Investigation of a Novel Intein-Based *Escherichia coli* Expression System for Human Methylmalonyl CoA Mutase

A thesis presented to Massey University in partial fulfillment of the requirements for the degree of

Master of Science in Biochemistry

Alice Rosemary Clark

2005

Abstract

Human methylmalonyl CoA mutase (hMCM) is a 78 kDa homodimeric mitochondrial matrix enzyme. hMCM catalyses the conversion of 2R-methylmalonyl CoA to succinyl CoA in the metabolism of propionyl groups, and requires the vitamin B_{12}-derived cofactor adenosylcobalamin (AdoCbl). The mechanism of catalysis involves homolytic cleavage of AdoCbl's unusual C-Co bond, to generate radicals. Dysfunctional hMCM results in the rare, potentially fatal metabolic disorder methylmalonic acidemia. An experimentally determined structure of hMCM would add to the understanding of both the mechanism of catalysis and the molecular basis of some of the mutations underlying methylmalonic acidemia. The structure of the bacterial orthologue from *Propionibacterium shermanii* has been solved by x-ray crystallography, enabling the development of structural models of hMCM. Critical differences, however, between these two enzymes, mean that some regions of the models could be inaccurate.

There is no x-ray crystal structure of hMCM. Purification of native hMCM for crystallization trials is complicated by ethical problems, low yields, and heterogeneity generated by the cofactor. To provide a more convenient source of pure, active human methylmalonyl CoA mutase for x-ray crystallography, an expression system for recombinant hMCM is required. Other researchers have expressed hMCM in *Escherichia coli* as (i) insoluble inclusion bodies, (ii) soluble fusion protein that cannot be separated efficiently from the fusion tag, or (iii) in low quantities.

This research aimed to develop an *E. coli* expression system for the production of active human methylmalonyl CoA mutase, to enable x-ray crystallography structural studies. Based on the results of previous expression systems, four novel expression vectors were developed utilising the maltose binding protein and thioredoxin as solubility tags. It was hoped that conventional protease cleavage, to remove these solubility tags, could be circumvented by the use of intein-mediated cleavage. Intein-mediated cleavage was successful, and soluble active hMCM was recovered in low yields from a C-terminal thioredoxin solubility tag construct. hMCM was insoluble when expressed with MBP at the C-terminus.

I am among those who think that science has great beauty.

A scientist in his laboratory is not only a technician: he is also a child placed before

natural phenomena which impress him like a fairy tale.

Marie Curie (1867 - 1934)

Acknowledgements

Firstly I would like to thank my supervisor, Dr Mark Patchett, for your enthusiasm, patience and encouragement. You made this MSc a wonderful experience that I will remember for the rest of my life.

I would like to thank Dr Kathryn Stowell, for allowing me work in her lab, and to everyone else in the lab, for all the fun times we had. I would especially like to thank Carole, Kei and Andy for their patience and endless help in the lab. Also, thank you to everyone else in IMBS for all their help and advice.

I would like to acknowledge the Palmerston North Medical Research Foundation for the financial support that made this project possible.

Thanks to my family, Di, John, Helen, Robin and Lily who helped me though the difficult times during my time at Massey University.

I am forever grateful to Sacha, Lia and Natisha who helped me with this MSc and supported me in so many different ways. You are beautiful girls and made me the happy person I am today. Thank you.

Finally, I would like to thank Scott - for your belief in my abilities, and your patience with my inabilities.

Table of Contents

List of Figures

List of Tables

List of Changes

- add, "cccDNA , covalently closed circular DNA" to the list **xiii**
- change "figure 4" to "figure 1.4" **4**
- change "compliment" to "complement" **10**
- insert "of" between "solubility" and "proteins" (third line from bottom) **11**
- change "Ion" to "Lon" **13**
- change "termini" to "terminus" **16**
- change "2001" in figure legend to "2002" **20**
- change "Evens" to "Evans" **22**
- change "tetramethylethylene di'amine" to "tetramethylethylenediamine" **25**
- change "sample to elucidate" to "sample. To elucidate" **27**
- change "NAHCO 3" to "$NaHCO_3$" **37**
- change "Lane 3 contain" to "Lane 3 contains" **65**
- change "appendix G" to "appendix F" **66**
- in line 11 insert "smaller size" after "This"
- change "change" to "changed" **80**
- change "contains rare codons within the hMCM gene" to "supplies extra tRNAs for the rare codons in the hMCM mRNA. These rare codons may" **88**
- on the second to last line: change "yield the " to "yield. The". Also change "decrease" to "decreases" **90**
- change "figure 5.12" to "figure 5.13" in the second to last line **91**
- change "containing rare codons" to "supplying extra tRNAs for the rare codons in the hMCM mRNA" **108**
- change "Hayes, 1992" to "Hayes, 1998" **108**
- change "*P. furiousis*" to "*P. furiosus*" **111**

List of Abbreviations

AdoCbl	Adenosylcobalamin
Amp	Ampicillin
BLAST	Basic local alignment search tool
BSA	Bovine serium albumin
CBD	Chitin binding domain
C-terminal	Carboxyl terminial
CoA	Coenzyme A
DNA	Deoxyribonucleic acid
dNTP	Deoxy-nucleotide tri-phosphate
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
g	Gravitational field. unit of
hMCM	Human methylmalonyl CoA mutase
IPTG	Isopropyl-β-D-thiogalactopyranoside
kb	Kilobase pairs (of DNA)
MalE	Gene encoding the maltose binding protein
MAP	Methionine aminopeptidase
MBP	Maltose binding protein
MCM	Methylmalonyl CoA mutase
MCS	Multiple cloning site
mRNA	messenger ribonucleic acid
NADH	Nicotin amide-adenine dinucleotide. reduced
N-terminal	Amino terminal
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SDS	Sodium dodecyl sulfate
TEMED	N, N, N`, N'-tetramethylethylenediamine
Trx	ORF encoding the thioredoxin protein with a His tag
Tris	Tris-(hydroxymethyl)-aminomethane
V	volts

1 Introduction

1.1 The Function of Methylmalonyl CoA Mutase (MCM)

Human methylmalonyl CoA mutase (hMCM) is a mitochondrial enzyme that catalyses the isomerisation of *R*-methylmalonyl CoA to succinyl CoA, and the succinyl group is ultimately oxidised in the citric acid cycle (figure 1.1). This reaction is required in the metabolism of propionyl CoA formed during the catabolism of odd chain fatty acids, cholesterol intermediates, thymine, uracil and the amino acids methionine, isoleucine and valine (Kolhouse *et al.*, 1988).

