191
6.19 A comparison between the ECDs in the closed conformation and *Hs*TTR as a homo-tetramer. 193

7.1 Hypothetical binding of a ligand to the ECD inducing a conformational change in TM 10 and blocking the pore. 197
7.2 The conserved salt bridge residues from bacteria to humans. 200
7.3 The alignment of the mammalian and bacterial peptide transporters used for the construction of the hybrid models. 201
7.4 Ramachandran plots of the hybrid models. 202
7.5 The hybrid models of *H. sapiens* PepT1 and PepT2. 203
7.6 The composition and results of mutant and chimeric PepT1 and PepT2 transporters constructed to investigate the role of the ECDs in the transport cycle. 206
7.7 The electrostatics of *Hs*PepT1 and *Hs*PepT2. 210
7.8 The structures of the pancreatic peptidases trypsin, α-chymotrypsin and elastase. 211
7.9 The blanked subtracted sensograms and equilibrium data fits for the ECDs with α-chymotrypsin and pepsin. 213
7.10 The blanked subtracted sensograms, equilibrium data fit and binding constants for trypsin and *Mm*PepT1^{ECD}, *Rn*PepT2^{ECD} and MBP. . . 215
7.11 Design of *Mm*PepT1^{ECD} mutants to localise trypsin binding site. . . . 216
7.12 An assessment of the purity of the *Mm*PepT1^{ECD} mutants. 218
7.13 The blanked subtracted sensograms and equilibrium data fits for the immobilised *Mm*PepT1^{ECD} mutants and trypsin. 219
7.14 A plot of the *Mm*PepT1^{ECD} mutant K_ds calculated from equilibrium data analysis. 220
7.15 Analysis of the *Mm*PepT1^{ECD} sequence and structure in the region of the H453A/D454A/E456A and D550A/E573A mutants. 221
7.16 A representation of how an interaction between PepT1^{ECD} and trypsin aids the uptake of di- and tri-peptides in the small intestine. 224
7.17 Possible conservation of trypsin binding interface between PepT1 and PepT2 ECD. 226

List of Abbreviations

1D	One dimensional
2D	Two dimensional
3D	Three dimensional
5-HIU	5-hydroxyisourate
ABC	ATP binding cassette
ASU	Asymmetric unit
ATP	Adenosine triphosphate
AUC	Analytical ultracentrifugation
BCA	Bicinchoninic acid
BSA	Bovine serum albumin
CC	Correlation coefficient
CCD	Charged-coupled device
CD	Circular dichroism
CNS	Central nervous system
CSF	Cerebrospinal fluid
CSS	Complexation significance score
DBC	Dimethyamine borane complex
DNA	Deoxyribonucleic acid
Dpp	Di-peptide permease
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
DTT	Dithiothreitol
ECD	Extra-cellular domain
EDTA	Ethylenediaminetetraacetic acid
ESRF	European Synchrotron Radiation Facility
FCCP	Carbonyl cyanide-4-(trifluoromethoxy)phenylhydrazone
FFT	Fast Fourier transform
FOM	Figure-of-merit
FSC	Fluorescence scanning calorimetry
FT	Flow through
FTF	Fast translation function
GF	Gel filtration
GST	Glutathione S-transferase
HA	Heavy atom
HLA	Hendrickson-Lattman

HRP	Horseradish peroxidase
HSQC	Heteronuclear single quantum coherence
HT	High-throughput
IBS	Irritable bowel syndrome
IO	Inward open
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
LB	Lysogeny broth
LLG	Log-likelihood gain
MBP	Maltose binding protein
MFS	Major facilitator superfamily
ML	Maximum-likelihood
MR	Molecular replacement
mRNA	Messenger ribonucleic acid
MW	Molecular weight
NCS	Non-crystallographic symmetry
NHE3	Na^+/H^+ exchanger
NMR	Nuclear magnetic resonance
NTA	Nitrilotriacetic acid
OB	Occluded bound
OD	Optical density
OO	Outward open
Opp	Oligo-peptide permease
OPPF-UK	Oxford protein production facility UK
PB	Power broth
PBS	Phosphate buffered saline
PBST	PBS, 0.05% v/v Tween 20
PCR	Polymerase chain reaction
PDB	Protein data bank
PMF	Proton motive force
POIB	Partially occluded inward bound
POT	Proton oligo-peptide transporter
PTR	Peptide transporter
PVDF	polyvinylidene fluoride
RMSD	Root-mean-square deviation
RP	Reverse purified
rpm	Revolutions per minute
RT	Room temperature
RU	Response units
SAD	Single wavelength anomalous dispersion
SAXS	Small-angle X-ray scattering
SBS	Short bowel syndrome

SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEM	Systematic error of the mean
SLC	Solute carrier
SnB	`Shake-n-bake`
SNR	Signal-to-noise ratio
SPR	Surface plasmon resonance
SRF	Self-rotation function
SSM	`Secondary-structure matching`
SUMO	Small ubiquitin-like modifier
SV	Sedimentation velocity
T4	Thyroxine
TAE	Tris-acetate-EDTA
TB	Terrific broth
TCA	Trichloroacetic acid
TEV	Tobacco etch virus
TEVC	Two electrode voltage clamp
TF	Trigger Factor
TFZ	Translation function Z-score
TLS	Translational/Librational/Skew
TM	Trans-membrane
Tpp	Tri-peptide permease
TRP	TTR-related protein
TRX	Thioredoxin
TTR	Transthyretin
WT	Wild type

I like the scientific spirit–the holding off, the being sure but not too sure, the willingness to surrender ideas when the evidence is against them: this is ultimately fine–it always keeps the way beyond open–always gives life, thought, affection, the whole man, a chance to try over again after a mistake–after a wrong guess. ***Whitman (1973)***

1

Introduction

1.1 Nutrient transport

The barrier all cells must overcome to uptake nutrients is the plasma membrane, which separates the cell from its surrounding environment. The membrane is predominantly populated by amphipathic phospholipids, composed of a hydrophilic polar head group and a hydrophobic acyl chain, which self assemble in a polar environment to form a planar bilayer (Bishop and Bell 1988). These acyl chains create an impermeable barrier for charged molecules and compounds preventing free diffusion. A simplified representation of cellular nutrient uptake is shown in Figure 1.1. Nutrients, such as fats which can form entropically favourable interactions with the lipid acyl chains, are able to diffuse across the membrane. Whereas polar nutrients, such as amino acids, require a carrier system which exists inside the membrane to allow uptake.

Amino acids are the basic units of proteins and are therefore essential nutrients to sustain

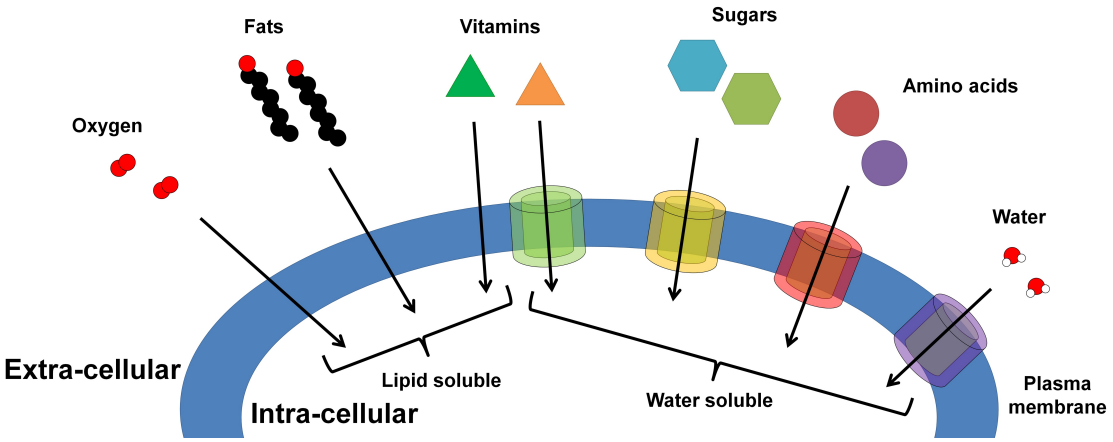


Figure 1.1: A simplified diagram of cellular nutrient uptake. The plasma membrane which divides the cell from its environment is shown in blue. For the cell to use the essential nutrients shown on the diagram; oxygen, fats, various minerals, sugars, amino acids and water, the nutrients have to be transported inside the plasma membrane. Oxygen, fats and some minerals can readily diffuse across the membrane, but sugars, amino acids and water cannot. Therefore the cell uses integral membrane protein transporters (shown as cylinders) to facilitate this process.

life. They also have other uses in the cell such as a dietary source of nitrogen and as a precursor for the keto acids utilised by the citric acid cycle. Amino acids are absorbed into the body predominantly in the form of di- and tri-peptides (described in § 1.2.1). The two proteins which are the principle subject of this thesis, PepT1 and PepT2 are integral membrane protein peptide transporters which facilitate the transport of di- and tri-peptides across the plasma membrane of cells. PepT1 is primarily expressed in the small intestine and is the sole route for peptide ingestion, whereas PepT2 is expressed in the kidneys, lung and Central nervous system (CNS). PepT2 potentially has several functions, one of which has been shown to be part of the innate immune system (discussed in § 1.3).

1.2 Mammalian di- and tri-peptide transport

One of the great successes of multicellular organisms is the division of cellular labour amongst tissue types and therefore the evolution of organs. This leap in complexity is comparable to the differences seen between eukaryotic and bacteria cells. Perhaps at the centre of the division is the digestive system. The development of cells to specifically absorb nutrients from the surrounding environment is likely to have been essential to the evolution of complex organisms. The evolution of the digestive system has seen the

development of a specific series of proteins which function in concert to efficiently break down and ingest the nutrients required to sustain life.

1.2.1 Dietary absorption of protein by the small intestine

In 1900, amino acids were discovered to make up the essential building blocks of proteins. This led to the view that proteins were completely digested in the stomach and small intestine into their constituent amino acids prior to absorption. However, doubt was cast on this theory by the time scale of protein digestion observed *in vitro*. Three peptidases; pepsin, trypsin and α-chymotrypsin could only achieve 90% liberation of amino acids after a 120 hr incubation (Frankel 1916; Dunn and Lewis 1921), whereas complete digestion of protein was observed in animals in a matter of hours. Therefore the idea that peptides could be the primary route of protein absorption was proposed as an explanation for the rapid absorption of protein observed *in vivo* (Fisher 1954).

The early studies into peptide and amino acid intestinal transport used an *in vitro* gut compartment technique. A section of the small intestine was removed from an animal and cut to make a single sheet of tissue. This sheet was manipulated to act as a barrier between two compartments. The contents of these compartments could be controlled, and therefore the transport of either peptide or amino acids could be detected. The di- and tri-peptides used in these experiments were GlyGly and GlyGlyGly (Agar, Hird, and Sidhu 1953; Wiggans and Johnston 1959). GlyGly was found to cross the basolateral membrane of the intestine confirming that peptide transport was possible. Both these studies reported extensive hydrolysis of experimental peptides; however, it was not clear where this hydrolysis occurred, in the intestinal lumen or epithelial tissue.

The gut compartment technique was refined to allow for a greater ease of experimental set up. The excised portion of intestine was sealed at each end to give a spherical sac of gut wall (Wilson and Wiseman 1954). An *in vivo* technique in dogs was also developed where the vein draining a section of intestine was cannulated to collect blood containing absorbed nutrients (Sheff and Smyth 1955). Only amino acids and GlyGly could be detected *in vivo*, either inside the luminal tissue or in the blood leaving the small intestine (Newey and Smyth 1962).

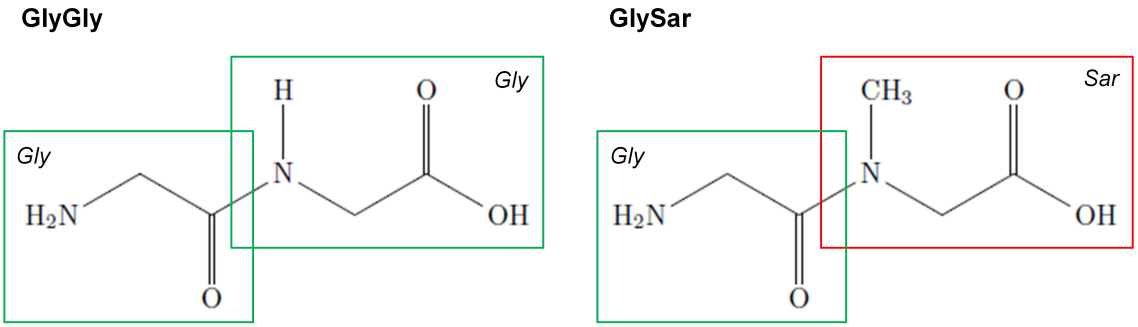


Figure 1.2: A comparison between GlyGly and the non-hydrolysable peptide glycyl-sarcosine. The structure of glycyl-sarcosine is similar to the di-peptide, GlyGly. The methyl group attached to the the nitrogen of sarcosine, makes the glycyl-sarcosine peptide bond resistant to conventional peptidase peptide hydrolysis.

These early studies were very limited by the methods and peptides available for experimentation. Since peptide transport and epithelial cell hydrolysis could not be distinguished from amino acid transport and luminal hydrolysis, extensive efforts were made to discover the optimal area of peptidase activity. This was found to be the gut epithelial tissues (Newey and Smyth 1962), suggesting that proteins were partially hydrolysed into small peptides and absorbed into the luminal cells where hydrolysis was completed.

The advance, which allowed the separation of peptide transport and hydrolysis was the use of the non-hydrolysable peptide glycyl-sarcosine (GlySar) (Figure 1.2). Sarcosine is composed of a glycine molecule with a methyl group attached to primary amine. When incorporated into the di-peptide GlySar, this methyl group makes the peptide amine proton resistant to nucleophilic attack, and therefore resistant to peptide hydrolysis. Investigations using GlySar as a substrate found that it built up in the luminal tissues (Cheeseman, Newey, and Smyth 1971; Matthews, Burston, and Addison 1972; Addison *et al.* 1975). This observation was later confirmed for hydrolysable peptides (Rubino, Field, and Shwachman 1971; Himukai, Suzuki, and Hoshi 1978). Only di- and tri-peptides appeared to be transported, and their transport appeared to be separate from amino acid transport. Amino acids did not inhibit peptide transport, and visa versa, whereas other di- and tri-peptides did inhibit peptide uptake.

A new technique in preparing the micro-villus membranes into vesicles (Forstner, Sabesin, and Isselbac 1968) allowed for much faster measurements of transport such that intra-lumen peptidase activity could be measured. Sigrist-Nelson (1975) found that 5 min after peptide absorption into the micro-villus membrane vesicles, 95% of peptide was hydrolysed.

This reinforced the idea that the gut wall was the final place of peptide hydrolysis. These experiments suggested that potentially a single transporter existed which transported all combinations of di- and tri-peptides (Payne and Gilvarg 1971).

> *In addition to this potential structural variation within any groups of dipeptides, tripeptides and so on, there is the variation in size to be encompassed if all peptides are to be classed together. A priori, therefore, it seems that the ideal way to ensure peptide transport resides in designing a minimum number of systems that possess requirements for only those unique structural features common to all peptides.* ***Payne and Gilvarg (1971)***

At this time work was also being done on the driving force behind nutrient uptake. General transport inhibition was observed in peptide uptake when stress conditions were applied to the gut wall such as; anoxia, 2,4-dinitrophenol and cyanide, which all inhibit the generation of Adenosine triphosphate (ATP) and so limit cellular energy production (Newey and Smyth 1962; Matthews, Burston, and Addison 1972; Addison *et al.* 1975). Inhibition was also observed for peptide and amino acid transport when the Na^+ ions in the medium were exchanged for either K^+ or Li^+ ions, although this inhibition was not consistent across all studies.

The idea that solute transport appeared to be driven by ion gradients was relatively new (1960). Investigations into peptide and amino acid absorption were happening in conjunction with other nutrients such as carbohydrates. The evidence that sugars were transported as monosaccharides was quite strong and their absorption was dependent upon the Na^+/K^+ ATPase (Crane, Forstner, and Eichholz 1965). These investigations had established that a potential difference existed across the gut wall and that sugar transport could generate a flow of current (Barry *et al.* 1964). This approach was then applied to amino acid and peptide transport, indicating both systems generated a current (Kohn, Smyth, and Wright 1968).

However, the more this was studied, the less it seemed that amino acid and peptide uptake were driven by the same processes. *In vitro* studies of amino acid and peptide transport were found to stimulate differing amounts of Na^+ absorption (Hellier, Thirumalai, and Holdsworth 1973). In large polypoid cells from the *Necturus maculosus* (mudpuppy), peptide was able to drive current in Na^+ free media (Boyd and Ward 1982). The Na^+/K^+

ATPase was proposed as a possible solution to this discrepancy. The ATPase resides in the basolateral membrane of the luminal epithelium where it simultaneously drives the efflux of Na^+ and uptake of K^+ ions, thereby maintaining an intra-cellular negative potential. Harmaline, an inhibitor of the ATPase, was observed to inhibit both amino acid and peptide uptake (Ganapathy and Radhakrishnan 1980), but the peptide inhibition was dependent upon the presence of Na^+ ions.

1.2.2 Peptide absorption in the kidney

The emerging model of gut peptide absorption was complicated by the discovery of peptide transport in the kidneys. An experiment looking into the clearance of peptide from the blood found that GlySar was not excreted in appreciable volumes and that it remained trapped in the kidney epithelium (Adibi, Krzysik, and Drash 1977). Peptide hydrolysis and amino acid uptake was known to occur in the kidney proximal tubule after blood filtration by the glomerulus (Ganapathy *et al.* 1980). However, it was not thought that peptide transport activity was present. It was found that both peptides and amino acids were transported by rabbit kidney brush border membrane vesicles but by separate mechanisms (Ganapathy *et al.* 1980; Ganapthy, Mendicino, and Leibach 1981). Only the amino acid transport appeared to be Na^+ dependent.

1.2.3 Peptide transport is driven by H^+ electro-chemical gradient

The driving force of peptide transport was eventually found to be the H^+ gradient (Ganapathy and Leibach 1983). In a series of experiments, the transport of peptides into intestinal and renal membrane vesicles were shown to be sensitive to extra-cellular pH. Transport could be stimulated by valinomycin, an ionophore, which allows K^+ ions to travel down their concentration gradient across the membrane, and thereby increasing the negative intra-cellular potential and stimulating a H^+ potential. Transport could be inhibited by application of Carbonyl cyanide-4-(trifluoromethoxy)phenylhydrazone (FCCP) which acts as a proton ionophore, dissipating the H^+ potential. Over thirty years of research were summarized into a very elegant diagram (Figure 1.3) of peptide transport (Ganapathy and Leibach 1985).

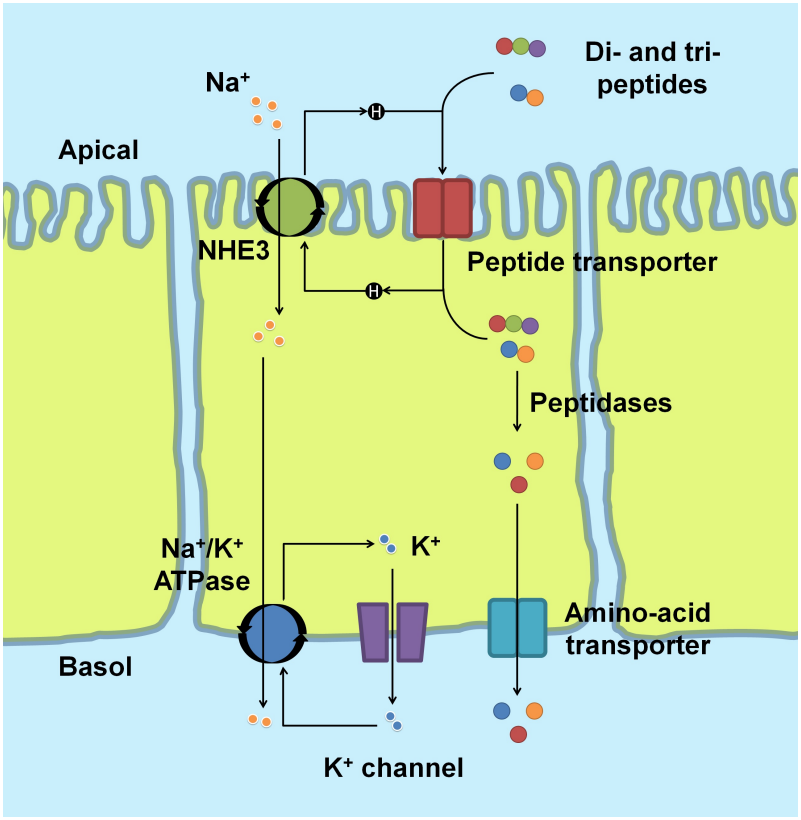


Figure 1.3: Representation of peptide transport in the intestinal and kidney brush boarder epithelia. The peptide transporter is driven by the apical H^+ gradient maintained by the NHE3. The intra-cellular Na^+ concentration is sustained by the K^+ channel and Na^+/K^+ ATPase. Peptides are transported by the peptide transporters and cleaved into amino acids by epithelial peptidases. Amino-acids are then transported across the basolateral membrane by amino acid transporters. This diagram has been adapted from Ganapathy and Leibach (1985) and Daniel *et al.* (2006).

The absorption of dietary nitrogen across the intestinal epithelia is maintained by three ion gradients; H^+, Na^+ and K^+. The peptide transporters specifically couple the apical membrane transport of di- and tri-peptides to the H^+ gradient. The Na^+/K^+ ATPase and K^+ channel maintain a low intra-cellular Na^+ concentration, which sustains the apical H^+ gradient driven by the Na^+/H^+ exchanger (NHE3). Peptides transported across the apical membrane are rapidly hydrolysed and the free amino acids are transported across the basolateral membrane by amino acid transporters.

1.3 Characterisation of the PepT1 and PepT2 genes

The cloning of the mammalian intestinal and renal peptide transporters, PepT1 (Fei *et al.* 1994; Liang *et al.* 1995) and PepT2 (Boll *et al.* 1996; Liu *et al.* 1995) respectively allowed

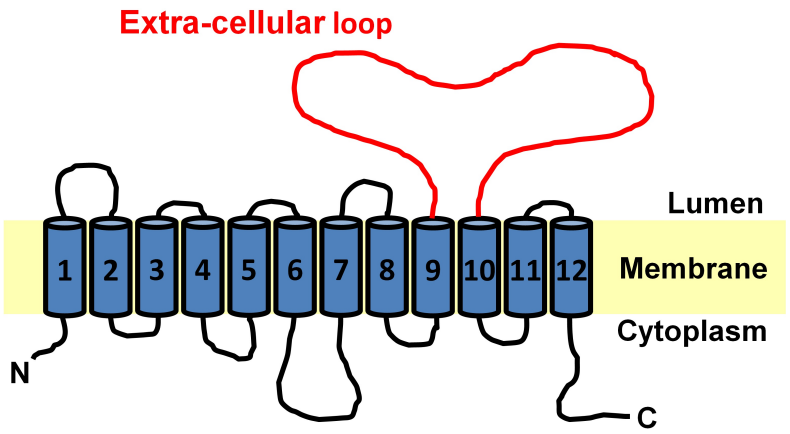


Figure 1.4: Peptide transporter topology. The 12 TM helices are shown in blue and the 20 kDa extra-cellular loop inserted between helices 9 and 10 is shown in red.

for a more stringent study into their similarities and differences. Hydropathy analysis indicated that the transporters were composed of 12 Trans-membrane (TM) helices (Figure 1.4) with a large extra-cellular loop inserted between TM helices 9 and 10. This extra-cellular loop is the principal study of this thesis and is explored further in § 1.6.

The sequence alignment of *H. sapiens* PepT1 and PepT2 indicated that they share 46% sequence identity. Transient expression of the transporters in *Xenopus laevis* oocytes suggested that, although the transporters appeared to share the same substrates, there was a difference in the apparent affinity (apparent K_m) of PepT1 and PepT2, 50 μM and 1000 μM respectively, for cefadroxil (peptide-mimetic substrate of PepT1 and PepT2 – discussed in § 1.3.2) (Fei *et al.* 1994; Boll *et al.* 1996). PepT1 appeared to be a high capacity/low affinity transporter and PepT2 a low capacity/high affinity transporter. This finding was in keeping with the idea proposed in § 1.1 that PepT1 is the primary route for the absorption of dietary protein.

1.3.1 PepT1 and PepT2 tissue localisation

The cloning of the mammalian genes allowed for a better appreciation of tissue type expression. PepT1 expression (discussed in § 1.2) was well documented in the small intestine but was not found in other parts of the gastrointestinal tract *i.e.* the oesophagus, stomach, or large intestine (Ogihara *et al.* 1996; Shen *et al.* 1999). PepT1 does not appear to be expressed in the colon of healthy individuals (Ogihara *et al.* 1996); however, PepT1 expression was observed in the colon of individuals with either Irritable bowel syndrome

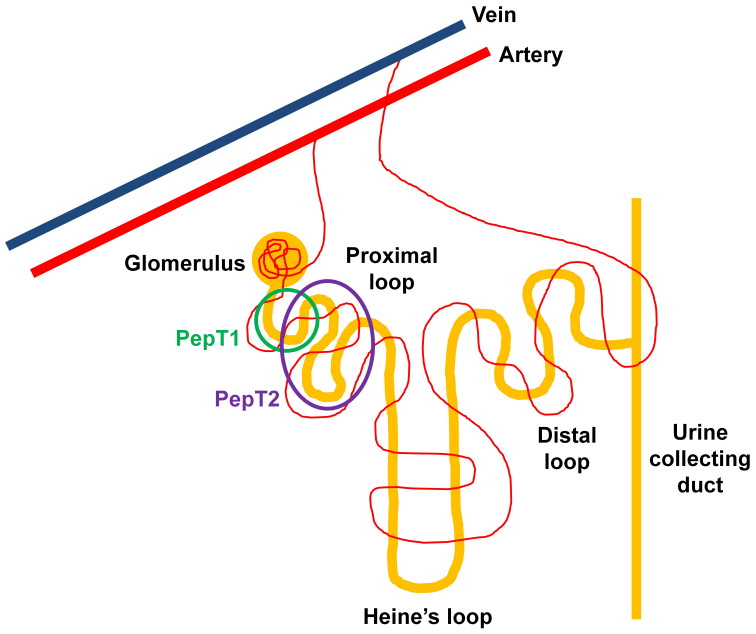


Figure 1.5: PepT1 and PepT2 localisation in the kidney nephron. Blood flows from the renal artery into the glomerulus, which acts as a 30 kDa filter of the blood. The proximal, Heine's and distal loops recover material still useful for the body. PepT1 and PepT2 expression in the proximal loop is shown by green and purple circle respectively.

(IBS) or Short bowel syndrome (SBS) (Merlin *et al.* 2001; Ziegler *et al.* 2002). IBS and SBS result in an inflamed small intestinal lumen or truncated small intestine respectively, and ultimately poorer uptake of dietary nutrients compared to Wild type (WT) individuals (Urban and Weser 1980). The alteration of PepT1 expression in IBS and SBS suggests, that PepT1 expression can be altered in the colon in response to poor nutrient absorption in the small intestine, and therefore the existence of a feed-back loop in PepT1 intestinal expression.

PepT1 expression has also been found in the kidney. PepT1 and PepT2 are both expressed in the kidney nephron proximal tubule with asymmetric localisation (Figure 1.5). PepT1 expression is confined to the *pars convoluta* whereas PepT2 is expressed predominantly in the *pars recta* (Smith *et al.* 1998; Shen *et al.* 1999). The reason for this dual asymmetric expression of PepT1 and PepT2 is not clear. PepT2 appears preferentially to transport GlySar over PepT1 in the proximal tubule (Silbernagl, Ganapathy, and Leibach 1987), suggesting that PepT2 is the dominant transport mechanism of peptides in the kidney. The observed expression of PepT1 and PepT2 in the kidney led to the hypothesis that the peptide transporters acts as a reabsorption mechanism for plasma peptides which pass into the nephron during blood plasma filtration (Figure 1.5).

PepT2 expression has been observed in the lung (Saito *et al.* 1996; Groneberg *et al.* 2001; Groneberg *et al.* 2002; Quarcoo *et al.* 2009; Li *et al.* 2013). Lung epithelial tissue was shown to contain PepT2 Messenger ribonucleic acid (mRNA), and white lung preparations are able to transport fluorophore conjugated peptide, indicating functional PepT2 expression. It is not clear what the primary function of PepT2 is in the lung. It is a possibility that PepT2 functions as part of the innate immune system (Swaan *et al.* 2008), as a similar system has been observed in the small intestine involving PepT1 (Merlin *et al.* 1998; Dalmasso *et al.* 2010). The peptidoglycan layer of Gram negative bacteria which line the mucosal epithelia releases immunogenic tri-peptides (Marasco *et al.* 1984) which can be absorbed by the PepT1 and PepT2. The internalised antigens are recognised by nucleotide binding oligomerisation domain receptors which oligomerise to propagate an inflammatory response (Merlin *et al.* 1998; Swaan *et al.* 2008; Dalmasso *et al.* 2010). PepT1 and PepT2 potentially function as an antigen gathering mechanism, sampling the epithelial mucosa for signs of pathogens.

PepT2 has also been found in the mammalian brain (Saito *et al.* 1996; Dringen, Hamprecht, and Bröer 1998; Wang *et al.* 1998) with homogeneous expression throughout, but significant expression in the choroid plexus epithelial cells (Berger and Hediger 1999). The choroid plexus acts as a barrier between the blood and the Cerebrospinal fluid (CSF) which surrounds the brain and contains neuronal nutrients and waste. PepT2 has been implicated in the clearance of peptides from the CSF, acting as an efflux pump (Teuscher, Keep, and Smith 2001; Shu 2002) and ensuring CSF homoeostasis (Ocheltree *et al.* 2005; Hu *et al.* 2007). The expression of PepT2 in the choroid plexus and the lung highlight the versatility of the transporters and the necessity for a deeper understanding of the transporters to assess the factors which influence their activity. This is also highly relevant to medicine, as PepT1 and PepT2 have been implicated in the transport and retention of peptidomimetic drugs (Brandsch, Knütter, and Bosse-Doenecke 2008).

1.3.2 Role in drug transport

PepT1 and PepT2 are thought to transport all combinations of di- and tri-peptides, which gives them a substrate pool of 8,400 compounds. Given this enormous range of substrates, it is perhaps not surprising that other, non-peptide based molecules fit the PepT1

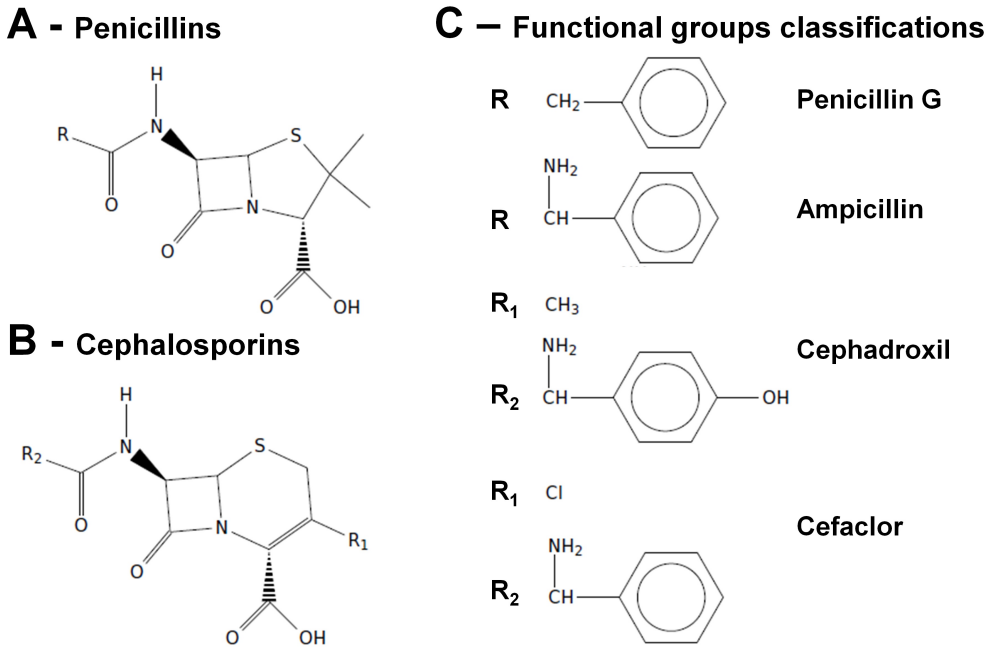


Figure 1.6: Two families of β-lactam antibiotics. The basic functional unit of the penicillins **(A)** and the cephalosporines **(B)**. **(C)** The R, R_1 and R_2 groups which are present in penicillin G, ampicillin, cefadroxil and cefaclor.

and PepT2 substrate pharmacophore. Transport of β-lactam antibiotics (Figure 1.6) by the small intestine was established using the everted sac method (Miyazaki *et al.* 1982; Nakashima *et al.* 1984) described in § 1.2.1. The link between β-lactam transport and PepT1 was identified in brush border membrane vesicles by titrating di-peptide into the reaction which inhibited β-lactam transport (Okano *et al.* 1986). The removal of the H^+ gradient by the addition FCCP also inhibited β-lactam transport (Inui *et al.* 1988). A similar link was also found between PepT2 and β-lactam antibiotic transport (Inui *et al.* 1984).

Heterogeneously expressed PepT1 and PepT2 in *X. laevis* oocytes confirmed these results (Fei *et al.* 1994; Boll *et al.* 1996). This firmly established the link between the peptide transporters and β-lactam antibiotics, and suggested that a drugs absorption and excretion could be directly mediated by PepT1 and PepT2 respectively. Hence a pharmacokinetic mechanism had been discovered which could now be used in the design of new therapies. Several β-lactam unrelated compounds have also been shown to be transported by PepT1; the anti-virals valacyclovir (Ganapathy *et al.* 1998) and oseltamivir (Poirier *et al.* 2012), and the anti-cancer agent ubenimex (bestatin) (Nakanishi *et al.* 1997) (Figure 1.7). It is not clear yet how these drugs are recognised by the peptide transporters, although it

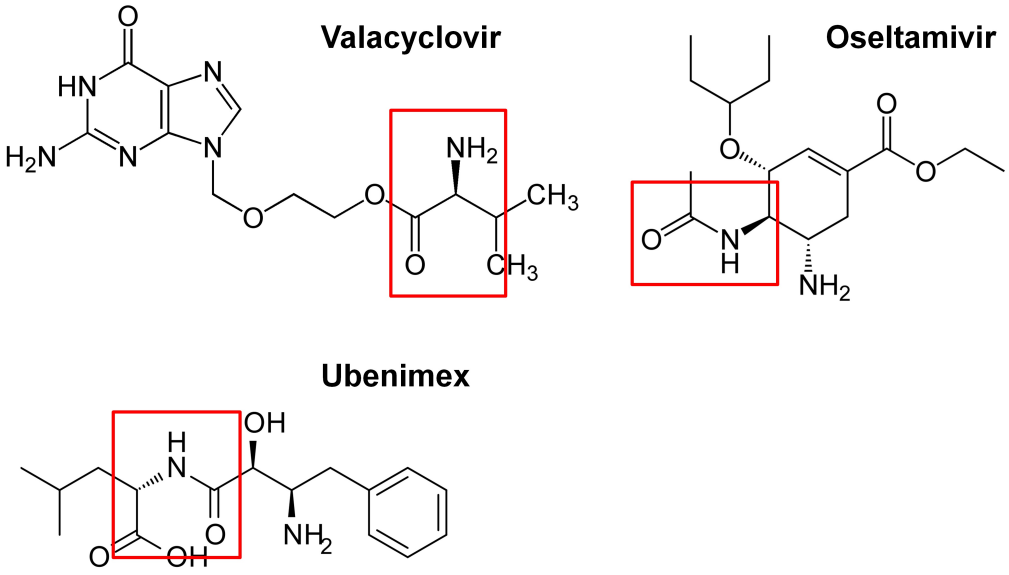


Figure 1.7: The structures of valacyclovir, oseltamivir and ubenimex. The peptide bond in each of the structures has been highlighted by a red square. It is not clear now these compounds are transported by PepT1, but the peptide bond is a common feature which may be recognised by the transporter.

seems likely that the peptide bond plays a role in substrate recognition.

1.3.3 Evolutionary conservation of the eukaryotic and bacterial peptide transporters

Peptide transporters are also observed in bacteria. Three independent peptide transport systems have been identified; the Oligo-peptide permease (Opp) (Barak and Gilvarg 1974), the Di-peptide permease (Dpp) (Ganapthy, Mendicino, and Leibach 1981) and the Tri-peptide permease (Tpp) (Barak and Gilbarg 1975; Naider and Becker 1975). The Opp system transports di- to penta-peptides and is composed of five genes organised onto a single operon in both *Escherichia coli* and *Salmonella typhimurium*; OppABCDF (Hogarth and Higgins 1983; Hiles *et al.* 1987). The genes work in concert, coupling the concentrative transport of peptide to ATP hydrolysis (Higgins *et al.* 1985), and are part of a wider family which is called the ATP binding cassette (ABC) transporter superfamily (Higgins *et al.* 1986; Hyde *et al.* 1990). Dpp is another ABC transporter which predominantly favours the transport of di-peptides (Abouhamad and Manson 1994).

Tpp, unlike the Opp and Dpp, was found to be a system in which a single protein coupled the transport of di- and tri-peptides to the Proton motive force (PMF) (discussed in

§ 1.4.1) (Hagting *et al.* 1994). Sequence alignments indicated an ancestral link between mammalian PepT1 and PepT2, and the bacterial Tpp (Figure 1.8). Steiner, Naider, and Becker (1995) proposed the creation of the Proton oligo-peptide transporter (POT) family. Paulsen and Skurray (1994) independently proposed Peptide transporter (PTR) family as it was not yet established that all the members were secondary active transporters driven by proton gradients (discussed in § 1.4.1). Given that POT better describes the family, this name is used throughout this thesis. However, both names pervade the literature so it is necessary to describe their origins. The POT family members are identified by 3 sequence motifs identified by Daniel *et al.* (2006); ExxERFxYYG (TM 1), PTR2 – 1 (TM 2 – 3) and PTR2 – 2 (TM 5).

Figure 1.8 shows a sequence alignment of POT family members from bacteria to higher eukaryae. As can be seen from the alignment, there is conservation in TM helices between the POT family members from different organisms, but little to none in the loops which connect the helices. The principal area which lacks sequence conservation is the extracellular loop region, between TM helices 9 and 10. This loop is absent in bacterial, single celled eukaryae and plant POT family members; the loop is only present in the animal peptide transporters. The reason for this is not currently clear.



Figure 1.8: A sequence alignment of a subset of POT family transporters. The alignment shows the conservation between a series of POT family proteins (The full names and Uniprot codes for each gene are listed in Table 2.1). Sequence conservation is highlighted in blue. The TM helices have been labelled 1 – 12 and the extracellular loop region is shown in green. The POT family motifs (Daniel *et al.* 2006); ExxERFxYYG, PTR2 – 1 and PTR2 – 2 are coloured in purple, cyan and orange respectively. Two extra TM helices observed in the bacterial POT members (discussed in § 1.4.2), HA and HB are coloured in yellow. The predominant areas of sequence conservation are TM helices 1 – 12 and the extra-cellular loop is only observed in animal sequences.

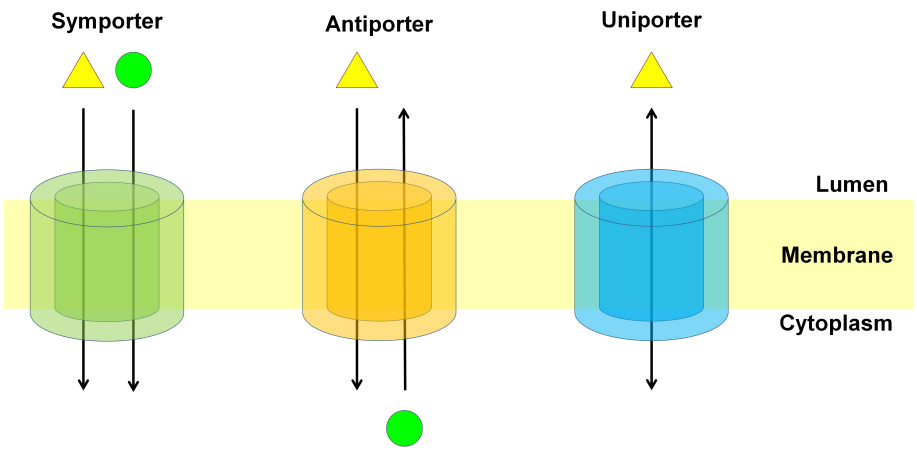


Figure 1.9: Secondary active transporter classifications. There are three types of secondary active transporters; symporters, antiporters and uniporters. Symporters and antiporters couple substrate transport to an ion gradient; whereas, uniporters allow substrate to cross the membrane maintaining a substrate equilibrium on either side.

1.4 Transporter classification and mechanism

1.4.1 Secondary active transport

The POT family are classed as secondary active transporters. They are 'active' because they drive the accumulation of substrate against its concentration gradient. The ABC transporters are primary active because they directly couple the transport of substrate to the hydrolysis of ATP. Secondary active transporters indirectly couple the accumulation of substrate to an energy donating molecule by an ion potential across the membrane (Forrest, Krämer, and Ziegler 2011). In the case of the bacterial peptide transporters, this is the PMF generated by the export of H^+ donated from protonated nicotinamide adenine dinucleotide and flavin adenine dinucleotide, accumulated during the citric acid cycle (Pao, Paulsen, and Saier 1998). Whereas in the case of the mammalian peptide or sugar transporters, the H^+ and Na^+ gradients are maintained by the Na^+/K^+ ATPase, an ion pump which exchanges Na^+ and K^+ ions powered by ATP hydrolysis.

Secondary active transporters are divided into three distinct classes (Figure 1.9); symporters, which couple solute transport to a parallel membrane potential (PepT1 and PepT2), antiporters which exchange membrane polarised substrates, and uniporters which allow the passage of substrate down a concentration gradient (Shi 2013). The predominant gradients used by symporters and anti-porters are the monovalent cations, H^+ and Na^+. The ion gradient broadly used by a given organism to drive solute transport is a

reflection of their environment. Gut pathogens such as *E. coli* have predominantly Na^+ coupled symporters, whereas autotrophic bacteria rely almost solely on the H^+ gradient of the PMF (Ren and Paulsen 2007).

Solute carrier classification

The sequencing of the human genome led to an effort to categorise all human membrane associated proteins. The Solute carrier (SLC) series of secondary active transporters are the largest transporter group (Fredriksson *et al.* 2008) and is composed of 52 protein families encompassing 418 genes collectively. PepT1 and PepT2 are in family SLC15 along with the peptide-histidine transporters, PhT1 and PhT2 (Smith, Clémençon, and Hediger 2013), which are proposed to be lysosomal transporters involved in removing degraded protein from the lysosome. In comparison with PepT1 and PepT2, little information is available on PhT1 and PhT2. The cloning of the genes indicated that both are able to transport di-peptides and histidine, although a complete substrate analysis has not been performed. Similar to PepT1 and PepT2, they appear to be asymmetrically distributed with PhT1 expression detected in the CNS, and PhT2 expression found in the lung, spleen and thymus (Sakata *et al.* 2001). The overall sequence identity between PepT1 and PepT2, and PhT1 and PhT2 is 20%; however, the POT motifs identified by Daniel *et al.* (2006) are present, strongly suggesting a functional link. The major sequence difference between the peptide and peptide-histidine transporters is the presence of the large extra-cellular loop observed between TM 9 and 10 in the former.

There are two interesting observations regarding PepT1 and PepT2 that can be made from inspection of the SLC tables. The first is the lack of multiple peptide transporter genes. The full roles played by PhT1 and PhT2 in bodily peptide transport are not understood. However, even counting their existence, there are only four transporters which appear to be able to transport peptides in the body. This is striking, considering that there are 7 amino acid and 26 sugar transporters. Given the substrate variability of di- and tri-peptides (discussed in § 1.3.2), this is highly atypical and suggests a promiscuous binding cavity within the transporters, perhaps similar to the multi-drug exporters (Nikaido 2011).

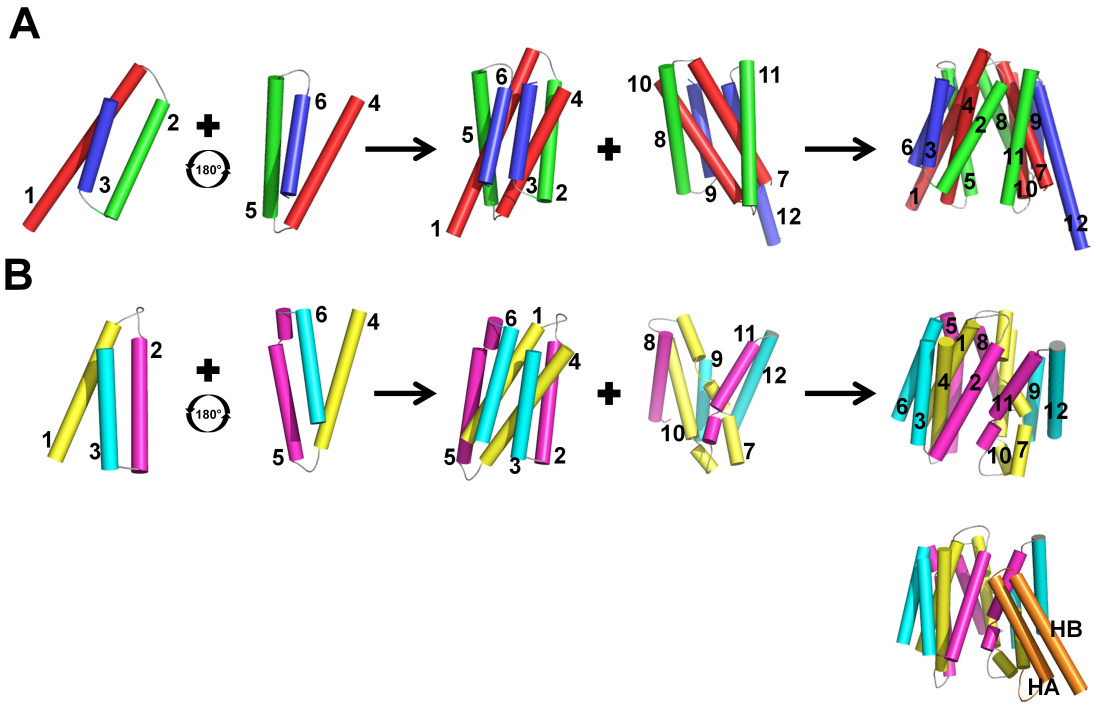


Figure 1.10: The MFS fold of GlpT and PepT$_{So}$. (A) and **(B)** show the organisation of the MFS fold in GlpT (Huang *et al.* 2003) (PDB: 1PW4) and PepT$_{So}$ (Newstead *et al.* 2011) (PDB: 2XUT) from *E. coli* and *S. oneidensis* respectively. Three helical bundles are inverted and repeated throughout the MFS fold. The pseudo symmetric helices are coloured red – green – blue and yellow – pink – cyan, in GlpT and PepT$_{So}$ respectively. The 2 extra TM helices of PepT$_{So}$, HA and HB are highlighted in orange. This figure was adapted from Shi (2013).

1.4.2 Major facilitator superfamily

The POT family is a member of the Major facilitator superfamily (MFS), a large ubiquitous superfamily accounting for approximately 50% of all known secondary active transporters (Pao, Paulsen, and Saier 1998; Chang *et al.* 2009; Reddy *et al.* 2012). Members of the MFS predominantly have 12 TM helices which are composed of two pseudo-symmetric 6-helical bundles with an intra-cellular N- and C-termini. There is an MFS sequence motif Gxxx(D/E)(R/K)xxG(R/K)(R/K) in the intra-cellular loop between TM helices 2 and 3, and to a lesser extent TM 8 and 9 (Maiden *et al.* 1987). However, this motif was first identified in the bacterial MFS sugar transporters and is only weakly observed in the mammalian peptide transporters. Two other common folds are observed in transporters, the LeuT fold and the NhaA fold (Shi 2013). No known peptide transporters have the LeuT or NhaA fold; however, they are of note when considering secondary active transporter mechanism.

The first crystal structures of the MFS fold, the *E. coli* lactose permease, LacY (Abramson *et al.* 2003), and glycerol-3-phosphate anti-porter, GlpT (Huang *et al.* 2003) highlighted the pseudo symmetrical architecture observed throughout the MFS. Figure 1.10A shows the structure of GlpT and highlights the construction of the MFS fold. The MFS fold is essentially multiple clusters of 3 helical repeats, where each adjacent repeat is related by a 180 ° rotation. In a 12 TM helical protein, this creates two 6-helical bundles joined by an intra-cellular linker referred to in this thesis as the N- and the C-terminal bundles. The majority of MFS transporters have 12 TM helices made up of two core 6-helical bundles. MFS transporters can deviate from 12 TM helices either by 2 TM helix insertions at either end of the bundle or by multiples of the 6-helical bundle, for example the nitrate transporter, NarK from *Paracoccus denitrificans* has 24 TM helices. The bacterial peptide transporters are a good example of the 2 TM helix insertion, although some do only have 12 TM helices. Figure 1.10B shows the MFS fold of $PepT_{So}$ from *Shewanella oneidensis* which is composed of 14 TM helices. As can be seen, the core transporter is built in the same manner as GlpT. The extra TM helices HA and HB (highlighted in Figure 1.8) are inserted in the linker between the N- and C-terminal bundles.

1.4.3 The alternating access mechanism of solute transport

The first explanation of a secondary active symporter mechanism, the alternating access model, was proposed by Jardetzky (1966). Essentially, the membrane embedded protein was proposed to have a substrate binding site which during the transport cycle would alternate access between the apical and basal membrane compartments. Since the solution of the first bacterial X-ray crystal structures, this model has been refined to a rocker switch (Figure 1.11) (Law *et al.* 2007). The rocker switch assumes that the two halves of the protein act in concert during the transport cycle. Solute and ion bind the protein in an Outward open (OO) conformation. The protein then undergoes a conformational change to an Occluded bound (OB) state by closing the extra-cellular gate. Another conformational change is required to open the intra-cellular gate to an Inward open (IO) state for solute and ion to leave the transporter. The transporter then cycles back to the OO conformation. Two intermediate states, are also known to exist between the open bound states and OB, called partially occluded, of which the crystal structure of $PepT_{So}$

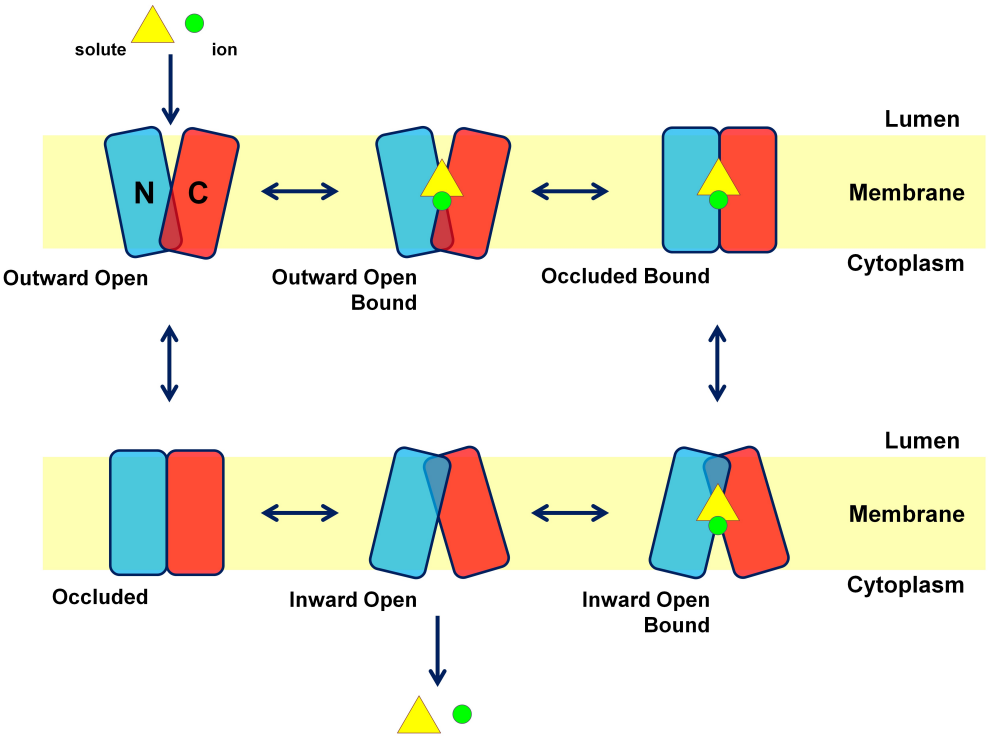


Figure 1.11: The alternating access rocker switch mechanism of MFS symport. The N- and C-terminal 6-helical bundles of the transporter are coloured in red and blue respectively. Solute and ion bind the symporter in an outward open state, where the intra-cellular gate of the symporter is closed. Upon binding of the substrates, the symporter undergoes a conformational change into an occluded bound state, closing the extra-cellular gate of the protein. The intra-cellular gate then opens whilst the extra-cellular gate remains closed, allowing for solute and ion to enter the cytoplasm of the cell. The transporter then cycles back into the outward open conformation. This figure was adapted from (Law *et al.* 2007).

is an example. This alternating access rocker switch mechanism appears to be consistent with the all known structures of secondary active transporters.

The MFS symporters are well described by this transport mechanism. The two halves of the protein hypothesised are easily definable in the N- and C-terminal 6-helical bundles. Figure 1.12 shows a selection of MFS symporter structures in the majority of conformations proposed by the alternating access rocker switch mechanism. However, there are gaps in the picture, as there are currently no available structures for the OB state. The structure of $PepT_{So}$ in the Partially occluded inward bound (POIB) conformation does show density in the binding site; however, it was not possible to correctly assign a molecule to it. The lack of the OB structure is not surprising, given that it represents a high energy state of the transporter (Law *et al.* 2007) and is therefore unlikely to crystallise.

A crystal structure of each transition through the transport cycle is not an impassable barrier to the study of MFS transporters. As the study of the bacterial peptide transporters has shown (discussed in § 1.4.4), insight into the transport cycle can be achieved from only having part of the model. Ultimately crystal structures themselves are models, which require validation by mutational studies to prove residue interactions. That said, structures are very useful guides when elucidating mechanisms. Trying to do so without a structural template, often leads to erroneous assumptions and studies of the mammalian peptide transporters are a good example of this.

1.4.4 Alternating access model in bacterial peptide transport

Since the sequencing of mammalian PepT1 and PepT2, a large amount of research has been performed to elucidate the transport mechanism. This was primarily done through mutational studies using heterologous expression transport assays which utilised either, tritium labelled substrate or a Two electrode voltage clamp (TEVC) method. It is hard to infer a mechanistic effect from these assays as no quality control of the protein can be performed. The gold standard method for the analysis of a membrane transporter is a liposome assay, where protein has been over-expressed, purified and then reconstituted in a liposome. The principle advantage of a liposome assay over a whole cell assay is that all variables can be controlled, and therefore a change in the independent variable can be inferred to the dependent variable. The principal disadvantage is that the protein has to be

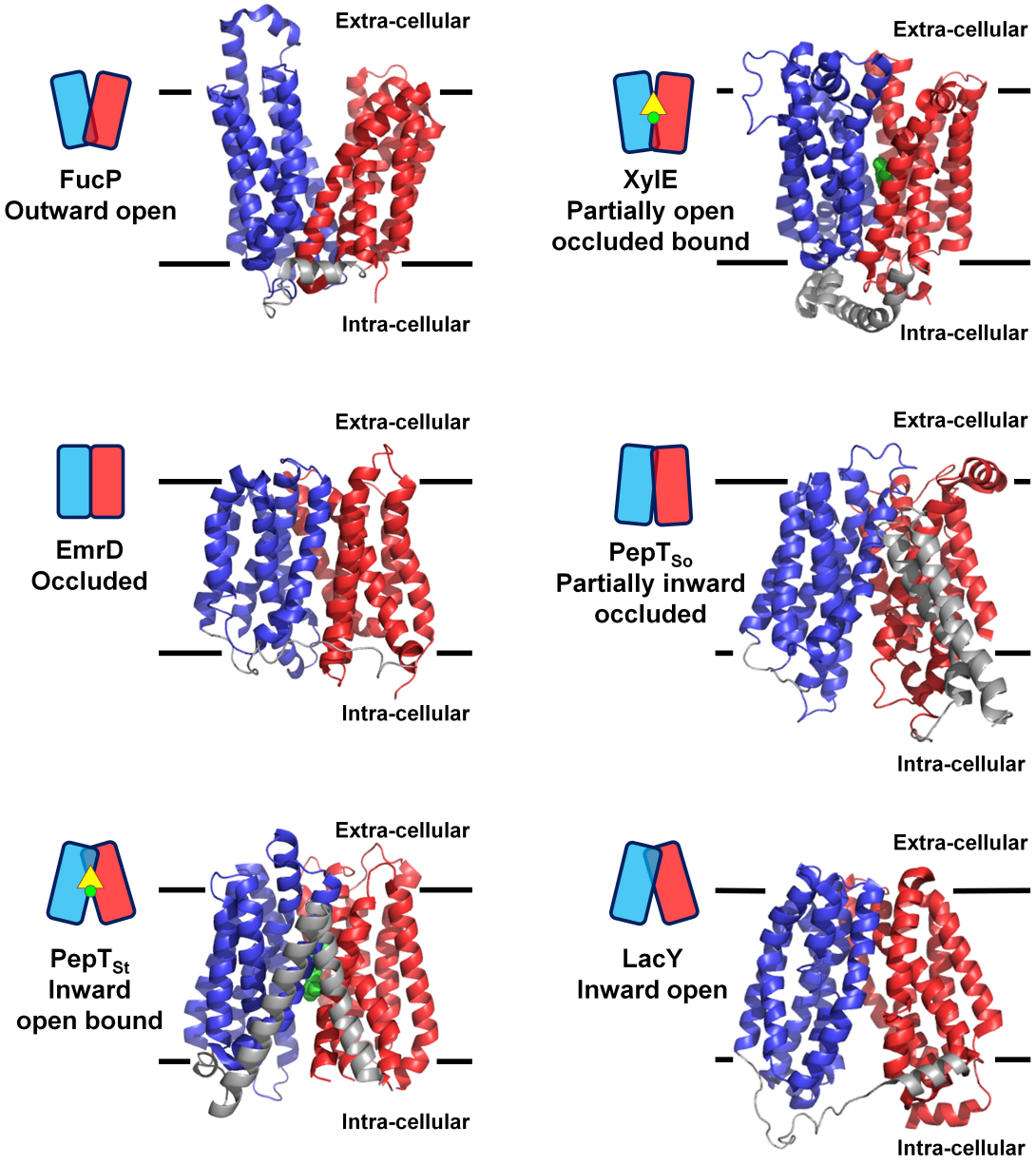


Figure 1.12: Rocker switch conformations of MFS symporters. A representative sample of the current MFS symporters in the proposed rocker switch conformations; FucP (PDB: 3O7Q), XylE (PDB: 4GBY), EmrD (PDB: 2GFP), $PepT_{So}$ (PDB: 2XUT), $PepT_{St}$ (PDB: 4D2D) and LacY (PDB: 1PV6). The N- and C-terminal 6-helical bundles have been coloured red and blue respectively and the linkers in grey. The partially occluded structures of *E. coli* XylE and $PepT_{So}$ are half-way between the full occluded and open states.

stable enough to be expressed and purified. No current protocol exists to over-express and purify mammalian PepT1 or PepT2; therefore only whole cell assays have been employed.

The other essential component in performing functional studies is a reasonable model of the system being investigated. Two models of PepT1 have been created to date. The first model was produced by Bolger *et al.* (1998) using the TM helical prediction proposed by Fei *et al.* (1994) and minimising the clashes between adjacent helices. The second was produced by Meredith and Price (2006) using the crystal structures of LacY (Abramson *et al.* 2003) and GlpT (Huang *et al.* 2003) as a template for a homology model. The crystal structures of $PepT_{So}$ and $PepT_{St}$ (Newstead *et al.* 2011; Solcan *et al.* 2012) which share 31% and 21% sequence identity respectively with *H. sapiens* PepT1 and PepT2, indicated both the Bolger *et al.* (1998) and Meredith and Price (2006) models were poor representations of peptide transporters. The purification of the bacterial homologues for the first time allowed for liposome assays to be performed on peptide transporters, and the generation of a template model for peptide transport.

The substrate channel salt bridges

A series of salt bridges, which flank the binding site of the transporter and the extra- and intra-cellular openings of the substrate channel, have been identified in the crystal structures of $PepT_{So}$ (Newstead *et al.* 2011), $PepT_{St}$ (Solcan *et al.* 2012; Lyons *et al.* 2014) and *Gk*POT (Doki *et al.* 2013). Figure 1.13A and B show the salt bridges identified in $PepT_{So}$. The bridges are numbered 1 – 4 from the extra- to the intra-cellular compartment and labelled N and C, denoting the N- and C-terminal bundles respectively. Figure 1.13C shows the comparable salt bridge residues in $PepT_{St}$ and *Gk*POT. Solcan *et al.* (2012) established two liposome based assays for $PepT_{So}$ and $PepT_{St}$, H^+-driven and counter-flow, which were subsequently repeated in Doki *et al.* (2013). The H^+-driven assay used an internal negative potential to drive the concentrative uptake of tritium labelled peptide into the liposome via the H^+ coupling. The counter-flow assay does not use an ion gradient, but fills the liposome with an excess of cold substrate, and surrounds the liposome with a fractional concentration of hot substrate. The purpose of the two assays was to distinguish between the different potential properties present for residue *i.e.* does the residue contribute to a transporter configuration which supports peptide or H^+ transport, both

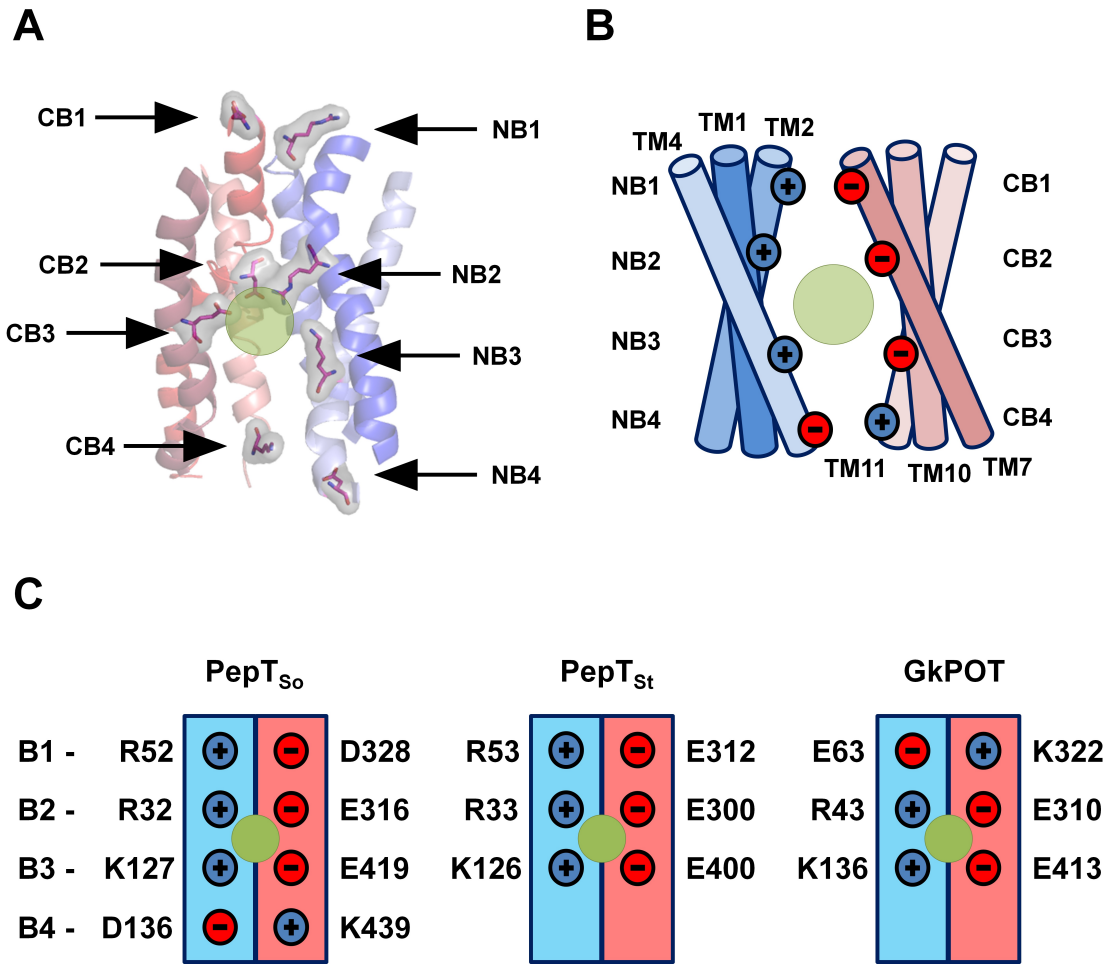


Figure 1.13: The salt bridges which line the substrate channel in the bacterial peptide transporters. The N- and C-terminal bundles have been coloured blue and red respective and the binding site is shown as a green circle. **(A)** and **(B)** show the salt bridges on the N- and C-terminal bundles of $PepT_{So}$ (PDB: 2XUT). The N- and C-terminal salt bridges, NB and CB respectively, line the substrate channel and have been numbered $1-4$ from the extra- and intra-cellular compartment. **(C)** shows the corresponding salt bridge residues from $PepT_{So}$ (Newstead *et al.* 2011), $PepT_{St}$ (Solcan *et al.* 2012; Lyons *et al.* 2014) and *Gk*POT (Doki *et al.* 2013).

or neither.

Solcan *et al.* (2012), Doki *et al.* (2013) and Solcan (2013) established that an alanine mutation in any bridge 3 residue completely impairs transporter function in the H^+-driven assay. In counter-flow however, only the CB3 residues removed transport. In counter-flow, the NB3 mutation reduced transport by approximately 50% in $PepT_{So}$ and *Gk*POT but had no effect in $PepT_{St}$. A similar effect was found in bridge 2. The CB2 mutations are not functional in either substrate driven or counter-flow. The NB2 mutations shows no reduction in transport in both H^+-driven and counter-flow for $PepT_{So}$ and *Gk*POT. However, in $PepT_{St}$ there is a large and moderate reduction in H^+-driven and counter-flow

respectively.

These results suggest that the negatively charged residues in bridges 2 and 3 on the C-terminal bundle have a key role in both peptide transport and H^+ trans-location, and the positively charged residues of NG3 also have a role in transport. The NB2 residues, even though they are well conserved across all POT members are not essential for transport. However, this result is perhaps explained by the close proximity of H61 (TM 2) in $PepT_{So}$, whose equivalent residue in GlpT acts as a proton donor (Law *et al.* 2008). There may be a comparable residue in *Gk*POT which aids the salt bridge formation.

Alanine mutations in bridge 1 residues reduce transport by approximately 70% compared to WT in $PepT_{So}$ and $PepT_{St}$ in the H^+-driven assay, but have no discernible effect in counterflow. The alanine mutations of bridge 4 residues of $PepT_{So}$ unlike the bridge 1 residues, completely abolish transport in both the H^+-driven and the counter-flow assays. This suggests the bridge 1 and bridge 4 residues play a part in the proton trans-location pathway and only the bridge 4 residues impact on substrate transport. Although only bridge 4 residues in bacterial are observed in $PepT_{So}$, they are of note as they are also present in animal PepT1 and PepT2 sequences.

The peptide binding site

Figure 1.14 shows the binding site of the bacterial peptide transporter crystal structures. Residues other than the ones shown in Figure 1.14 undoubtedly play a part in creating the binding cavity, but these are the residues as proposed by Solcan *et al.* (2012). The binding site is composed of residues from TM helices 1, 2, 4, 7, 8 and 10 with TM 1 making the largest contribution. The residues of TM 1 are part of the POT sequence motif, ExxERFxYYG. The K127/126/136/121 on TM 4 and the E419/400/413/402 on TM 10 ($PepT_{So}$/$PepT_{St}$/*Gk*POT/$PepT_{So2}$) are salt bridge 3 discussed in Figure 1.4.4 and the Y68/68/78/62 and W312/296/306/268 make up part of a hydrophobic binding pocket (Newstead *et al.* 2011; Lyons *et al.* 2014).

AlaAlaAla and AlaPhe, and alafosfalin have been crystallised in the binding cavities of $PepT_{St}$ (Lyons *et al.* 2014) (Figure 1.14B and C), and *Gk*POT (Doki *et al.* 2013) (Figure 1.14D) and $PepT_{So2}$ (Guettou *et al.* 2013) (Figure 1.14E) respectively. All the struc-

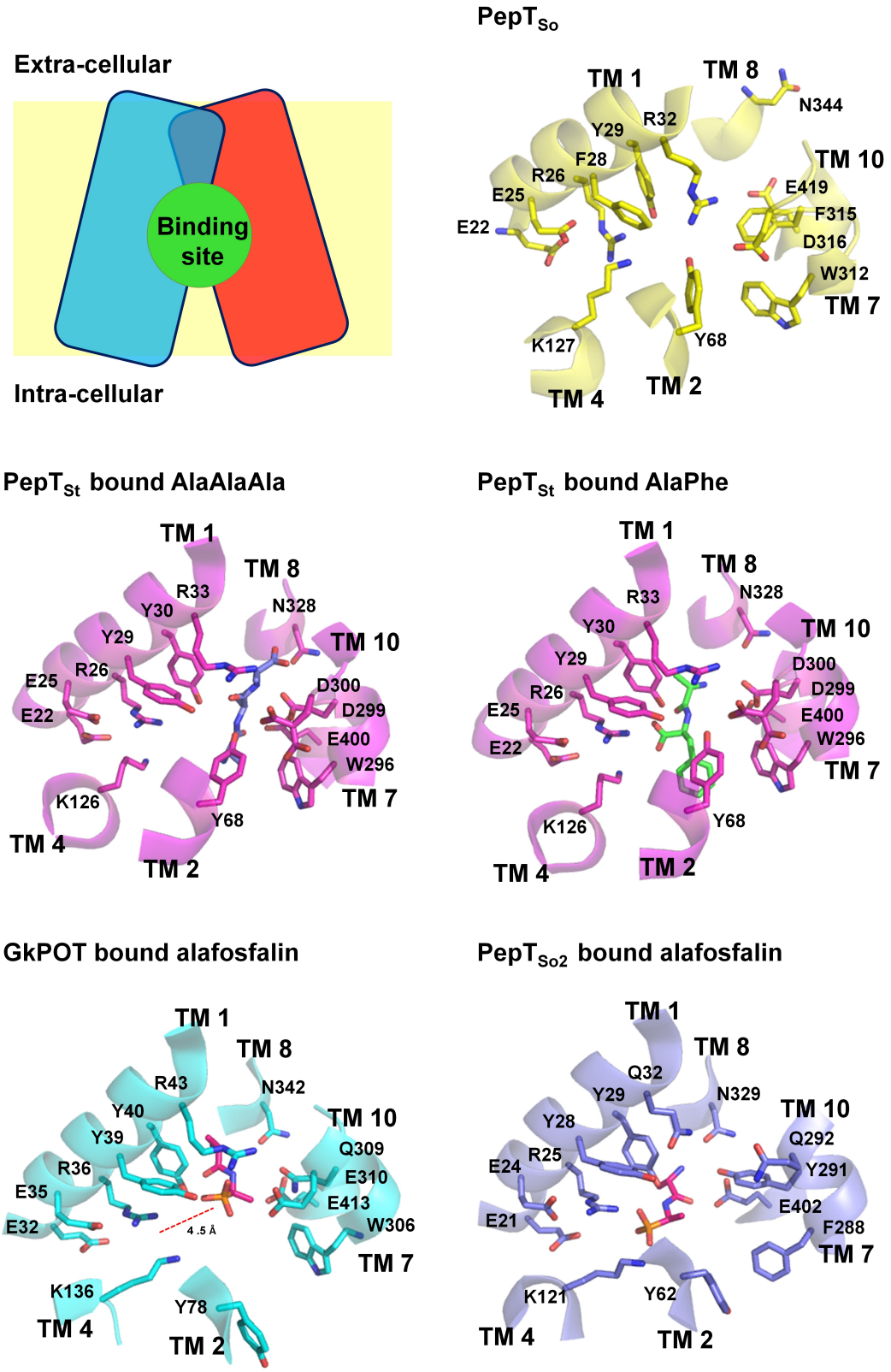


Figure 1.14: The substrate binding site of the bacterial peptide transporters. The binding site of $PepT_{So}$ (Newstead *et al.* 2011) (PDB: 2XUT), $PepT_{St}$ with bound AlaAlaAla (PDB: 4D2D) and bound AlaPhe (PDB: 4D2C) (Lyons *et al.* 2014), *Gk*POT with bound alafosfalin (PDB: 4IKZ) and $PepT_{So2}$ with bound alafosfalin (PDB: 4LEP). This figure was adapted from Solcan *et al.* (2012).

tures with bound substrate are in the IO conformation, but different substrate binding conformations are observed. The phosphate group of the alafosfalin is coordinated by Y39, Y40, R43 and E410 in the *Gk*POT crystal structure. However, in the $PepT_{So2}$ crystal structure the alafosfalin has shifted down the binding site by 4.5 Å, coordinating the phosphate with R25 and K121. The two peptides bound into the structure of $PepT_{St}$ also show different binding conformations. The AlaAlaAla is bound parallel to the membrane normal, orientated with the C-terminus close to R33. The AlaPhe is orientated perpendicular to the membrane normal with the carboxyl group interacting with R26 and K126. The multiple binding conformations observed in these structure are a representation of the possible substrate binding modes and demonstrate the poly specificity in the binding site (Lyons *et al.* 2014).

Solcan *et al.* (2012), Doki *et al.* (2013) and Solcan (2013) have performed transport assays on the majority of the residues displayed in Figure 1.14B, C and D. The alanine mutations of the charged residues in the ExxERFxYYG motif completely remove observable transport in the H^+-driven assay, whereas only a modest reduction in transport was observed in counter-flow in $PepT_{So}$ and $PepT_{St}$. This suggests a role in H^+ trans-location rather than peptide transport. The crystal structures of AlaPhe and AlaAlaAla in $PepT_{St}$ support this result, as only R26 is involved in substrate carboxyl coordination, and this may not be essential for substrate transport. However, in *Gk*POT the E32Q and E35Q mutations remove transport in both assays, implying a role in substrate and H^+ trans-location. The crystal structure of *Gk*POT therefore does not support this result as the substrate makes no contact with the charged residues of the ExxERFxYYG motif.

Alanine mutations in F28/Y29 ($PepT_{So}$/$PepT_{St}$) have only a modest effect on H^+-driven and counter-flow, whereas alanine and phenylalanine mutations in Y29/30 ($PepT_{So}$/$PepT_{St}$) have an effect on H^+-driven and counter-flow. However, the effect on counter-flow for the phenylalanine mutation is modest in $PepT_{St}$. This result implicates the hydroxyl group of the Y29/30 as being important in H^+ trans-location and the phenyl ring system in substrate transport. The same effect was observed for Y68 on TM 2 in both $PepT_{So}$ and $PepT_{St}$. This result is well supported by the crystal structures. Y30/40/29 ($PepT_{St}$/*Gk*POT/$PepT_{So2}$) makes contributions to substrate binding in all four structures, and Y68 in $PepT_{St}$ is part of the hydrophobic pocket which encloses the phenyl ring

in the AlaPhe structure.

The C-terminal bundle residues appear to play less of a role in the binding site than the N-terminal bundle. The largest contribution comes from TM 7 (the repeat swap of TM 1). Of the residues examined by Solcan *et al.* (2012), Solcan (2013) and Lyons *et al.* (2014) mutations in N344/328 and E316/D300 ($PepT_{So}$/$PepT_{St}$) both completely remove transport in both assays. A similar result was seen for E310Q in *Gk*POT (Doki *et al.* 2013). However, neither alanine or glutamine mutations in N342 had an observable effect on transport. These asymmetric effects observed in the N344/328/342 ($PepT_{So}$/$PepT_{St}$/*Gk*POT) mutations are interesting because they mirror the opposite results in the charged residues of the ExxERFxYYG motif. These slight changes in residue contributions probably reflect the fact that $PepT_{So}$, $PepT_{St}$, *Gk*POT and $PepT_{So2}$ are different proteins with less than 50% sequence identity. Therefore definitive functional residue assignments should be made cautiously between different proteins.

1.4.5 Mechanistic model of bacterial peptide transport

There are many unanswered questions still to be determined with regard to the peptide/H^+ symport mechanism, for example the exact number of trans-located H^+ per peptide. However, given the salt bridge and binding cavity data, a model was proposed by Lyons *et al.* 2014 based on the bacterial structures and the transport assay data. Figure 1.15 shows the trans-location of a di- and tri-peptide in horizontal and vertical conformations respectively. A H^+ binds E300 (CB 2) and activates the transporter. Upon peptide binding, the C-terminal carboxyl group or the N-terminal amino group of a di- or tri-peptide respectively interrupts the salt bridge 3 stimulating the conformational change to the IO state. The hydrophobic cavity acts as a side chain well in the intermediary OB conformation. The peptide and H^+ are released into the intra-cellular compartment, and the protein cycles back to the OO state. The mechanism for the last step is still not fully understood.

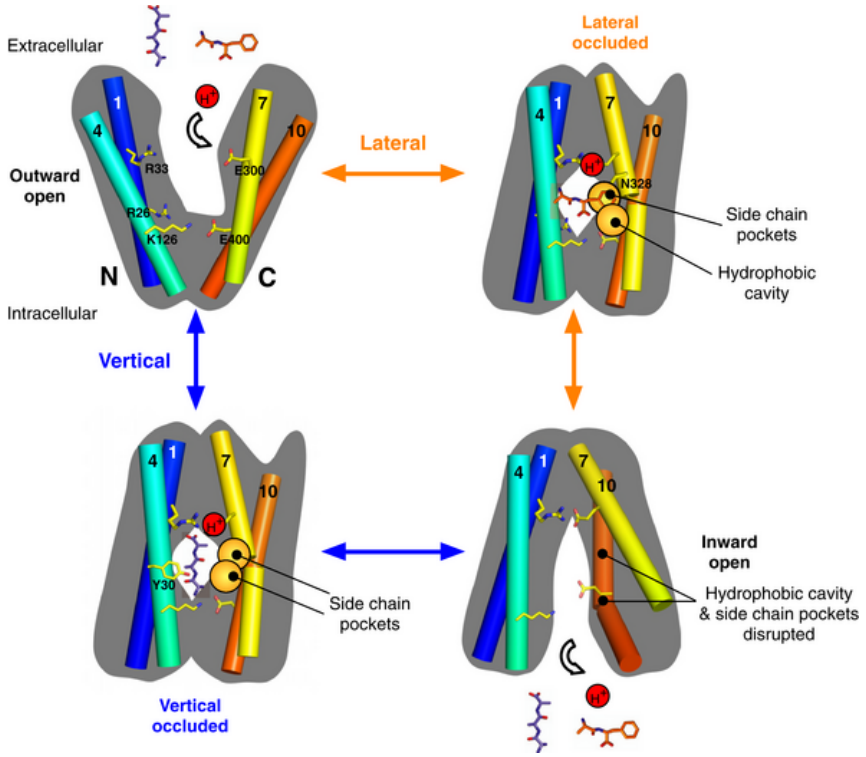


Figure 1.15: A model of bacterial peptide transport. TM helices 1 (blue), 4(cyan), 7 (yellow) and 10 (orange) are the principal gating helices in the alternating access rocker switch mechanism and the red H^+ represents an undetermined number of protons. The diagram shows the two potential binding modes of peptide transport proposed by the X-ray crystal structures of di- and tri-peptides bound in the binding cavity of the bacterial peptide transporters, lateral and vertical respectively. The B2 and B3 salt bridge residues indicate the top and bottom of the binding cavity, and the side chains of both di- and tri-peptides can be accommodated in pockets inside the cavity. This figure is based on the $PepT_{St}$ structure and was taken from Lyons *et al.* (2014).