In prokaryotes MCM has a different role; it is involved in the terminal fermentation pathway of succinyl CoA to propionate, and catalyses the conversion of succinyl-CoA to methylmalonyl CoA (Zagalak and Retey, 1974).

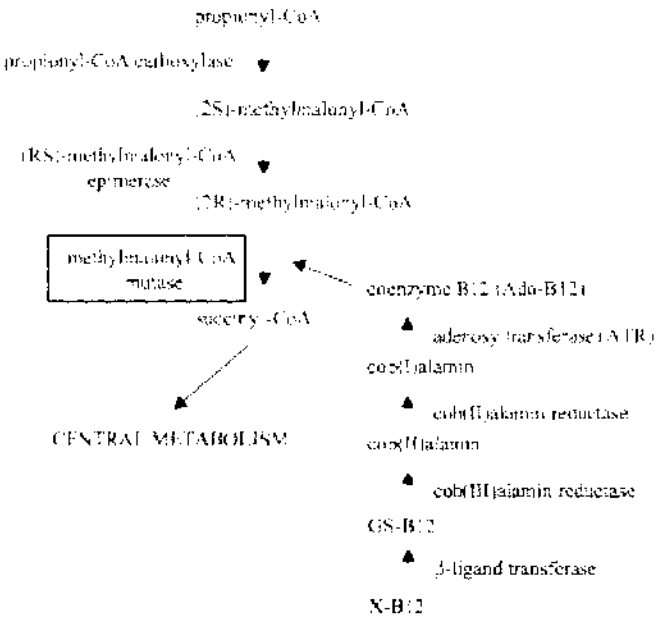

Figure 1.1 Propionyl CoA Metabolism in Humans. Showing the formation of the hMCM substrate methylmalonyl CoA and synthesis of the cofactor coenzyme B_{12} from inactive B_{12} precursors (XB12).

(Figure adapted from Leal *et al.*, 2003)

1.2 The Cofactor and Catalytic Mechanism

MCM belongs to a group of prokaryotic and animal enzymes that require the vitamin B_{12}-derived cofactor adenosylcobalamin (AdoCbl), also known as coenzyme B_{12} and of this group only MCM is found in both prokaryotes and animals (Banerjee, 1997). AdoCbl-dependent enzymes catalyse carbon skeleton rearrangements via a radical mechanism involving the carbon-cobalt bond in the AdoCbl cofactor, (figure 1.2), (Taoka *et al.*, 1997; Banerjee and Vlasie, 2002).

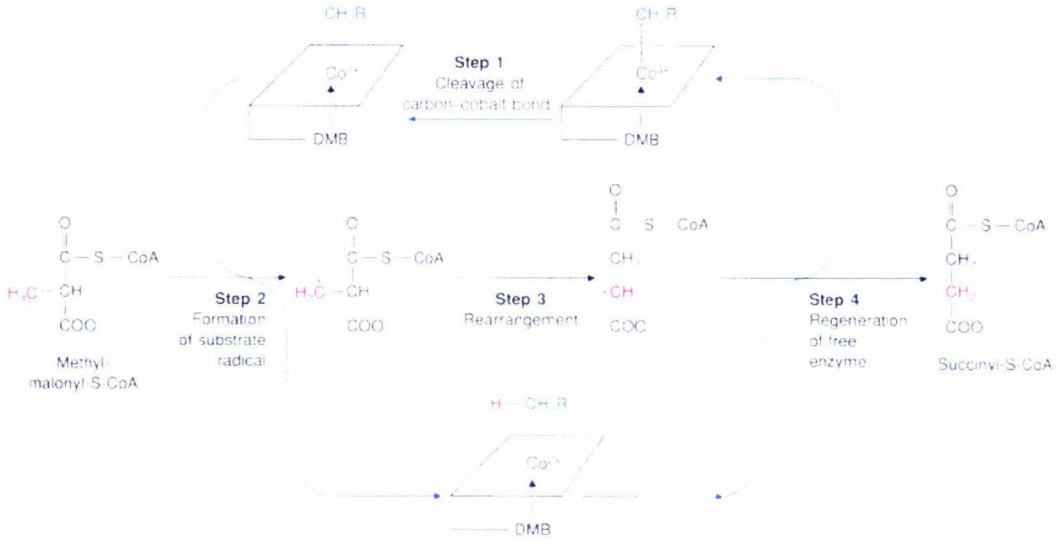

Figure 1.2 The Reaction of Methylmalonyl CoA Mutase. Showing the radical generation by the adenosylcobalamin cofactor (shown as the square containing Co), then the formation of succinyl CoA and the regeneration of the free enzyme.

(Figure adapted from Matthews *et al.*, 1999)

AdoCbl comprises a modified corrin ring surrounding a cobalt atom (figure 1.3). A protein histidine side chain acts as a sixth ligand binding to the cobalt atom in the holoenzyme (Mancia *et al.*, 1996). The adenosyl moiety is attached to the cobalt though a covalent cobalt-carbon bond. This carbon-cobalt bond has special reactivity, and is also very rare; there are only two known enzymes in animals that employ carbon-cobalt bonds in reaction mechanisms. Cytoplasmic methionine synthase uses methylcobalamin and mitochondrial MCM uses adenosylcobalamin (Kolhouse and Allen, 1977). The reaction catalysed by hMCM is also unusual for another reason; the enzyme breaks and reforms a carbon-carbon bond in the carbon skeleton of methylmalonyl CoA.