1.5 Conservation of peptide transport model from bacteria into mammalian PepT1 and PepT2

Putting this bacterial model in the context of mammalian PepT1 and PepT2 is not straight forward. As seen with $PepT_{St}$ and *Gk*POT, residue conservation is not a guarantee of functional conservation and, as can been seen from Figure 1.8, a significant proportion of PepT1 and PepT2 remains uncharacterised. The extra-cellular loop inserted between TM 9 and 10 is approximately 20 kDa and no function has been investigated or hypothesised. That said, the salt bridges which straddle the binding site (B2 and B3) are well conserved across the POT family, and Xu *et al.* (2009) found that a E595C mutation in *Hs*PepT1 (homologous residue to CB2) removed observable transport in an transiently transfected oocyte based assay. The TM 1, POT family motif, ExxERFxYYG, makes up a large proportion of the binding site, also suggesting a conservation of binding mechanism.

A large amount of mutational analysis has been conducted on mammalian PepT1 and PepT2; however, as highlighted in § 1.4.4 an accurate model of PepT1 and PepT2 has still to be generated to properly interpret these data. That said, a series of PepT1/PepT2 chimeric studies were performed which are worth mentioning here, as their results do inform upon the individual characteristics of PepT1 and PepT2 and the function of the extra-cellular loop.

1.5.1 Chimeric studies on mammalian PepT1 and PepT2

Döring *et al.* (1996), Fei *et al.* (1998), Terada *et al.* (2000a), and Döring *et al.* (2002) performed chimeric studies on the mammalian peptide transporters, exchanging different regions of PepT1 and PepT2 to create a series of chimeric proteins. Based on these chimeric studies and others (Boll *et al.* 1996; Terada *et al.* 2000b), there appear to be three parameters to experimentally distinguish mammalian PepT1 and PepT2. PepT2 activity appears to be influenced by an extra-membranous pH, with a 5-fold reduction in activity from pH 6.0 – 8.0, whereas PepT1 over the same range shows only a 2-fold reduction. PepT1 activity also appears to be approximately 10-fold greater than PepT2. Finally, due to the limitations of the current transport assays, it is not possible to calculate a given substrate affinity; however, it appears that PepT1 and PepT2 have different apparent K_ms

for various substrates. Table 1.1 shows the constitution of each chimera from the studies carried out by Döring *et al.* (1996), Fei *et al.* (1998), Terada *et al.* (2000a), and Döring *et al.* (2002), and how the exchanges effected the transport properties.

The results from Table 1.1 should be interpreted with caution as, although the transporters do share a high percentage of their primary sequence, the exchange of helices from one transporter to the other is likely to introduce confounding variables into the experiment. Few of the chimera studies undertook experiments to determine all three properties for each mutant. Of the chimeras which were assessed for their pH dependence and activity, HR2, HR3/N1C2, CH1 and N2C1, all appear to indicate that the N-terminal bundle residues mediate these characteristics (Döring *et al.* 1996; Fei *et al.* 1998; Terada *et al.* 2000a).

The residues which determine the substrate specificity also appear to reside in the N-terminal bundle. However, there is no consensus for the substrates which determine a PepT1 or PepT2 'like' transport, so this conclusion should be treated with caution. That said, if these results are accurate then it seems probable that N-terminal bundle of PepT1 and PepT2 is more important than the C-terminal bundle in determining the characteristics of the transporters. Which implies that the extra-cellular loop region is not important for individual characteristics of PepT1 and PepT2. The loop may still have a functional role in transport, but this still needs to be experimentally determined.

1.6 Characterisation of the extra-cellular loop

1.6.1 SLC protein extra-cellular loops

Large intra-helical extra-cellular loops are very uncommon in integral membrane transporters. In all of the human SLC families, only four have members which contain intra-helical extra-cellular sections which are greater than 100 residues in length (Table 1.2). Interestingly, the most common place of loop insertion appears to be between TM 9 and 10. The sequence alignment of these extra-cellular loops yields no sequence identity suggesting, but not proving, no conservation of function. That said, Table 1.2 does show that the extra-cellular loop of the peptide transporters is not wholly unique.

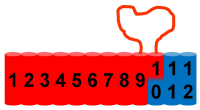
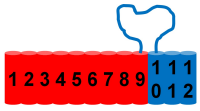
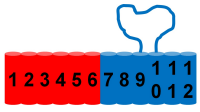
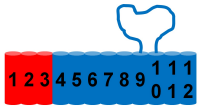
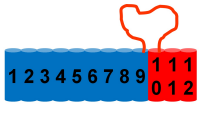
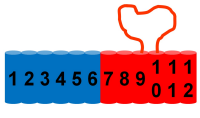
Mutant	Chimera	pH dependence	Activity	Substrate Specificity
HR1		-	-	PepT1
HR2		-	PepT1	PepT1
CH2			Not functional	
HR3/N1C2		PepT1	PepT1	PepT2
CH4		-	-	Unclear
CH3		-	-	Unclear
HR4		-	-	PepT2
HR5		-	-	PepT2
RH1		-	-	PepT2
RH2		-	-	PepT2
CH1		PepT2	PepT2	PepT2
N2C1		PepT2	PepT2	PepT2

Table 1.1: Chimeric study mutant constitution and results. The table was constructed based upon Döring *et al.* (1996), Fei *et al.* (1998), Terada *et al.* (2000a), and Döring *et al.* (2002). The red or blue helices indicate they belong to either PepT1 or PepT2 respectively. A consistent substrate was not used by all of the studies, therefore it is not clear which parts of the transporter are responsible for the different substrate specificities observed between PepT1 and PepT2. The results of the chimera studies do suggest that the N-terminal bundle of the transporter is responsible for the difference in pH dependence and transporter activity, apparent between PepT1 and PepT2.

SLC Family	Gene	Fold	Function	Insertion Helices
SLC2	GLUT10	MFS	Uniporter	9 – 10
	GLUT12	MFS	Uniporter	
SLC15	HMIT	MFS	Myo-inositol/H^+ symporter	
	PepT1	MFS	Peptide/H^+ symporter	
	PepT2	MFS		
SLCO	All members	-	Organic anion exchanger	
SLC22	All members	-	Organic cation symporter	1 – 2
SLC24	NCKX1	-	Na^+, Ca^{2+} and K^+ exchanger	1 – 2, 5 + 6
SLC34	All members	-	Na^+, $HPO_4{}^{2-}$ symporter	4 – 5

Table 1.2: The members of the SLC families which have large intra-helical extra-cellular sections. The 418 *H. sapiens* members of the 52 SLC families catalogued by Hediger *et al.* (2004), were subjected to hydropathy analysis using `TMHMM` (Krogh *et al.* 2001), to discover SLC genes which also possess large intra-helical extra-cellular regions. 100 residues was taken as a likely length for a self contained domain and all SLC members fitting this criteria are shown in this table. Based on this analysis, members of the MFS, appear to have a common place insertion between TM helices 9 and 10.

It is not clear what role these extra-cellular loops play in the function of the respective transporters. The bacterial homologous X-ray crystal structures of the SLC2 and SLC22, XylE and PipT respectively have been solved (Sun *et al.* 2012; Pedersen *et al.* 2013). The structures indicate that like the peptide transporter homologues, the extra-cellular loop observed in the human transporters appear to be separate from the TM domain, suggesting that the loop may not be essential for the function of the transporter.

1.6.2 Structure prediction analysis

To gain more information on the peptide transporter loop regions, the loops were put through a series of *in silico* based prediction programs, to predict the loop folding propensities and their secondary structure. Figure 1.16 shows the `Jpred` (Cole, Barber, and Barton 2008) secondary structure predictions for the *H. sapiens* PepT1 (391 – 580) and PepT2 (410 – 607) loops. `Jpred` predicts that the loops both contain a large amount of secondary structure and therefore they are both likely to be folded. `Jpred` also suggests that both loops are all β proteins, and given the complementarity of β sheets 1, 2, 11, 13 and 14, it is possible that both loops have the same overall fold.

`PONDR-FIT` (Xue *et al.* 2010) was used to assess the likelihood that the loops are ordered protein. Figure 1.17 shows the disorder plots for *H. sapiens* PepT1 and PepT2. The plots suggests that both loop regions are predominantly stable, with disorder propensities of

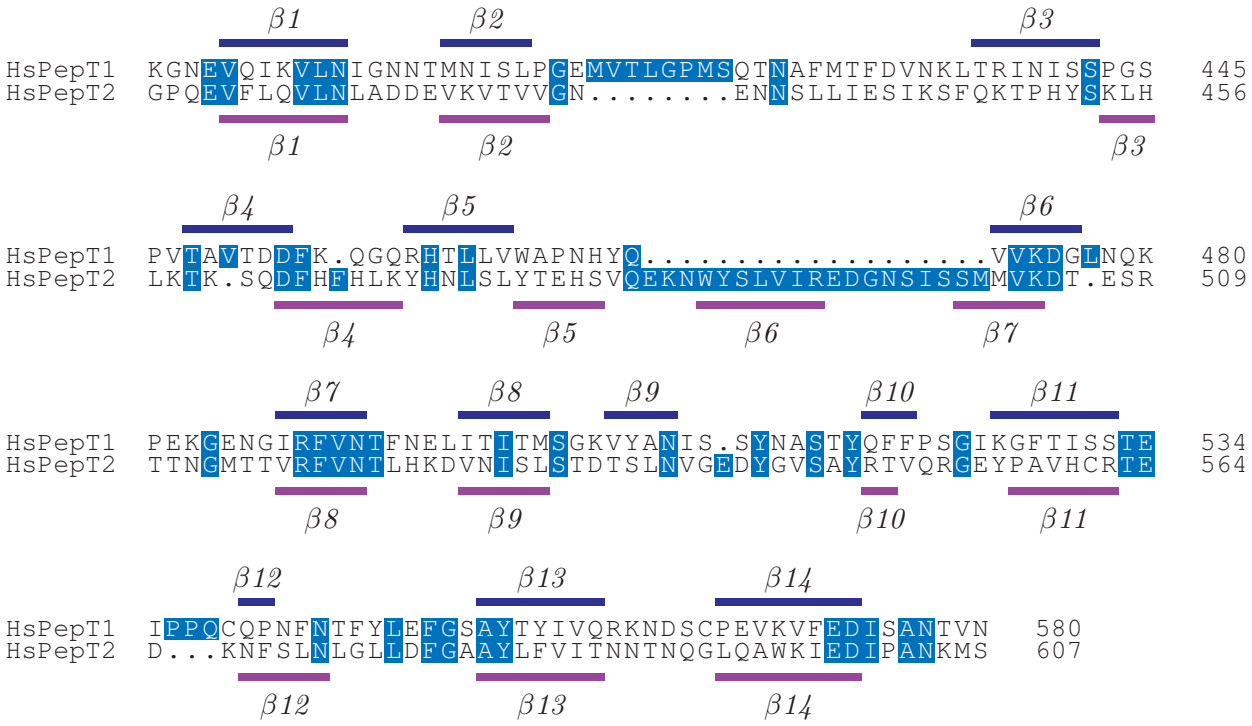


Figure 1.16: Secondary structure predictions of the *H. sapiens* PepT1 and PepT2 extra-cellular loop. Secondary structure predictions were performed using `Jpred` (Cole, Barber, and Barton 2008). Only β-sheets were predicted and these are shown as blue and purple bars above and below the aligned *Hs*PepT1 and *Hs*PepT2 sequences respectively. The full names and Uniprot codes for each gene are listed in Table 2.1.

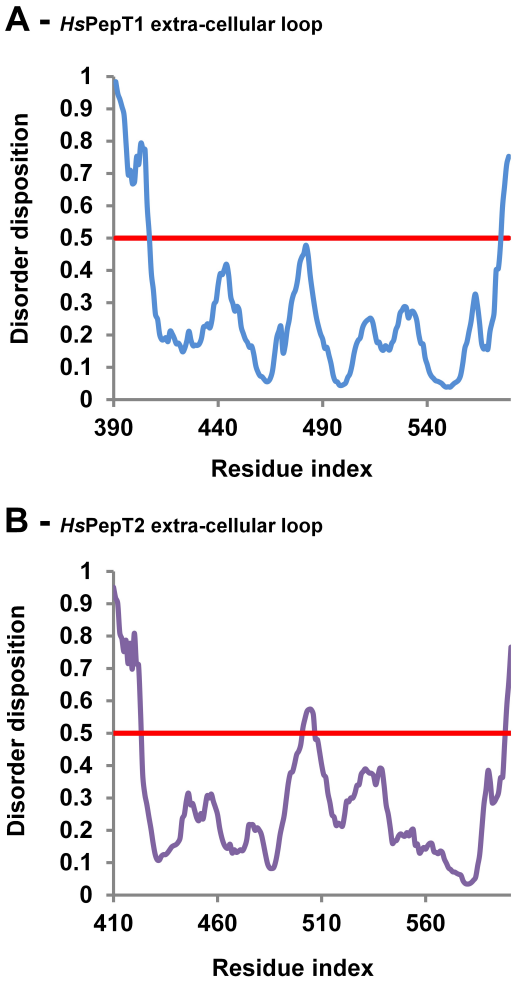


Figure 1.17: Residue disorder predictions for the extra-cellular loop regions of *H. sapiens* PepT1 and PepT2. The residue disorder predictions for *Hs*PepT1 and *Hs*PepT2 are shown in **(A)** and **(B)** respectively. The red line delineates a 0.5 probability of disorder. The plots were generated using PONDR-FIT (Xue *et al.* 2010). The full names and Uniprot codes for each gene are listed in Table 2.1.

less than 0.5 for the majority of the region, although both appear to be more disordered in the N-terminal region.

Bioinformatic predictions should be interpreted with a degree of caution, as these predictions are based upon the secondary structure libraries, which do not necessarily sample the fold of the protein of interest. However, given the number of structures now deposited in the Protein data bank (PDB) it is unlikely that a large number of unique folds still exist. So based on the Jpred and PONDR-FIT predictions the loop region of PepT1 and PepT2, is a self-contained folded Extra-cellular domain (ECD), and is referred to as such for the remained of this thesis.

1.7 Aims

PepT1 and PepT2 represent important protein targets in terms of understanding their native function; the regulation of di- and tri-peptide transport in the body, and as a mechanism for drug delivery and retention. The bacterial structures have given, and will continue to give, significant insight into the likely mechanism of the human transporters. However, as has been highlighted in § 1.4.5, the ECD remains an unknown variable and so requires further study to better understand its structure and function.

The bacterial structures suggested that the ECDs remain separate from the TM domains; the bioinformatic analysis indicates that both ECDs contained secondary structural elements and were likely to be folded. If the ECDs did form a self-contained domain, it was considered probable that the PepT1 and PepT2 ECDs could be heterologously expressed and could prove amenable to X-ray crystallography, or Nuclear magnetic resonance (NMR). Once the structures of the ECDs had been solved, homology models of the entire PepT1 and PepT2 transporters could be constructed. These could then be used to interpret the function data from both bacterial and mammalian peptide transport studies.

Thus the principal aims of this thesis were to solve the structures of the PepT1 and PepT2 ECDs from a mammalian organism. The structures of the ECDs would, at the very least, provide enough structural information to construct a homology model of *H. sapiens* PepT1 and PepT2 and, at the most, give an insight into the function of these domains. To complete these aims, the following approach was taken:

1. Perform high throughput cloning and over-expression of the loop regions from a range of organisms at the Oxford protein production facility UK (OPPF-UK) and assess their amenability for structural characterisation (Chapter 3).
2. Solve the crystal structures of ECDs from PepT1 (Chapter 4) and PepT2 (Chapter 5).
3. Interpret the structures for a probable function and likely conformation (Chapter 6).
4. Build a homology model of *H. sapiens* PepT1 and PepT2 (§ 7.3).
5. Investigate the function of the loop region (Chapter 7).

2

Materials and methods

2.1 Summary

In this chapter, the materials and methods used in this thesis are described, saving the methods pertinent to X-ray crystallographic diffraction experiments, which are discussed in Chapters 4 and 5. This chapter is designed to be used as a reference.

2.2 Materials

2.2.1 Instruments

- Veriti® thermal cycler
- 96R SPRIPlate (Agencourt)

- Permeable block seal AB-0718 (ABgene)
- NanoDrop 8000 UV-Vis spectrophotometer (Thermo Scientific)
- Innova® 44 shaker incubator (New Brunswick Scientific)
- Avanti™ J-20 centrifuge (Beckman Coulter)
- Biorobot 8000 (Qiagen)
- XCell4 Surelock™ Midi-Cell (Invitrogen)
- 96 well CrystalQuick™ plate (Greiner Bio-one)
- Matrix Hydra II (Thermo Scientific)
- Microsys 4000 (Cartesian)
- 4 and 20 °C Rock Storage and Vis-UV imaging System (Formulatrix)
- iCycler™ thermal cycler (Bio-Rad)
- G:BOX Chemi XX6 (Syngene)
- Spectramax M3 UV-Vis spectrophotometer (Molecular Devices)
- Allegra X-22 fitted with swinging-bucket rotor SX4250 (Beckman-Coulter)
- Polyvinylidene fluoride (PVDF) western blot membrane (Millipore)
- Photographic film (Thermo Scientific)
- Avanti™ J-20 centrifuge (Beckman Coulter)
- TS Series Benchtop E1061 Cell Disruptor (Constant Systems) with Neslab ThermoFlex 1400 recirculating chiller (Invitrogen)
- Peristaltic pump P-1 (GE Healthcare)
- 5 kDa molecular weight Centrifugal Filter (Vivaspin)
- AKTA purifier system with Box-900, UPC-900, P-900 (GE Healthcare)
  - Superdex S75 16/60 Gel filtration (GF) column (GE Healthcare)
  - Superdex S200 10/300 GF column (GE Healthcare)
  - Superdex S75 10/300 GF column (GE Healthcare)
- 5 mL Q Trap anion exchange column (GE Healthcare)
- MRC 2 Well crystallisation plate (Swissci)
- Mosquito® Crystal robot (ttplabtech)
- Incubator (New Brunswick Scientific)
- Seed bead (Hampton Research)
- 24-well hanging drop crystallisation plates (Hampton Research)
- Square siliconized glass cover slides (Hampton Research)
- 600 MHz NMR magnet (Oxford Magnets)
- NMR tube (Shigemi)
- Mx3005p™ real time PCR machine (Stratagene)

- J-815 CD Spectrometer (Jasco)
- MicroCal iTC200 isothermal titration calorimeter (Malvern)
- Series S, CM5 chip (GE Healthcare)
- Biacore™ T200 (GE Healthcare) Surface plasmon resonance (SPR) instrument

2.2.2 Bacteria strains, media, chemicals and kits

- In-fusion® DNA ligase (Clonetech® Laboratories)
- Phusion-Flash DNA polymerase (Thermo Scientific)
- KOD Xtreme™ DNA polymerase (Novagen)
- AMPure Magnetic Beads (Beckman-Coulter)
- Agarose (Sigma)
- SYBR® Safe DNA Gel Stain (Life Technologies)
- One Short® OmniMAX™ 2 T1® *E. coli* strain (Life Technologies)
- Power Broth™ (PB) (Molecular Dimensions)
- Lysogeny broth (LB) (Sigma)
- Terrific broth (TB) (Sigma)
- Carbenicillin (Formedium)
- Chloramphenicol (Sigma)
- Mini-prep Kit (Qiagen)
- Rosetta-gami™ (DE3)pLysS/pRARE (Novagen)
- Overnight Express™ instant terrific broth medium (TBonex)
- NuPAGE® 10% Bis-Tris Midi gel (Novex)
- InstantBlue (Expedion)
- Floppy-Choppy limited proteolysis kit (Jena Biosciences)
- KOD Hot Start DNA Polymerase (Novagen)
- PageRuler pre-stained protein ladder (Thermo Scientific)
- GeneJET™ gel extraction kit (Thermo Scientific)
- All restriction enzymes were bought from Fermentas
- Quick-Stick Ligase (Bioline)
- Mini-prep kit (Thermo Scientific)
- Gel extraction fit (Thermo Scientific)
- C43 (DE3) *E. coli* competent cells (Lucigen)
- BugBuster™ (Novagen)

- Anti-His antibody (Sigma)
- Anti-*M. musculus* Horseradish peroxidase (HRP) conjugate antibody (Sigma)
- ECL Plus^{TM} western blotting detection kit (Amersham)
- Ni-sepharose (Qiagen)
- Amylose resin (GE Healthcare)
- His-Trap HP (GE Healthcare)

2.3 Alignments and figures

All alignments in this thesis were calculated using `ProCons` (Do *et al.* 2005) and viewed and manipulated in `Jalview` (Waterhouse *et al.* 2009). All of the sequences and Uniprot codes used in this thesis are shown in Table 2.1. Sequence alignments and topology figures were constructed using TEXshade (Beitz 2000a) and TEXtopo (Beitz 2000b).

Protein representations were created using `PyMOL` (Schrodinger 2010).

2.4 Gels and other quantification techniques

2.4.1 SDS-PAGE gels

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was conducted, unless otherwise stated, using manually constructed Tris-Gly gels at either 10, 12, or 15% w/v acrylamide, depending on the separation required. Table 2.2 gives the volumes of the components used to construct these gels with different acrylamide concentrations. The separating gel mixture was prepared first, poured into a 1.0 mm gel cassette (Novex®) and covered with 70% v/v ethanol. Once set, the ethanol was removed and the stacking gel mixture prepared, poured onto the top of the set separating gel and a comb inserted to create the sample wells. Gels were stored at 4 °C until required.

Protein samples were mixed in a 1:1 ratio with loading buffer (8 M Urea, 200 mM Tris pH 8.8, 1 mM Ethylenediaminetetraacetic acid (EDTA), 1.5% w/v Dithiothreitol (DTT), trace bromophenol blue) and incubated for 5 min at 95 °C. Samples were then loaded into the Tris-Gly gel immersed in running buffer (125 mM Tris, 1.25 M Glycine, 0.5% v/v SDS) and PageRuler^{TM} (Thermo Scientific) was used as a molecular weight marker. A

Organism	Gene	Uniprot Code
Homo sapiens	PepT1	P46059
	PepT2	Q16348
Macaca mulatta	PepT1	Q6WFZ8
	PepT2	Q6WFZ7
Sus scrofa	PepT1	Q7YSA7
	PepT2	A5A4L7
Bos Taurus	PepT1	A5D7E5
	PepT2	F1MPK7
Equus ferus caballus	PepT1	F6SG69
	PepT2	F6R282
Canis familiaris	PepT1	Q8WMX5
	PepT2	E2QWX1
Oryctolagus cuniculus	PepT1	P36836
	PepT2	P46029
Rattus norvegicus	PepT1	P51574
	PepT2	Q63424
Mus musculus	PepT1	Q9J1P7
	PepT2	Q9E507
Gallus gallus	PepT1	F1P0K5
	PepT2	U5TSZ8
Danio rerio	PepT1	Q75YE4
	PepT2	Q63424
Anolis carolinensis	PepT1	G1KST4
	PepT2	G1KDE4
Xenopus laevis	PepT2	Q7SY87
Caenorhabditis elegans	OPT3	O01840
	OPT2	Q06593
	OPT1	Q17758
Arabidopsis thaliana	NRT1.1	Q05085
Shewanella oneidensis	PepT	Q8EKT7
	$PepT_2$	Q8EHE6
Saccharomyces cerevisiae	PTR2	P32901
Streptococcus thermophilus	PepT	Q5M4H8
Geobacillus kaustophilus	PepT	GK2020
Escherichia coli	DtpA	P77304
Escherichia coli	LacY	C9QQT5
Escherichia coli	GlpT	C6E9W1
Homo sapiens	Transthyretin	Q53WY6
Mus musculus	Transthyretin	P07309
Xenopus laevis	Transthyretin	P07309
Homo sapiens	Trypsin	P07477
Bos Taurus	Trypsin	P00760
Mus musculus	Trypsin	P07146
Homo sapiens	α-chymotrypsin	Q6GPI1
Bos Taurus	α-chymotrypsin	P00766
Mus musculus	α-chymotrypsin	Q9CR35
Homo sapiens	Elastase	P08861
Bos Taurus	Elastase	Q29461
Mus musculus	Elastase	P05208

Table 2.1: A list of the genes used for alignments and bioinformatic analysis in this thesis.

Separating gel			
	Acrylamide concentration (% w/v)		
	10	**12**	**15**
30% w/v Acrylamide	2.00 mL	2.40 mL	3.00 mL
H_2O	2.44 mL	2.04 mL	1.44 mL
1.5 M Tris pH 8.8		1.50 mL	
10% w/v Ammonium Persulfate		0.06 mL	
Tetramethylethylenediamine		4.00 µL	
Stacking gel			
30% w/v Acrylamide		333 µL	
H_2O		1.13 mL	
0.5 M Tris pH 6.8		500 µL	
10% w/v Ammonium Persulfate		20 µL	
Tetramethylethylenediamine		3 µL	

Table 2.2: Components and volumes used for Tris-Gly SDS-PAGE gels.

potential difference of 180 V was applied to the gel, and the protein samples typically took 1.3 hr to run through the gel. Protein gels were stained with InstantBlue^{TM} (Expedeon) and photographed using a G:BOX Chemi XX6 (Syngene).

2.4.2 Trichloroacetic acid precipitation

On the occasions when the concentration of a protein sample was too low to visualise (§ 6.5.1) via SDS-PAGE, a Trichloroacetic acid (TCA) precipitation was performed. A TCA stock was prepared composed of 100% w/v TCA and 2% w/v sodium deoxycholate. 1 mL of TCA stock was added to 0.5 mL of protein solution, and incubated for 10 min at Room temperature (RT). The solution was centrifuged at 16,000 × g for 5 min. The pellet was then washed twice by cycles of cold acetone removed by centrifugation. Residual acetone was removed by incubating the sample at 95 °C for 10 min. The pellet was then re-suspended in loading buffer and run on an SDS-PAGE gel.

2.5 Bicinchoninic acid assay

To measure the concentration of all protein samples, a Pierce® Bicinchoninic acid (BCA) assay (Thermo Scientific) was performed. The BCA assay, is a protein detection assay based upon the selective quantification of the protein dependent reduction of Cu^{2+} to Cu^{+}

(Smith *et al.* 1985). A standard was constructed based on known volumes of Bovine serum albumin (BSA) with a working range of between 0.125−2.000 mg/mL, giving Equation 2.1.

$$[\text{Protein}\,(\text{mg/mL})] = A_{562} \times 2.49 - 0.0866 \tag{2.1}$$

Unknown protein samples were diluted into the working range of the assay. BCA reagent A was mixed with reagent B to create a working stock with a ratio of 50:1 A:B. 10 µL of sample was mixed with 200 µL of the BCA working stock and mixed thoroughly with a pipette. Samples were then incubated at 37 °C for 30 min. Samples were left to cool to RT for 5 min and the absorbance of each sample was measured at A_{562} using a Spectramax M3 (Molecular Devices) microplate reader. The sample concentration could then be inferred using Equation 2.1.

2.6 Cloning and test expression of ECD constructs

2.6.1 OPPF-UK high-throughput cloning and test-expression

The OPPF-UK system uses High-throughput (HT) technologies which allow for 48 constructs to be cloned into expression vectors concordantly (see § 3.3).

OPPF-UK primer design

The OPPF-UK vectors (Table 3.1) are ligated to Polymerase chain reaction (PCR) products by a homologous recombination event catalysed by In-fusion® Deoxyribonucleic acid (DNA) ligase. Table 2.3 shows the sequences that are added to the 5′ end of the forward and reverse primers which are homologous to the pOPIN linearised vectors. The priming sequences used for the specific genes are given in Table 2.4. The OPPF-UK developed program, `Opiner`, was used to design these primers.

OPPF-UK polymerase chain reaction protocol

PCRs were performed using either Phusion-Flash (Tables 2.5A) or KOD Xtreme™ (Table 2.5B) high fidelity DNA polymerases in 96-well PCR plates. Once prepared the plates

pOPIN Vector		pOPIN homologous sequence
E	For	aggagatataccatg
	Rev	gtgatggtgatgttt
*	For	aagttctgtttcagggcccg
	Rev	atggtctagaaagcttta

Table 2.3: The homologous sequences used for the OPPF-UK pOPIN vectors. The '*' indicates PCR-products suitable for pOPIN vectors: F, HALO, J, MSYB, NUSA, M, S3C, TF, TRX (described in § 3.3).

Gene	Primer (For/Rev)
*Hs*PepT1	AAAGGAAACGAAGTCCAAATTAAAGTTTTG/ GTTAACTGTGTTGGCTGAAATATCTTCAAAC
*Hs*PepT2	GGTGGAAATCAAGTCCAAATTAAGGTCTTG/ AACCAGGGCATATTGTGGTAGCTGCC
*Mm*PepT1	GGTGGAAATCAAGTCCAAATTAAGGTCTTG/ CACAGTGTTGGGTGGGATGTCTTC
*Mm*PepT2	ATAAAAATAAATGGAATGATCCATCCTCAGC/ GGATAGTTTATTGGCTGGAATGTCTTCTGCC
*Dr*PepT1	TCAACCTCTCAAACCCAGTTGAAGTTCC/ CAAAGCTCTCGCTGCAAGTTTGTCC
*Oc*PepT1	CCCAAAGCCAATGAAGTCCAAATTAAAGTTC/ CATTGTGTTGGGCGGAATATCTTCAAAC
*Ce*OPT2	GCCTTTTTGATCACTGGATTTGTTCAACTTC/ TCCCTGCTTAACTCTGTGAAAACCGTAATTTG
*Ce*OPT2	CCTTCCCTTGGACGAATTTATCTGCAAC/ CACAGAGTTGGGGCGTACCAACAAGTTG

Table 2.4: Priming sequences for the peptide transporter ECDs used in the OPPF-UK cloning. The sequences listed here were added to the 3' end of the pOPIN specific primer in Table 2.3. The full names and Uniprot IDs for each gene are listed in Table 2.1.

A

Component	Volume (μL)
Phusion-Flash Master Mix	25
Forward Primer (10 nmol)	2
Reverse Primer (10 nmol)	2
Template cDNA (100 nmol)	2
ddH_2O	19

B

Component	Volume (μL)
KOD Xtreme™	1
KOD 2 × Buffer	25
dNTP (2 μmol)	10
Forward Primer (10 nmol)	2
Reverse Primer (10 nmol)	2
Template cDNA (100 nmol)	2
ddH_2O	8

Table 2.5: PCR reaction volumes used in OPPF-UK cloning. The volumes for the respective components of a 50 μL Phusion-Flash **(A)** or KOD Xtreme™ **(B)** PCR reaction.

were centrifuged briefly, using a JS-5.3 (Beckman-Coulter) swinging bucket rotor to remove air bubbles. The PCRs were performed with a Veriti® thermal cycler using the reaction times for the specific polymerases shown in Table 2.6. The accuracy of the PCRs were checked by running 5 μL of the PCR products through a 1.5% w/v Agarose, 1 × Tris-acetate-EDTA (TAE) gel with a potential difference of 100 V for 1 hr.

	Polymerase			
	Phusion–Flash		**KOD Xtreme™**	
	Temp (°C)	**Time** (s)	**Temp** (°C)	**Time** (s)
Activation	94	120	94	120
Denaturation	94	30	94	30
Annealing	70	20	65	20
Extension	72	15	72	60
	Cycle to Denaturation ≈ 30 times			
Final Extension	72	30	72	120

Table 2.6: Reaction times and temperatures for PCR amplifications using Phusion Flash or KOD Xtreme™ DNA polymerases.

Component	Mass (ng)	Typical Volume (μL)
TAE Buffer		0.5
Purified PCR–Product	100	4.5
Linearised vector	100	0.5
ddH_2O		5.0

Table 2.7: The composition of the In-fusion® reaction mixture.

OPPF-UK PCR purification protocol

The PCR-products were purified by DNA affinity chromatography using AMPure magnetic beads (Beckman-Coulter). The parent DNA was digested by the addition of 1 μL of *Dpn*I restriction nuclease, incubated for 1 hr at 37 °C. 90 μL of AMPure magnetic beads were incubated with the PCR products for 10 min at RT. The 96-well PCR plate was then placed on a magnetic plate for 1 min allowing the magnetic beads to separate from the solution. The clear solution was then aspirated and the beads washed twice in 200 μL of 70% v/v Ethanol and then allowed to dry for 10 min at RT to remove any excess ethanol. The PCR products were finally eluted from the magnetic beads by the addition of 35 μL of Tris-EDTA buffer.

OPPF-UK In-fusion PCR-product-vector ligation protocol

The PCR products and vectors were ligated by homologous recombination catalysed by In-fusion® DNA ligase. The reaction was prepared using the components shown in Table 2.7. This volume was used to hydrate 1 unit of dry In-fusion® enzyme. The mixture was incubated at 42 °C for 30 min and then quenched in 40 μL of Tris-EDTA buffer.

OPPF-UK transformation protocol

Cells were defrosted on ice 15 min prior to transformation. 10 μL (approximately 100 μg) of each of the 48 quenched In-fusion® reactions were used to inoculate 48 units of OmniMAX™ 2 competent *E. coli* cells. The cells were then incubated on ice for 30 min. Cells were heat shocked at 42 °C for 30 s and then incubated on ice for a further 5 min. 150 μL of Power broth (PB) was then added to the cells and they were incubated at 37 °C. After 1 hr, 10 μL of this culture was used to inoculate 1.5% w/v Agar LB

containing 0.1 mg/mL X-gal, 50 μM Isopropyl β-D-1-thiogalactopyranoside (IPTG) and 100 μL ampicillin, in a 24-well cell culture plate. The plate was then incubated for 16 hr at 37 °C to allow for colony growth. The addition of X-gal and IPTG allows for blue/white colony picking.

OPPF-UK HT plasmid isolation protocol

A 96-deep well block containing 2 mL of PB, 100 μg/mL ampicillin per well was prepared and a single white colony per construct from the 24-well transformation plate was used to inoculate two wells. The block was covered with an air permeable seal and incubated at 37 °C/280 Revolutions per minute (rpm) for 16 hr. Cells were pelleted by centrifugation at 5000 × g for 10 min and a Bio-robot 6000 (Qiagen) was used to perform 96 mini-preps (Qiagen) on the 96 pellets. Typical DNA concentrations, measured by absorption at A_{260} using a NanoDrop 8000 UV-Vis spectrophotometer (Thermo Scientific), were ≈ 50 ng/μL.

OPPF-UK PCR conformation protocol

To confirm the presence of the correct insert, a PCR was performed using the pOPIN specific forward and reverse primers, upstream and downstream respectively of the insert. Phusion-Flash polymerase (described in § 2.6.1) was used with a fifth of the volumes specified in Table 2.5A and parameters in Table 2.6, but with an increased extension time of 30 s. 10 μL of the completed PCRs were run through a 1.5% w/v Agarose, 1 × TAE gel to check for correct construct lengths (Figure 3.4).

OPPF-UK HT small scale test expression protocol

The 48 test constructs were transformed (described in § 2.6.1) into B834(DE3) and Rosetta-gami^{TM} (DE3) pLysS/pRARE competent cells. The Rosetta-gami^{TM} also require 34 μg/mL chloramphenicol in addition to 50 μg/mL carbenicillin in the agar due to the pLysS plasmid. A 96-deep well block with 2 mL of LB plus antibiotics per well was prepared and each well was inoculated with a single colony of transformant. The block was then covered with an air permeable seal and incubated at 37 °C/280 rpm. After 16 hr, 16 24-deep well blocks were prepared with 4 mL of media plus antibiotic per well; 8

blocks contained LB and 8 blocks contained Overnight Express^{TM} (TBonex) auto-inducing medium.

A single well of LB and TBonex was inoculated with 50 µL of each starter culture from a test construct, transformed into either B834(DE3) or Rosetta-gami^{TM} cells. The 24-deep well blocks were then sealed and incubated at 37 °C/280 rpm. Cultures were grown until an Optical density (OD) λ = 600 nm of 0.5, and then induced. The LB cultures were induced by the addition of IPTG until a final concentration of 0.5 mM, whereas TBonex is an auto-induction media. The growth temperature of the LB and TBonex cultures were reduced to 25 °C and 20 °C, and left to express for 16 and 20 hr respectively.

Cells were harvested by transferring 1 mL of each culture to a 96-deep well block and cells pelleted by centrifugation at 5,000 × g for 20 min at 4 °C. Construct expression was determined by mini Ni-Nitrilotriacetic acid (NTA) purification of the 6-His tagged constructs using a BioRobot 6000 Qiagen protocol. The purified test samples were run through a NuPAGE® 10% Bis-Tris Midi-gel and visualised using InstantBlue. The gels from the test expressions are shown in Figure 3.5.

2.6.2 Cloning and test expression of pLou3 and pEHisTev ECD constructs

PCR and PCR-product purification protocol

The primers used for the pLou3 and pEHisTev constructs are shown in Table 2.8. PCRs at Oxford Biochemistry were conducted using KOD Hot Start DNA polymerase. The PCR reaction volume was constructed as detailed in Table 2.9A and the thermocycler parameters are given in Table 2.9B. The entire PCR product was loaded into a 1.5% w/v Agarose, 1 × TAE gel and purified using a Thermo Scientific gel band extraction kit.

pLou3 and pEHisTev ECD ligation protocol

Purified PCR-products and expression vectors were digested by their appropriate restriction enzymes (Table 2.8) for 3 hr at 37 °C using the volumes given in Table 2.10. Cut fragments were purified using the Thermo Scientific gel extraction kit (binding buffer was

Gene	RE	Primer (For/Rev)
*Hs*PepT2	Nco1 EcoR1	ATATCCATGGCTggtccccaggaggttttcc/ ATAGAATTCTTAaatgtcttcaatcttccaggcc
*Rn*PepT2	Nco1 EcoR1	ATATCCATGGCTgcttcccaagagatatttttacaag/ ATAGAATTCTTAaatgtcctctgccttccagg
*Xl*PepT2	Nco1 EcoR1	ATATCCATGGCTgtatccgcagccaaggagag/ ATAGAATTCTTAactgttggcagtgacatcttc
*Dr*PepT2	Nco1 EcoR1	ATATCCATGGCTgccaaagagagccttgttcag/ ATAGAATTCTTAgatgtcctccatctttttagcg
*Ce*OPT1	Nco1 EcoR1	ATATCCATGGCTatggcaaatgaggcacatctaac/ ATAGAATTCTTAggatacgagagactcggttg
*Ce*OPT3	Nco1 EcoR1	ATATCCATGGCTccttcccttggacgaatttatctg/ ATAGAATTCTTAcacagagttggggcgtacc
*Mm*PepT1	Nco1 Sal1	TTACCATGGGTggaaatcaagtccaaattaag/ ATTGTCGACtcacacagtgttgggtgggatg

Table 2.8: Primers used for the cloning of the ECDs into the pLou3 and pE-HisTev expression vectors. The capital letters indicate the primer sections containing the restriction sites in the For and Rev primers, and lower case letters show the PCR priming sequence. **RE** = Restriction enzyme. The full names and Uniprot codes for each gene are listed in Table 2.1.

added to the digest and the kit protocol followed as normal). 25 ng of linearised vector was incubated with 75 ng of cut PCR-product and one unit of Quick-Stick ligase (Bioline) at RT for 10 min. The entire reaction was then used to transform defrosted OmniMAX^{TM} 2 competent cells (described in § 2.6.2). pLou3 and pEHisTev have an ampicillin and kanamycin resistance cassette respectively.

pLou3 and pEHisTev transformation protocol

All transformations not conducted at the OPPF-UK were transformed using the following protocol. Defrosted competent cells were inoculated with 100 ng of plasmid DNA. Cells were then incubated on ice for 30 min, heat shocked at 42 °C for 30 s before a further incubation on ice for 2 min. For vectors containing an ampicillin resistance cassette, the cells were immediately plated onto 1.5% w/v Agar LB containing 100 μg/mL ampicillin, for all other vectors, 300 μL of LB was added to the cells and they were incubated at 37 °C/200 rpm. After 1 hr, 150 μL of the culture was used to inoculate 1.5% w/v Agar, containing appropriate antibiotic, and incubated for 16 hr to allow for colony growth.

A

Component	Volume (μL)
KOD Hot Start Polymerase	1
KOD 10X Buffer	5
dNTP (2 μmol)	5
$MgSO_4$ (2 μmol)	3
100% w/v Dimethyl sulfoxide	4.5
Forward Primer (10 nmol)	2
Reverse Primer (10 nmol)	2
Template cDNA (100 nmol)	2
ddH_2O	25.5

B

	KOD Hot Start	
	Temp (°C)	**Time (s)**
Activation	94	120
Denaturation	94	30
Annealing	55	20
Extension	72	60
	Cycle to Denaturation ≈ 30 times	
Final Extension	72	120

Table 2.9: Volumes and parameters used for KOD Hot Start PCRs. (A) Volumes of components per 50 μL reaction. **(B)** Times and temperatures used for the PCR amplification.

Component	Volume (μL)
Vector/PCR Product	25
Enzyme Buffer	5
Enzyme 1	1.5
Enzyme 2	1.5
ddH_2O	17

Table 2.10: The composition of a restriction enzyme digest.

pLou3 and pEHisTev cloning cleavage conformation protocol

A single colony from the transformation plate was used to inoculate 3.5 mL of LB plus antibiotic, and incubated at 37 °C/200 rpm for 16 hr. Cells were pelleted by centrifugation at 3900 × g. The Thermo Scientific mini-prep kit was used to isolate the plasmids. Typical DNA concentrations, measured at A_{260} were 200 ng/μL. To confirm the presence of the insert a test cleavage was performed using a 1/5 of the volumes given in Table 2.10. Plasmids which gave the correct cleavage fragments, as assessed by an agarose gel were sent for sequencing.

pLou3 and pEHisTev ECD construct test expression protocol

The PepT2^{ECD} constructs in the pLou3 and pEHisTev vectors were transformed (§ 2.6.2) into C43(DE3) *E. coli* cells. A single colony was used to inoculate 3.5 mL of LB plus antibiotic and was incubated at 37 °C for 16 hr. This starter culture was used to inoculate 3.5 mL of TB, which was grown at 37 °C/280 rpm until an OD_{600} of 0.5. The culture was then induced with IPTG to a final concentration of 0.5 mM and grown at 25 °C/280 rpm.

After a 16 hr incubation, a final OD_{600} reading was taken (typically about 4.0), and the cells were diluted to an OD_{600} of 1.0 in BugBuster^{TM} (Sigma) to a final volume of 100 μL. The cells were then freeze/thawed and vortexed for 20 min at RT. Soluble and insoluble material were separated by centrifugation at 16,000 × g for 30 min. The supernatant was then removed and the pellet was re-suspended in 100 μL of 1 × Phosphate buffered saline (PBS). The supernatant and re-suspended pellet were run though a 12% acrylamide Tris-Gly gel, in order to perform an anti-His western blot.

Western blot protocol

An SDS-PAGE gel was run containing the samples of interest. A polyvinylidene fluoride (PVDF) membrane was soaked in 100% v/v methanol for 30 s and then rinsed with ddH_2O. The stacking gel and foot were removed from the SDS-PAGE gel, and then the SDS-PAGE gel was placed in contact with the activated PVDF membrane sandwiched between 2 pieces of card and four pieces of sponge soaked in cold Transfer buffer (25 mM Tris, 192 mM Gly, 20% v/v Methanol, pH 8.3). The sandwich was loaded into the transfer

container (Invitrogen) and filled with cold Transfer buffer. This container was loaded into a gel running chamber (Invitrogen) and the outer section was filled with cold water. A potential difference of 30 V was applied to the blot for 90 min.

After transfer, the PVDF membrane was removed and blocked for 30 min in 10 mL of 1 × PBS, 0.05% v/v Tween 20 (PBST) with 5% w/v milk at RT with gentile agitation. This solution was discarded and replaced with 5 mL of 1 × PBST, 5% w/v milk with primary antibody (*M. musculus* anti-6-His (Sigma)) for 90 min at RT with gentile agitation. Unbound primary was removed by washing the membrane 3 times with 10 mL of 1 × PBST for 10 min at RT with gentile agitation. The membrane was then incubated with secondary antibody (anti-*M. musculus* Horseradish peroxidase (HRP) conjugate (Sigma)) in 5 mL of 1 × PBST for 30 min, at RT with gentile agitation. Unbound secondary was removed by washing the membrane 3 times with 10 mL of 1 × PBST for 10 min at RT with gentile agitation. Excess PBST was removed by blotting the membrane on a piece of tissue paper. The HRP was activated using the ECL Plus^{TM} (GE Healthcare) western blotting kit, the membrane was wrapped in cling film and exposed to photographic film for $1 - 5$ s[1]. The film was then fixed using a Compact X4 (X-ograph Healthcare Ltd).

2.6.3 Large scale expression and cell lysate preparation.

Large scale expressions were conducted by transforming (§ 2.6.2) the construct into an appropriate *E. coli* expression strain. A single colony from the transformation plate was used to inoculate 50 mL of LB plus antibiotic and incubated at 37 °C/280 rpm for 16 hr. This culture was used to inoculate 2 L of TB plus antibiotic in 2 2.5 L Tunair® flasks and incubated at 37 °C/280 rpm. Cultures were grown to an OD_{600} of 0.5, and then induced by addition of IPTG to a final concentration of 0.5 mM. Growth was continued at 20 °C/280 rpm for 18 hr until the cells were harvested by centrifugation at 6,000 × g/4 °C/20 min in a JLA 8.1000 (Beckman-Coulter) rotor. Pellets were removed and stored at at −80 °C.

At the start of the purification, the frozen cell pellets were re-suspended in 150 mL of 1 × PBS, 20 mM imidazole pH 8.0, 150 mM NaCl and 10% v/v glycerol, and lysed by a

[1]Varied depending on the amount of bound antibody. The anti-His detection was very strong and the blot only needed to be exposed for a short time.

Cell Disruptor (Constant systems); two passes of 151 and 165 MPa successively. From the time at which the cells were broken to purified protein, all solutions connected to the purification were kept on ice or at 4 °C. Unbroken cells and cell debris were removed by centrifugation; 30,000 × g in a JLA-16.250 (Beckman-Coulter) for 30 min. Cell membranes were removed by ultra-centrifugation; 131,000 × g in a Ti-45 (Beckman-Coulter) for 1 hr. The ultra-centrifugation supernatant was used as the basis for all purifications.

2.6.4 ECD-3CX construct cloning

The 3C protease cut site is L-E-V-L-F-Q-/-G-P (cleavage site indicated by the '/'). In *Mm*PepT1^{ECD} the linker residues (N-Q-K-P-E-K-G) are relatively exposed to solution in either lobe conformation. The new linker (*Mm*PepT1$^{\text{ECD-3CX}}$) exchanged the Q-K-P-E for the 3C cut site (N-L-E-V-L-F-Q-G-P-K-G). In *Rn*PepT2^{ECD} the linker residues (D-T-G-I-K-P-G) appear less exposed, therefore the cut site residues were inserted into the middle of the linker residues. The new linker (*Rn*PepT2$^{\text{ECD-3CX}}$) sequence was D-T-G-I-K-L-E-V-L-F-Q-G-P.

Overlapping PCR was used to construct the *Mm*PepT1$^{\text{ECD-3CX}}$ and *Rn*PepT2$^{\text{ECD-3CX}}$ constructs. The forward and reverse primers from Table 2.8 were used in conjunction with the 3CX linker primers from Table 2.11. Two PCRs were prepared using KOD Hot Start Polymerase (Table 2.9A) and the primer couplets, ECD specific forward/linker reverse, and linker forward/ECD specific reverse, using the PCR parameters from Table 2.9B. The completed PCRs were run through a 1.5% w/v Agarose, 1 × TAE gel, then the desired product was excised and cleaned using a Thermo Scientific gel extraction kit. A second PCR was performed, using the purified PCR products from the first 2 reactions with the ECD specific forward and reverse primers, and using the same polymerase and parameters. The PCRs were run through a 1.5% agarose gel and cleaned by gel band purification using a Thermo Scientific gel extraction kit. The purified overlapping PCR fragments (Figure 2.1A and B) were cloned into the pLou3 expression plasmid using the cleavage and ligation protocol outlined in § 2.6.2. The ligated overlapping PCRs and vectors were transformed into OmniMAX^{TM} 2 competent cells. Single colonies of from the transformation plates were picked to inoculate 3.5 mL cultures of LB and left shaking at 37 °C for 16 hr. The plasmid DNA was isolated from these cultures using a mini-prep kit

Gene	Mutation	Primer (For/Rev)
*Mm*PepT1	Linker 1	ctggaagttctgttccagggtccgaaaggggagaacggaatcagg/ cggaccctggaacagaacttccaggttaagaccatctttaccac
*Rn*PepT2	Linker 2	ctggaagttctgttccagggtccggggatggcagccatcagg/ cggaccctggaacagaacttccagtttgattcctgtatccttgaccag
*Mm*PepT1	D476A-3CX	cgtgtggtaaaagctggtcttaacctg/ caggttaagaccagcttttaccacacg
*Rn*PepT2	D505A-3CX	gcatgctggtcaaggccacaggaatcaaactgg/ ccagtttgcttcctgtggccttgaccagcatgc
*Mm*PepT1	D476A	ccgtgtggtaaaagccggtcttaaccaaaagc/ gcttttggttaagaccggcttttaccacacgg
*Mm*PepT1	D574A	gaaggaatttgaagccatcccacccaacac/ gtgttgggtgggatggcttcaaattccttc
*Rn*PepT2	D505A	gcatgctggtcaaggccacaggaatcaaaccagc/ gctggtttgattcctgtggccttgaccagcatgc
*Rn*PepT2	D600A	gcctggaaggcagaggccatttaagaattcgg/ ccgaattcttaaatggcctctgccttccaggc

Table 2.11: A list of the primers used in the lobe conformation mutation studies. Linker 1 and 2 refer to the primers used to substitute the flexible linker between the lobes for a 3C protease cut site (3CX). The suffix -3CX indicate a specific primer for the 3C protease linker cut site. The full names and Uniprot codes for each gene are listed in Table 2.1.

(Thermo Scientific). The 3CX construct DNA was screened for the correct size insertion by a test cleavage (as described in § 2.6.2) (Figure 2.1C). The transformant DNA which gave cleavage bands of the correct size were sent for sequencing to confirm the presence of the 3C protease site.

2.6.5 ECD-3CX aspartate mutant cloning

The pLou3 *Mm*PepT1$^{\text{ECD-3CX}}$ D476A and D574A mutants were created using site-directed mutagenesis. The primers used are detailed in Table 2.11. The PCRs were prepared as described in Table 2.9A using KOD Hot start DNA polymerase. The PCR thermocycler parameters are detailed in Table 2.12. 2 µL of *Dpn*I were added to the completed plasmid amplifications and incubated at 37 °C for 3 hr. 10 µL of this solution was used to immediately transform OmniMAX^{TM} 2 competent cells using the transformation protocol described in § 2.6.2. Plasmid DNAs for the resulting colonies were isolated using the Thermo Scientific mini-prep protocol and sent for sequencing to confirm the presence of

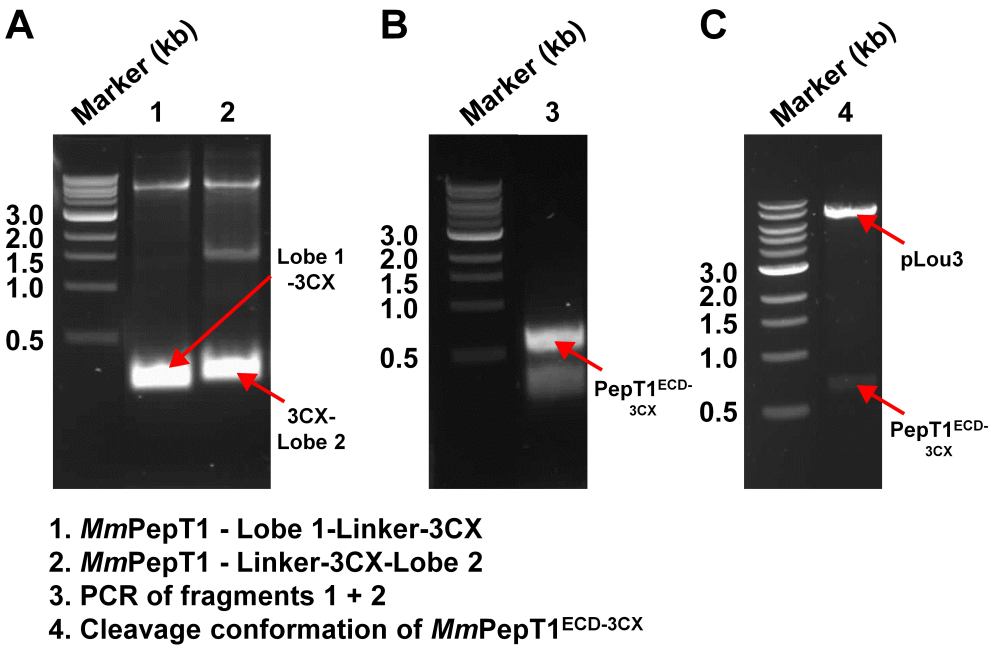


Figure 2.1: The 1.5% w/v Agarose, 1 × TAE gels from the cloning of the *Mm*PepT1$^{\text{ECD-3CX}}$ constructs. **(A)** Initial forward and reverse PCRs. **(B)** Combined PCR. **(C)** Cleavage conformation (described in § 2.6.2) of the *Mm*PepT1$^{\text{ECD-3CX}}$ construct in the pLou3 expression vector, using the restriction enzymes *Nco*I and *Hind*III.

the correct mutation.

2.6.6 *Rn*PepT2^{ECD} D505A and D600A mutant cloning

The D505A and D600A *Rn*PepT2^{ECD} constructs in the pLou3 vector were mutated using site-directed mutagenesis using the same method as § 2.6.5, but with the D505A and D600A mutant primers detailed in Table 2.11.

	Temp (°C)	**Time** (s)
Activation	94	120
Denaturation	94	60
Annealing	55	60
Extension	68	420
	Cycle to Denaturation ≈ 17 times	
Final Extension	68	420

Table 2.12: Thermocycler parameters for KOD Hot Start site directed mutagenesis PCR.

Screen
PACT premier™
JCSG+
Wizard™ I+II
Wizard™ III+IV
PEG/Ion HT
Index
Morpheus™
Midas™

Table 2.13: A list of the sparse-matrix screens use during the crystallisation.

2.7 Crystallisation

2.7.1 96-well sitting drop experiments

Sparse matrix crystal screening was performed at the OPPF-UK. The screens used are shown in Table 2.13. The screens were stored at 4 °C in 96-deep well blocks. Prior to crystallisation, crystal screens were equilibrated to RT and the well conditions transferred from the deep-well blocks to 96-well Greiner sitting drop plates by a Matrix Hydra II (Thermo Scientific) liquid handling robot. Each well contained 100 μL of a particular precipitant. A Microsys 4000 (Cartesian) robot was used to place drops of protein solution and well condition onto the well stage in the ratios (in nL) 100:100, 100:200 and 200:100 (Figure 4.2A.1). The plates were then sealed using ClearSeal Film™ (Hampton Research) and incubated in a Rock Imager (Formulatrix) at either 4 or 20 °C. Photographic time points were taken by the hotel at: 3, 5, 15, 35, 55, 75, 95, 115, 135, 175, 215 hr, and 12, 15, 18, 25, 38, 52, 78, 132, 185 days.

2.7.2 24-well hanging drop experiments

Individual well solutions were mixed, passed through a 0.22 μm filter and equilibrated to the experimental temperature. The edges of the crystallisation wells were coated in vacuum grease and also left to equilibrate. 0.5 mL of well condition was pipetted into each well. Siliconized glass cover slides were cleaned with pressurised CO_2, and a drop of protein solution (typically 0.5 – 5 μL) was pipetted onto the surface of the cover slip. A comparable volume of the precipitant was then mixed with the protein solution using a pipette tip. The cover slip was then inverted and sealed onto the well by the vacuum

grease (Figure 4.4). The plate was then incubated at either 4 or 20 °C. The plate was checked every day for the first week of the experiment and then every week there after.

2.7.3 Reductive methylation

Reductive methylation was used to try and increase the propensity of the *Mm*PepT2^{ECD} 410 – 609 construct to crystallise (§ 5.2.2). 1 mL of 5 mg/mL protein solution was dialysed into 50 mM HEPES pH 7.5, 150 mM NaCl to remove any primary amines from the solution. 20 µL of cold Dimethyamine borane complex (DBC) was added to the protein and incubated at 4 °C for 2 hr with gentle agitation. After the 2 hr, 40 µL of 1 M formaldehyde was added to the mixture and incubated at 4 °C for 2 hr with gentle agitation. A further 10 µL of DBC was then added, and the solution was incubated for 16 hr at 4 °C with gentle agitation. The reaction was quenched by the addition of 1 mL of 20 mM Tris pH 7.5, 150 mM NaCl. The sample was then run over a S75 16/60 GF column equilibrated in the crystallisation buffer.

2.7.4 SeMet derivatisation

SeMet derivatised protein was produced using an Auto-inducing SeMet medium PASM-5052 (Table 2.14) (Studier 2005). 2 L of the PASM-5052 were used for the *Rn*PepT2^{ECD} SeMet derivatisation.

2.7.5 Sparse matrix micro-seeding protocol

Sparse matrix micro-seeding was conducted as outlined in Ireton and Stoddard (2004) and D'Arcy, Villard, and Marsh (2007). Fresh crystals were grown and harvested mid growth time. 1 – 3 crystals were looped from the drop and placed in 50 µL of the reservoir solution at 4 °C containing a seed bead (Hampton Research). This solution was vortexed for 3 min, placing the solution on ice every 30 s for 10 s. This created the 'seed stock'. Sparse matrix screens were equilibrated to RT in 96-well Swissci plates with 50 µL of condition per well. A Mosquito® robot was used to mix 10 mg/mL protein, seed stock and well condition, in a ratio of 100:50:150 nL respectively. Crystal plates were then stored at either 4 or 20 °C.

Component	Quantity
Na_2HPO_4	50 mM
KH_2PO_4	50 mM
$(NH_4)_2SO_4$	25 mM
$MgSO_4$	2 mM
Glycerol	0.5% v/v
Glucose	0.05% w/v
α-lactose	0.2% w/v
17 amino-acids (No Cys, Tyr, Met)	200 µg/mL
Met	10 µg/mL
SeMet	125 µg/mL
Vitamin B_{12}	100 µg/mL
Fe^{3+}	10 µM
Ca^{2+}	4 µM
Mn^{2+}	2 µM
Zn^{2+}	2 µM
Co^{2+}	0.4 µM
Cu^{2+}	0.4 µM
Ni^{2+}	0.4 µM
Mo^{4+}	0.4 µM
Se^{4+}	0.4 µM
B^{3+}	0.4 µM

Table 2.14: The composition of PASM-5052 auto-inducing SeMet *E. coli* growth media.

2.8 Biophysics

2.8.1 Fluorescence scanning calorimetry

To quantify the effect of different buffers on the stability of the *Mm*PepT2^{ECD}, Fluorescence scanning calorimetry (FSC) was performed. 220 µL of each trial condition was prepared. Each condition typically contained 20 mM buffer, 150 mM NaCl, 0.1 mg/mL protein and 2 × Sypro orange (Sigma). 50 µL of each condition was pipetted into 4 wells of an opaque 96-well real-time quantitative PCR plate and sealed with a transparent seal. The plate was briefly centrifuged to remove bubbles in a JS-5.3 swinging bucket rotor. The plate was then placed in a Mx3005p^{TM} (Stratagene) real time PCR machine. The temperature of the plate was increased at a rate of 1 °C/min. At 1 min intervals the fluorescence of the Sypro orange was monitored using excitation and extinction wavelengths of 488 and 600 nm respectively. The results were analysis using the the `Excel` spread-sheets developed by Niesen, Berglund, and Vedadi (2007) and `Prism` (Graph Pad). The fluorescence intensity changes were modelled as a Boltzmann sigmoidal function, where the V_{50}

Component	Quantity
Na_2HPO_4	50 mM
KH_2PO_4	50 mM
$^{15}NH_4Cl$	50 mM
Na_2SO_4	5 mM
$MgSO_4$	2 mM
Glycerol	0.5% v/v
Glucose	0.05% w/v
α-lactose	0.2% w/v
Fe^{3+}	10 µM
Ca^{2+}	4 µM
Mn^{2+}	2 µM
Zn^{2+}	2 µM
Co^{2+}	0.4 µM
Cu^{2+}	0.4 µM
Ni^{2+}	0.4 µM
Mo^{4+}	0.4 µM
Se^{4+}	0.4 µM
B^{3+}	0.4 µM

Table 2.15: The composition of N-5052 auto-inducing ^{15}N *E. coli* growth media.

= T_m. The final T_m was calculated as the average of 4 readings within a confidence range of 95%.

2.8.2 Heteronuclear single quantum coherence NMR

Protein was expressed in 1 L of the auto-inducing ^{15}N medium N-5052 (Table 2.15) (Studier 2005) and purified as described in Chapter 3 at a yield of 10 mg/L. The final GF column step was equilibrated in 20 mM Na Acetate pH 5.5, 150 mM NaCl. The peak fraction was concentrated to approximately 40 mg/mL and back diluted in 20 mM Na Acetate pH 5.5, to 6.73 mg/mL (300 µM) and a NaCl concentration of 11 mM. 285 µL of 300 µM protein was mixed with 15 µL of 100% v/v D_2O, 50 µM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) and loaded into a Shigemi tube. The sample was degassed and loaded into a 600 MHz magnet with a GE consul and triple resonance probe. The experiment was performed at 20 °C, with sweep width of the 2 k and referenced against DSS. The data was processed in `NMRpipe` (Delaglio *et al.* 1995) and viewed in `NMR Analysis` (Fogh *et al.* 2005; Vranken *et al.* 2005) with the help of Dr Jolyon Claridge (Deparment of Biochemistry, University of Oxford).

Standard	MW (kDa)	V_E (mL)
Blue Dextran	2000	8.08
Albumin	66	9.70
Carbonic Anhydrase	29	11.93
Cytochrome C	12.4	13.67

Table 2.16: The globular protein MW standard V_es from the Superdex S75 10/300 GF column. The V_e/V_0 was plotted against log(MW) to construct the calibration curve in Figure 2.2.

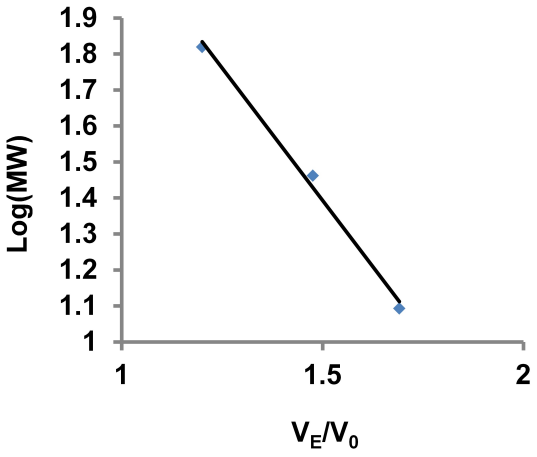


Figure 2.2: The calibration curve for the Superdex S75 10/300 GF column based on the V_e of the globular protein standards shown in Table 2.16

2.8.3 S75 10/300 globular protein calibration

To construct the S75 10/300 standard curve 3 proteins; albumin, carbonic anhydrase and cytochrome C, were passed down the GF column equilibrated in 20 mM Tris pH 7.5, 150 mM NaCl. Blue dextran was used to calculate the void volume (V_0). Each standard was diluted to a concentration of 10 mg/mL and passed concordantly down the S75 GF column, and the peak V_e recorded (Table 2.16). The V_e/V_0 was plotted against the log Molecular weight (MW) (Figure 2.2) and the regression line of the points is given in Equation 2.2.

$$\log(\mathrm{MW}) = 3.6 - 1.5\left(\frac{V_e}{V_0}\right) \tag{2.2}$$

2.8.4 Analytical ultracentrifugation

Samples were prepared for the Analytical ultracentrifugation (AUC) pH screening experiment by GF. A superdex S75 10/300 GF column was equilibrated in either: 20 mM

Na Acetate pH 5.5, 20 mM Tris pH 7.5, 20 mM Tris pH 8.0 or 20 mM Bicine pH 9.5 (all with 150 mM NaCl). 100 µL of 20 mg/mL either *Mm*PepT1^{ECD}, *Rn*PepT2^{ECD}, *Rn*PepT2^{ECD}-D505A or *Rn*PepT2^{ECD}-D600A was loaded onto the equilibrated column. The peak fraction was collected and concentrated to approximately 2 mg/mL in a 600 µL volume using a 5 kDa MW centrifugal filter.

Buffer densities and viscosities were measured using a Lovis 2000 M Microviscometer (Anton Paar) and the protein partial specific volumes were estimated using `Sednterp` (Laue *et al.* 1992). Buffer blanks were created by using the void volume from the GF column prior to elution of the protein sample. Sedimentation velocity (SV) experiments were conducted at 20 °C, at 40,000 rpm on a ProteomeLab XL-1 (Beckman-Coulter) using an An-50 Ti (Beckman-Coulter) rotor. The resulting Rayleigh interference profiles were analysed using `SEDFIT` (Schuck 2000; Brown and Schuck 2006) (Appendix B.1) c(s) distributions. The $s_{20,w}$ values were calculated for each concentration in the pH series (Appendix B.2). The $s^0_{20,w}$ values were calculated by extrapolating the zero concentration point from the $s_{20,w}$ values for three different concentrations per pH point (Figure 2.3).

2.8.5 Circular dichroism protocol

Where it was possible protein samples were diluted into buffers with no NaCl, ensuring a NaCl concentration of less than 3 mM. Where it was not possible the NaCl concentration was reduced as much as possible, whilst still obtaining a Circular dichroism (CD) trace and a high tension voltage of less than 700 mV.

Thermal melt

The sample holder inside the spectrometer was set to 20 °C. 200 µL of buffer was pipetted into a clean cuvette, placed into the spectrometer and allowed to equilibrate. 3 'buffer blank' spectrums were recorded and averaged between wavelengths 185 and 260 nm with a 0.5 nm step. The average buffer blank was then subtracted from all following sample spectrums. A dilution series was then conducted from 0.5 mg/mL to 0.1 mg/mL, to find a concentration which maximised the signal and ensured a final data point with a high tension voltage of less than 700 mV. Generally, the sample concentration was

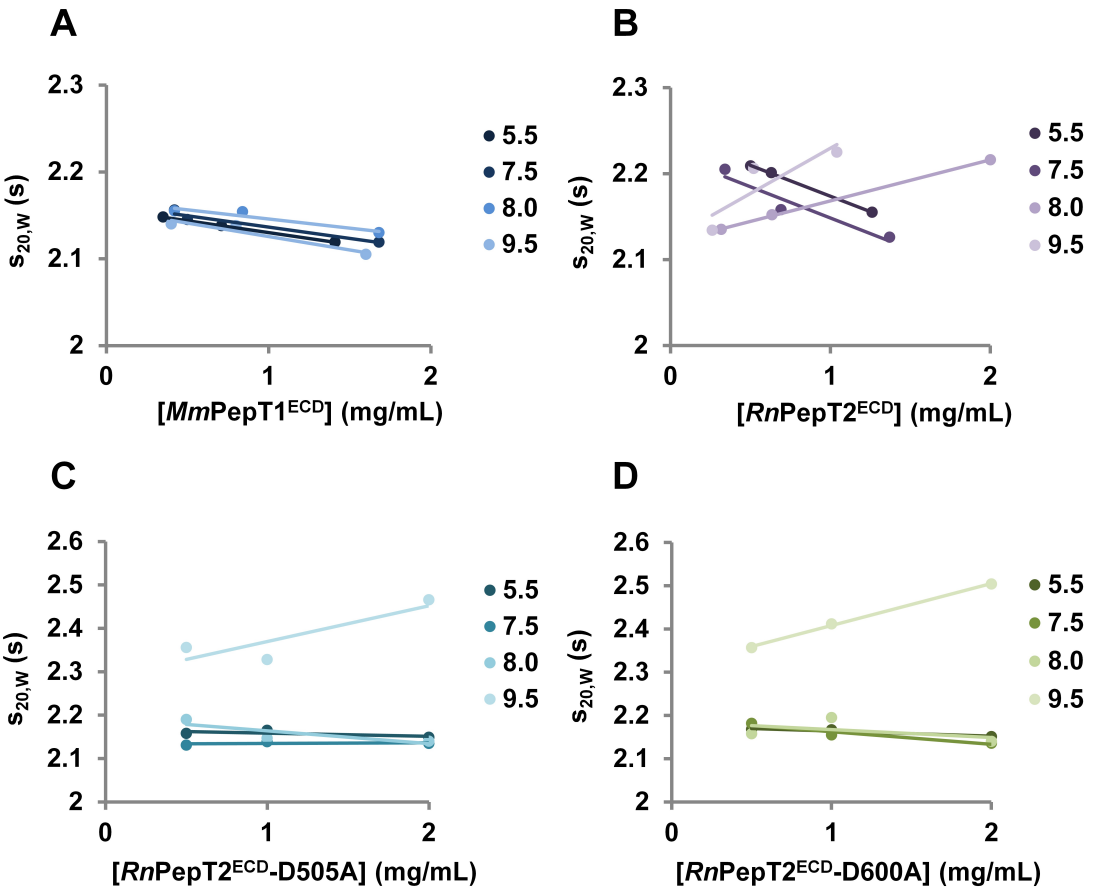


Figure 2.3: The $s^0_{20,w}$ calculation plots for *Mm*PepT1^{ECD} (A), *Rn*PepT2^{ECD} (B), D505A (C) and D600A (D). The $s_{20,w}$ values calculated by the c(s) distribution (Appendix B.2) are plotted against sample concentration. The $s^0_{20,w}$ are calculated using the y intercept of a linear regression of the concentration series.

Mutation	Primer (For/Rev)
D475A	caggtggtaaaggctggtcttaaccag/ ctggttaagaccagcctttaccacctg
D573A	aaggtgtttgaagctatttcagccaac/ gttggctgaaatagcttcaaacacctt

Table 2.17: *Hs*PepT1 site directed mutagenesis primers used during the oocyte assays.

approximately 0.2 mg/mL. During the thermal melt, one scan was taken per 2 °C increase in temperature from 20 – 80 °C. The melt transition was recorded by plotting the mdeg observed at 210 nm against temperature. The transition was then be fit to a Boltzman sigmoidal function using `Prism`.

2.8.6 Two electrode voltage clamp *X. laevis* oocyte assays

Cloning

The 8 constructs used in the TEVC oocyte assays are described in Figure 7.6. The initial cloning of the WT *Hs*PepT1 and *Hs*PepT2 genes into a pBF *X. laevis* oocyte expression vector, with a FLAG tag epitope engineered onto the C-terminus, was performed by Dr Jo Parker (Department of Biochemistry, University of Oxford). Dr Parker also constructed the ECD deleted and swapped mutants. The *Hs*PepT1 and *Hs*PepT2 ECDs were assumed to extend between residues 391 – 578 and 411 – 600 respectively. The mutants were created by overlapping PCR in a similar method to the one described in § 2.6.4. The D573A and D475A/D573A mutants were created by site directed mutagenesis using the same protocol as described in § 2.6.5 and the primers are shown in Table 2.17.

X. laevis oocyte assays

None of the TEVC *X. laevis* oocyte assays detailed in this thesis were performed by the author. The assays were conducted by Anne Barrett (Department of Biochemistry, University of Oxford) in the lab of Dr Stephen Tucker (Department of Physics, University of Oxford). The peptide transporter pBF vector mRNA was created using a AmpliCap SP6 High Yield Message Maker kit (CELLSCRIPT). The *X. laevis* oocytes were injected with 50 ng of the mRNA and incubated for 3 – 4 days at 17 °C. Oocytes were bathed

in ND96 buffer (95.4 mM NaCl, 2 mM KCl, 1.8 mM $CaCl_2$, 5 mM HEPES pH 7.5) and voltage clamped at −50 mV with a voltage step consisting of a 300 ms pulse from −160 mV to +60 mV in 10 mV increments. GlySar concentrations of 0.1, 0.3, 0.5, 0.7, 1.0, 2.0, 3.0, 4.0, 5.0, 7.0 and 10.0 mM were then incubated in the ND96, and the effect on the current was measured.

The graphs shown in Figure 7.6 were created by subtracting the buffer associated current, and normalised against the maximum current observed (10 mM GlySar). Data was analysed using `Clampfit 10.4.0.36` (Axon Instruments) and fit to the Michaelis-Menton equation (Equation 2.3) using `Prism`. Current (I) was used instead of transport velocity to calculate an apparent K_m for each oocyte. Five oocyte K_ms were averaged and displayed as mean ± Systematic error of the mean (SEM).

$$I = \frac{I_{MAX}[GlySar]}{K_m + [GlySar]} \tag{2.3}$$

2.8.7 Surface plasmon resonance protocol

Cloning

Table 2.18 shows the primers used in the creation of the *Mm*PepT1^{ECD} mutant constructs detailed in Figure 7.12A. Site directed mutagenesis (as described in § 2.6.5) was used to mutate the *Mm*PepT1^{ECD} pLou3 construct. Mutants were expressed and purified using the same protocol as the WT *Mm*PepT1^{ECD} (described in § 3.4.1).

Immobilisation

To assess the binding of the proteases, trypsin, α-chymotrypsin and pepsin, to the ECDs and *Mm*PepT1^{ECD} mutants, an SPR binding assay was performed. The ECDs were used as the immobilised ligand and the proteases were used as the analyte. Freshly purified ECD was prepared for amide immobilisation by passing the protein down an S75 16/60 GF column equilibrated in 1×PBS. The ECDs were immobilised onto a lane of a Series S, CM5 chip (GE Healthcare) using an Biacore^{TM} T200 (GE Healthcare) SPR processing unit, aiming for a final Response units $(RU)_{MAX}$ of 1000 RU. Table 2.19 shows the final im-

Mutation	Primer (For/Rev)
H453A/D454A/E456A	gtcaccacagtagctgctgcttttgcgcagggtcaccggc/ gccggtgaccctgcgcaaaagcagcagctactgtggtgac
H453E/D454R/E456R	gtcaccacagtagctgatgcttttgcgcagggtcaccggc/ gccggtgaccctgcctaaaagcatcagctactgtggtgac
R472A/K475A	ccccagtcaatacgctgtggtagcagatggtcttaacc/ ggttaagaccatctgctaccacagcgtattgactgggg
R472E/K475A	ccccagtcaatacgaagtggtagaagatggtcttaacc/ ggttaagaccatcttctaccacttcgtattgactgggg
N515A	tgtcaccagtcacgccgccagcggctacc/ ggtagccgctggcggcgtgactggtgaca
E509R	gagtgggaaagtatatagaaatgtcaccagtcac/ gtgactggtgacatttctatatactttcccactc
D550A	gattttaaatcttccaaccttgcctttggcagcgcgtatac/ gtatacgcgctgccaaaggcaaggttggaagatttaaaatc
E573A	gaagtgaaggaatttgcagacatcccacccaac/ gttgggtgggatgtctgcaaattccttcacttc
E425A	caaatgtctcagacagccacgttcatgactttc/ gaaagtcatgaacgtggctgtctgagacatttg
E568A	gcgagtgatggctgcctggcagtgaaggaatttgaagac/ gtcttcaaattccttcactgccaggcagccatcactcgc

Table 2.18: *Mm*PepT1^{ECD} mutant site directed mutagenesis primers used to elucidate the trypsin binding site.

mobilised protein levels for the ECDs and mutants used in the experiments. Ethanolamine blanks were immobilised on the adjacent reference cell.

Protease assay

Prior to the experiment lyophilised trypsin (Sigma) and α-chymotrypsin (Sigma), purified from bovine pancreas, was re-suspended in 25 mM MES pH 6.5, 100 mM NaCl, 10 mM $CaCl_2$, 8 mM Pefabloc® (Roche) to a final concentration of 2 mM. The trypsin and α-chymotrypsin solutions were incubated for 30 min at RT to allow the Pefabloc® to inactivate the enzymes. Pepsin was just re-suspended in SPR running buffer (25 mM MES pH 6.5, 100 mM NaCl, 10 mM $CaCl_2$, 0.005% v/v Tween 20). A serial dilution was then performed in the SPR running buffer to give experimental concentrations of 1000, 500, 250, 125, 62.5, 31.25, 15.625 and 7.8125 µM. These conditions, including a buffer blank, were passed over the immobilised ECDs and blank cell at a flow rate of 90 nL/s for 15 s. Analyte was allowed to dissociate for 30 s and the chip was regenerated in 2 M NaCl, passed over the flow cells at a flow rate of 30 nL/s for 30 s. Each assay was performed in

ECD	Immobilised RU_{MAX} (RU)
*Mm*PepT1^{ECD}	846
*Rn*PepT2^{ECD}	1104
MBP	834
*Mm*PepT1^{ECD}-H453A/D454A/E456A	1121
*Mm*PepT1^{ECD}-H453E/D454R/E456R	1168
*Mm*PepT1^{ECD}-R472A/K475A	1095
*Mm*PepT1^{ECD}-R472E/K475A	1172
*Mm*PepT1^{ECD}-N515A	1481
*Mm*PepT1^{ECD}-E509R	1337
*Mm*PepT1^{ECD}-D574A	1081
*Mm*PepT1^{ECD}-D550A/E573A	1232

Table 2.19: Final immobilisation levels of the ECDs and mutants used in the SPR protease assays.

triplicate.

Data analysis

The blank sensograms were subtracted from the experimental data. The equilibrium binding points, were taken from each member of the dilution series, buffer subtracted, averaged and fit using `Prism` to a one-site, specific binding curve (Equation 2.4).

$$Y = \frac{B_{MAX}[Analyte]}{K_d + [Analyte]} \tag{2.4}$$

3

Cloning, expression and purification of the PepT1 and PepT2 ECDs

3.1 Summary

In this chapter, the cloning and expression of the PepT1 and PepT2 ECD constructs used in this thesis are explained. It was decided to use the OPPF-UK HT system to clone and assess the expression of the ECDs in *E. coli*. *Mm*PepT1^{ECD} 391 – 580 expressed in pOPINM, which has an N-terminal Maltose binding protein (MBP) tag and purified yielding 10 mg per 1 L of TB. This *Mm*PepT1^{ECD} construct was used in all subsequent crystallisation and biophysical experiments. *Mm*PepT2^{ECD} 410 – 409 was expressed and purified in pOPINS3C, an N-terminal Small ubiquitin-like modifier (SUMO) tag vector. This construct later proved incompatible with crystallisation exper-

iments (as discussed in Chapter 5). A truncated construct, 410 – 601 from both *Mm*PepT2^{ECD} and *Rn*PepT2^{ECD} could be expressed and purified in the pLou3 expression vector, a different N-terminal MBP vector, yielding 10 mg per 1 L of TB. The *Rn*PepT2^{ECD} 410 – 601 was used for crystallisation and biophysical experiments described in Chapter 5 and 6.

3.2 Heterologous protein expression

Crystallography and protein biophysics are reductionist techniques which minimise experimental variables. So when preparing protein for such methodologies it is necessary to ensure that the protein is pure and mono-dispersed. Even given the advances in signal sensitivity which allow for less sample to be used to accurately measure a biophysical outcome, large quantities of pure protein are required for modern biophysics. Therefore, in order to characterise the ECDs structurally and biophysically, it is necessary to recombinantly express the ECDs in an heterologous organism to study them.

Major advances in recombinant protein expression have been made over the past 30 years and have seen the development of five well defined systems for heterologous protein expression: bacterial, mammalian, baculovirus, yeast and cell free synthesis. The primary organism is *E. coli* and accounts for 88% of deposited structures in the PDB (Nettleship *et al.* 2010). The size of this figure is due to the ease and speed of screening constructs in *E. coli*, and the development of fusion protein tags, such as MBP, Glutathione S-transferase (GST) and SUMO (Malhotra 2009; Young, Britton, and Robinson 2012). Fusion tags can enhance expression by aiding the translation of mRNA not otherwise translated by the bacterial ribosome. Solubility can also be increased by the presence of a very stable protein attached to the protein of interest.