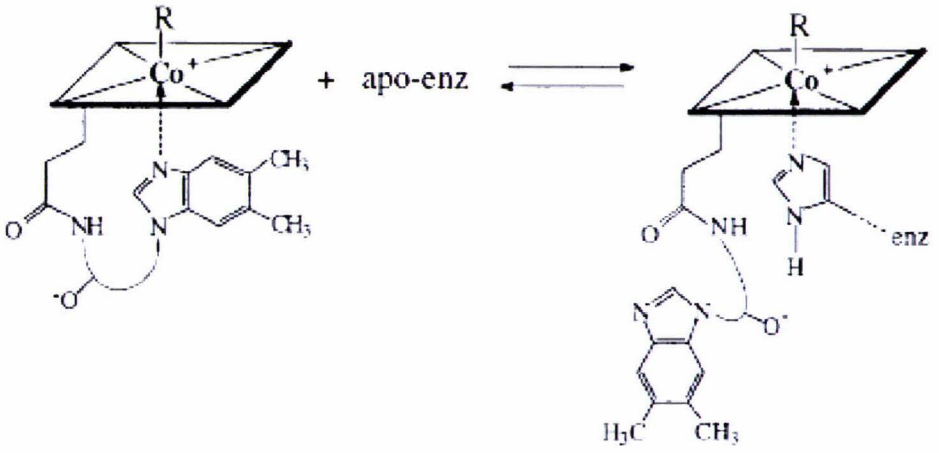

Figure 1.3. Adenosylcobalamin Binding to Bacterial *P. shermanii* MCM. Left: 'base on' AdoCbl, separate from the enzyme. Right: 'base off' AdoCbl, bound to the enzyme through a histidine residue. The adenosyl moiety (R) provides the carbon atom for the carbon-cobalt bond. The four nitrogen's of the corrin ring coordinate the cobalt (centre), and the histidine of MCM (below) displaces the lower axial ligand of cobalt.

(Figure adapted from Tollinger *et al.*, 2001)

1.3 The Human Methylmalonyl CoA Mutase Gene

The *mut* locus on chromosome 6 contains 13 exons, and (in total) the gene is 35 kbp long (Ledley *et al.*, 1988). hMCM is constitutively expressed and encodes a protein with a 32 amino acid mitochondrial leader sequence that is cleaved to produce the mature 78 kDa peptide. (Nham *et al.*, 1990). The active human enzyme purified from liver is an α_2 homodimer of approximately 150 kDa, and each subunit binds a molecule of adenosylcobalamin, the cofactor. Adenosylcobalamin is synthesised from vitamin B_{12} *in vivo* as shown in figure 1.1 (Fenton and Roseburge, 1995; Banerjee and Chowdury, 1999).

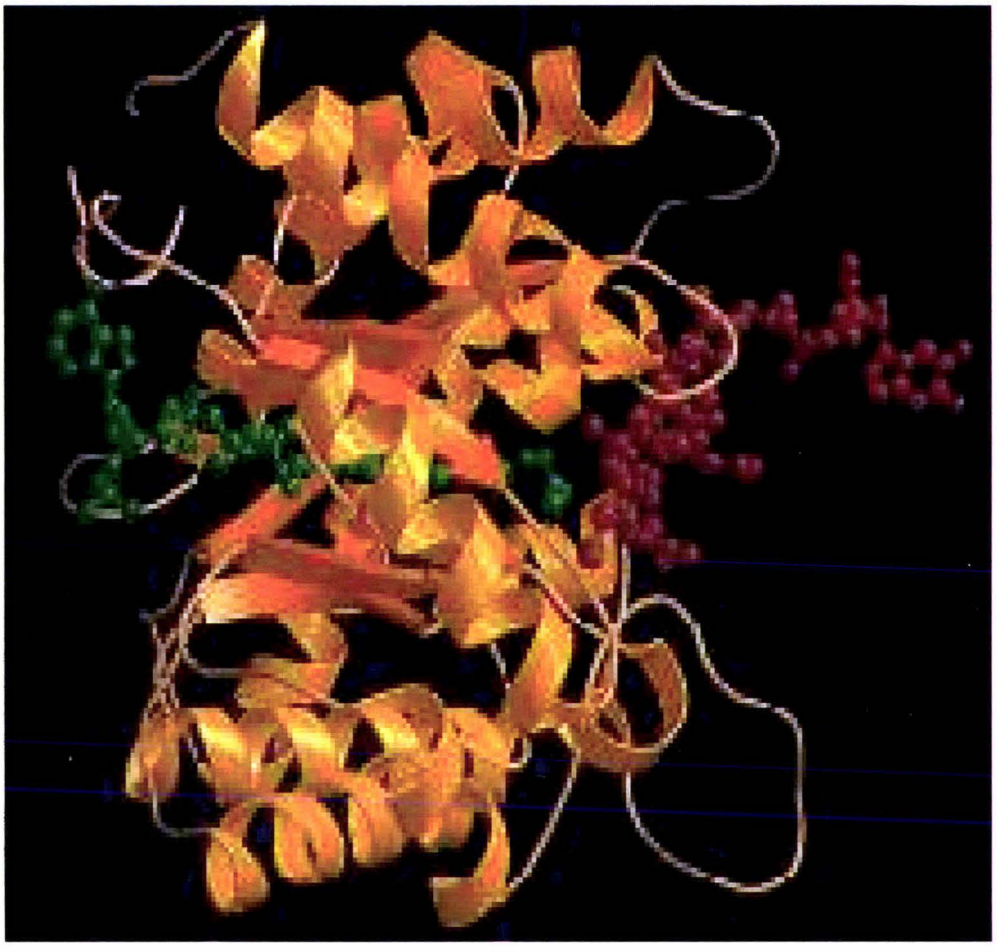

Figure 1.4 Structure of the *Propionibacterium shermanii* methylmalonyl CoA mutase.
The substrate, methylmalonyl CoA (green) and the co-factor, adenosylcobalamin (red), are shown bound in the active site.

(Mancia *et al*, 1996)

The precursor protein sequence contains a C-terminal cobalamin binding domain in residues 578-750. The $(\beta\alpha)_8$ barrel formed by residues 87-416 contains a methylmalonyl CoA binding site. Residues at the N-terminus may be important in the dimerisation of the two subunits (figure 1.5). A loss in homology between the bacterial and human enzymes makes predictions about the structure and function of this region less certain (Thoma and Leadley, 1996). The MCM $\alpha\beta$ heterodimer of the *Propionibacterium shermanii* homologue has been cloned and expressed in *E. coli* (McKie *et al.*, 1990), and the crystal structure has been solved to 2Å resolution, see figure 4 (Mancia *et al*, 1996). The α subunit of *P. shermanii* MCM shows a remarkably high (61%) amino acid identity to the α subunit of human MCM (Leadlay and Ledley,

1989). However the β subunit, while clearly related to the α subunit shares only 34% identity with the amino acid sequence of the human gene, and its role remains enigmatic (figure 1.5).