Recombinant protein expression also allows for the incorporation of affinity purification tags, such as 6-His and FLAG (Malhotra 2009; Young, Britton, and Robinson 2012). 6-His tags have a high affinity for metal (II) bound NTA moieties. This has an advantage over FLAG based methods as it is cheap and robust. The latter qualities are important factors when considering expression systems. *E. coli* are cheap, fast and robust, and unless particular qualities unique to the other systems are required, such as glycolysation, *E. coli*

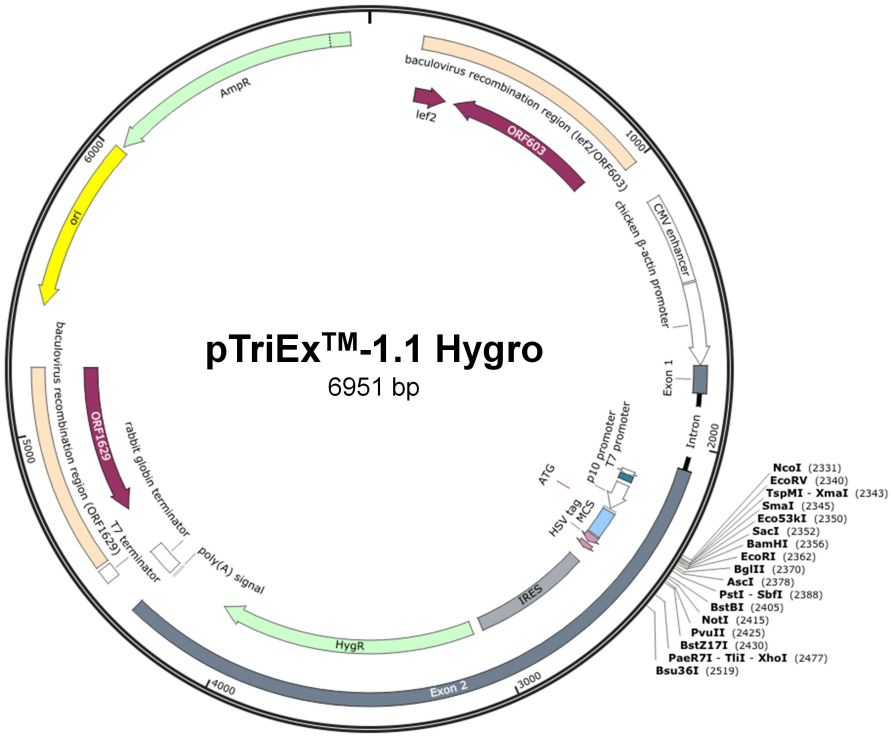


Figure 3.1: A vector map of the pTriEx^{TM}-1.1 Hygro vector. The pTriEx vector is the basis of the pOPIN vector suit and allows for *E. coli*, HEK293 and baculovirus intra-cellular expression. The image was generated by `SnapGene`.

are the first option in heterologous expression organism.

Given that no prior structural information on the ECDs exists it was appropriate to first try recombinant *E. coli* expression. HEK293 cells and baculovirus infected *Spodoptera frugiperda* (*Sf*9) cells can be undertaken if *E. coli* expression fails to give results. The OPPF-UK is a world renowned centre for HT protein expression. They have developed an expression vector library (pOPIN suite) based on the pTriEx^{TM}-1.1 (Novagen) plasmid (Figure 3.1), which allows for *E. coli*, HEK293 and baculovirus cloning and expression (Bird 2011). Therefore initial cloning and test expression of the PepT1 and PepT2 ECDs was undertaken by the author at the OPPF-UK.

3.3 OPPF-UK high throughput cloning and test expression

The advantage of the OPPF-UK HT *E. coli* expression system is that it allows for 48 constructs to be cloned rapidly. Expression of each construct is assessed in two *E. coli* strains and media/induction types; B834(DE3) and Rosetta-gami^{TM}(DE3)pLysS/pRARE

pOPIN vector	Fusion Protein
E	Insert-6His
F	6His-3C-Insert
M	6His-MBP-3C-Insert
J	6His-GST-3C-Insert
TRX	6His-TRX-3C-Insert
S3C	6His-S3C-3C-Insert
NUSA	6His-NUSA-3C-Insert
HALO	6His-HALO-3C-Insert
MSYB	6His-MSYB-3C-Insert
TF	6His-TF-3C-Insert

Table 3.1: A list of the OPPF-UK vectors used for the *E. coli* intra-cellular expression. All of the vectors have a 6-His tag to allow for Ni-NTA affinity purification. The fusion tags, shown in blue, increase the chance of soluble expression and can also be used as purification tools. All constructs apart from pOPINE have a 3C protease cut site inserted in a linker between the protein of interest and the fusion tag.

cells, and LB/IPTG induction and Overnight Express™ (TBonex) auto-inducing media (Studier 2005). The expression of each construct is therefore assessed in 4 different ways and done concordantly using HT technologies (Bird 2011).

At the beginning of this project 5 peptide transporter genes were available: *Hs*PepT1, *Hs*PepT2, *Mm*PepT1, *Mm*PepT2 and *Dr*PepT1 (see Table 2.1). Upon consultation with Dr Ray Owens and Dr Louise Bird, of the OPPF-UK it was decided to used the pTriEx™-1.1 based HT *E. coli* expression protocol. If this failed to give results, the intra-cellular HEK293 and *Sf*9 systems could be utilised without the need to repeat any cloning. After these original clones were tested, 4 more cDNAs became available: *Hs*PepT1-NX, which has the potential N-glycan sites removed, *Oc*PepT1, *Ce*OPT1 and *Ce*OPT3.

The 10 intra-cellular pOPIN vectors used in the HT test expression are shown in Table 3.1. The tags incorporated into the pOPIN suite have all been shown to be effective in particular instances, at encouraging expression of a stable, folded protein of interest. It is not possible to determine in advance what qualities a tag will require in order to do this; therefore it is good practice to try as many as possible to increase the probability of success. The pOPIN suite fusion tags include MBP, GST, Thioredoxin (TRX), SUMO (S3C), NusA, Halo Tag®, MsyB and Trigger Factor (TF).

Gene	Construct
*Hs*PepT1	391 – 579
*Hs*PepT2	410 – 609
*Hs*PepT1-NX	391 – 579
*Mm*PepT1	391 – 580
*Mm*PepT2	410 – 609
*Dr*PepT1	396 – 590
*Ce*OPT3	430 – 698
*Ce*OPT1	380 – 576
*Oc*PepT1	390 – 578

Table 3.2: The genes and construct lengths used in the OPPF-UK test expressions. The *Hs*PepT1-NX refers to the a *H. sapiens* construct which has had the probable N-Glycans sites in the ECD mutated to alanine (N439A and N509A). The full names and Uniprot codes for each gene are listed in Table 2.1.

3.3.1 ECD construct design

The crystal structure of $PepT_{So}$ (Newstead *et al.* 2011) allows for an accurate identification of the TM boundaries of the ECD. A sequence alignment of *H. sapiens* and *M. musculus*, PepT1 and PepT2, against $PepT_{So}$ is shown in Figure 3.2. The limits of TM helix 9 and 10 indicate the probable residues which mark the N- and C-terminus of the ECD in PepT1 and PepT2. Figure 3.3A shows the residues in the linker between TM helix 9 and 10 in $PepT_{So}$. Figure 3.3B and C, highlight the sections of the multiple vertebrate PepT1 and PepT2 sequences where the N- and C-terminal residues of the ECDs are likely to reside.

There is a well conserved proline at the end of TM 9 in both PepT1 and PepT2, therefore the start of the ECD constructs was assumed to be after P391 and P409 in PepT1 and PepT2 respectively (Figure 3.3B.1 and 3.3C.3). The probable C-terminal residue however, was not clear. For $PepT1^{ECD}$, it was decided to finish the construct at the end of a reasonably well conserved triplet P/A576-N577-T578 (Figure 3.3A.2). The following residues are well conserved so this probably marked the start of TM helix 10. Residues 574 – 587 in PepT2 (Figure 3.3C) all show high conservation. It was therefore decided to terminate $PepT2^{ECD}$ before the W610 as this was likely to mark the start of the TM helix 10 (Figure 3.3C.4). The construct lengths used in the OPPF-UK HT *E. coli* expression screen are listed in Table 3.2.

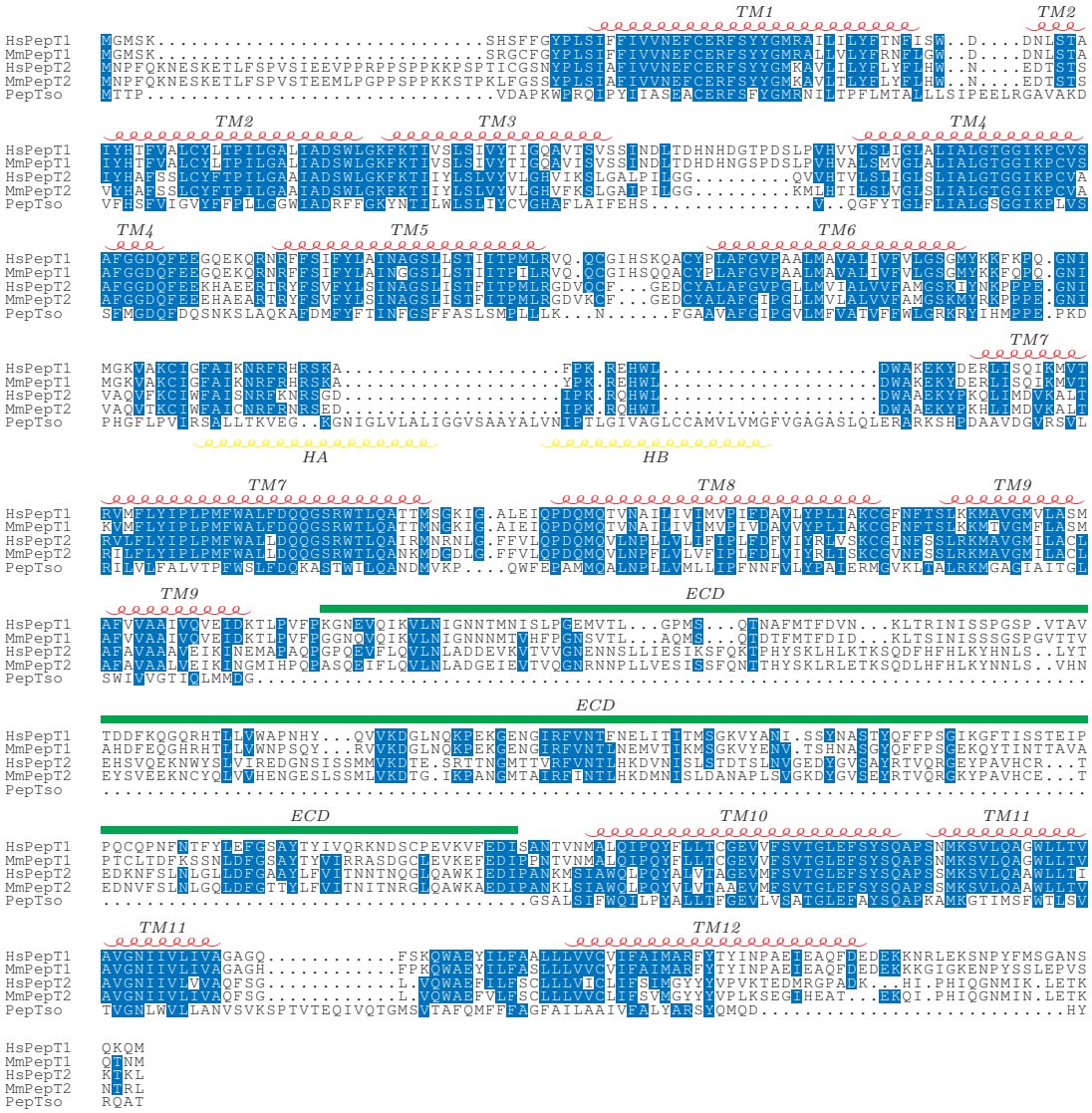


Figure 3.2: A sequence alignment of *H. sapiens* and *M. musculus*, PepT1 and PepT2 against $PepT_{So}$. The TM helices of $PepT_{So}$ have been marked onto the alignment, 1 – 12 in red above the sequences, and HA and HB below in yellow. The ECD sequences are highlight by the green bar. Figure 3.3 gives a more detailed analysis of the ECD termini based on this alignment.

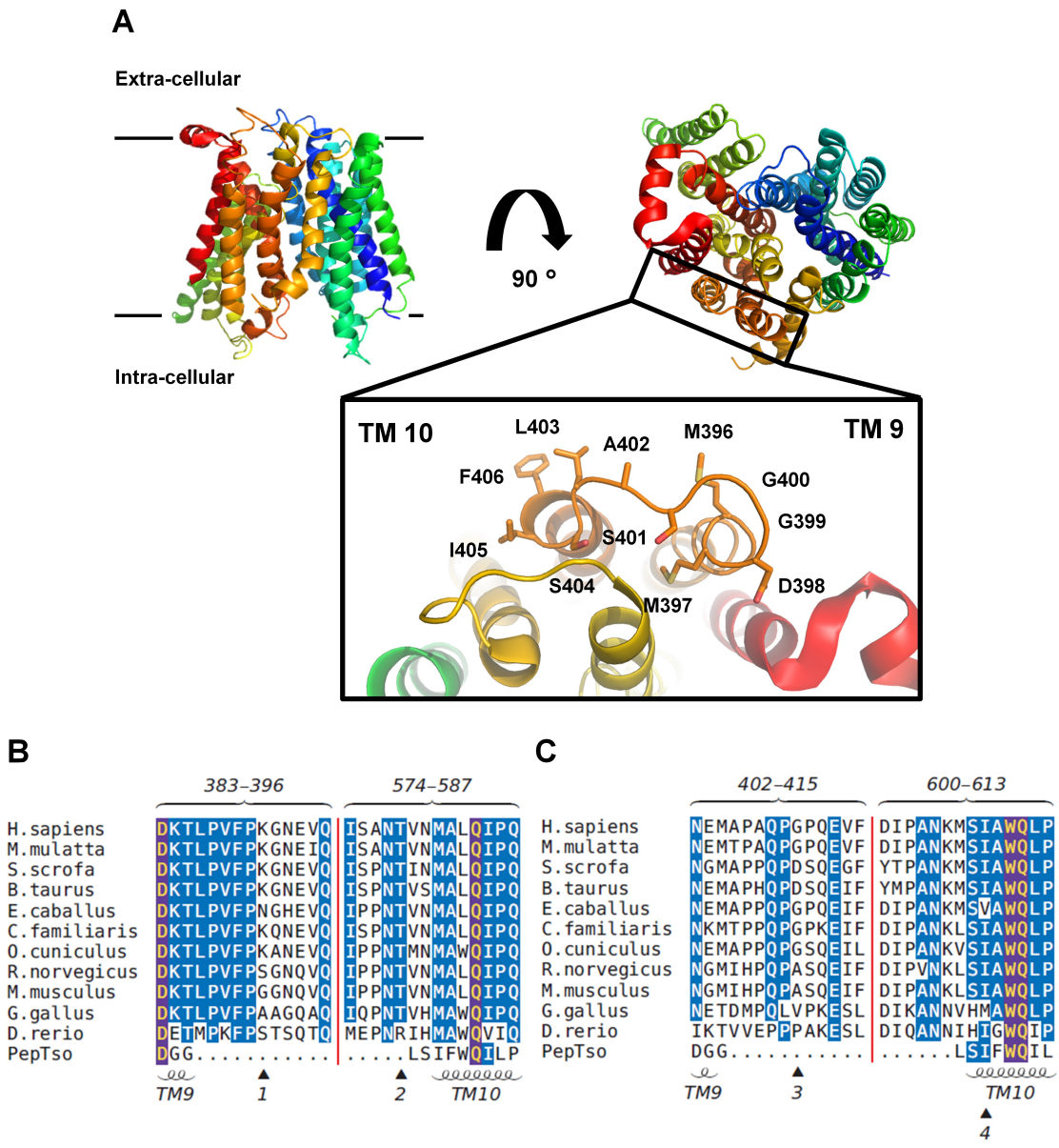


Figure 3.3: The N- and C-terminal locations of the ECDs from PepT1 and PepT2. (A) The extra-cellular linker connecting TM helix 9 and 10 in $PepT_{So}$ (Newstead *et al.* 2011) (PDB: 2XUT). **(B)** and **(C)** show alignments of vertebrate PepT1 and PepT2 respectively in the ECD N- and C-terminal regions against $PepT_{So}$. The full names and Uniprot codes for each gene are listed in Table 2.1. The numbers above the braces refer to the *H. sapiens* sequences. The proposed N- and C-terminus of the ECDs from PepT1 and PepT2 have been labelled 1 and 2, and 3 and 4, respectively. TM helices 9 and 10 are shown on the alignment.



Figure 3.4: PCR insert conformations of the ECDs in the pOPIN vectors. (A) A 1.5% w/v Agarose, 1 × TAE gel showing 48 PCR conformations of *Hs*PepT1, *Hs*PepT2, *Mm*PepT1, *Mm*PepT2 and *Dr*PepT1 ECDs in the pOPIN vectors detailed Table 3.3. The well numbers refer to the clone numbers, which are given in Table 3.3. **(B)** The DNA ladder used in **(A)**.

3.3.2 Cloning of ECDs into the pOPIN suite

The OPPF-UK cloning protocols are outlined in § 2.6.1. As explained in Bird (2011) the HT methodology is carried throughout the cloning and expression protocol. The PCR methodology described in § 2.6.1 is for 48 reactions performed concordantly. Phusion-Flash was the first choice DNA polymerase, which yielded PCR products with the correct MW in approximately 95% of reactions. For the 5% of PCRs which failed, a second reaction was performed using KOD Xtreme^{TM} to completed the 48 PCRs. PCR products were purified by a *Dpn*I digest, followed by DNA affinity exaction. The purified PCR products were ligated into the pOPIN suite (described in § 2.6.1) and transformed into OmniMAX^{TM} 2 competent cells. Plasmid DNA was purified from the transformants, and a PCR conformation was performed to screen for correct construct insertion.

Figure 3.4 shows the PCR conformation of constructs 1 – 48 run on a 1.5% w/v Agarose, 1 × TAE gel. The majority of cloned constructs gave an approximately correct MW, and were sent for sequencing at Source BioSciences using pOPIN specific forward and reverse primers. Constructs 7, 14, 38 and 47 gave an incorrect MW band and so were re-cloned. Constructs which had the correct sequence were taken forward for HT test expression.

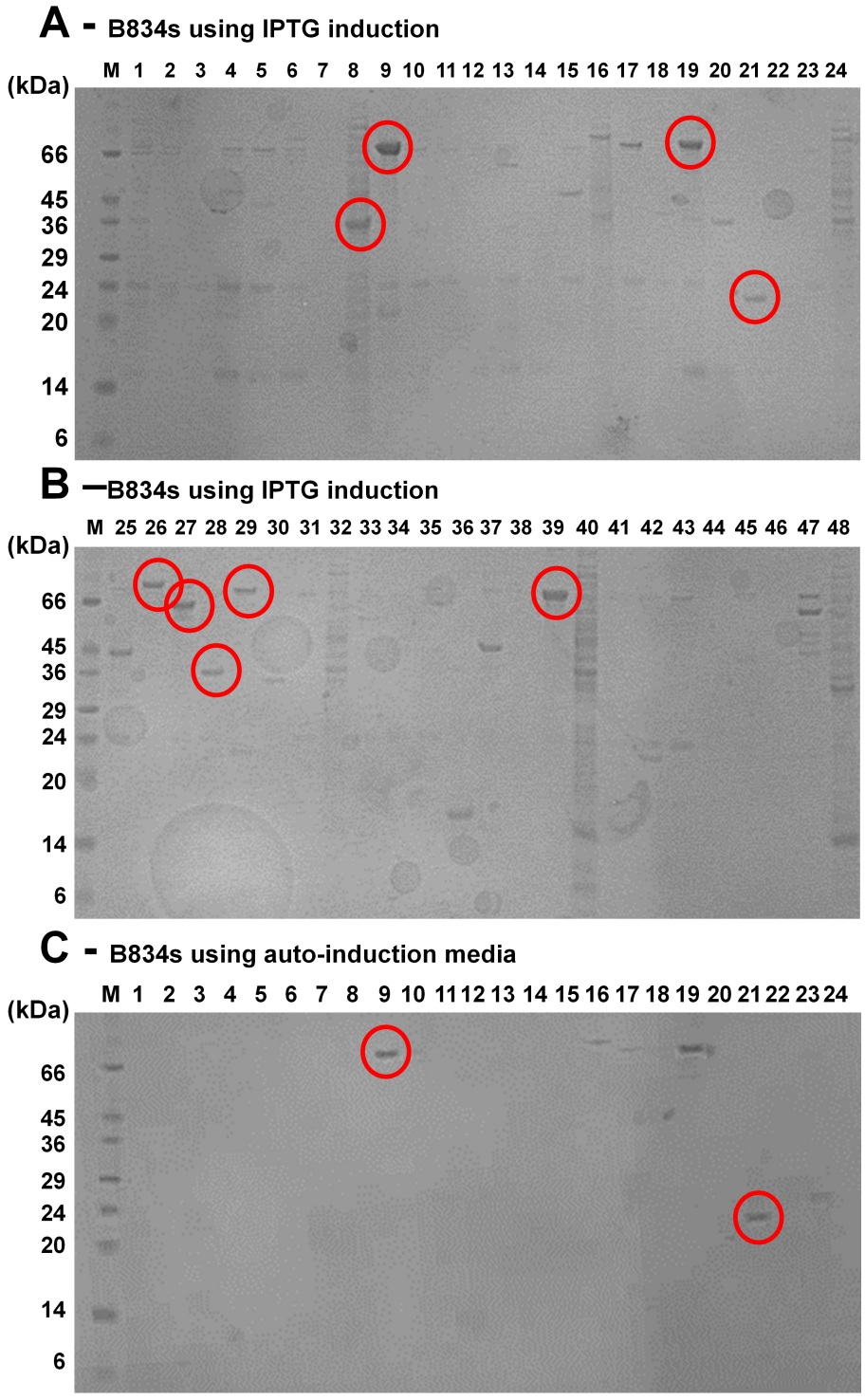
In total 69 different peptide transporter ECDs were cloned into the OPPF-UK pOPIN vectors (detailed in Table 3.3). The OPT2 construct failed to give any constructs from the cloning. It was later discovered that the parent gene had an insertion in the region of the ECD and therefore was not viable.

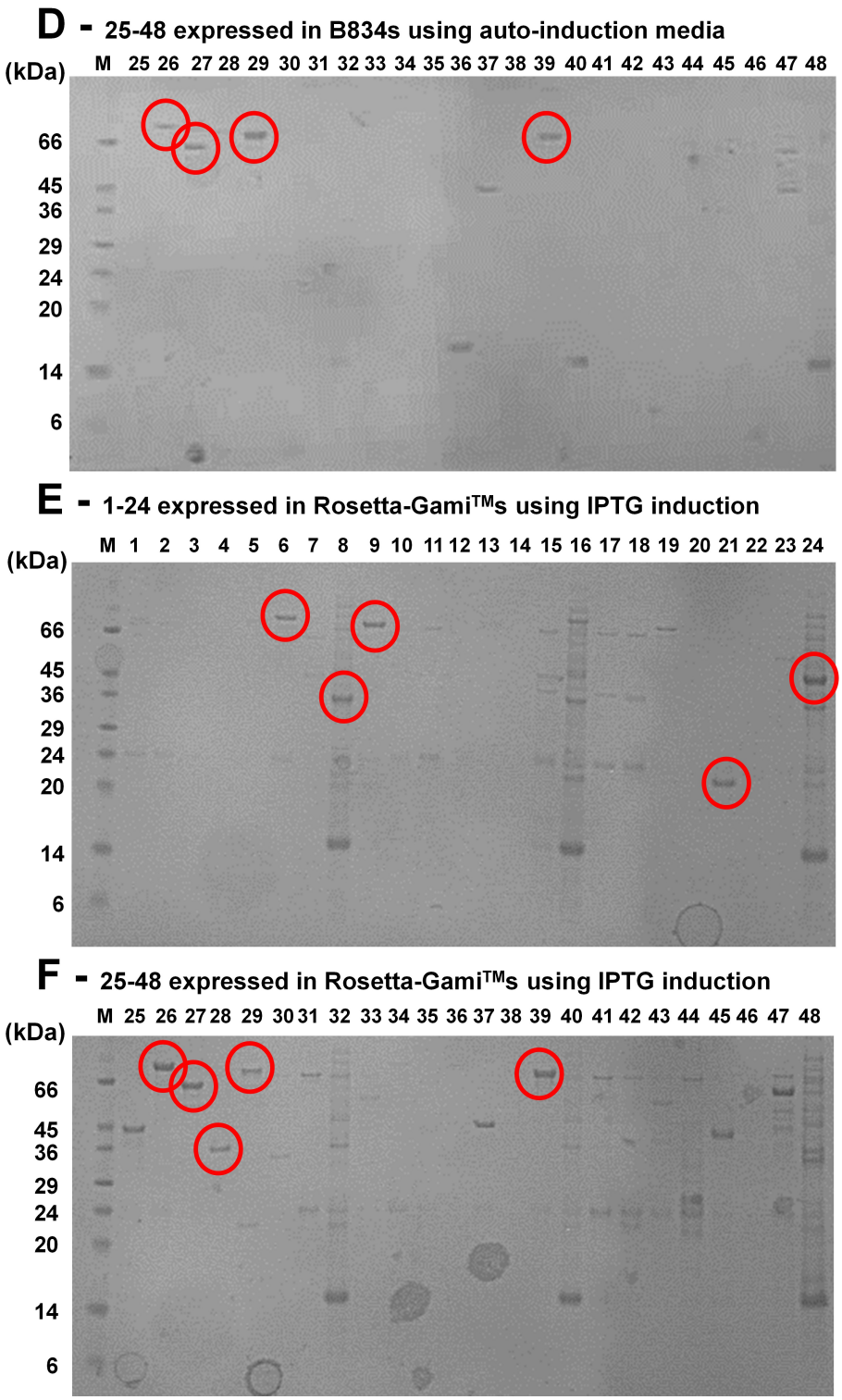
3.3.3 OPPF-UK HT small scale test expression of ECDs in pOPIN vectors

The HT small scale expression of the pOPIN vectors at the OPPF-UK is documented in Bird (2011) and is summarized in § 2.6.1. Each construct is expressed in B834(DE3) and Rosetta-gami^{TM}(DE3) pLysS/pRARE cells and in either LB with IPTG induction or TBonex with auto-induction. In total the protocol allows for the simultaneous assessment of 192 expression conditions.

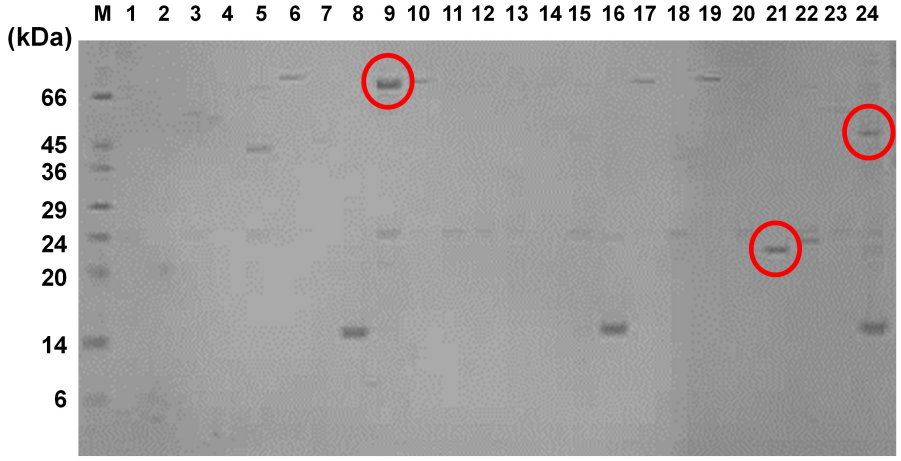
Two test expressions were performed at the OPPF-UK totalling 69 different constructs. The raw protein gels from the test expressions are shown in Figure 3.5. Table 3.3 tabulates the results from the test expressions. The test expression of the *Mm*PepT2 constructs were repeated as an error was found in the cloning. The constructs which showed bands at approximately the correct MW were flagged (highlighted with red circles in the Figure 3.5) and ranked by size (Table 3.3).

Figure 3.5: 12 reducing NuPAGE 10% Bis-Tris Midi gels from the OPPF-UK HT *E. coli* intra-cellular test expression of the ECD pOPIN constructs. Two test expression were conducted and are shown on gels **A-H** and **I-L** respectively. Gels **A-H** show the results from the 1st test expression and are labelled above each gel with their expression vector and mode of induction. Gels **I-L** show the results from the 2nd test expression, and are labelled above each gel with induction type, with the specific cell types labelled on the gel. The red circles indicate bands of the correct molecular weight. Clone numbers and results from the test expressions are detailed in Table 3.3.

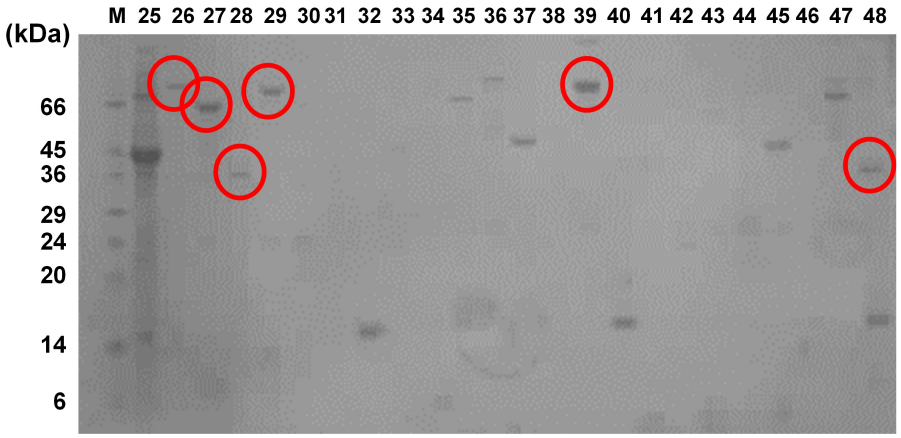



D - 25-48 expressed in B834s using auto-induction media
(kDa)
M 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
66
45
36
29
24
20
14
6
E - 1-24 expressed in Rosetta-Gami™s using IPTG induction
(kDa)
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
66
45
36
29
24
20
14
6
F - 25-48 expressed in Rosetta-Gami™s using IPTG induction
(kDa)
M 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
66
45
36
29
24
20
14
6

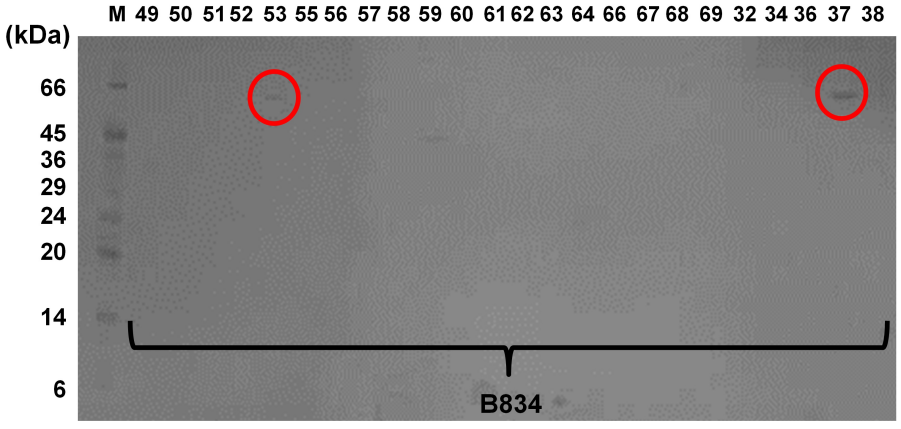
G - 1-24 expressed in Rosetta-Gami™s using auto-induction media



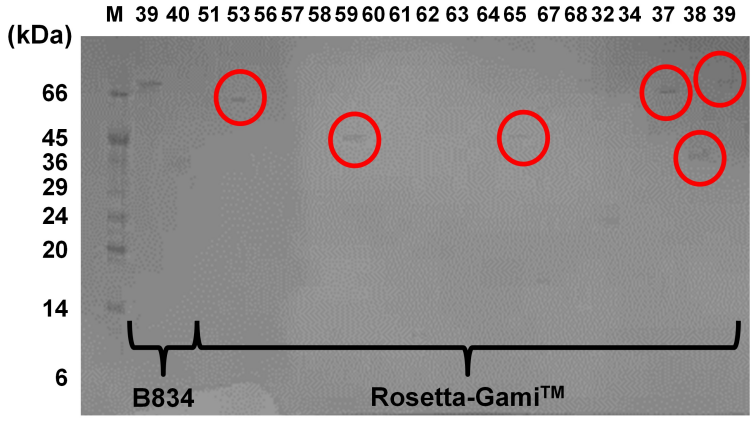
H - 25-48 expressed in Rosetta-Gami™s using auto-induction media



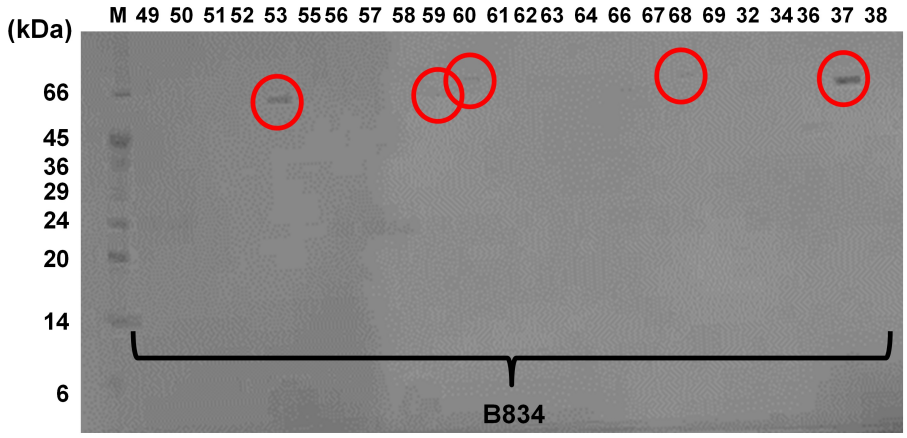
I - IPTG induction



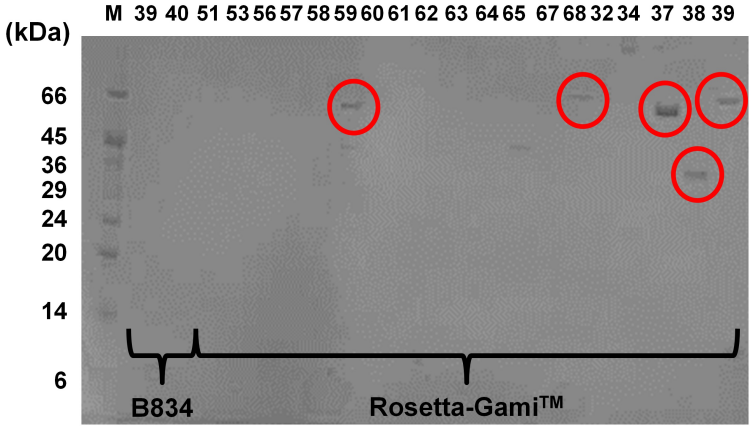
J - IPTG induction



K - Auto inducing media



L - Auto inducing media



			Expression Results			
			IPTG		Auto	
Clone No.	**Gene**	**Vector**	**B834**	**Rose***	**B834**	**Rose***
1	*Hs*PepT1^{ECD}	E	-	-	-	-
2		F	-	-	-	-
3		HALO	-	-	-	-
4		J	-	-	-	-
5		MSYB	-	-	-	-
6		NUSA	-	-	+	-
7		M	-	-	-	-
8		S3C	+	-	+	-
9		TF	+++	+	+	+++
10		TRX	-	-	-	-
11	*Hs*PepT2^{ECD}	E	-	-	-	-
12		F	-	-	-	-
13		HALO	-	-	-	-
14		J	-	-	-	-
15		MSYB	-	-	-	-
16		NUSA	-	-	-	-
17		M	-	-	-	-
18		S3C	-	-	-	-
19		TF	++	-	-	-
20		TRX	-	-	-	-
21	*Mm*PepT1^{ECD}	E	+	+	+	+
22		F	-	-	-	-
23		HALO	-	-	-	-
24		J	-	-	++	+
25		MSYB	-	-	-	-
26		NUSA	+	+	++	+
27		M	++	+	++	++
28		S3C	+	-	+	+
29		TF	+	++	+	+
30		TRX	-	-	-	-
31	*Mm*PepT2^{ECD}	E	-	-	-	-
32		F	-	-	-	-
33		HALO	-	-	-	-
34		J	-	-	-	-
35		MSYB	-	-	-	-
36		NUSA	-	-	-	-
37		M	-	-	-	-
38		S3C	-	-	-	-
39		TF	++	+	++	++
40		TRX	-	-	-	-
41	*Dr*PepT2^{ECD}	E	-	-	-	-
42		F	-	-	-	-
43		J	-	-	-	-
44		NUSA	-	-	-	-
45		M	-	-	-	-

Continued on next page

Table 3.3 – *Continued from previous page*

			Expression Results			
			IPTG		Auto	
Clone No.	**Gene**	**Vector**	**B834**	**Rose***	**B834**	**Rose***
46		S3C	-	-	-	-
47		TF	-	-	-	-
48		TRX	-	-	-	+
49		F	-	-	-	-
50		HALO	-	-	-	-
51		J	-	-	-	-
52	*Hs*PepT1^{ECD}-NX	MSYB	-	-	-	-
53		M	+	+	+	+
54		TF	-	-	-	-
55		TRX	-	-	-	-
56		F	-	-	-	-
57		HALO	-	-	-	-
58		J	-	-	-	-
59	*Ce*OPT3^{ECD}	M	-	-	+	+
60		TF	-	-	+	-
61		TRX	-	-	-	-
62		F	-	-	-	-
63		HALO	-	-	-	-
64		J	-	-	-	-
65		MSYB	-	-	-	-
66	*Oc*PepT1^{ECD}	M	-	-	-	-
67		S3C	-	-	-	-
68		TF	-	-	+	+
69		TRX	-	-	-	-
32		F	-	-	-	-
34		J	-	-	-	-
36		MSYB	-	-	-	-
37	*Mm*PepT2^{ECD} repeat	M	++	+	++	+++
38		S3C	-	+	-	+
39		TF	+	+	-	+
40		TRX	-	-	-	-

Table 3.3: Results from the OPPF-UK intra-cellular *E. coli* test expression of the ECD pOPIN constructs. These results are based on Figure 3.5. (+) ≈ 1 mg/L, (++) ≈ 5 mg/L, (+++) ≈ 10 mg/L. '*' = -gami^{TM}.

Table 3.4 shows a summary of the test expression results from constructs 1 – 69. TF appeared to be a good candidate fusion tag showing good expression for 7 different constructs. MBP and SUMO also showed expression in 3–4 constructs. Large scale expression trials were planned for the constructs in Table 3.4.

Clone No.	Gene	Vector	Score
9	*Hs*PepT1^{ECD}	TF	+++
37	*Mm*PepT2^{ECD}	M	+++
19	*Hs*PepT2^{ECD}	TF	++
27	*Mm*PepT1^{ECD}	M	++
24	*Mm*PepT1^{ECD}	J	++
26	*Mm*PepT1^{ECD}	NUSA	++
29	*Mm*PepT1^{ECD}	TF	++
39	*Mm*PepT2^{ECD}	TF	++
48	*Dr*PepT1^{ECD}	TF	++
6	*Hs*PepT1^{ECD}	NUSA	+
8	*Hs*PepT1^{ECD}	S3C	+
38	*Mm*PepT2^{ECD}	S3C	+
21	*Mm*PepT1^{ECD}	E	+
28	*Mm*PepT1^{ECD}	S3C	+
53	*Hs*PepT1^{ECD}-NX	M	+
60	*Ce*OPT3^{ECD}	M	+
61	*Ce*OPT3^{ECD}	TF	+
68	*Oc*PepT1^{ECD}	TF	+

Table 3.4: Summary of the ECD pOPIN *E. coli* intra-cellular test-expression results. This table is based upon Table 3.3. (+) ≈ 1 mg/L, (++) ≈ 5 mg/L, (+++) ≈ 10 mg/L.