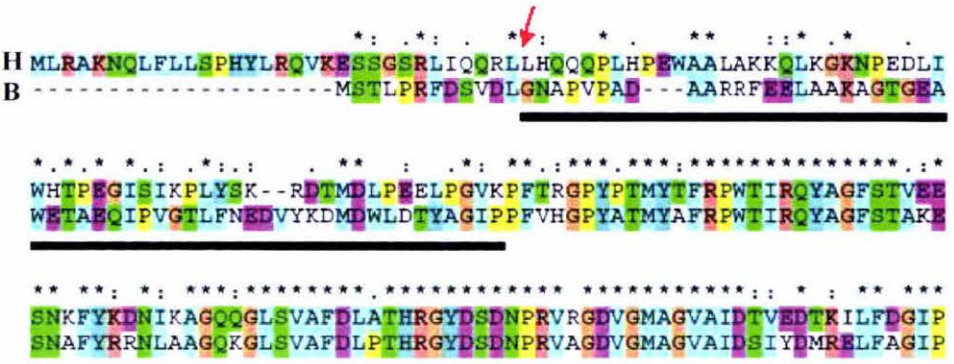

Figure 1.5 An Alignment of Human & Bacterial Amino Acids at the N Termini of MCM. There is 61% identity overall between the α subunits of the bacterial and human enzymes, however the identity in the 59 N-terminal amino acids drops to 22%. Red arrow indicates the cleavage site of the mitochondrial leader sequence within the human enzyme. The 59 N terminal amino acids are shown here with a black line. **H** is the human MCM sequence, and **B** the bacterial sequence (of the α subunit from *P. shermanii*)

Mutations in the methylmalonyl CoA mutase gene (*mut*) cause *mut* methylmalonic acidemia (*mut* MMA). The disease *mut* MMA has an estimated incidence of 1/30000 to 1/50000 live births, and symptoms include metabolic acidosis, lethargy, dehydration, vomiting, and neurological problems (Fuchshuber *et al.*, 2000). *Mut* MMA patients can be distinguished from those suffering from other forms of MMA by their non responsiveness to B_{12} therapy; other forms of MMA are often due to mutations in enzymes involved in the synthesis of coenzyme B_{12} and so can be corrected by B_{12} therapy. There are two types of *mut* MMA, mut^0 and mut^-. The mut^- form of MMA is milder, and is characterised by a reduced hMCM activity. Most of the mutations associated with mut^- MMA are found in the AdoCbl binding domain, and so by increasing the AdoCbl concentration some of the enzyme activity is recovered (Adjalla *et al.*, 1998). Mut^0 MCM, however, is completely inactive and this form of *mut*MMA is severe and often fatal (Peters *et al.*, 2002). So far 81 different mutations in the hMCM 'structural gene' have been identified. Three are found at higher frequencies in some populations (~1% of births are carriers) E117X in Japanese, G717V in African

Americans and N219Y in a French/Turkish population (Aquaviva *et al.*, 2001, Aquaviva *et al.*, 2005).

The known mutations in hMCM causing *mut*MMA are more commonly missense or nonsense nucleotide substitutions, and many unidentified mutations may occur in both the structural gene and promoter region. (Peters *et al.* 2002). Cobalamin affinity and enzyme kinetics has been characterised for most of the known mutations from *mut* MMA patients, has lead to a better understanding of the structure-function relationship of the human enzyme (Janata *et al.*, 1997; Andrews *et al.*, 1993; Crane *et al.*, 1991).

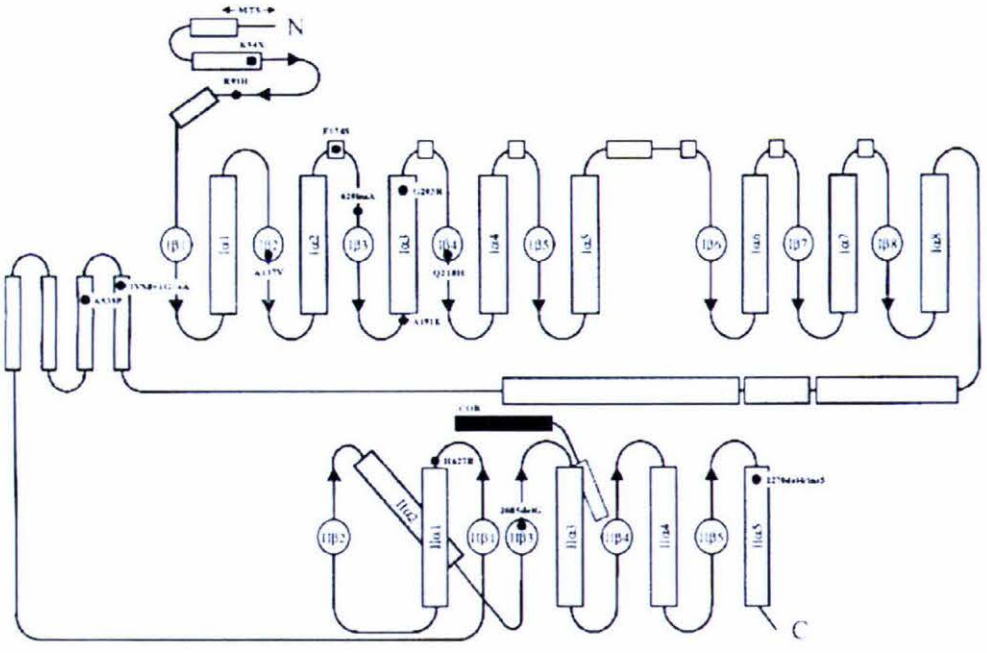

Figure 1.6 Proposed Topology Model of hMCM. The positions of 30 disease causing mutations are indicated as black filled circles. Solid black bar is the cobalamin co-factor.

(Adapted from Fuchshuber *et al.*, 2000)

The topology model in figure 1.6 shows the known mutations in the C-terminal cobalamin binding region, G623R, G626C, G630E, G703R are substitutions of glycines thought to be interacting with the cofactor. These residues surround the H627 that binds as the sixth ligand to the cobalt of the AdoCbl. Mutation at G648 or G717 also disrupts the structure orientation of the H627, affecting the binding of the co-factor, and consequently these are both highly conserved residues. Some other mutations are found

at the putative interface of the two dimers at the N-terminal domain (R93H), and in the channel to the active site (W105R).