3.4 Large scale expression of OPPF-UK constructs

Large scale test expression and purification of the *Hs*PepT1^{ECD} was attempted in the TF and MBP constructs. However, the ECD was not stable post fusion tag cleavage. In contrast the *Mm*PepT1^{ECD}, residues 391 – 580, was found to be stable. The optimised purification is described in § 3.4.1.

3.4.1 *Mm*PepT1^{ECD} pOPINM purification

*Mm*PepT1^{ECD} was transformed into Rosetta-gami^{TM} competent cells. 2 L of *E. coli* were grown and cell lysate prepared as described in § 2.6.3. The cell lysate was incubated with 10 mL of Ni-NTA sepharose equilibrated in Lysis buffer (see Buffers 3.1) for 1 hr/150 rpm. The lysate bead mixture was poured through a gravity column and the unbound material eluted. The sepharose was washed with Lysis buffer and bound protein was eluted in 50 mL of Ni-Elution buffer. This was immediately loaded onto a gravity column containing 20 mL of amylose resin equilibrated in Lysis buffer. The amylose resin was washed in 50 mL of Am-Wash buffer and the MBP-ECD fusion protein was eluted in 60 mL of Am-Elution

Buffers 3.1: *Mm*PepT1$^{\mathbf{ECD}}$ and *Mm*PepT2$^{\mathbf{ECD}}$ 410 – 609 purification.

- Lysis:
  - 1 × PBS
  - 20 mM Imidazole pH 8.0
  - 150 mM NaCl
  - 10% v/v Glycerol
- Ni-Elution:
  - 1 × PBS
  - 250 mM Imidazole pH 8.0
  - 150 mM NaCl
  - 10% v/v Glycerol
- Am-Wash:
  - 20 mM Tris pH 7.5
  - 150 mM NaCl
- Am-Elution:
  - 20 mM Tris pH 7.5
  - 150 mM NaCl
  - 10 mM Maltose
- Dialysis:
  - 20 mM Tris pH 7.5
  - 150 mM NaCl
  - 10 mM Imidazole pH 8.0
- GF:
  - 20 mM Tris pH 7.5
  - 150 mM NaCl

buffer. Figure 3.6 shows the gels and GF trace from a *Mm*PepT1^{ECD} purification.

Fusion protein concentration was approximated with a UV-spectrometer at A_{280}. 1 mg of 3C protease was added per 50 mg of fusion protein. The cleavage reaction was dialysed overnight into 2 L of Dialysis buffer at 4 °C for 16 hr with gentle agitation.

The cleaved *Mm*PepT1^{ECD} was passed through a 0.22 µm filter and Reverse purified (RP) using a 5 mL His-Trap HP (GE Healthcare) and a 20 mL amylose gravity column equilibrated in Dialysis buffer. The His-Trap was washed in 25 mL of Dialysis buffer post binding and the Flow through (FT) and wash steps were pooled. The pooled fractions were passed through an amylose gravity column which was subsequently washed with 50 mL of Dialysis buffer. These fractions were then pooled and concentrated in a 5 kDa molecular weight centrifugal filter.

The *Mm*PepT1^{ECD} was concentrated to 2 mL and passed through a Superdex S75 16/60 GF column equilibrated in GF buffer (Figure 3.6). The fractions containing the ECD were pooled and concentrated to 20 mg/mL using a 5 kDa molecular weight centrifugal

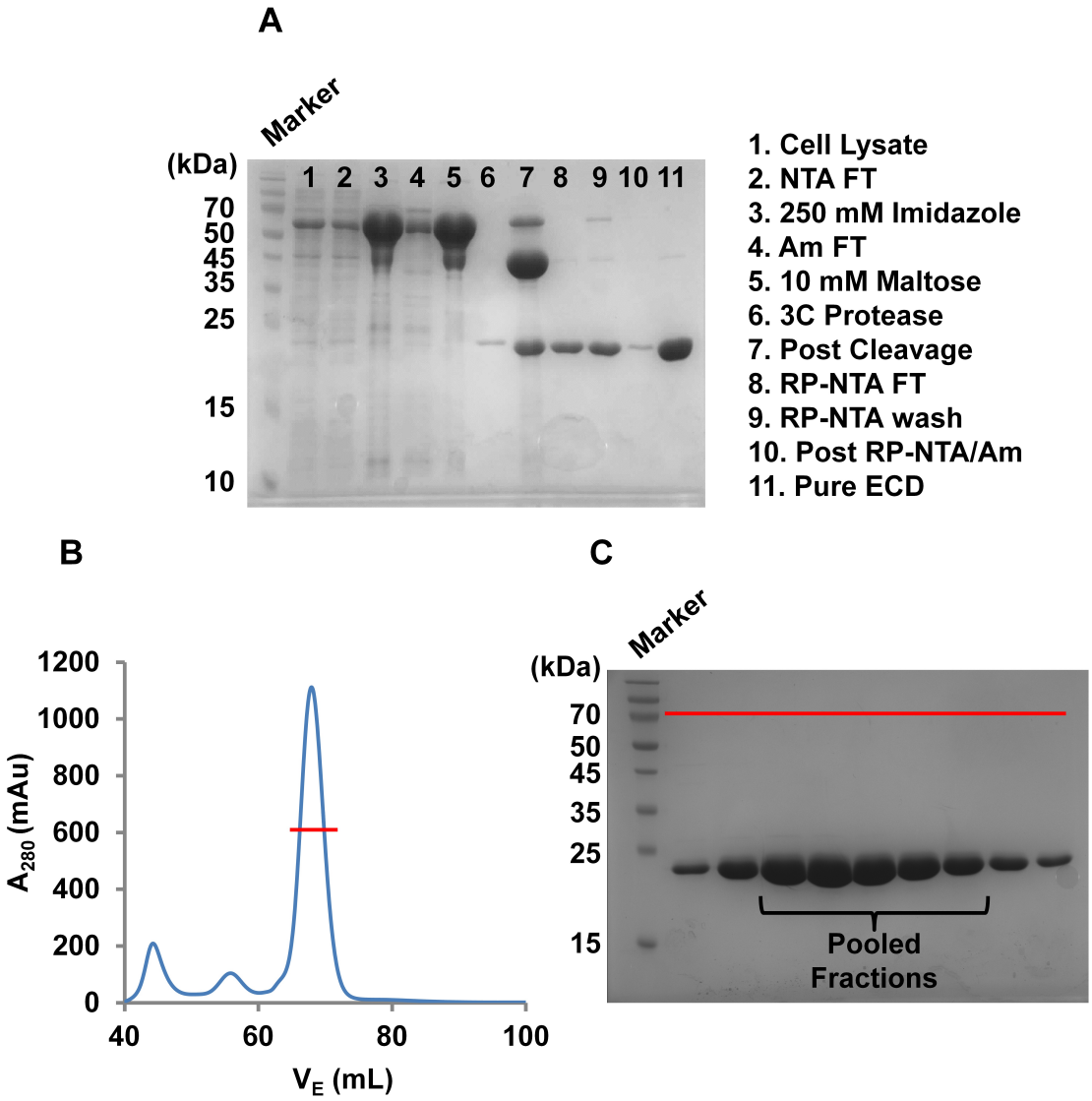


Figure 3.6: The *Mm*PepT1^{ECD} pOPINM purification. **(A)** A reducing 15% Tris-Gly gel showing an overview of the complete purification. **(B)** The final S75 16/60 GF trace of the concentrated pooled fractions from the RP-NTA/Am fractions. The red line indicates the GF fractions run on a reducing 15% Tris-Gly gel **(C)**.

filter, flash frozen in liquid nitrogen and stored at −80 °C. Yields were typically 10 mg of *Mm*PepT1^{ECD} per 1 L of culture.

3.4.2 *Mm*PepT2^{ECD} 410 – 609 pOPINS3C purification

The *Hs*PepT2^{ECD} and *Mm*PepT2^{ECD} TF constructs were assessed in large scale purifications. *Hs*PepT2^{ECD}, like *Hs*PepT1^{ECD}, appeared not to be stable post fusion tag cleavage, and the *Mm*PepT2^{ECD} could not be separated from the TF tag post cleavage. This was also the case the the MBP construct. However, a large scale test expression of the *Mm*PepT2^{ECD} (residues 401 – 609) pOPINS3C showed good expression at approximately 15 mg/L of fusion protein (Figure 3.7A). The construct was transformed into Rosetta-gami^{TM} competent cells. 2 L of *E. coli* were grown and lysate prepared as described in § 2.6.3. The buffers used in the purification were the same as those used for the purification of *Mm*PepT1^{ECD} (Buffers 3.1).

The cell lysate was incubated with 10 mL of Ni-NTA sepharose equilibrated in Lysis buffer for 1 hr/150 rpm. The lysate bead mixture was poured through a gravity column and the unbound material eluted. The sepharose was washed with Lysis buffer and bound protein was eluted in 50 mL of Ni-Elution buffer. The fusion protein concentration was approximated with a UV-spectrometer at A_{280}. 1 mg of 3C protease was added per 50 mg of fusion protein. The cleavage reaction was dialysed overnight into 2 L Dialysis buffer at 4 °C for 16 hr with gentle agitation.

The cleaved *Mm*PepT2^{ECD} was passed through a 0.2 µm filter and reverse purified using a 5 mL His-Trap HP equilibrated in Dialysis buffer. The His-Trap was washed in 25 mL of Dialysis buffer post binding and the FT and wash steps were pooled. These pooled fractions were concentrated in a 5 kDa molecular weight centrifugal filter to 2 mL.

The concentrated *Mm*PepT2^{ECD} was passed through a Superdex S75 16/60 GF column equilibrated in GF buffer (Figure 3.7C). The fractions containing the ECD were pooled and concentrated to 20 mg/mL using a 5 kDa molecular weight centrifugal filter, flash frozen in liquid nitrogen and stored at −80 °C.

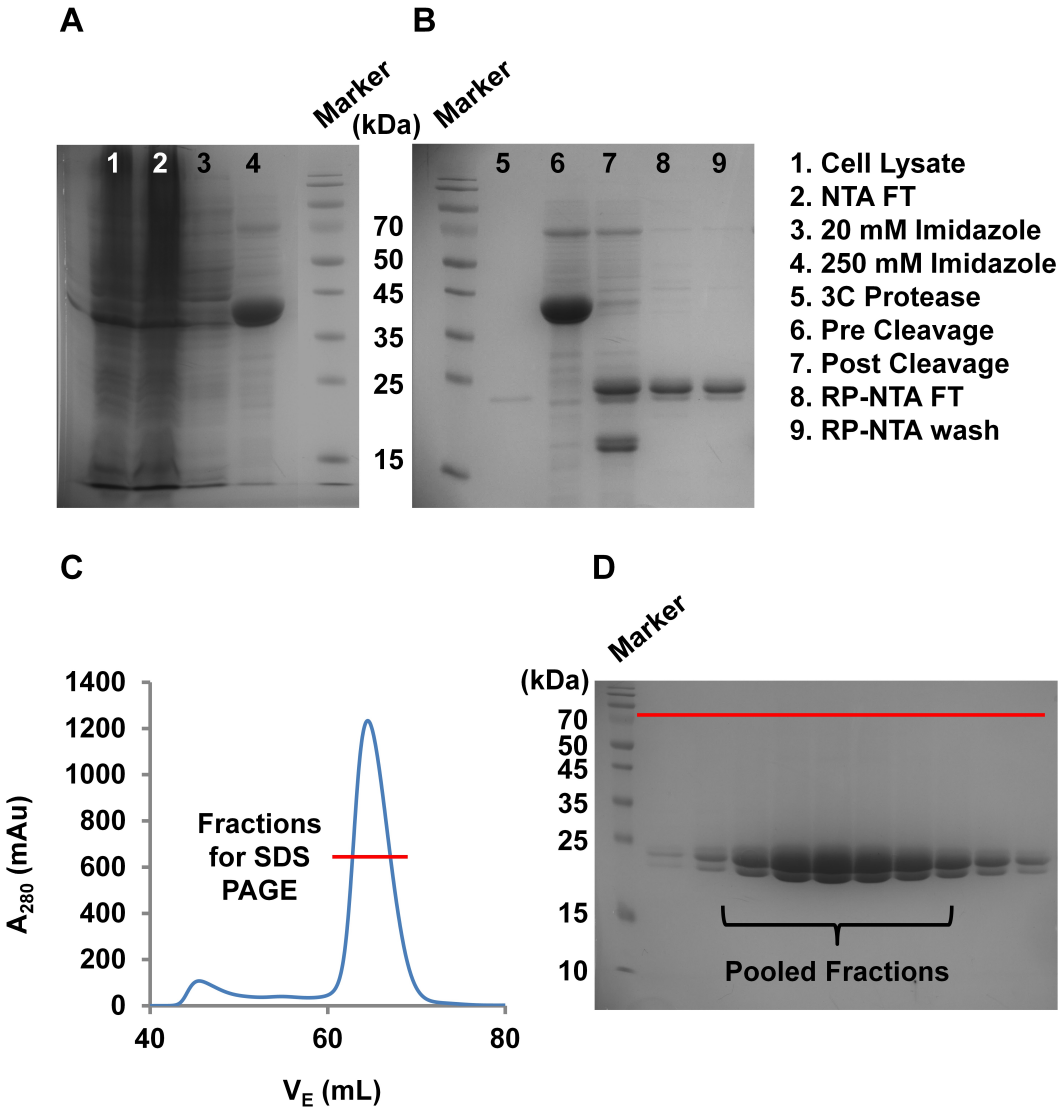


Figure 3.7: The *Mm*PepT2^{ECD} pOPINS3C purification. Two reducing 12% Tris-Gly gels, **(A)** and **(B)**, showing the capture and RP of the ECD. **(C)** The final S75 16/60 GF trace of the concentrated pooled fractions from the RP-NTA. The red line indicates the GF fractions run on a reducing 15% Tris-Gly gel **(D)**. Mass spectrometry analysis of the multiple bands present in the GF peak fractions confirmed all were full length *Mm*PepT2^{ECD} 610 – 609.

Gene	Construct
*Ce*OPT3^{ECD}	379 – 573
*Ce*OPT1^{ECD}	411 – 664
*Hs*PepT2^{ECD}	410 – 601
*Mm*PepT2^{ECD}	410 – 601
*Rn*PepT2^{ECD}	410 – 601
*Xl*PepT2^{ECD}	362 – 555
*Dr*PepT2^{ECD}	372 – 559

Table 3.5: Genes and constructs used in the Oxford Biochemistry *E. coli* ECD test expression. The full names and Uniprot codes for each gene are listed in Table 2.1.

3.5 Optimised PepT2^{ECD} construct cloning and test expression

It was possible to crystallise the *Mm*PepT1^{ECD} construct (described in Chapter 4). However, the *Mm*PepT2^{ECD} 410 – 609 construct, purified in § 3.4.2 failed to give any crystallisation hits (described in § 5.2). Therefore new genes were purchased to search for other stable homologues which might have a greater propensity to crystallise. The *C. elegans* OPT^{ECD} genes were also re-screened.

The new genes and construct lengths are shown in Table 3.5. As described in Chapter 5 the PepT2^{ECD} constructs were truncated from residues 410 – 609 to 410 – 601. The new constructs were cloned and screened at Oxford Biochemistry using the pLou3 (based on pMAL^{TM}) and pEHisTev vectors (described in § 2.6.2). pLou3 and pEHisTev are expression vectors with an N-terminal 6-His-MBP tag and an N-terminal 6-His tag respectively, both linked to the protein of interest with a Tobacco etch virus (TEV) protease site.

3.5.1 Cloning of the new PepT2 constructs into the pLou3 and pEHisTev expression vectors

The cloning procedure for the new PepT2^{ECD} constructs is outlined in § 2.6.2. Figure 3.8 shows the test cleavage (described in § 2.6.2) for the *Ce*OPT1^{ECD}, *Hs*PepT2^{ECD} and *Rn*PepT2^{ECD} for both the pLou3 and pEHisTev vectors. The isolated transformed colony plasmids were cleaved with *Nco*I and *Hind*III restriction nucleases which excises the entire insert in both the pLou3 and pEHisTev vectors, and should give a band of 600 bp and 700 bp for *Hs*PepT2^{ECD} and *Rn*PepT2^{ECD}, and *Ce*OPT1^{ECD} respectively. The sequencing of

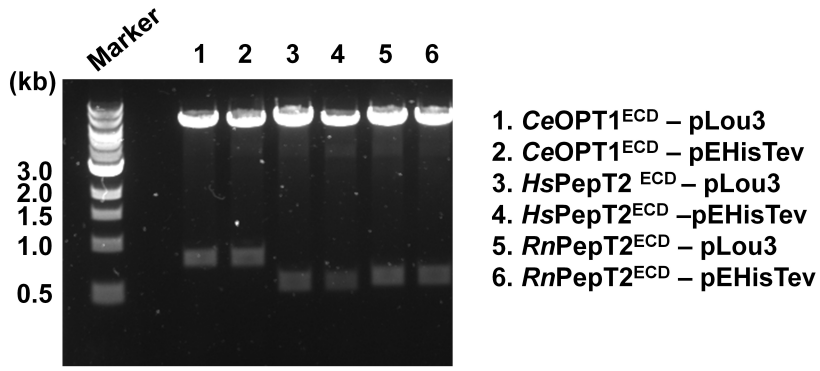


Figure 3.8: A 1.5% w/v Agarose, 1 × TAE text cleavage gel of the PepT2^{ECD} constructs in the pLou3 and pEHisTev vectors. The isolated cloned plasmids were cleaved using the restriction enzymes *Nco*I and *Hind*III. The liberated DNA fragments corresponded to the correct sizes for the expression plasmids and ECD inserts.

these constructs indicated that the cloning was successful.

3.5.2 Test expression of the PepT2^{ECD} pLou3 and pEHisTev constructs

The test expression procedure for the PepT2^{ECD} constructs in pLou3 and pEHisTEv vectors is outlined in § 2.6.2 and Figure 3.9 shows a test expression for some of these constructs. The *H. sapiens*, *R. norvegicus* and *X. laevis* PepT2 and *Ce*OPT3 constructs appear to have expressed well in the pLou3 vector. Large scale test expressions and purifications of the four constructs were performed. The *Mm*PepT2^{ECD} and *Rn*PepT2^{ECD} pLou3 410 – 601 constructs expressed well in large scale and could be purified (described in § 3.5.3).

3.5.3 Purification of *Mm*PepT2^{ECD} and *Rn*PepT2^{ECD} 410 – 601 pLou3 constructs

The purification protocol for the *Mm*PepT2^{ECD} and *Rn*PepT2^{ECD} 410 – 601 pLou3 constructs was the same. The 410−601 constructs were transformed into C43(DE3) competent cells (as described in § 2.6.2) and 2 L of *E. coli* were grown and the cell lysates prepared as described in § 2.6.3.

The cell lysate was incubated with 10 mL of Ni-NTA sepharose equilibrated in Lysis buffer (Buffers 3.2) for 1 hr/150 rpm. The lysate bead mixture was poured through a gravity

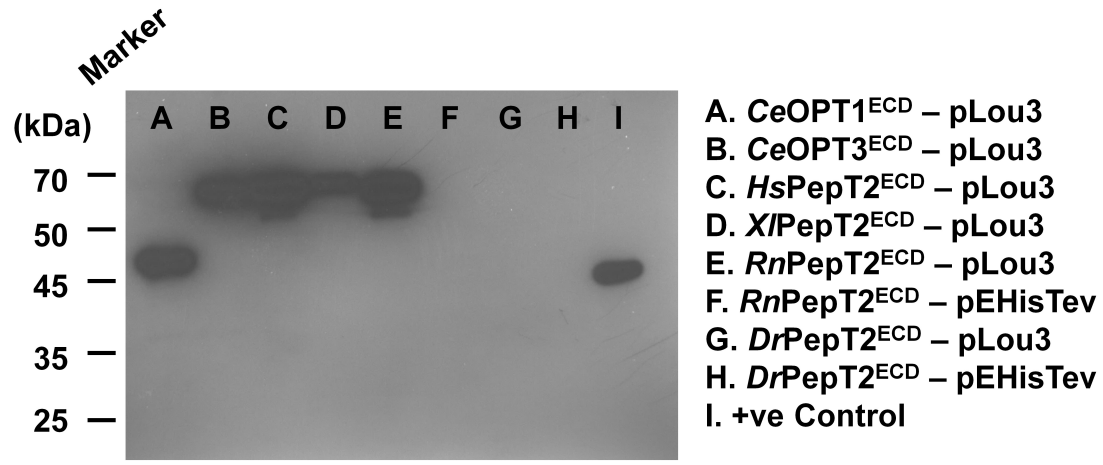


Figure 3.9: An anti-His western blot from a PepT2^{ECD} pLou3 and pEHisTev test expression using TB media. The blot shows good expression of the *Ce*OPT1^{ECD}, *Ce*OPT3^{ECD}, *Ce*OPT1^{ECD}

column and the unbound material eluted. The sepharose was washed with Lysis buffer and bound protein was eluted in 50 mL of Ni-Elution buffer. This was immediately loaded onto a gravity column containing 20 mL of amylose resin equilibrated in Lysis buffer. The amylose resin was washed in 50 mL of Am-Wash buffer and the MBP-ECD fusion protein eluted was in 60 mL of Am-Elution buffer. Figure 3.10 shows the gels and traces from a *Rn*PepT2^{ECD} purification.

Fusion protein concentration was approximated with a UV-spectrometer at A_{280}. 5 mg of TEV protease were added per 50 mg of fusion protein. The cleavage reaction was dialysed overnight into 2 L of Dialysis buffer at 4 °C for 16 hr with gentle agitation.

The cleaved 410 – 601 was passed through a 0.2 µm filter and RP using a 5 mL His-Trap HP and a 20 mL amylose gravity column equilibrated in Dialysis buffer. The His-Trap was washed in 25 mL of Dialysis buffer post binding and the FT and wash steps were pooled. The pooled fractions were passed through the amylose gravity column which was subsequently washed with 50 mL of Dialysis buffer. These fractions were then pooled and concentrated in a 5 kDa molecular weight centrifugal filter to 5 mL.

The concentrated 5 mL fraction was passed down a 5 mL Q-Trap (GE Healthcare) equilibrated in Dialysis buffer. The flow rate was set to 5 mL/min. The eluent was passed over the column and after 15 mL of buffer eluted, the NaCl concentration was gradually increased to 20% over 50 mL. The 410 – 601 elutes first from the column at a NaCl con-

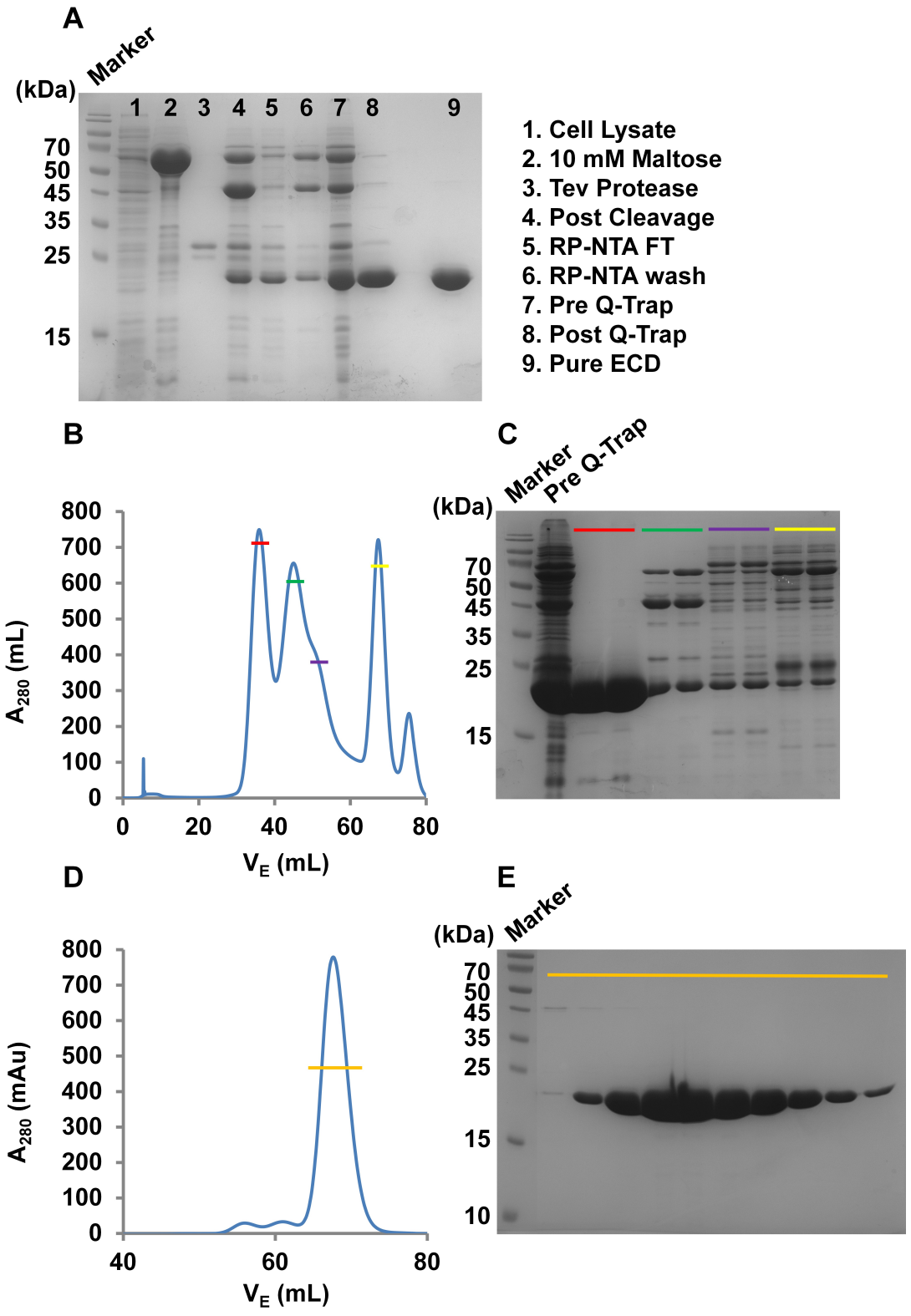


Figure 3.10: A *Rn*PepT2^{ECD} 410 – 601 pLou3 purification. **(A)** A reducing 15% Tris-Gly gel showing an overview of the complete purification. **(B)** The Q-Trap trace of the concentrated FT and wash steps from the RP-Amylose. **(C)** The reducing 15% Tris-Gly gel showing the peak fractions highlighted from the the Q-Trap. **(D)** The final S75 16/60 GF trace of the concentrated pooled Q-Trap peak (red). **(E)** The reducing 15% Tris-Gly gel of GF fractions of the highlighted peak from **(D)**.

Buffers 3.2: *Mm*PepT2^{ECD} and *Rn*PepT2^{ECD} purification.

- Lysis:
  - 1 × PBS
  - 20 mM Imidazole pH 8.0
  - 150 mM NaCl
  - 10% v/v Glycerol
- Ni-Elution:
  - 1 × PBS
  - 250 mM Imidazole pH 8.0
  - 150 mM NaCl
  - 10% v/v Glycerol
- Am-Wash:
  - 20 mM Tris pH 7.5
  - 150 mM NaCl
- Am-Elution:
  - 20 mM Tris pH 7.5
  - 150 mM NaCl
  - 10 mM Maltose
- Dialysis:
  - 20 mM MES pH 6.5
  - 20 mM Imidazole
- GF:
  - 20 mM Na Acetate pH 5.5
  - 150 mM NaCl

centration of approximately 100 mM. The peak was concentrated to 2 mL using a 5 kDa molecular weight centrifugal filter and passed through a Superdex S75 16/60 equilibrated in GF buffer. The fractions containing the ECD were pooled and concentrated to 20 mg/mL using a 5 kDa molecular weight centrifugal filter, flash frozen in liquid nitrogen and stored at −80 °C.

Yields for the *Rn*PepT2^{ECD} and *Mm*PepT2^{ECD} 410 – 601 constructs were typically 10 mg of ECD per 1 L of culture. The *Rn*PepT2^{ECD} 410 – 601 construct proved amenable to crystallography which is described in Chapter 5.

3.6 Discussion

The OPPF-UK system gave a good appreciation of the expression of the animal peptide transporter ECDs. The protocol proved very successful with *Mm*PepT1^{ECD}. A complete presentation of the crystallisation of *Mm*PepT1^{ECD} is given Chapter 4. The time taken from primers to crystals was 6 months. Given that no prior knowledge of the ECD existed

prior to this thesis, this was a good achievement and due credit goes to the advances made in *E. coli* heterologous expression systems.

The strength of the OPPF-UK pOPIN suite is that there is inbuilt versatility between expression systems due to the pTriEx vector. Fortunately the ECDs could both be expressed in *E. coli* but, if *E. coli* had not been applicable HEK293 and *Sf*9 insect cells could have both been attempted without extra cloning. However, there are weakness in the HT protocol. The use of 24-deep well blocks for the test expression meant that induction could not be tailored to the individual conditions. All 24 conditions were induced at the same time, irrespective of the individual OD_{600}. Therefore the assessment for all 192 conditions was not a fair test and typically meant that only the constructs which expressed well and were stable, were observed. This is not a major limitation as generally these are the conditions and constructs of interest. However, difficult proteins which are not stable and do not express well are likely to be over looked, and it is these proteins which are becoming more typical in biophysical and crystallographic analyses.

Another weakness of the OPPF-UK system is the potential for false positives. Large fusion tags, such as TF, are likely to pull through constructs into solution, even when the construct is unstable once cleaved. So time can be wasted trying to optimise purification protocols for constructs which are not stable. This was observed for the ECDs in the *H. sapiens* ECD constructs, which both expressed with a TF tag, but proved to be unstable once cleaved.

As it turned out, the pOPIN suite may not have been necessary for the ECDs, as both *Mm*PepT1^{ECD} and *Rn*PepT2^{ECD} expressed well in MBP tagged vectors such a pLou3. However, it was not possible *a priori*, to determine the system and conditions which would ultimately lead to expression of stable, folded protein; therefore, the use of the OPPF-UK system was a useful exercise.

4

Crystal structure of *M. musculus* PepT1^{ECD}

4.1 Summary

In this chapter, the crystallisation and X-ray structure determination of the *Mm*PepT1^{ECD} is explained. Crystallographic theory pertinent to the solving of the PepT1 and PepT2 ECD X-ray crystal structures is also discussed. The crystals used for data-collection were grown using hanging drop vapour diffusion against mother liquor containing 20% v/v PEG 6,000, 0.1 M MES pH 6.0, 0.2 M Ammonium Chloride. Crystals were cryo-protected in mother liquor containing 25% v/v glycerol. The phase problem was solved using the single wavelength anomalous dispersion method. Three datasets were used in the structure determination: N1 (native low), Hg-SAD and N2 (native high) processed to a resolution of 2.38, 2.84 and 2.10 Å respectively. The model was built into the electron density calculated from the 2.10 Å dataset to a final

R_{work} and R_{free} of 0.194 and 0.241 respectively. The model shows that the ECD has tertiary structure and that it forms a self contained extra-cellular domain of the native PepT1 transporter. The structure indicates that the ECD could have a functional role in the transport cycle.

4.2 Crystallisation

4.2.1 Sparse matrix Screening

The *Mm*PepT1^{ECD} (residues 391 – 580) protein, purified as described in § 3.4.1 and confirmed by mass-spectrometry, was put through a series of sparse matrix crystallisation screens (Table 2.13). These sparse matrix crystallisation experiments were prepared using a Microsys 4000 (Cartesian) crystallisation robot in 96-well sitting drop Greiner plates (§ 2.7.1).

Prior to crystallisation, the protein was passed through a Superdex S75 GF column (Figure 4.1) equilibrated in 20 mM Tris pH 7.5, 150 mM NaCl, to ensure a homogeneous mono-dispersed sample (D'Arcy 1994; McPherson 1990). The GF trace (Figure 4.1A) and SDS-PAGE gel of the peak fractions (Figure 4.1B) show that the protein was mono-dispersed and pure. Fractions at the centre of the GF peak were pooled and concentrated to 10 mg/mL. The protocol used for 96-well sitting drop crystallographic screening is described in § 2.7.1.

Sitting drop crystallisation vapour diffusion was first proposed by McPherson (1976). The layout of a single well from a 96-well sitting drop plate is shown in Figure 4.2. In a vapour diffusion experiment, a volume of precipitant is mixed with protein (Figure 4.2A.1.), placed inside a chamber containing a reservoir of precipitant (Figure 4.2A.3.), and sealed (Figure 4.2A.1.). The sealed chamber allows the equilibration of the chemical potentials for each component of the precipitant and protein buffer. As these potentials equilibrate, aqueous water, and volatile buffer components, evaporate and exchange between the well and the drop. Crystal growth can be viewed by a phase diagram (Figure 4.2B) (Chayen and Saridakis 2008). The exchange of solvent between drop and well slowly concentrates the protein (Figure 4.2B the marker arrow shows a gradual increase in protein concentration) into a supersaturated solution where the protein starts to precipitate in a highly

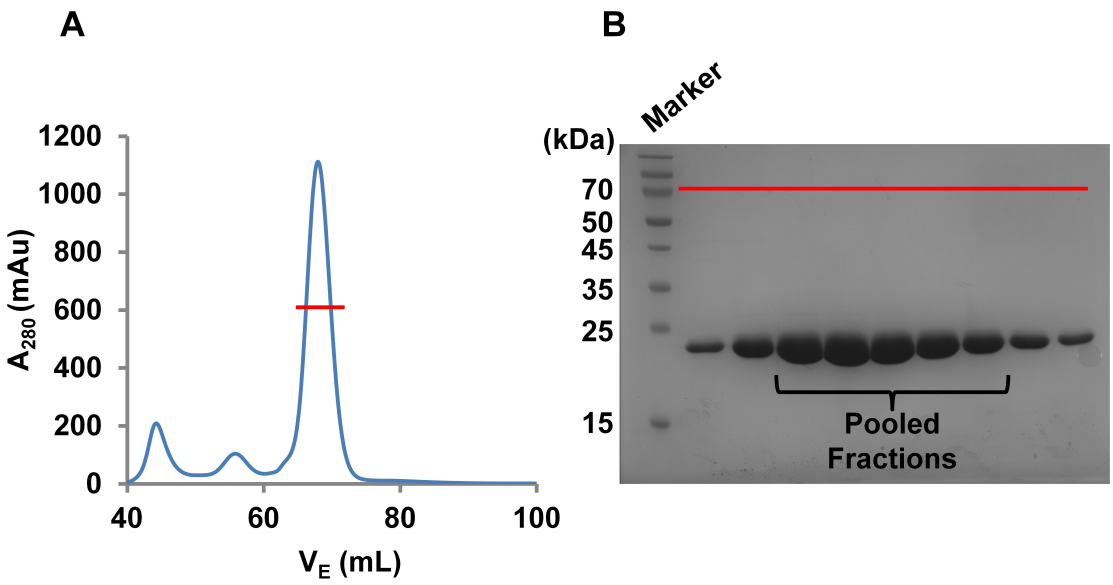


Figure 4.1: *Mm*PepT1^{ECD} GF trace and gel. (A) A Superdex S75 16/60 GF trace of *Mm*PepT1^{ECD} in 20 mM Tris pH 7.5, 150 mM NaCl. The red band indicates the fractions run on a reducing 15% Tris-Gly SDS-PAGE gel **(B)**.

ordered regular Three dimensional (3D) arrangement (nucleation) *i.e.* a crystal. As the protein concentration in solution falls the drop will enter a metastable zone where crystal growth is optimal.

A range of *Mm*PepT1^{ECD} concentrations were screened in the PEG/Ion matrix; 5.0, 7.5, 10.0, 15.0, 20.0 and 30.0 mg/mL. The optimum *Mm*PepT1^{ECD} concentration was found to be 10.0 mg/mL for PEG/Ion, which showed approximately 30% of drops exhibiting precipitation. This concentration was then used for the other screens. *Mm*PepT1^{ECD} gave an initial hit from Wizard^{TM} III+IV condition 82 (Emerald Biosciences); 20% v/v PEG 6,000, 0.1 M MES pH 6.0, 0.2 M Ammonium Chloride, at a protein concentration of 10 mg/mL, and 4 °C incubation temperature (Figure 4.3). Crystal nucleation was first observed on day 7 and continued for a further 31 days.

4.2.2 Crystal optimisation

The first step of crystal optimisation was to recreate the crystal hit. The same Wizard^{TM} III+IV screen was prepared using the same batch of protein used for the initial experiment. However, crystals were not observed in these repeat experiments.

Crystal optimisation was performed in 96-well sitting drop (Figure 4.4C) and 24-well

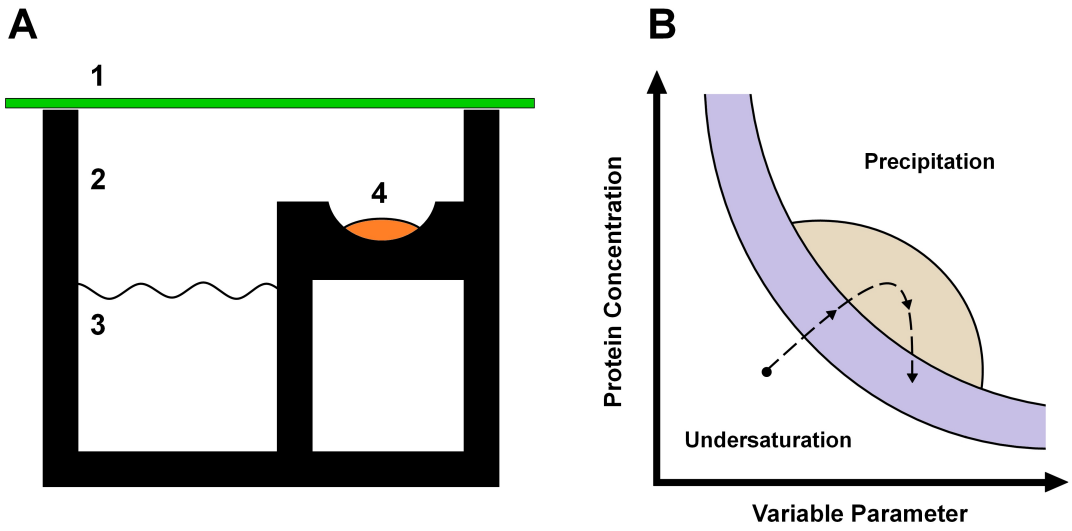


Figure 4.2: Sitting drop vapour diffusion crystallography. (A) A diagram of a sitting drop vapour diffusion crystallisation experiment. 1. Well seal. 2. Diffusion cavity. 3. Crystallisation precipitant. 4. Well stage containing a mixture of protein solution and precipitant. **(B)** A crystallisation phase diagram adapted from Chayen and Saridakis (2008). The dotted line shows the potential path of vapour diffusion drop from start to crystallisation. The beige area denotes the crystal nucleation zone. The purple region is the metastable zone where growth is likely to occur.

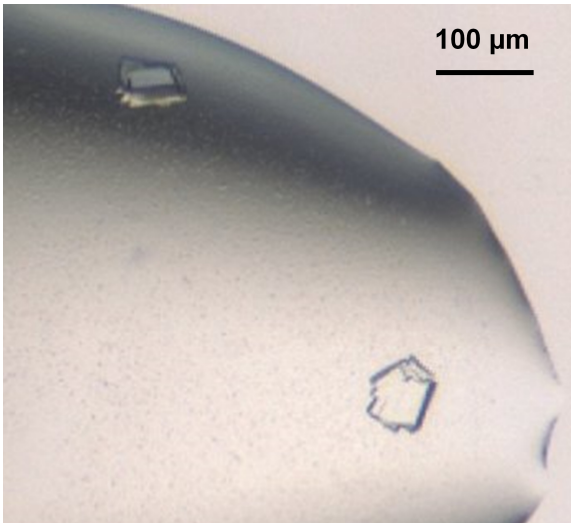


Figure 4.3: Crystal hit of *Mm*PepT1$^{\mathbf{ECD}}$. The hit grew in Wizard^{TM} III+IV condition 82 after 38 day incubation at 4 °C. The crystal had a box morphology of with approximate crystal dimensions of 100 × 100 × 25 µm.