1.4 The Novel Gene MMAA

In 2002 a methylmalonic academia patient with a novel mutation was reported. The mutation was in a novel gene that was called MMAA (**m**ethyl**m**alonic **a**cademia linked to the cbl**A** complementation group). The gene was mapped to chromosome 4q31.1-2, and comparison of the genomic and cDNA sequences revealed 7 exons, and a mitochondrial signal sequence. It was a highly conserved protein between organisms and homologues are found in archea, eubacteria and eukaryotes. The gene was assigned a putative function as a transport protein for vitamin B_{12}, based on sequence alignments and analysis of prokaryotic gene arrangements (Dobson *et al.*, 2002). Subsequently a different group has backed up these findings, with the discovery of seven novel mutations in MMAA, all in patients responsive to B_{12} therapy (Yang *et al.*, 2004).

In contrast to these findings, mutation of a gene with very high sequence homology (*meaB*) in the bacteria *Methylbacterium extorquens* AM1, caused a loss of MCM activity. MCM activity was not recoverable by the addition of B_{12} which would be expected if indeed *meaB* encoded a protein for a B_{12} transport. In pull down assays a complex of MCM and the *meaB* protein formed. *meaB* was also found to be required for MCM activity *in vitro*. The authors suggest that the function of the *meaB* protein is to prevent the inactivation of MCM during catalysis. This may arise by a stabilising effect on the dimer form of the mutase, or protection of the enzyme from attack by oxygen, water or highly reactive radical intermediates (Korotkova and Lidstrom, 2004). Whether this gene *meaB* is a functional homologue of *MMAA* remains unclear, more detailed biochemical analysis is required.

1.5 Past Recombinant Expression Systems of hMCM

Earlier attempts to express hMCM in *E. coli* resulted in the production of inclusion bodies, which are insoluble aggregates of misfolded protein. This is common in *E. coli* with highly expressed recombinant proteins (reviewed by Baneyx and Mujacic, 2004).

Refolding of these inactive inclusion bodies from *E. coli* has been attempted with the use of rapid dilution, dialysis, detergent assisted refolding, size exclusion chromatography and use of chaperonin-assisted refolding. Refolding efficiency reached 1% soluble active enzyme, using the rapid dilution technique. This suggests that the polypeptide produced is folded incorrectly, but otherwise fully active when expressed in *E. coli*, (Hayes. 1998). This is not always so, some recombinant proteins may require posttranslational modification that the *E. coli* is not able to perform. Eukaryotic proteins may also expressed as soluble, but not biologically active in *E. coli*.

In an attempt to correct the misfolding *in vivo* of hMCM expressed in *E. coli*, cells were co-transformed with hMCM expression plasmid and pGroESL, a pACYC-derived plasmid that overexpresses the *E. coli* chaperonins GroEL and GroES chaperones (Goloubinoff *et al* 1989). However the co-expression of chaperonins had little effect on the misfolding of the protein (Janata *et al.*, 1997).

There is an increase in yield of soluble hMCM, observed when expressed with a Trx tag, in *E. coli* (Janata *et al.*, 1997). This soluble protein is not being recovered from the insoluble pool of protein, but rather appears to be from a reduction in protein being degraded (figure 1.7).

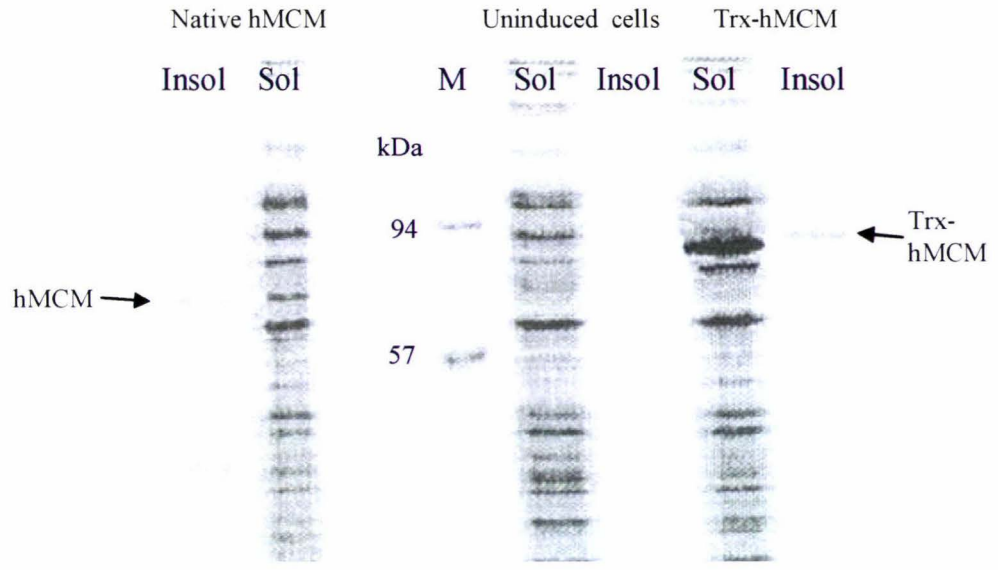

Figure 1.7 SDS PAGE Gel Showing the Expression Systems of Janata et al, (1995). hMCM expression without a solubility tag (native hMCM) is compared to expression with a Trx tag (Trx hMCM). The Trx solubility tag drastically increases the yield of soluble protein without reducing the amount of insoluble protein. Insol : Insoluble fraction. Sol: Soluble fraction. M: Molecular size markers

(Figure adapted from Janata *et al.*, 1997)

The increase in yield recombinant protein with a Trx tag, (figure 1.7), could be a result of two changes that the Trx makes to the polypeptide. One possibility is increasing the half-life due to N-terminal amino acid change. It is also possible that instead of, or in addition to, the Trx, may also reduce the degradation of aberrantly folded protein. The stability of the hMCM is improved thus reducing the aberrantly folded hMCM, and so also reduces the protease degradation of this aberrantly folded hMCM. The authors do not comment on either of these possibilities (Janata *et al.* 1997).

In humans the hMCM pre-protein is targeted to the mitochondria where the signal sequence removed, to produce an N terminal Leu on the polypeptide (table 1.1, protein 1). In *E. coli,* however, a leucine at the N-termini of a polypeptide has a very short half life of 2 mins (Tobias *et al.* 1991). Expression of the native hMCM polypeptide, even with a Met before the Leu, in *E. coli* would be expected to result in the production of a short-lived protein. This high turnover rate, ultimately leads to a lower yielding *E. coli* expression system. Janata *et al.* (1997) expressed the hMCM in two systems, without

and with the Trx tag (table 1.1, proteins 2 and 3 respectively) showing a remarkable increase in protein yield with the Trx tag.