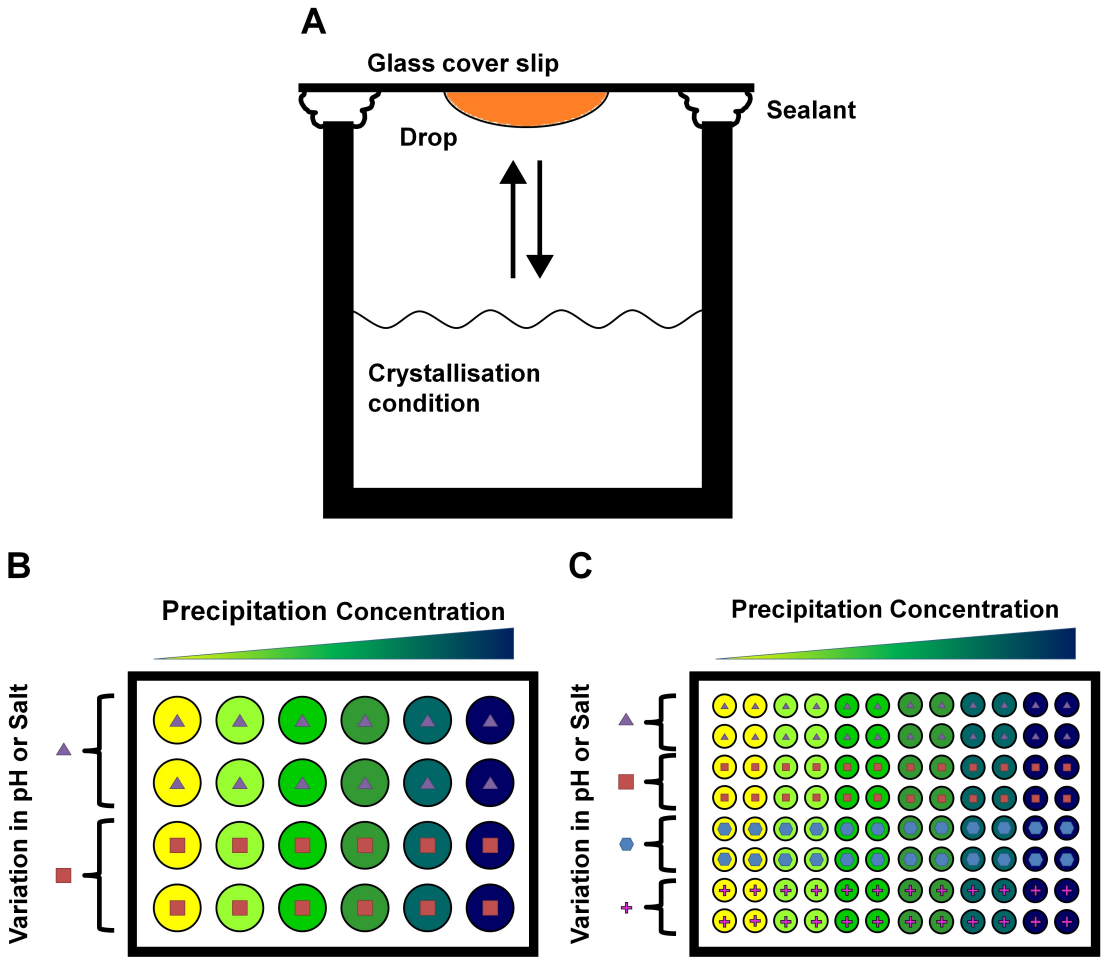


Figure 4.4: Crystal hit optimisation. **(A)** A diagram of a hanging drop vapour diffusion crystallisation experiment. The drop of protein and precipitant mixture is suspended from a glass cover slip which is sealed to the well with vacuum grease. Once sealed, the drop and condition equilibrate through vapour diffusion encouraging protein crystallisation. **(B)** and **(C)** show plans of 24 and 96-well crystal optimisation plates. Optimisation experiments were designed by increasing precipitant concentrations along adjacent columns and varying buffer, pH or salt concentrations down the rows.

hanging drop vapour diffusion plates (Figure 4.4B). The 24-well hanging drop plates use vapour diffusion, like the 96-well sitting drop (Figure 4.4A) but, due to the increase in well dimensions, the precipitant volume, the internal space and the drop volume can all be larger. These changes mean that the optimal conditions for crystal growth differ in the 96 and 24-well plates but, the advantage of the 24-well format is that larger crystals are more likely to grow.

In the 24-well optimisations a PEG 6,000 concentration range of 16 – 26% v/v with 2% increments was used. The buffer and salt from the Wizard^{TM} III+IV *Mm*PepT1^{ECD} hit were 0.1 M MES pH 6.0 and 0.2 M $(NH)_4Cl$ respectively. Therefore, different combinations of

0.1 or 0.2 M MES at pH's 5.0, 5.5, 6.0, 6.5 and 7.0, and either 0.1 or 0.2 M $(NH)_4Cl$ experimental conditions were prepared. The preparation of a 24-well crystallisation experiment is described in § 2.7.2.

The 96-well optimisations were prepared using a Mosquito® (ttplabtech) crystallisation robot in Swissci sitting drop plates. A similar protocol as detailed in the § 2.7.1 was used, with the exception that the mother liquor was pipetted manually and a 50 μL well volume was used. The practical reason for the change in robots and plates was that crystal screening was conducted at the OPPF-UK, and the optimisation was performed in Oxford Biochemistry. The OPPF-UK uses the Cartesian with Greiner plates and Oxford Biochemistry uses a Mosquito® with Swissci plates. The decision was made to do crystal screening experiments at the OPPF-UK, as it had 4 and 20 °C hotels which were not available in Oxford Biochemistry at that time. The screening process can be very long, involving many thousands of drops. Having an image record of every well in every experiment is a good way of keeping track of potential hits, and also seeing how the any crystals grow.

Initially it proved difficult to reproduce the crystal hit in either the 96 or 24-well plates. Since the crystal hit had grown at 4 °C, the precipitant and materials were equilibrated to 4 °C prior to the optimisation experiment. However, these experiments grew no crystals. Therefore it was decided to recreate the hit using bought reagents from Emerald Biosciences but this again ended in failure.

The optimisation drops were clear at PEG concentrations from 16 – 22%, at pH 5.5 and 6.0 in 0.1 M $(NH)_4Cl$. The bought condition drops also did not precipitate. When the experiments were prepared at 20 °C, precipitation was observed after 24 hr. The temperature, as has been shown for other proteins (Christopher, Phipps, and Gray 1998), changed the solubility of the *Mm*PepT1^{ECD} dramatically. Therefore to increase the nucleation, optimisation experiments in 24-well plates were prepared at 20 °C and moved to 4 °C to stabilise crystal growth.

The drops in these temperature experiments showed no obvious change from the other optimisation experiments. However, the wells with bought condition gave a wide range of results over a 24 hr 4 °C incubation; from amorphous precipitant to hundreds of micro crystals. The drops with the least nuclei were observed after very short 20 °C incuba-

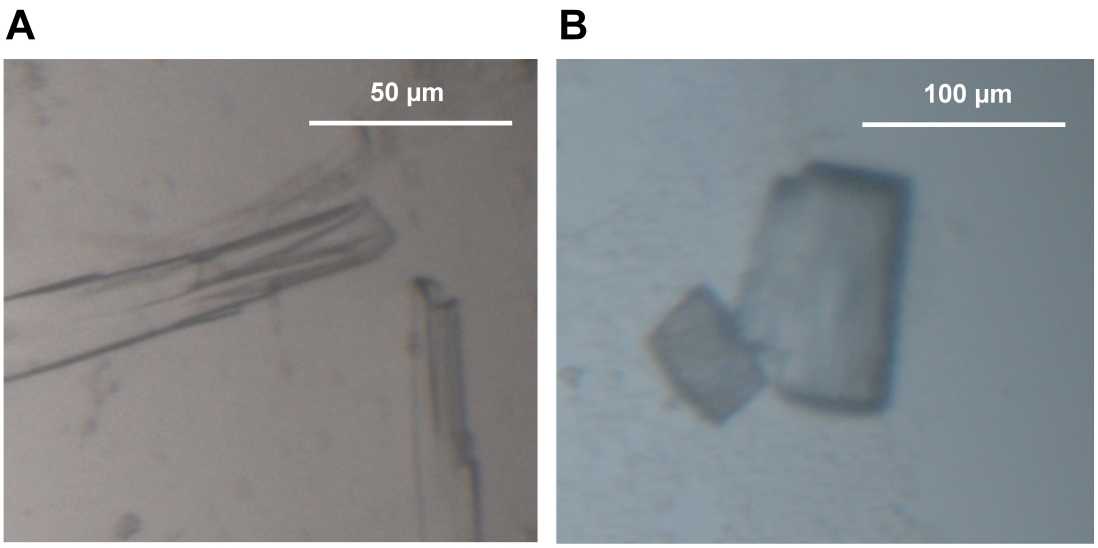


Figure 4.5: The two crystal forms of the optimised *Mm*PepT1^{ECD} crystals. **(A)** The rod morphology with cell dimensions of $25.0 \times 25.0 \times 75.0$ µm. **(B)** The box morphology with cell dimensions of $50.0 \times 100.0 \times 100.0$ µm.

tions (approximately 5 min). Generally, for 24-well plates, $1 - 3$ drops can be prepared simultaneously. This was increased to 6 drops to decrease the time the plates spent at 20 °C.

Occasionally when mixing a 24-well plate drop on the glass cover slip, an air bubble is introduced into the drop. In this event the bubble can be removed using a pipette with the loss of some protein and condition. It is good practice to spot this extra portion next to the drop. It was observed in these spots that the crystals were larger than their compatriots in the main drop. It was therefore decided to experiment with the ratio of protein solution to precipitant. Experiments were prepared with the bought condition at these protein to condition volume ratios (in µL); 0.50:1.00, 0.75:1.00, 1.00:0.50, and 1.00:0.75.

The change in drop ratio produced crystals large enough to be taken to the beamline. Crystals had a rod-like morphology with a protein solution to precipitant ratio of 0.75 µL:1.00 µL. The plate was prepared at 20 °C in a time of approximately 5 min. The plate was then transferred to 4 °C and crystal growth stopped after 24 hr. The resulting crystal had average dimensions of $25.0 \times 25.0 \times 75.0$ µm (Figure 4.5A).

It should be noted that the rod crystals could only be grown in the condition bought from Emerald Biosciences. When the precipitant solution was made in Oxford Biochemistry,

only crystals with a box-like morphology could be grown (Figure 4.5B). The diffraction of the box-like crystals was never beyond 3 Å, whereas the rod crystals diffracted to 1.5 Å. All crystals and diffraction experiments discussed further in this chapter had a rod morphology.

4.3 Data Collection

4.3.1 Bragg Diffraction

A protein crystal is a regular arrangement of protein molecules in 3D. The smallest translatable unit of this arrangement is termed the unit cell. As these units stack together in 3D, they create a 3D lattice in real space (xyz). Each atom in each protein molecule has a cloud of electrons with a density, ρ, the 3D crystal lattice is therefore a 3D array of equivalent points of electron density. As X-rays interact with the electron density of the crystal, diffraction ensues.

Figure 4.6A shows a photon scattered from two points in a crystal and the Bragg condition for constructive interference. The X-ray radiation is scattered by the whole crystal, although the majority of the radiation simply passes straight through the crystal. The incident wave $\mathbf{S_0}$ is scattered by the points of electron density into the scattered wave $\mathbf{S_1}$. Constructive interference will only occur when $n\lambda = 2d\sin\theta$ and $n \in \mathbb{N}$ (Bragg condition).

The directions of constructive interference can be visualised geometrically using the Ewald sphere construction (Figure 4.6B). A section of a cubic reciprocal lattice associated with a cubic crystal is shown in green. The reciprocal lattice (hkl) is a mathematical construction to predict and catalogue the directions of constructive interference. The scattering vector $\mathbf{S}$ has a magnitude of $2\sin\theta/\lambda$ when the radius of the Ewald sphere is $1/\lambda$, and the reciprocal lattice vector intersects the circumference of the sphere. This is very useful construction as it determines which directions will originate diffracted rays as the crystal is rotated.

The diffraction pattern is related to the real space electron density by the Fourier transform of the cell (Equation 4.1a). Each structure factor ($\mathbf{F_{hkl}}$) (Equation 4.1b) is the resultant diffracted wave from the entire crystal and is defined by an amplitude ($|\mathbf{F_{hkl}}|$) and phase (φ).

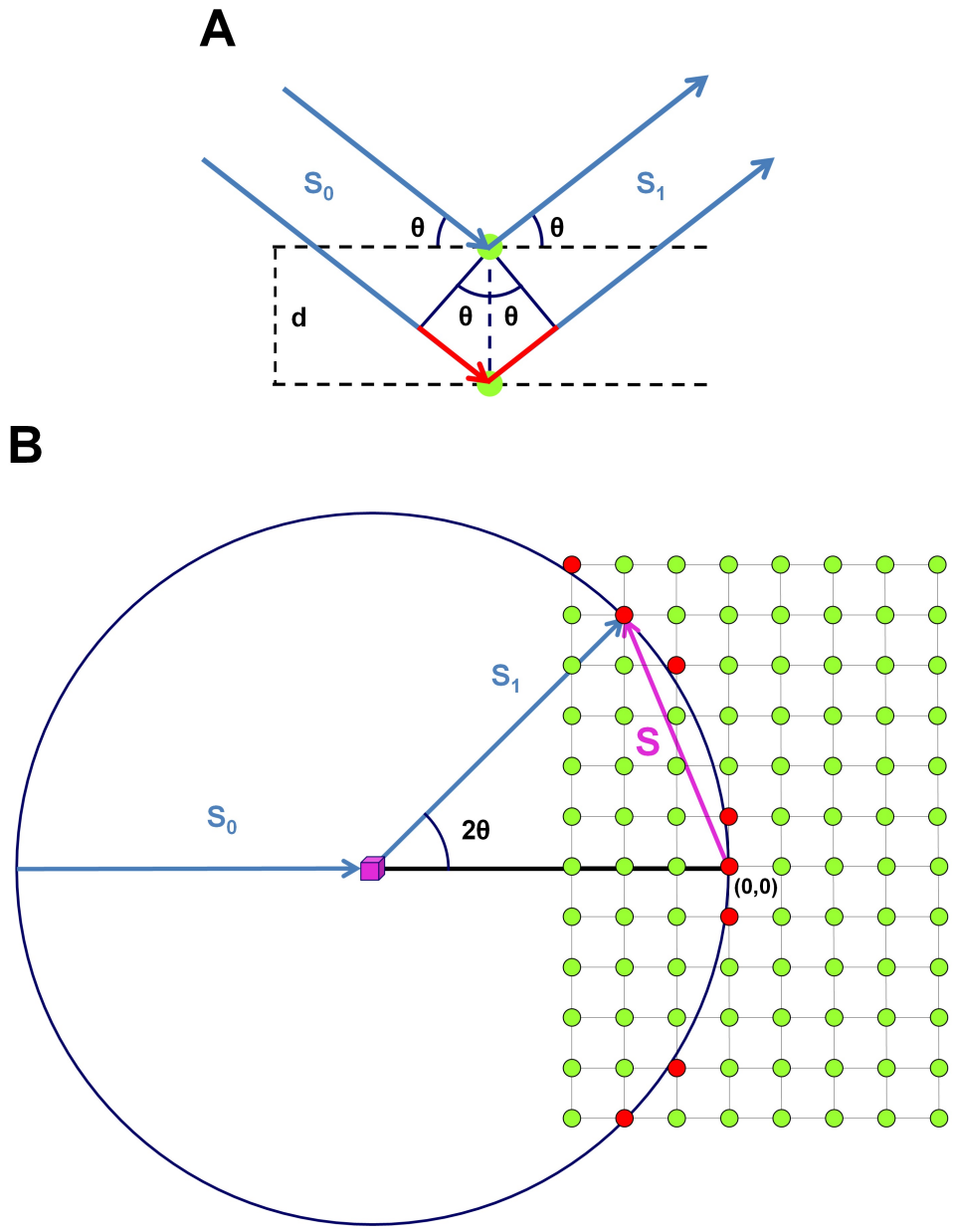


Figure 4.6: Bragg diffraction. (A) A diagram showing an incident wave $\mathbf{S_0}$ interacting with two points of electron density (shown as green circles) separated by a distance d. $\mathbf{S_1}$ is the resultant wave. The wave vector highlighted in red shows the extra distance travelled by the bottom wave with respect to the top. **(B)** The Ewald sphere. A geometric construction which allows the prediction of the lattice points which fulfil the Bragg condition and therefore leads to the constructive interference of rays. The purple crystal, in this case is cubic, which produces a cubic reciprocal lattice shown as a network of green dots. Red dots indicate reciprocal lattice points which touch the surface of the Ewald sphere of radius $1/\lambda$.

$$\mathbf{F}_{\boldsymbol{hkl}} = \frac{1}{V}\int_0^1\int_0^1\int_0^1 \rho_{(x,y,z)} \cdot e^{(2\pi i(hx+ky+lz))}\,\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{4.1a}$$

$$\mathbf{F}_{\boldsymbol{hkl}} = |\mathbf{F}_{\boldsymbol{hkl}}| \cdot e^{(i\varphi_{(h,k,l)})} \tag{4.1b}$$

The solving of an X-ray crystallographic structure involves taking the inverse Fourier transform of Equation 4.1a. A series of $|\mathbf{F}_{\boldsymbol{hkl}}|$ and their associated phases are assigned to each $\mathbf{F}_{\boldsymbol{hkl}}$. These $\mathbf{F}_{\boldsymbol{hkl}}$ are summed over the hkl volume, and a map of the electron density is calculated in real space. A protein structure is built into this density. The accuracy of that model is measured, by the agreement of the model to prior knowledge of protein structure, and errors associated with the X-ray crystallographic data collection and processing.

4.3.2 Radiation damage and cryo-protection

Radiation damage is a significant problem during X-ray crystallographic data collection and is principally caused by the non-elastic scattering of X-rays and the absorption of photons by the crystal (Holton 2009). Although both elastic and non-elastic scattering deposit energy into the crystal, the absorption of photons by the crystal results in the breakage of covalent bonds and formation of free radical species (Sliz, Harrison, and Rosenbaum 2003). Once created these free radicals diffuse throughout the crystal causing a chain reaction of further damage. Damage to the crystal can be seen in specific effects, such as the breakage of di-sulphide bonds, and in global effects, such as unit cell shifting.

To minimise the effects of radiation damage, data-collection is generally undertaken at −170 °C. Since protein crystals are filled with aqueous solvent, cooling the crystal to −170 °C results in the formation of ice. Ice formation inside the crystal destroys the integrity of the crystal and causes protein precipitation. To avoid ice formation crystals are cryo-protected and flash cooled in either liquid nitrogen or cold nitrogen gas.

To assess the efficacy of different cryo-protectants, the first crystals taken to the beamline were either cryo-protected in 20, 25% v/v glycerol or 30% v/v PEG 400. The 20% v/v glycerol crystals showed some browning, indicative of inadequate cryo-protection. No obvious difference was observed between the 25% v/v glycerol and 30% v/v PEG 400, so

25% v/v glycerol was used for subsequent data-collections. Significant radiation damage was not observed during the processing of the data.

4.3.3 Indexing for data collection

A crystal can be thought of as unit regularly translated in 3D. The principal pieces of information to gather at the start of an experiment are the dimensions and orientation of this unit cell. Seven types of unit cells are possible, giving rise to 14 Bravais lattices. The process for calculating the dimensions and orientation of the unit cell, and thereby assigning each reflection to a reciprocal lattice point, is called indexing. A complete index of all spots is performed after a dataset is collected. However, knowledge of the unit cell vectors, crystal orientation and mosaicity, a measure of the alignment of adjacent unit cells with respect to each other, is required in order to collect a complete dataset.

Figure 4.7 shows the diffraction observed from a *Mm*PepT1^{ECD} crystal. The fine spacing of the lunes of Bragg maxima indicates a relatively large unit cell, which is indicative of protein rather than of a small molecule. A combination of `iMosflm` (Powell, Johnson, and Leslie 2013) and `EDNA` (Incardona *et al.* 2009) were used to calculate the probable unit cell and orientation of the crystal. `iMosflm` has an extremely efficient algorithm for calculating the probable unit cell based on a One dimensional (1D) Fast Fourier transform (FFT) (Powell 1999), and proposes the most likely crystal orientation and unit cell vectors based on the 14 different Bravais lattices, and the smallest associated penalty. The penalty is calculated for each possible solution from the disagreement between the spot predictions and observations.

`EDNA` is a wrapper for `Mosflm`, `BEST` (Bourenkov and Popov 2010) and `RADDOSE` (Murray, Garman, and Ravelli 2004; Paithankar, Owen, and Garman 2009) which are combined to calculate the optimal data collection parameters for a given crystallographic experiment. `EDNA` uses `Mosflm` to calculate the probable unit cell parameters and the mosaicity. `BEST` and `RADDOSE` are then used to calculate the ideal beam flux, exposure time, $\Delta\phi$ per image, and the maximum resolution to give the highest Signal-to-noise ratio (SNR) per reflection without extensively damaging the crystal.

The majority of data collection experiments are performed upon crystals where the exper-

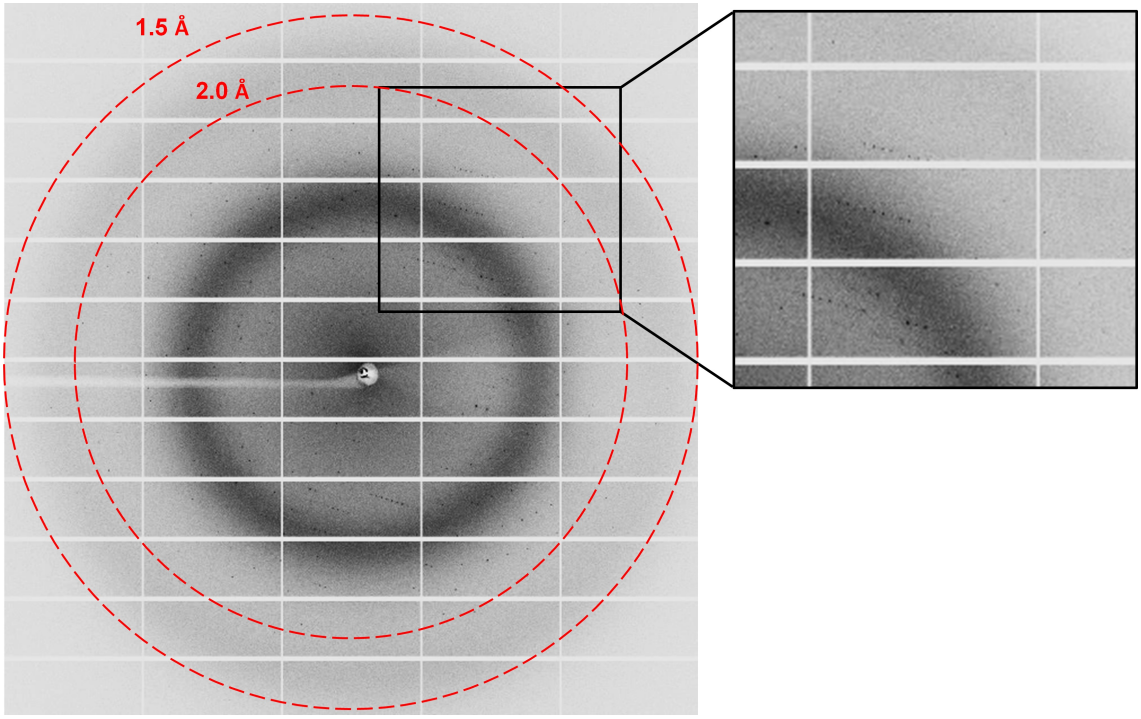


Figure 4.7: An image of the diffraction from a *Mm*PepT1^{ECD} crystal. This image was taken at a detector distance of 275.26 mm, 0.2 ° oscillation, λ 0.968 nm and 0.2 s exposure. The 1.5 and 2.0 Å resolution shells are shown in red.

imenter has some knowledge of the unit cell parameters, and how the crystals behave in the X-ray beam. In the case of unknown crystals good estimates of the unit cell, crystal orientation and mosaicity are essential for a successful data collection.

The indexing indicated that the most probable Bravais lattice was primitive orthorhombic (Figure 4.8) with unit cell dimensions a, b, and c, of 52.8, 70.0 and 110.8 Å respectively at 90 °. As stated above, the unit cell defines the minimal unit from which the crystal can be generated purely by translation. However, inside that unit cell, rotational symmetries are possible, which therefore allow for a simplification of the unit cell into an Asymmetric unit (ASU). Orthorhombic lattices have three non-equivalent perpendicular 2-fold axes. These 2-fold axes can be screw. The possible combinations of 2-fold or 2-fold screw axes in a primitive orthorhombic unit cell are either P222 (no screw), P$2_1$22 (one screw axis in a, b, or c), P$2_1 2_1 2$ (two screw axis in a, b, or c) or P$2_1 2_1 2_1$; the '$_1$' indicates a translation of half the axis distance. These different ways of constructing the unit cell from the ASU are called space-groups. It is not possible to determine the correct space-group without determining the content of the unit cell. However, at this stage, an unambiguous assignment of the space-group is not necessary. A dataset of an orthorhombic lattice requires at least a 90

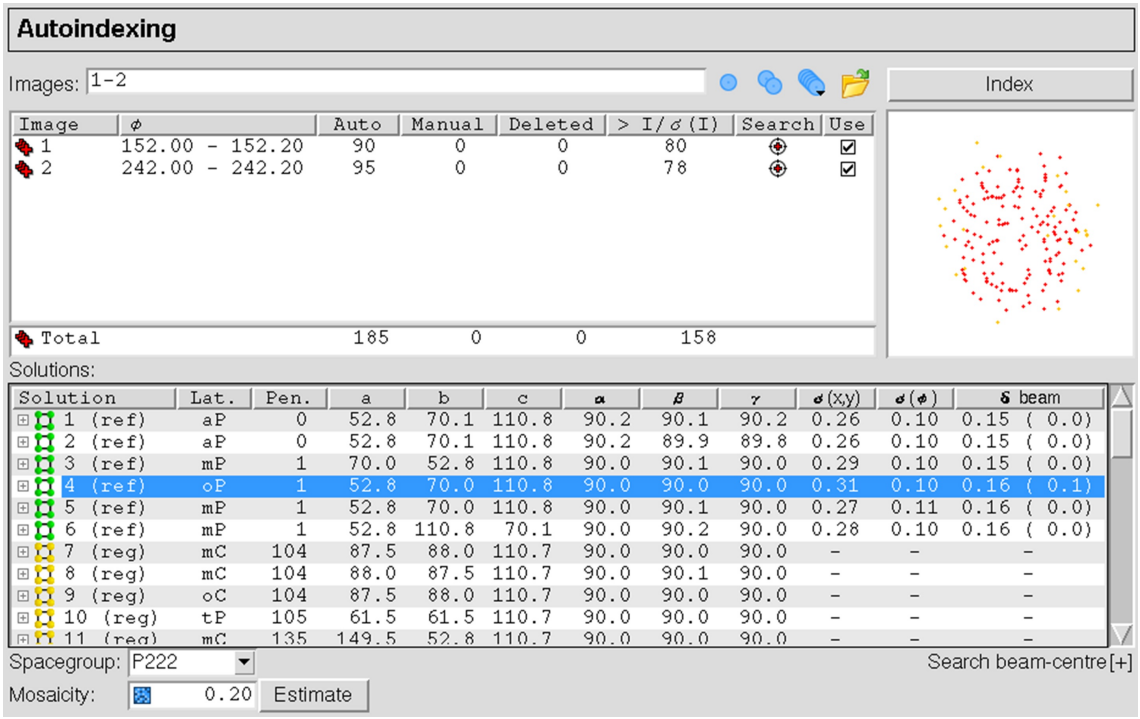


Figure 4.8: A screen shot created by `iMosflm` (Powell, Johnson, and Leslie 2013) using two images from a *Mm*PepT1^{ECD} crystal taken 90 ° apart. The solution with highest symmetry and lowest penalty has been highlighted.

° ϕ wedge to sample the entire reciprocal space.

4.3.4 Data-collection Parameters

Diffraction data from *Mm*PepT1^{ECD} crystals were collected at Diamond Light Source beamlines I04 and I24. Three datasets were used to solve the structure of the *Mm*PepT1^{ECD}; N1 (native low resolution), Hg-SAD which allowed the phase calculation (full details in § 4.6) and N2 (native high resolution). Full data collection parameters are shown in Table 4.1. The N2 dataset was collected to improve the resolution. The first wedge (80 ° $\leq \phi \leq$ 180 °) unfortunately did not contain a complete dataset. Therefore a second wedge was collected to ensure the entire reciprocal lattice was sampled (discussed in § 4.9).

4.3.5 Detectors

N1 and N2 were collected using a Pilatus 6M-F (Dectris) hybrid-pixel detector. Hg-SAD was collected using an Quantum 315 (ADSC) Charged-coupled device (CCD) detector. The practical difference between the two detectors is the type of data-collection strategy

	Dataset		
	N1	**Hg-SAD**	**N2**
Experiment type	Native (low)	Hg (SAD)	Native (high)
Date	06/10/2011	04/11/2011	21/03/2012
Beamline	I24	I04	I24
λ (Å)	0.968	1.0069	0.9686
Temperature (°C)	−170	−170	−170
ϕ (°)	0 – 120	320 – 319	0 – 30, 80 – 180
$\Delta\phi$ (°)	0.2	1.0	0.2
Exposure time (s)	0.2	1.0	0.2
Transmission (%)	40	50	40
Detector Distance (mm)	369.65	281.71	298.49
Detector	Pilatus 6M-F	Quantum 315	Pilatus 6M-F

Table 4.1: Data collection parameters from the data collection of *Mm*PepT1$^{\mathbf{ECD}}$ crystal datasets. Complete crystallographic tables are shown in Table A.1.

favoured. CCD detectors convert the incoming X-ray photons into visible light using a phosphor. This visible light is then focused onto the CCD plate which converts the visible light into a voltage proportional to the intensity of the original X-ray photon. A hybrid-pixel detector can directly convert the X-ray photons into a proportional voltage (Helliwell 2005).

The properties of each detector dictate a different experimental strategy. The phosphor photon conversion in the CCD detector requires a shutter between images and has a readout time in ms. Due to this data acquisition time, experiments are usually conducted with a $\Delta\phi \approx 1.0\,^{\circ}$. Bragg candidates are generally sampled as complete reflections as the goniometer rotates through the exposure. The hybrid-pixel detector was designed to allow for fine ϕ-slicing ($0.05\,^{\circ} \leq \Delta\phi \leq 0.5\,^{\circ}$) diffraction experiments. Bragg candidates tend to only be partially sampled in each image and a complete spot profile is constructed through successive images. However, it should be noted that the sampling of Bragg candidates as either full or partial reflection is dependent upon the mosaicity of the crystal.

4.4 Data reduction

Data reduction is the general name given to the processes of indexing, integrating, scaling and merging. Indexing, as described in § 4.3.3, is the assignment of each reflection to its reciprocal lattice point based on the crystal unit cell dimensions and orientation. Inte-

grating involves the calculation of the reflection intensity at each reciprocal lattice point. Scaling attempts to take into account relative differences between images in the dataset, and reduces all intensities to a common scale, which normalises intensity measurements between different times of the experiment. Finally merging is the comparison of symmetry equivalent reflections and assessment of the quality of a dataset.

N1 and Hg-SAD were processed using the automated data-processing software `xia2 -3dii` (Winter 2009). `xia2 -3dii` utilised `Distl`, `Labelit`, `XDS`, `Pointless`, `Scala` and the CCP4 suite (Zhang *et al.* 2006; Sauter, Grosse-Kunstleve, and Adams 2004; Kabsch 2010; Evans 2006; Winn *et al.* 2011) to index, integrate, scale and merge the data into space-group $P2_12_12_1$. `xia2` has a number of different options depending on the software packages to be used. The steps that follow are discussed in the context of `xia2 -3dii`.

4.4.1 Indexing

`XDS` is a suite of programmes designed to index, integrate and scale a dataset. The indexing difference vector based algorithm (Kabsch 1993) is less efficient than the FFT employed by `Mosflm` but equally accurate. Difference vectors from pairs of neighbouring reflections are plotted on a histogram. A solution which best explains the difference between clusters is found and refined. The unit cell constants are then refined against the 65 enantiomorphic space-groups for macromolecular protein crystals, and the solution with the highest symmetry is selected. For *Mm*PepT1^{ECD}, the `XDS` assessment agreed with the initial `Mosflm` calculation and the inflections were assigned to orthorhombic point group symmetries and a primitive lattice.

In using this method, it is good practice to include all reflections in the histogram which means using the entire dataset for the calculation and hence a reduced efficiency compared to `Mosflm`. However, given modern processing power, this is not a limitation. There are potential issues with selecting the Bravais lattice with the highest symmetry. Crystals with pseudo translations, twinning or monoclinic angles close to 90 ° can give rise to intensities which appear to belong to lattices with higher symmetry. However, at this stage of the data processing the highest symmetry compatible with the data is a reasonable assumption.

4.4.2 Integration

The principal reason in using `XDS` in preference to `Mosflm` for indexing and subsequent integration is that `XDS` uses 3D profile fitting for reflection integration (Mueller, Wang, and Schulze-Briese 2012). This is particularly necessary for fine ϕ-slicing experiments. The reflection is viewed as a 3D Gaussian and therefore a more accurate determination of the reflection intensity can be calculated. The reflection background is assumed as the area surrounding the Two dimensional (2D) spot profile on the detector and is subtracted from the calculated intensity.

4.4.3 Scaling

Scaling in `xia2 -3dii` is performed by `XDS Xscale`; `Scala` and now `Aimless` utilises a similar method as the one outlined for `Xscale` (Kabsch 2010). Scaling is performed in two steps. The first attempts to account for experimental differences across the dataset. The second brings the entire dataset to a global common scale and requires the calculation of an isotropic B factor (B_{iso}) and associated scale factor ($1/k$) (1 per frame). B_{iso} is a measure of the falloff in intensity as a function of resolution. The B_{iso} and $1/k$ are approximated by Equation 4.2. The dataset is divided into resolution bins and plotted against $\ln(\bar{I}/\bar{f}^2)$ where $\bar{f}^2$ is the intensity of the average protein. Figure 4.9A shows a Wilson temperature factor plot of a *Mm*PepT1^{ECD} crystal. The slope and y-intercept of a linear regression line is a reasonable approximation of B_{iso} and $1/k$ respectively. $1/k$ is an powerful approximation reducing datasets from different crystals, collected using different experimental conditions, to the same relative scale.

$$\ln\left(\frac{\bar{I}}{\bar{f}^2}\right) = \ln k - \frac{2B_{iso}\sin^2\theta}{\lambda^2} \tag{4.2}$$

4.4.4 Merging

Merging a dataset requires the assignment of the integrated and scaled intensities to a particular space-group. It is not possible to correctly assign the space-group until the phases have been estimated. `xia2` uses `Pointless` (Evans 2006) to make a reasonable

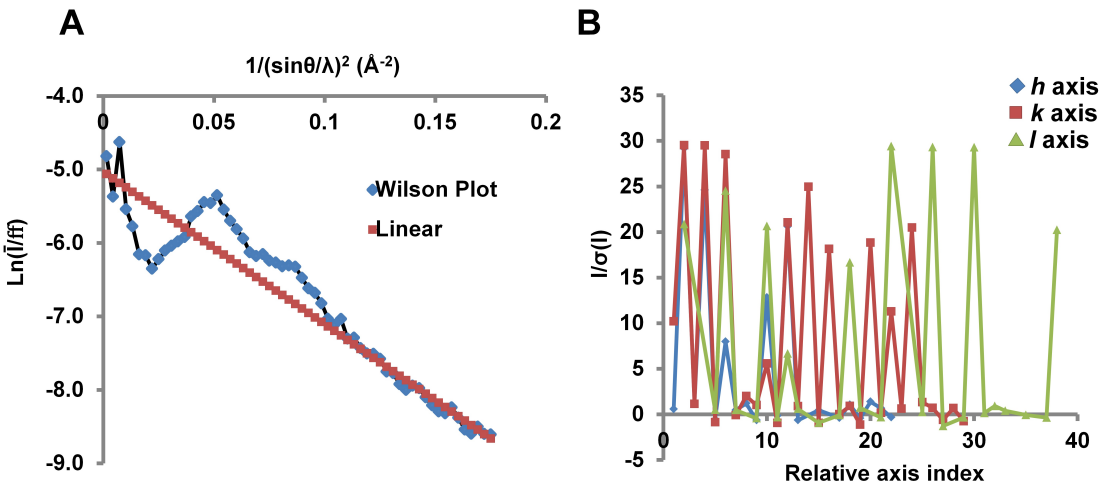


Figure 4.9: Temperature factor and space-group assignment for the *Mm*PepT1^{ECD} crystal. **(A)** A Wilson temperature factor plot generated by `CTruncate` (Winn *et al.* 2011). B_{iso} and $1/k$ were calculated to be 41.5 Å and 148.4 respectively. **(B)** A plot of $I/\sigma(I)$ against the h, k, and l indices generated by `Pointless` (Evans 2006). A 2-fold screw axis shows zero intensity every odd index.

judgement on the evidence for a particular space-group. The Bravais lattice, assuming no twinning or pseudo symmetry, should be correct. Therefore *Mm*PepT1^{ECD} crystals should belong to either P222, P222$_1$, P2$_1$2$_1$2 or P2$_1$2$_1$2$_1$ (given the a, b, and c assignments). The presence of a screw axis can be detected as a series of systematic reflection absences about an axis. Figure 4.9B shows an $I/\sigma(I)$ plot against the h, k and l axes indices. Absences are observed every odd index. Therefore it is likely that all three axes have a 2-fold skew axis, so P2$_1$2$_1$2$_1$ was selected by `xia2` for merging.

Once a space-group is selected, the data is merged using `Scala` (Evans 2006) or `Aimless` (Evans and Murshudov 2013), and `CTruncate` (Winn *et al.* 2011) converts the intensities into $|F_{hkl}|$. Statistical analysis is performed to assess the success of the data reduction and the overall quality of the dataset.

The power of `xia2` is that for non-complicated datasets it does a comparable job to the manual processing done by the experimenter in 1 – 5 min depending on the software used. Beamlines such as the ones at Diamond Light Source have an automated `xia2` pipeline built into the data-collection software. As the next crystal is being mounted onto the beamline, statistics are available on the last dataset collected. This gives the modern experimenter an enormous advantage during data collection, as they can quickly determine how a dataset processes under standard conditions and decide if another needs to be taken.

However, it is essential that the output metrics are correctly assessed. Table 4.2 shows the data processing statistics for the three datasets used to solve the structure.

The resolution limit was determined by examination of the $\langle |I|/\sigma(I) \rangle$, R_{meas} and $CC^{1}/_{2}$. The $\langle |I|/\sigma(I) \rangle$ is effectively the signal-to-noise (Equation 4.3a). R_{meas} is based on R_{merge} (Arndt, Crowther, and Mallett 1968), which is a measure of the variation in intensity observations of a given reflection and its symmetry equivalents. However, R_{merge} does not account for the multiplicity of the dataset (Diederichs and Karplus 1997), therefore R_{meas} was proposed to take this into account (Equation 4.3b). $CC^{1}/_{2}$ was proposed to give a resolution limit for the presence of useful information in the dataset (Karplus and Diederichs 2012) and is the linear correlation coefficient between reflections in random half datasets (Evans 2006; Evans and Murshudov 2013).

$$\langle |I|/\sigma(I) \rangle = \frac{1}{N} \sum_{\mathbf{h}}^{N} \frac{|I_{(\mathbf{h})}|}{\sigma(I_{(\mathbf{h})})} \tag{4.3a}$$

$$R_{meas} = \frac{\sum_{\mathbf{h}} \left(\frac{N}{N-1}\right)^{1/2} \sum_{i=1}^{N} |I_{(i\mathbf{h})} - \bar{I}_{(\mathbf{h})}|}{\sum_{\mathbf{h}} \sum_{i=1}^{N} I_{(i\mathbf{h})}} \tag{4.3b}$$

There are still no definitive rules for the resolution limit of a dataset other than the detector (Evans and Murshudov 2013). However, a general agreement of an $\langle |I|/\sigma(I) \rangle \cong$ 2.0, $R_{meas} \cong$ 1.0 and $CC^{1}/_{2} \cong 0.5$ was used as a cut-off for useful information. This was achieved by `xia2` for the N1 and Hg-SAD dataset. It should be noted here that, since the development of Maximum-likelihood (ML) refinement programs (discussed in § 4.7), which weight the errors associated with the model during refinement, there is a strong argument for not excluding any intensity measurement regardless of the error associated with it.