Source	N terminal amino acids	After Met processing	Halflife in *E. coli*
1. Native hMCM:		Leu His ~	
2. Janata, hMCM	Met Leu His ~	Leu His ~ (15%)	2 mins
3. Janata, thio hMCM	Met Ser ~T~ Leu His ~	Ser TAG Leu His ~(84%)	10 hours

Table 1.1 N terminal amino acid and the turnover rate of proteins in *E. coli.* From left, the source, the sequence of the first amino acids, the sequence after the N-terminal Met is removed and the predicted half life in *E. coli.*.

The effect of a post induction temperature, on hMCM solubility with and without a Trx fusion on the N-terminus was critical (Janata *et al.*, 1997). With the Trx-fusion at 37 °C most of the protein was in the form of inactive inclusion bodies, however at 12 °C the majority of the protein is correctly folded, soluble and fully active. The Trx-fusion expression system in *E. coli* GI698 at 12 °C, resulted in high levels of soluble recombinant protein, however the Trx tag was not able to be separated from the target protein. Enterokinase cleavage was inefficient, but at higher protease concentrations and longer reaction times, degradation of the hMCM was observed. A specific activity of 23-26 U/mg was recovered despite the Trx fusion remaining on the hMCM.

Human fibroblasts and *S. cerevisiae* have been used to express mainly mutant enzymes, for use in enzyme activity studies of the mutations. Low activity can be detected using a ^{14}C-propionate incorporation assay (Aquaviva *et al.*, 2001, Crane *et al.*, 1992, Peters *et al.*, 2002). This requires very little active enzyme and is an effective method for diagnoses and study of *mut* MMA patients. In *S. cerevisiae* the protein was soluble; but not active and was unable to complement *mut* fibroblast cells. However, mouse MCM was expressed in the same system, and produced an active enzyme, able to compliment a *mut* deficiency in human cells (Andrews *et al.*, 1993).

1.6 Enhancing solubility of Recombinant Proteins in *E. coli.*

1.6.1 Background

If the aggregates are forming during cell lysis, changing the lysis buffer could prevent partitioning of the recombinant protein into the insoluble fraction (Bondos *et al.* 2003). Recombinant expression of hMCM in *E. coli* results in the formation of aggregates *in vivo:* and the inclusion bodies can be observed by a light microscope (Mark Patchett. *pers. Comm.*). Consequently the expression conditions must be altered to solublise the hMCM protein in *E. coli* (Baneyx and Georgiou 1990).

The formation of expressed protein aggregates is often rapid and irreversible. The kinetics of the aggregation may be significantly faster than the folding kinetics to the native state, as once the aggregation process is initiated, the aggregates will form rapidly (Cellmer *et al.* 2005 and Goldberg *et al.* 1991). Solubilising the aggregates is accompanied by a new problem: proteases now have access to the recombinant protein as is it now not sequestered in the inclusion bodies (Cheng et al., 1981). To prevent inclusion bodies from forming and degradation of the soluble recombinant protein, some expression conditions can be changed: e.g. the rate of expression can be slowed, decreasing the local concentration of folding intermediates. The redox potential of the *E. coli* cytoplasm can be changed or the protein can be targeted to the periplasm which contains many chaperones. Solubility tags can be used or a protease deficient *E. coli* cell line used.

1.6.2 Periplasmic Expression

Secretion of recombinant proteins to the periplasm in *E. coli* has been show to improve the solubility proteins. A signal sequence is placed in the gene, most commonly the signal sequence from the MBP which is located in the periplasim of *E. coli* (Loo et al, 2002; Witholt *et al.*, 1976).

1.6.3 Temperature

The macromolecule concentration in cells is in excess of 300 mg/ml, so a newly synthesised protein will not fold in isolation, but rather while interacting with a number of other molecules that may influence the way the protein folds (Dobson, 2001). Decreasing the temperature of the *E. coli* culture during induction and expression has been used extensively to slow the recombinant protein expression rate, enhancing solubility of recombinant proteins. This is thought to be due to the reduction in the concentration of partially folded proteins at any given time, reducing the chance of two partially folded proteins associating and initiating aggregation (Schein and Noteborn, 1988; Janata *et al.*, 1997).

The bacterial methylmalonyl CoA mutase (of *P. shermanii*) was expressed in *E. coli* K38, in a vector containing the two MCM enzyme subunits. Expression at 37 °C resulted in insoluble protein, but when induction temperature was lowered to 30 °C, resulting in the expression of soluble active methylmalonyl CoA mutase (McKie *et al.*, 1990

1.6.4 *E. coli* Strains

E. coli BL21 (DE3) is deficient in the Lon and OmpT proteases. Lon is a cytoplasmic protease responsible for the degradation of proteins possessing a non native structure (Gottesman, 1989). OmpT is an outer membrane protease that cleaves dibasic amino acid sequences (Grodberg and Dunn 1988). These proteases can contribute to the degradation of some recombinant proteins during expression and purification (Baneyx and Georgiou 1990). This strain is also lysogenic for a λ prophage that contains an IPTG inducible T7 RNA polymerase gene, allowing expression from vectors containing a T7 promoter (Studier, *et al.*, 1990).

E. coli strain Rossetta-gami 2 contains mutations in genes for thioredoxin reductase and glutathione oxidoreductase (*trxB* and *gor*). Thioredoxin reductase is part of an active process preventing disulfide bonds from forming in the cytoplasm of *E. coli*. Glutathione oxidoreductase is involved in the oxidative stress response of *E. coli* and loss of activity changes the redox potential of the cytoplasm (Becker-Hapak and

Eisenstark 1995; Davis *et al.*, 1982). These mutations improve the folding of many eukaryotic proteins expressed in *E. coli* in particular proteins containing disulfide bonds (Derman, *et al.*, 1993, Cassland *et al.*, 2004).

The *E. coli* strain Rosetta-gami 2 also contains a mutation in the lac permease *(lacY)* gene, giving the Tuner™ genotype. These cells allow a more even entry of IPTG through the cell membrane. More even IPTG entry causes induction to occur in an IPTG concentration-dependent manner throughout the culture, allowing stricter control of promoter induction, and the use of very low levels of IPTG to slow expression of recombinant protein.

Amino acids can be encoded for by more than one codon in the mRNA. Different organisms will favour the use of some codons for an amino acid, resulting in a bias in the levels of tRNA for particular codons (see the appendix). The codons used infrequently in a certain organism are called rare codons, and are translated inefficiently, due to the low level of corresponding tRNA. The *E. coli* strain Rosetta Rosetta-gami 2 have extra copies of tRNA genes *argU*, *argW*, *ileX*, *glyT*, *leuW*, *proL*, *metT*, *thrT*, *tyrU*, and *thrU* on pRARE, a pACYC184-derived plasmid (Novy *et al.*, 2001). These genes encode the tRNA for codons not abundantly found in *E. coli* mRNA; AUA, AGG, AGA, CUA, CCC, and GGA allowing the ribosome to translate the eukaryote mRNA more efficiently. Translation pausing of the ribosome, at regions of mRNA that contain rare codons may result in premature translation termination, translational stalling, translation frameshifting and misincorporation of amino acids. These events can lead to low yield and misfolding of recombinant proteins (Looman *et al.* 1987; Makhoul and Trifonov, 2002; Kurland and Gallant, 1996; McNulty *et al.*, 2003). The codons of the hMCM sequence were analysed at the database on the website http://www.kazusa.or.jp/codon/ (see appendix). A total of 25 codons are present in hMCM that the tRNA is found below 0.5% frequency in *E. coli.*

E. coli ER2566 is an expression strain provided by New England Biolabs with the vectors pTYB4 and pTYB11, and is the recommended expression strain for both these vectors. (Lüttkopf *et al.*, 2001). The cells *E. coli* ER2566 cells are deficient in Ion and OmpT proteases and contain the λ prophage with an IPTG inducible T7 RNA polymerase gene.

1.6.5 Solubility Tags

Many proteins formerly used as affinity purification or detection tags were also found to enhance expression levels and solubility of the target protein, so are now also referred to as solubility tags. A solubility tag is a second protein fused in the same open reading frame as the recombinant target protein (figure 1.8). The open reading frames of both will now be expressed as one polypeptide, and after folding and purification the polypeptide can be separated to yield two polypeptides i.e. the target protein and the solubility tag see section 1.7 on 'purification and processing of recombinant proteins'. Sometimes (rarely) the target protein may aggregate after cleavage even if it is correctly folded as a fusion. (Sati *et al.*, 2002).

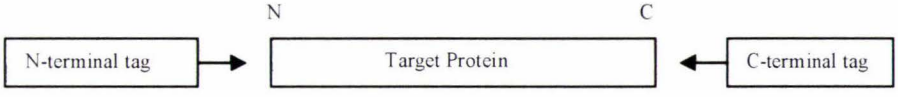

Figure 1.8 Showing Possible Positions of a Solubility Tags Relative to the Target Protein. The tag can be expressed on the C or N termini of the target protein. The N-terminus of proteins is the most common position for a solubility tag.

To be a suitable solubility partner a protein must be highly soluble in the *E. coli* cytoplasm. However, not all highly soluble proteins are affective solubility tags (Kapust and Waugh, 1999). It is also unlikely that all solubility tags act in the same way, different target and solubility tag combinations will affect the solubility of the target protein differently (Kapust and Waugh, 1999). Although solubility tags are now widely used in recombinant systems, the mechanisms by which some of them prevent or reduce aggregation are not well understood due to the lack of understanding about the formation of protein aggregates in cells.

The 40 kDa maltose binding protein (MBP) (Pryor and Leiting, 1997) has been shown to be a very effective solubility tag. This was demonstrated in a study of the effect three different solubility tags had on six different target proteins, all insoluble and inactive when expressed in *E. coli*. The three solubility tags tested were Trx, MBP and glutathione S-transferase. The effectiveness of the solubility tag was measured by the solubility and activity of each target protein. The MBP was consistently found to be the

most effective tag at solubilising the target protein (Kupust and Waugh, 1999). These results have been supported by other labs (Goulding and Perry, 2003). This led the authors to suggest that the MBP is a passive general chaperone, binding to unfolded proteins to promote folding, perhaps by preventing self association, (figure 1.9).

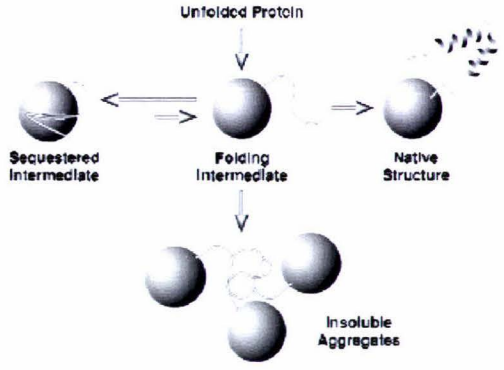

Figure 1.9 A Model of the Proposed Mechanism of MBP. The sphere represents the MBP and the line and the helix represents the unfolded and folded protein, respectively. (Figure adapted from Kapust and Waugh, 1999)

In support of this mechanism is the ineffectiveness of the MBP as a solubility tag when fused to the C-terminus of the target protein. This arrangement results in the target protein being synthesised and folding earlier as it comes of the ribosome, (Gilbert *et al.* 2004), suggesting that the MBP must be fully folded to be an effective solubility tag (Sachdev and Chirgwin, 1998). The MBP has also been shown to interact with unfolded polypeptides *in vitro*, promoting folding of these unfolded proteins (Richarme and Caldas, 1997). Using the maltose sugar immobilised on resin, the MBP is also affective as an affinity purification system, and as an affinity tag, the MBP is typically placed at the C-termini of the target protein (Maina *et al.*, 1988; Hennig and Schafer, 1998).

The 11 kDa thioredoxin protein is a commonly used solubility tag and dramatically increases the solubility of many recombinant proteins expressed in the *E. coli* cytoplasm (LaVallie *et al.*, 1993). It can be effective placed at the N or C terminus of the target protein, varying from case to case, depending on the target protein. However it is most commonly used as a solubility tag at the N-termini of the target protein. Because thioredoxin lacks a convenient immobilisable small ligand, it is unsuitable for an 'affinity purification system', therefore thioredoxin is often used in conjunction with a

small affinity tag such as the His_6 tag, and this thioredoxin-His construct is referred to as the Trx tag.

1.7 Purification and Processing of the Recombinant Protein

1.7.1 Poly His Tag Purification

A poly-histidine tag is a sequence of four to ten contiguous histidines within a protein, which selectively binds to a Ni column via the co-ordination of the histidine imidazole groups to the Ni metal. The protein is eluted by changing the pH, or adding free imidazole to the chromatography buffer (Hochuli *et al.*, 1987). The chromatography resin has an attached nitrilotriacetic acid (NTA) group, ideal for binding metal ions with co-ordination numbers of six, such as Ni and Co (Hochuli *et al*, 1988). The His tag can be placed at the N or C termini of the protein, depending on the structure of the target protein. Purification of his-tagged proteins can also allow purification under denaturing conditions. However, proteins with a metal cofactor bound are not recommended, as the potential cofactor- NTA resin interaction may interfere with the stability of the protein (Reviewed by Terpe, 2003).