The multiplicity is the average number of observations per unique reflection. This was mainly of consideration for the Hg-SAD dataset, as a high multiplicity was essential for an accurate calculation of anomalous difference between the Friedel pairs (explained fully in § 4.6).

Completeness is the percentage of the reciprocal lattice which has been sampled. This was of particular importance for the dataset N2. Due to an error during data collection,

only 80% of the reciprocal lattice was sampled. Therefore the N2 dataset is a combination of two wedges $80^\circ \leq \phi \leq 180^\circ$ and $0^\circ \leq \phi \leq 30^\circ$. These were separately indexed and integrated using `XDS`, and scaled and merged together in `Aimless` (discussed in § 4.9).

		N1	Hg-SAD	N2
Unit Cell	a, b, c (Å)	53.07, 70.41, 111.30	53.55, 70.37, 111.60	53.45, 70.12, 111.19
Space-group $P2_12_12_1$	α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution Range (Å)		70.41 – 2.37 (2.47 – 2.37)	70.37 – 2.85 (3.05 – 2.85)	43.62 – 2.10 (2.16 – 2.10)
Measured Reflections		74130 (8384)	144862 (26394)	119801 (9797)
Unique Reflections		17446 (1929)	10365 (1823)	25021 (2027)
Multiplicity		4.2 (4.3)	14.0 (14.5)	4.8 (4.8)
Mosaicity (°)		0.114	0.14	0.09
Average B ($Å^2$)		56.9	61.43	43.84
Overall $\langle \lvert I \rvert / \sigma(I) \rangle$		12.8 (2.2)	15.1 (4.1)	15.9 (2.7)
Overall Completeness (%)		99.4 (99.7)	100.0 (100.0)	99.5 (99.0)
R_{merge}		0.072 (0.736)	0.172 (0.756)	0.050 (0.798)
R_{meas}		0.092 (0.946)	0.178 (0.782)	0.062 (0.988)
$CC^1/_2$		0.998 (0.737)	0.997 (0.910)	0.999 (0.480)

Table 4.2: Data processing statistics of the *Mm*PepT1^{ECD} datasets. The values in parentheses are for the highest resolution shell and complete crystallographic tables are shown in Table A.1.

Mol/ASU	V_m (Å^3/Da)	Solvent (%)	P|2.38 Å
1	5.00	75.43	0.01
2	2.50	50.87	0.99
3	1.67	26.30	0.00
4	1.25	1.73	0.00

Table 4.3: A table showing the V_m probability distribution for dataset N1. Generated by the CCP4 program `Matthews_coef` (Winn *et al.* 2011).

4.5 Asymmetric unit analysis

The unit cell is the minimum translational unit inside the crystal. However, because of potential rotational symmetries present in the unit cell, it might not be the smallest unit of unique information. In the case of *Mm*PepT1^{ECD}, the likely space-group is $P2_12_12_1$, this means there are 4 copies of a unique ASU related by rotational screw axes.

Prior to attempting to calculate the phases of the $|\mathbf{F}_{hkl}|$, it is possible to infer some information about the content of the ASU. Given the unit cell dimensions, the space-group, the resolution and the size of the protein, a reasonable estimate of the number of monomers and solvent content of the ASU can be made (Matthews 1968; Kantardjieff and Rupp 2003). The CCP4 program `Matthews_coef` (Winn *et al.* 2011) was used to calculate the probability distribution for the Matthews coefficient (V_m) as a function of solvent content. The results are shown in Table 4.3. The mode of the V_m distribution for structures with a solvent content of 47% (Kantardjieff and Rupp 2003) solved up till 2013 is 2.34 Å^3/Da. Therefore it is a reasonable conclusion that two monomers of the *Mm*PepT1^{ECD} exist in the ASU.

The direction of the Non-crystallographic symmetry (NCS) 2-fold axis relating the monomers in the ASU can also potentially be viewed in a Patterson map Self-rotation function (SRF). The Patterson map (P_{uvw}) is a inverse Fourier transform of the $|\mathbf{F}_{hkl}|^2$ over Patterson space (uvw) (Equation 4.4). This is a convolution of the reciprocal lattice with itself and generates an inter-atomic vector map in the space of the difference vectors uvw. Since this map shows the frequency of inter-actomic vectors inside the ASU, it is also possible to directly calculate the atomic positions in real space (direct methods see § 4.6.1). It is generally not feasible to interpret protein Patterson maps due to the large number of atoms in the protein and therefore the large number of peaks in the Patterson (for a molecule of N

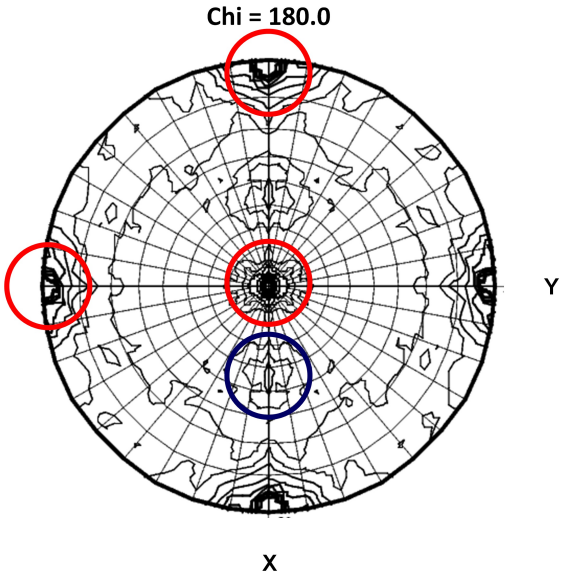


Figure 4.10: The $\chi = 180°$ SRF section of the *Mm*PepT1^{ECD} dataset N1. This image was generated in MOLREP (Vagin and Teplyakov 1997). The 3 peaks which indicate the orthorhombic 2-fold axis are shown by red circle. The peak which potentially shows a 2-fold rotational axis in the ASU is highlighted by a blue circle.

atoms, $N(N-1)$ peaks in the Patterson). However, rotational symmetries are still visible in the Patterson. The SRF identifies rotational angles which produce coincidental peaks in the Patterson about the origin. These coincidental peaks should identify any axis with rotational symmetry and potentially any non-crystallographic rotational symmetry in the ASU.

$$\mathrm{P}_{uvw} = \frac{2}{V} \sum_{h} \sum_{k} \sum_{l} |\mathbf{F}_{\boldsymbol{hkl}}|^2 \cos 2\pi(hu + kv + lw) \tag{4.4}$$

Figure 4.10 shows the $\chi = 180°$ section of the N1 dataset SRF. An orthorhombic crystal has 3 orthogonal 2-fold axes, therefore 3 large peaks are visible in the SRF which correspond to the crystal axis (Figure 4.10 red circles). A smaller peak is also highlighted by a blue circle. This could be noise from the 2-fold related vectors along the x-axis, or it could be an indication of NCS relating the contents of the ASU. A 2-fold axis in the ASU is supporting evidence for the existence of two *Mm*PepT1^{ECD} monomers in the ASU.

4.6 Experimental Phasing

The equation which allows the reconstitution of the ASU electron density for a crystal is given in Equation 4.5. This is the inverse Fourier transform of the $\mathbf{F}_{hkl}$ equation given in Equation 4.1a. The equation states that the sum of the $|\mathbf{F}_{hkl}|$ with their associated phases (φ) over their reciprocal space (h, k, l) indices will construct the electron density of the atoms (x, y, z) in the crystal. The reciprocal space indices catalogue the $|\mathbf{F}_{hkl}|$ which are the square roots of the collected intensities. However, the phases must be estimated and refined using other information.

$$\rho = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |\mathbf{F}_{hkl}| \cdot e^{(i\varphi_{(h,k,l)})} \cdot e^{(-2\pi i(hx+ky+lz))} \tag{4.5}$$

4.6.1 Phase calculation

Analysis of the content of the ASU is useful when starting to calculate the structure factor phases. Phase information is lost during the data collection due to the impracticability of constructing a high resolution X-ray lens. Currently the only methods which exist to solve the phase problem involve knowledge of the ASU content. The most prevalent of these, in the era of a PDB with over 100,000 submitted structures, is Molecular replacement (MR) (Scapin 2013). The MR method involves creating an ASU search model based on a known homologous structure with a sequence identity of ≥ 25 %. This was not possible when solving the structure of the *Mm*PepT1^{ECD} as no close homologous structures were known. MR is discussed more fully in the structure determination for the PepT2^{ECD} in § 5.10.3.

For very high resolution (≤ 1.2 Å) small protein structures (≤ 200 residues), it is possible to directly (direct methods) calculate the relative atomic positions from the Patterson map, as the inter-atomic vectors can be logically interpreted; although it should be noted the first direct methods programs did not use Patterson maps in the calculation. The *Mm*PepT1^{ECD} crystals diffracted to 2.1 Å; therefore the resolution is not high enough to produce a readily interpretable Patterson map. The ultimate goal of experimental phasing methods is to produce a simplified ASU and therefore a simplified Patterson map which can be interpreted.

4.6.2 Isomorphous replacement

There are now a number of methods to experimentally phase macromolecular protein structures and all involve the incorporation of Heavy atom (HA) bound into a protein crystal, or the utilisation of HAs already inside the protein structure. The first method to be developed was isomorphous replacement which is outlined in this section. The premise behind isomorphous replacement is shown in Equation 4.6, where $\mathbf{F}_{\mathrm{P}}$, $\mathbf{F}_{\mathrm{PA}}$ and $\mathbf{F}_{\mathrm{A}}$ are the structure factors of the protein crystal, the protein crystal with bound HAs and the HA crystal sub-structure respectively.

$$\mathbf{F}_{\mathrm{PA}} = \mathbf{F}_{\mathrm{P}} + \mathbf{F}_{\mathrm{A}} \tag{4.6}$$

HAs have a large number of electrons which means they scatter X-rays significantly more than the generally light atoms (C, O, N, S and H) present in protein. In a HA phasing experiment, a relatively small number of HAs are incorporated into a protein crystal ($\mathbf{F}_{\mathrm{PA}}$). A difference Patterson map is calculated between $\mathbf{F}_{\mathrm{P}}$ and $\mathbf{F}_{\mathrm{PA}}$ which contains the inter-atomic vectors between the HA and the protein. However, due to the electron dense HA, the principal vectors present in the difference Patterson are between the HAs. The inter-HA vector peaks therefore present as a simplified difference Patterson map of a simplified ASU (the HA sub-structure). The phases of the HA sub-structure (φ_A) are estimated using the HA coordinates calculated in the Patterson, and these are used to estimate the phases of the protein structure (φ_P) (Figure 4.11).

A problem still remains, as once $\mathbf{F}_{\mathrm{A}}$ is known that leaves two possible solutions for the phases of $\mathbf{F}_{\mathrm{P}}$; these solutions are known as hands. Traditional isomorphous replacement structures solved this problem by using 2 or more HA derivatised crystals. More modern phasing methods, such as Single wavelength anomalous dispersion (SAD) also have this problem, but it tends to be solved by systematic solvent flattering (discussed in § 4.6.4).

A series of crystal soaking experiments were performed similar to those in Dauter, Dauter, and Rajashankar (2000). *Mm*PepT1^{ECD} crystals were soaked in mother liquor containing either 10 μM $K_2Pt(NO_2)_4$ or Hg Acetate for a period of 30 s, 2 min and 10 min. The HA soaked crystals were then back washed in mother liquor for several seconds to remove un-bound HAs, cryo-protected in mother liquor laced with 25% v/v glycerol and cryo-

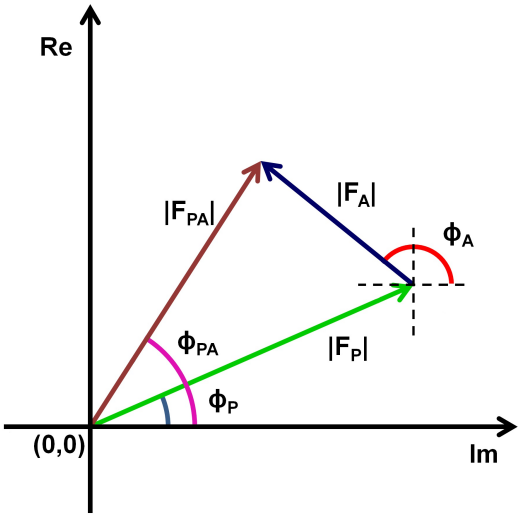


Figure 4.11: An Argand diagram showing the relative amplitudes and phase components of $\mathbf{F_{PA}}$, $\mathbf{F_P}$ and $\mathbf{F_A}$. Each structure factor is composed of an amplitude ($|\mathbf{F}|$) and a phase (φ). The phases of any structure factor are known, they can be used to calculate the phases of other structure factors.

cooled. The 10 min Hg soaks proved fatal for the integrity of the crystals which started to disintegrate during the soak. The 30 s and 2 min crystals survived the soaking experiment and were taken to the beamline.

To assess the presence of any HA in the crystal, fluorescences scans of the HA trial crystals were conducted. Figure 4.12 shows the peak wavelength dependent factors f' and f'' computed from a fluorescence scan of a *Mm*PepT1^{ECD} crystal soaked in Hg (II) acetate, illuminated at I04, and giving f' and f'' of −12.39 and 10.18 respectively. The fluorescence scan confirms the presence of Hg. The diffraction from the 2 and 10 min HA soaks crystals showed attenuated diffraction. The fluorescence scan from the 30 s $K_2Pt(NO_2)_4$ soaked crystal did not show any significant absorption on the Pt edge. The Hg-SAD dataset was collected from a 30 s Hg Acetate soaked crystal.

4.6.3 Single wavelength anomalous dispersion

The isomorphous replacement phasing method requires native and HA derivatised crystals in order to calculate the difference Patterson and solve the phase problem. An explicit requirement for isomorphous replacement is that the two or more crystals used to create the difference Patterson are isomorphous *i.e.* the diffraction patterns and crystal parameters do not differ appreciably between the native and HA derivatised crystal. If the derivatised

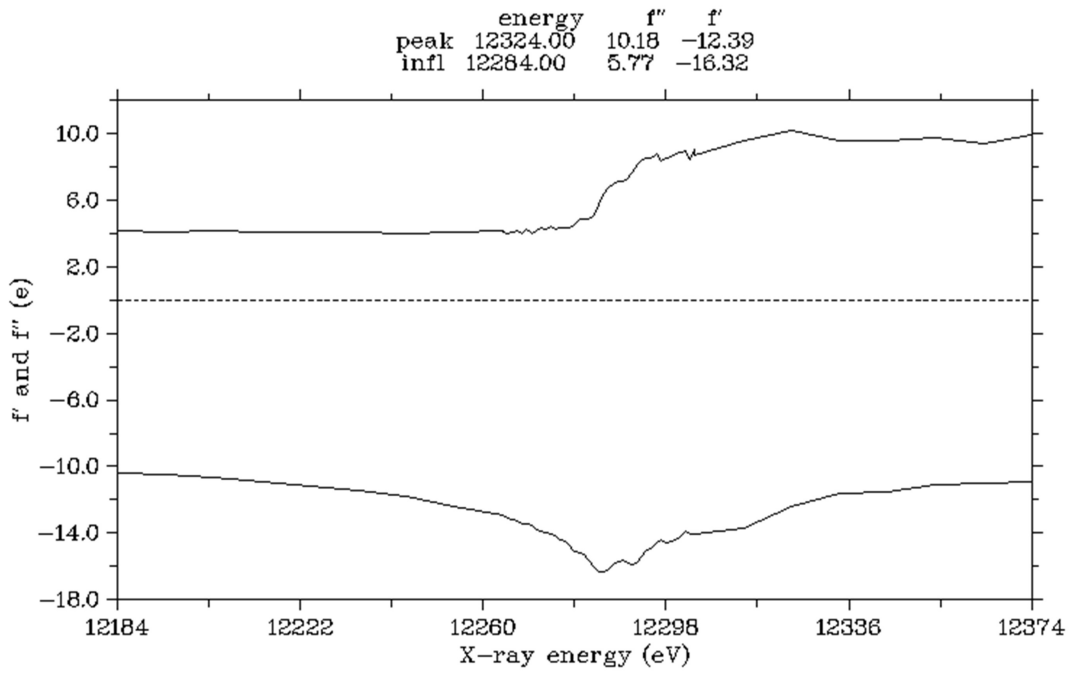


Figure 4.12: The peak wavelength dependent factors f' and f'' calculated by `Chooch` (Evans and Pettifer 2001) from a fluorescence scan collected at Diamond Light Source I04 over the Hg (II) absorptive edge on a MmPepT1^{ECD} crystal soaked in Hg (II) Acetate.

and native crystals are non-isomorphous, Equation 4.6 does not hold, and the phases cannot be solved. Due to the problem of non-isomorphism, isomorphous replacement method has given way to phasing methods which utilise the anomalous scattering and can be performed on a single crystal derivatised with a HA.

The SAD and multi-wavelength anomalous dispersion methods of phasing exploit the violation of centrosymmetry observed in reciprocal space introduced by anomalous scatters. Each structure factor, $\mathbf{F}_{\boldsymbol{hkl}}$ has a conjugate mate, $\mathbf{F}_{\boldsymbol{-h-k-l}}$. Friedel's law states that, in the absence of anomalous signal the $|\mathbf{F}_{\boldsymbol{hkl}}| = |\mathbf{F}_{\boldsymbol{-h-k-l}}|$ and $\varphi_{hkl} = \varphi_{-h-k-l}$. Anomalous signal is generated by the absorption of the incoming photon by a HA and the excitation of an electron from a HA core shell. The absorption changes the contributions to the scattering made by f' and f'', which respectively reduces the total scattering and shifts the phases by 90 °. To calculate the anomalous differences between the $\mathbf{F}_{\boldsymbol{hkl}}$ and $\mathbf{F}_{\boldsymbol{-h-k-l}}$ accurately a dataset with a larger multiplicity is required *i.e.* the Hg-SAD dataset has a 360 ° ϕ wedge (Table 4.1). The phase calculation is similar to the one described in § 4.6.2. The anomalous differences allow for the simplification of the ASU Patterson. Like

A

Resolution (Å)	12.12	8.83	7.71	7.00	6.50	6.12	5.81	5.56	5.35	5.16	5.00
N(data)	205	185	194	187	180	180	183	178	164	191	173
$\langle \lvert I \rvert / \sigma(I) \rangle$	39.2	35.2	30.9	29.0	24.3	26.1	26.8	26.9	26.2	26.0	28.1
Complete (%)	94.0	97.9	99.0	98.9	99.4	99.4	98.9	100.0	99.4	99.5	100.0
d''/σ	1.93	1.53	1.73	1.45	1.28	1.34	1.19	1.18	1.20	1.15	0.90

B

x	y	z	**Height**
0.58571	0.78203	0.22312	99.90
0.02071	0.34330	0.18362	97.92
0.96430	0.05103	0.15179	69.65
0.21584	0.78962	0.17721	43.77
0.55943	0.91147	0.01438	41.19

Table 4.4: The Friedel mate difference intensities and initial HA positions for the Hg-SAD dataset. Tables were generated by SHELXD (Schneider and Sheldrick 2002). **(A)** The relative number of reflections N(data), $\langle \lvert I \rvert / \sigma(I) \rangle$, Completeness and d"/$\sigma$ plotted against resolution shell. The d''/σ is a measure of the difference in intensity observed between Friedel mates. **(B)** The top 5 relative peak heights of the HA sites found by SHELXD.

isomorphous replacement, the only factor which is not resolved is the handedness of the calculated phases.

The HA sub-structure of the *Mm*PepT1^{ECD} crystals, containing 3 Hg atoms, was solved using SHELXC/D (Usón and Sheldrick 1999) in space-group $P2_12_12_1$. The SHELXD algorithm is a direct methods, dual-space, Patterson-seeded program (Schneider and Sheldrick 2002) based on the Shake-n-bake (SnB) method (Miller *et al.* 1994). The strongest peaks are selected from the Patterson, and a series of vectors (seeds) are generated. The HA positions which relate to these seeds are calculated and refined in real-space. Structure factors for the refined HA positions are calculated, and a minimization is performed between the $|\mathbf{F}_{\mathrm{calc}}|$ and $|\mathbf{F}_{\mathrm{obs}}|$ Patterson peaks. The methods have been improved by the removal of a random 30% of $|\mathbf{F}_{\mathrm{obs}}|$ which seems to help the recycling of incorrect solutions (Schneider and Sheldrick 2002). Two metrics are available; CC and CC_{weak}. The former is the Correlation coefficient (CC) between $|\mathbf{F}_{\mathrm{calc}}|$ and $|\mathbf{F}_{\mathrm{obs}}|$ and the latter is the CC between the 30% excluded data. For SAD phasing, a $CC > 30$ suggests a correct solution.

The input data used for the phase calculation were the unmerged intensities from the Hg-SAD dataset. Table 4.4A shows the calculated intensity differences between the $|\mathbf{F}^{+}|$ and $|\mathbf{F}^{-}|$ in the Hg-SAD dataset. The data was cut to 6.0 Å to remove noise. SHELXD found 17 potential HA sites with a CC and CC_{weak} of 54 and 21 respectively. The HA

Bin	D_{min}	D_{max}	φ–power	R_{cullis}	FOM
1	59.52	14.04	3.152	0.433	0.59695
2	14.04	10.07	2.240	0.561	0.52296
3	10.07	8.26	1.974	0.562	0.57817
4	8.26	7.17	2.021	0.587	0.54693
5	7.17	6.42	1.745	0.627	0.50327
6	6.42	5.87	1.524	0.685	0.48665
7	5.87	5.44	1.189	0.768	0.44650
8	5.44	5.09	0.983	0.783	0.40694
9	5.09	4.80	1.025	0.866	0.35177

Table 4.5: The acentric phasing statistics calculated during HA sub-structure refinement of the Hg-SAD dataset. The statistics were generated by SHARP (Bricogne *et al.* 2003).

peak height dropped after the first two sites which indicated a correct solution. The top 5 sites were used for refinement and phase calculation (Table 4.4B).

The 5 HA positions were refined using SHARP (de La Fortelle and Bricogne 1997; Bricogne *et al.* 2003), a ML (expanded in § 4.7) phase determination program. The relative log-likelihood gain converged to 4.06×10^{-7} over 3 cycles of HA parameter refinement. Two HA sites were dropped during refinement which corresponded to the lower peak heights in Table 4.4B. The final phasing statistics for the dataset are shown in Table 4.5. The phasing power remained above 1.0 over the majority of the acentric reflections indicating a good agreement between the observed and calculated anomalous differences. The R_{cullis} remained below 1.0 over all acentric reflections indicating phase information in the dataset. The Figure-of-merit (FOM) or mean φ-error was reasonable.

4.6.4 Density modification

Once a correct HA substructure has been found and refined, the φ_A and $|\mathbf{F}_A|$ allow for 2 possible phase solutions. This phase ambiguity is solved during density modification of the dataset (Wang 1985). A significant proportion of a protein crystal are channels of bulk solvent. A solvent mask is constructed around these channels which leads to a significant sharpening of the observed protein density as the phases are modified to enforce it. The phase ambiguity is generally solved by determining which phase enantiomorph has the largest electron density correlation ($|E|^2$), although this assumption can be wrong and may require refinement.

Solvent Content	Solomon Contrast	$\|E\|^2$ Correlation
0.60	0.41	0.55
0.55	0.41	0.56
0.50	0.38	0.63
0.45	0.39	0.64
0.40	0.47	0.68

Table 4.6: The `Solomon` contrast and $|E|^2$ values for different solvent contents of the ASU. The values were generated in `Solomon` (Abrahams and Leslie 1996) after 50 rounds of refinement excluding reflections below 5.0 Å resolution.

The correct hand assignment and solvent flattening was performed in `Solomon` (Abrahams and Leslie 1996). Based on the V_m and SRF, two monomers were assumed to occupy the ASU, with a probable water content of approximately 50%. A series of `Solomon` density modifications were performed for different starting ASU solvent contents: 60, 55, 50, 45 and 40%. Table 4.6 shows the the results of the solvent content analysis. A solvent content of 0.50 was selected for continued density modification, due to the low `Solomon Contrast`, high $|E|^2$ and agreement with the V_m. Inspection of the electron density from the 50 and 45% solvent content maps indicated well defined solvent channels.

Another density modification method is histogram matching, proposed by Lunin (1988) and Zhang and Main (1990). The gradients of electron-density probability histograms for a well refined protein at a given resolution are independent of individual protein differences. Therefore the electron-density probability histogram of a poorly refined protein can be matched against the histogram of well refined protein structures and thus improve the poorly refined phases and density.

`DM` (Cowtan 1994) was used to histogram match the $|\mathbf{F}|$ as the resolution was extended to the limit of the dataset (2.84 Å) and the size of the solvent sphere was reduced to 3.0 Å. This decreased the `Solomon` contrast to 0.35, without changing the $|E|^2$ correlation. The final maps generated had improved secondary structure resolution. However, the secondary structural elements (β-sheet) lacked overall connectivity (Figure 4.13).

4.7 Structure building/refinement

The initial map phases were determined to a resolution of 2.9 Å from the Hg-SAD dataset. Secondary structural elements were clearly visible in the map but the connectivity was

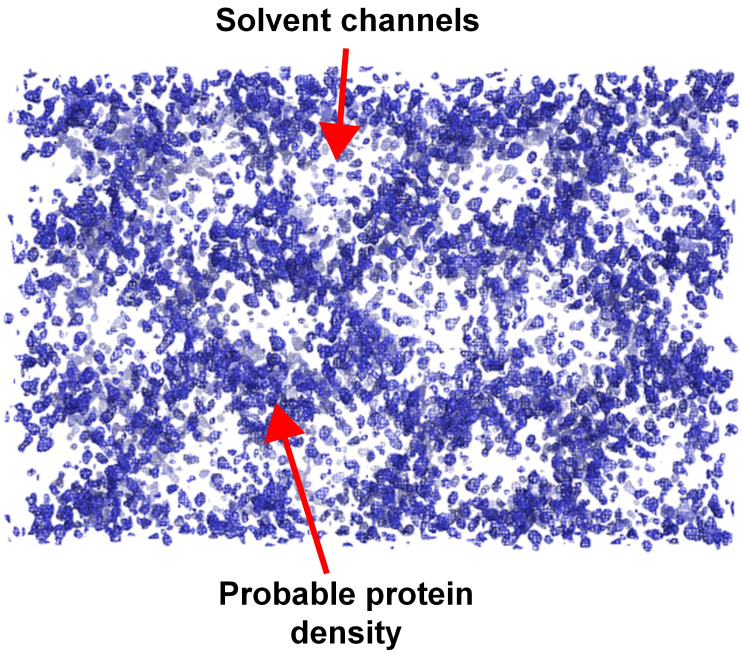


Figure 4.13: An electron density map generated from the Hg-SAD $|\mathbf{F}_{hkl}|$ and `SHARP` refined phases. Protein density is observable and well defined solvent channels, strongly indicative of a correct phase solution.

generally poor. To improve the density maps, more phase information was calculated by model building and refinement.

There are two metrics which help determine the accuracy of the model; the R_{work} and the R_{free} (Equation 4.7). The R_{work} and R_{free} are linear residuals between the $|\mathbf{F}_{obs}|$ and the model structure factors $|\mathbf{F}_{calc}|$. The R_{free} was proposed as a means of cross validation (Brünger 1992) of the R_{work} to ensure that models are not over parametrised and therefore over fit. 5% of the reflections were removed prior to refinement and excluded from any structure calculation.

$$R_{factor} = \frac{\sum_{hkl} ||\mathbf{F}_{obs}| - |\mathbf{F}_{calc}||}{\sum_{hkl} |\mathbf{F}_{obs}|} \tag{4.7}$$

The majority of refinement programs utilise ML functions to minimise the difference between the $|\mathbf{F}_{obs}|$ and $|\mathbf{F}_{calc}|$ (Murshudov, Vagin, and Dodson 1997). ML refinement programs iteratively maximise the likelihood of the model given the data. `Buster` (Blanc *et al.* 2004), `Phenix.refine` (Afonine *et al.* 2010) and `Refmac5` (Murshudov, Vagin, and Dodson 1997) all utilise ML functions during a combined real and reciprocal space

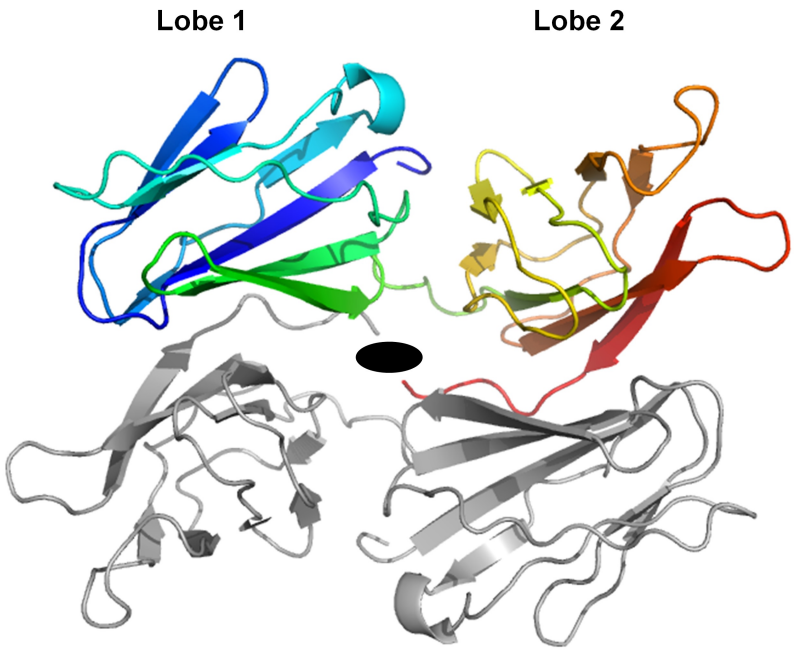


Figure 4.14: The asymmetric unit of the *Mm*PepT1^{ECD} crystal containing two ECD monomers. One monomer has been rainbow highlighted from the N- to the C-terminus. The 2-fold NCS axis at the centre relates the two monomers. The sub-domains have been labelled lobe 1 and lobe 2.

refinement of $|\mathbf{F}_{obs}|$ and $|\mathbf{F}_{calc}|$.

Model building was performed in `Coot` (Emsley and Cowtan 2004) and the majority of the refinement was done in `Buster`. The R_{free} reflections were assigned by `CTruncate` (Winn *et al.* 2011) upon $|\mathbf{F}_{hkl}|$ calculation for the Hg-SAD dataset. The same R_{free} test set was transferred to the the N1 and N2 datasets. Secondary structure was built and rigid-body refined until gradually a complete Cα trace had been built. The phases were parameterized as phase-probabilities (Hendrickson-Lattman (HLA) coefficients), using `DM` (Cowtan 1994) and the experimental HLAs were included into the N1 .mtz file using `CAD` (Winn *et al.* 2011).

The two monomers in the ASU were related by a 2-fold rotational axis (Figure 4.14). The `Secondary-structure matching` (SSM) (Krissinel and Henrick 2004) program was used to find the NCS and translational vectors which related the monomers. The NCS operators were used to average the NCS density using `Dmmulti` (Cowtan 1994). The density was then improved enough for an automated building programme to produce a partial thread of one monomer.

		Final Model
R_{work}		0.193
R_{free}		0.242
Refined Residues		376
Refined H_2O		150
B factor (Å^2)		41.52
RMSD Bond Length (Å)		0.01
RMSD Bond Angle (°)		1.01
Ramachandran Plot (%)	Favoured	97.38
	Allowed	2.62
	Disallowed	0.00
Rotamer Outliers (%)		0.88
`Molprobity` (Percentile)	Clashscore	98
	Overall	99

Table 4.7: Refinement statistics for the *Mm*PepT1^{ECD} model.

The automated building program used was `Buccaneer` (Cowtan 2006; Cowtan 2012). `Buccaneer` uses the FFT and a Bayesian probability function to calculate likely seed positions and chain extensions. Chains are subsequently joined to create complete Cα-traces and R-groups are assigned based on a library of protein fragments. Using `Coot`, the complete primary sequence of one monomer was built and superimposed onto the density of the second monomer using `SSM`. This allowed for a complete trace of both monomers.

The complete structure was refined against the intensities of the highest resolution dataset, N2 in `Buster`. Inspection of the *Mm*PepT1^{ECD} monomer showed that it could be divided into two sub-domains, 'lobes' connected by a flexible linker (Figure 4.14). During refinement the lobes were classified as Translational/Librational/Skew (TLS) domains (Schomaker and Trueblood 1968), NCS was assumed between the monomers (Smart *et al.* 2012) and the disulphide bond was modelled as both reduced and oxidised. Waters were added and pruned manually. Three areas of non-protein/water density were observed which were consistent with glycerol molecules. Glycerol was built into the density and refined. `Molprobity` (Chen *et al.* 2010) was used to aid refinement and validate the structure. Hydrogens were added to refine side-chain clashes and Ramachandran plots (Figure 4.15) were used to aid Phi/Psi angle refinement. The final rounds of refinement were performed using `Phenix.refine` (Afonine *et al.* 2010) to enforce good stereo-chemistry. Table 4.7 shows the model statistics for the crystal structure.

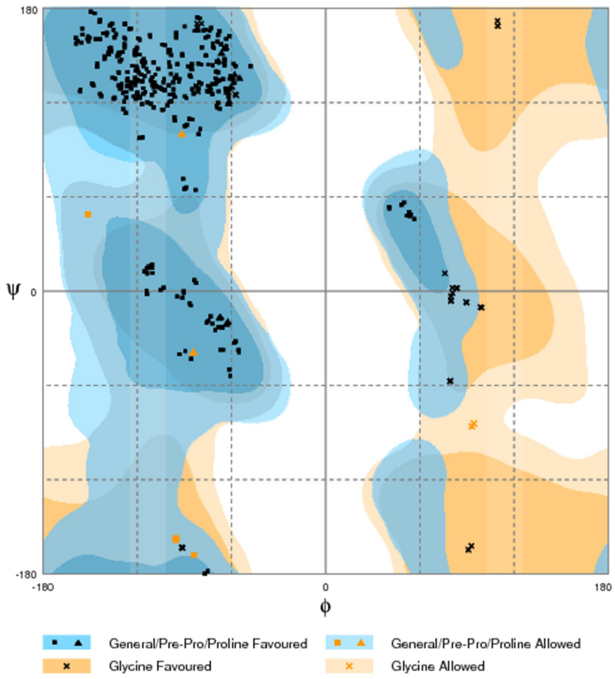


Figure 4.15: A Ramachandran plot of the *Mm*PepT1^{ECD} ASU calculated using `Rampage` (Lovell *et al.* 2003).

4.8 *Mm*PepT1^{ECD} ASU Model

The ASU of the crystal was occupied by two monomers of the *Mm*PepT1^{ECD} related by a 2-fold rotational axis (Figure 4.16A). This dimerization appears to be crystallographic as subsequent biophysical analysis indicated the ECD was monomeric in solution (discussed in Chapter 6). Figure 4.16A shows the ECD is composed of two lobes, which are connected by a flexible linker. The dimerization interface is almost exclusively between lobe 1 and lobe 2 on the adjacent chains.

The lobes themselves are composed of a 4-sided β-sandwich. An `SSM` (Krissinel and Henrick 2004) alignment of the lobes (Figure 4.16B) gives a Cα Root-mean-square deviation (RMSD) $\approx$ 1.5 Å. The reason for the structural similarity can be seen in the topology (Figure 4.16C). The secondary structure and connectively between lobe 1 and lobe 2 are identical apart from the 3_{10}-helix in the D-E loop and the disulphide bond between the M-N and O-P loops.

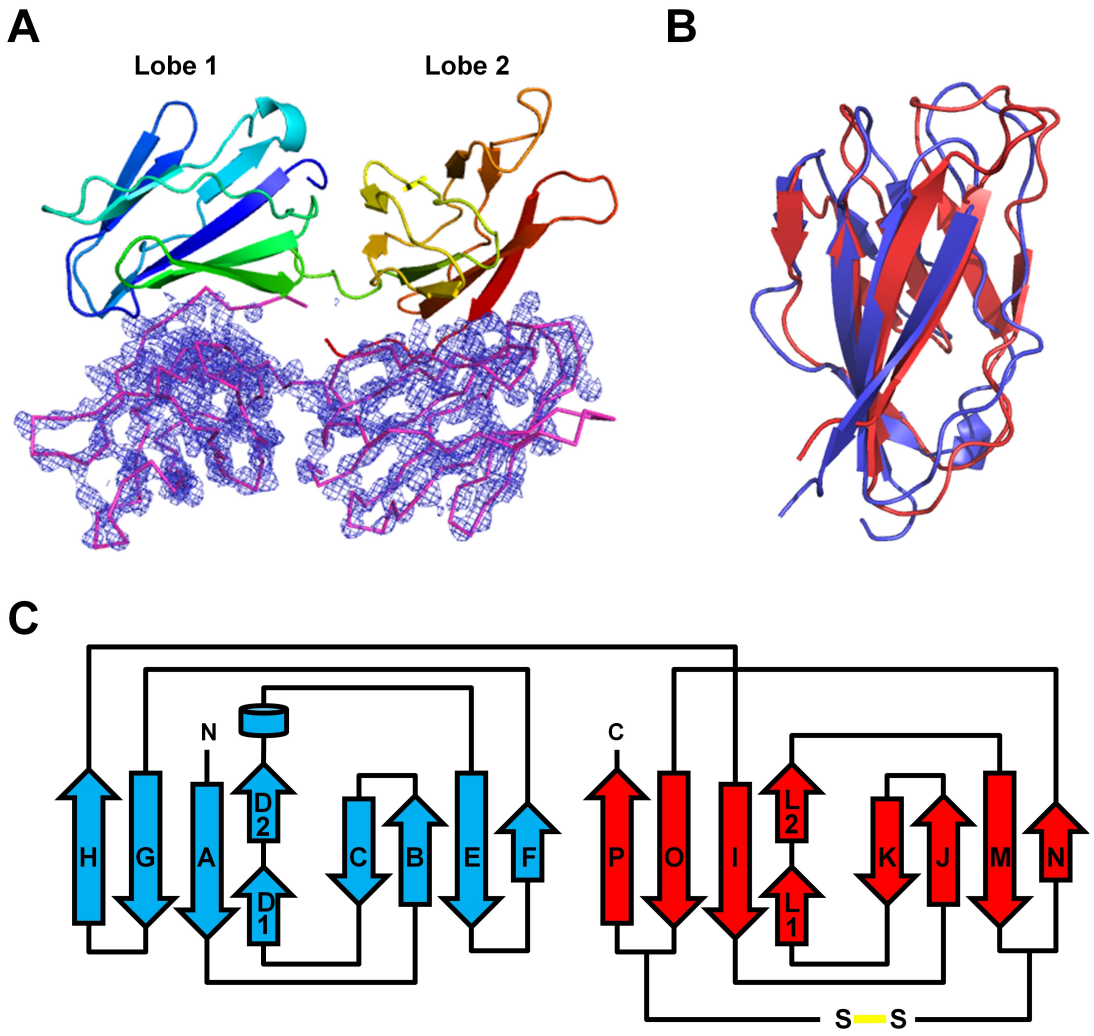


Figure 4.16: Characteristics of the *Mm*PepT1^{ECD}. **(A)** The asymmetric unit of the *Mm*PepT1^{ECD} crystal containing two 2-fold NCS related ECD monomers. One monomer has been rainbow highlighted from the N- to the C-terminus. The other is shown as a purple ribbon with the refined $2F_0$-F_C electron density map overlaid. **(B)** A Cα alignment of lobe 1 (blue) and lobe 2 (red) using SSM (Krissinel and Henrick 2004). The lobes align with a Cα RMSD of 0.5 Å. **(C)** A topology diagram of *Mm*PepT1^{ECD}. Lobe 1 and lobe 2 are coloured red and blue respectively.