1.7.2 The Chitin Binding Domain

The 5 kDa chitin-binding domain is the 51 C-terminal amino acids of chitinase A1 from *Bacillus circulans* WL-12 (Watanabe *et al.*, 1994). The chitin binding properties of the domain are exploited, and purification is achieved by using a column with immobilised chitin (Chong *et al.*, 1997).

1.7.4 Site specific Proteases

Site specific proteases recognise a certain sequence of amino acids, cleaving the polypeptide. This has been exploited to allow the separation of affinity/solubility tags from the target protein, (figure 10).

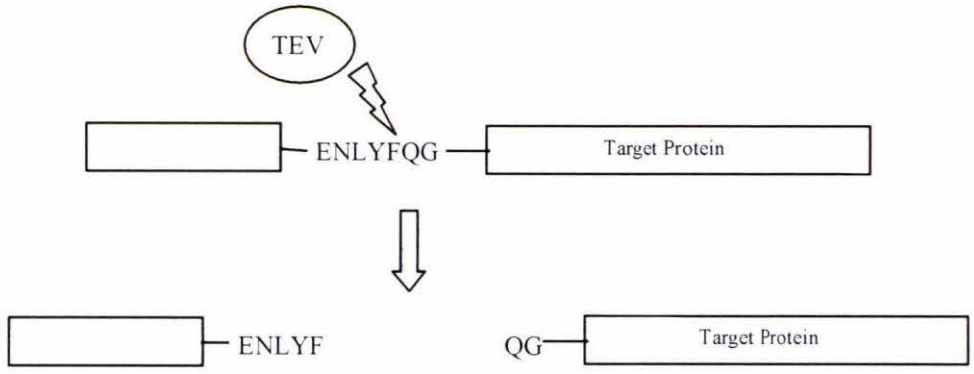

Figure 1.10 Protease Cleavage of Solubility Tags. A sequence of amino acids that are the target for a protease (in this example TEV) can be inserted between the target protein and the tag. Incubation with the protease separates the tag from the target protein. The protease must have access to the sequence so it must be on the exterior of the proteins structure.

Enterokinase is a site specific protease that cleaves after lysine at its recognision site Asp-Asp-Asp-Asp-Lys, and is frequently used to separate tags from the target protein (Sun *et al.*, 2005). In some target protein-tag fusions the enterokinase cleavage site is not accessible and cleavage can be inefficient or completely prevented. In the expression system of hMCM-Trx, inefficient enterokinase cleavage occurred, suggesting that the protease was unable to access the cleavage site (Janata *et al.*, 1997).

The Tobacco Etch Virus Protease (TEV) protease is a 48 kDa site specific protease, from the tobacco etch virus (Carington and Dougherty, 1988). His-tagged TEV can be expressed and purified in *E. coli* making it a cheap and convenient source of protease. A TEV site, (ENLYFQG), has been engineered into the pET32a vector (Rosemary Brown, *Pers. Comm.*), to allow removal of the Trx solubility tag by digesting the purified protein complex with the recombinant TEV protease. This can be used as an alternative to the enterokinase; the differences between the two proteases may allow the TEV to effectively cleave the Trx tag. In addition to changing the protease, the sequence immediately either side of the site is also different which may allow better access for the TEV.

1.7.5 Inteins

The use of inteins eliminates the need to incubate the recombinant protein with a protease to cleave the target protein from purification and/or solubility fusion tags. Conventional proteases can have drawbacks as seen in the earlier section 1.3.3 'site specific proteases'. Access to the recognition site, protease sites in the target protein and further purification are required. Often addition or change of amino acids at the N or C terminus is necessary to generate the target sequence of the protease, resulting in final protein product containing several additional or changed amino acids and this can affect the protein activity and folding. The protease may require incubation at elevated temperatures for activity, which can degrade or denature the target protein (Janata *et al.*, 1997; Humphries *et al* 2002).

Inteins perform a post translational modification, excising an internal polypeptide sequence (intein), and ligating the flanking sequences (exteins) with a peptide bond (figure 1.11). Key conserved residues have been identified by amino acid sequence alignment and then functional studies of the effects of point mutations on cleavage to determine their role in the splicing reaction (Xu and Perler, 1996).

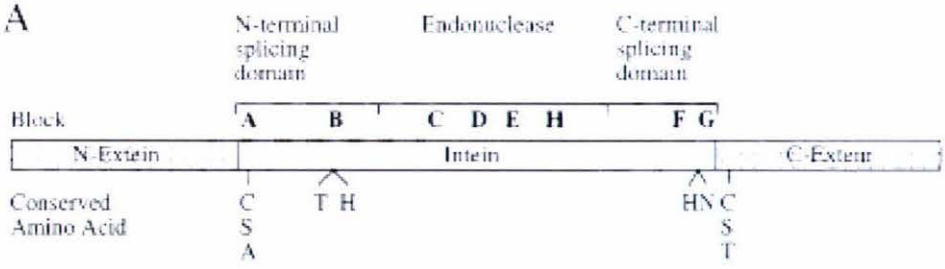

Figure 1.11 Domains and Conserved Residues in Inteins. Blocks A and B are important in the N-terminal splicing, and blocks F and G are involved in the C-terminal splicing. The region C to H can contain an endonuclease in some inteins. Also note the arrangement of the intein, N-extein and C-extein relative to each other. Shown below are the highly conserved amnio acids.

(Figure adapted from Evans and Xu, 2001)

The mechanism involves a peptide bond within the protein being attacked and broken resulting in the formation of a thioester intermediate. This mechanism is supported by the formation of a branched intermediate with two N-termini (figure 1.11) (Xu *et al.*

1993). Intein activity may be compared to that of the spliceosome that removes introns from RNA (Evans and Xu, 2002). Comparisons can also be made with protease activity, as the peptide bond is broken in a similar way, using a mechanism not unlike that of the 'catalytic triad' that many proteases use. The reaction is also analogous to the native peptide ligation reaction, used for assembling small synthetic peptides into larger synthetic proteins (Dawson *et al.*, 1994)

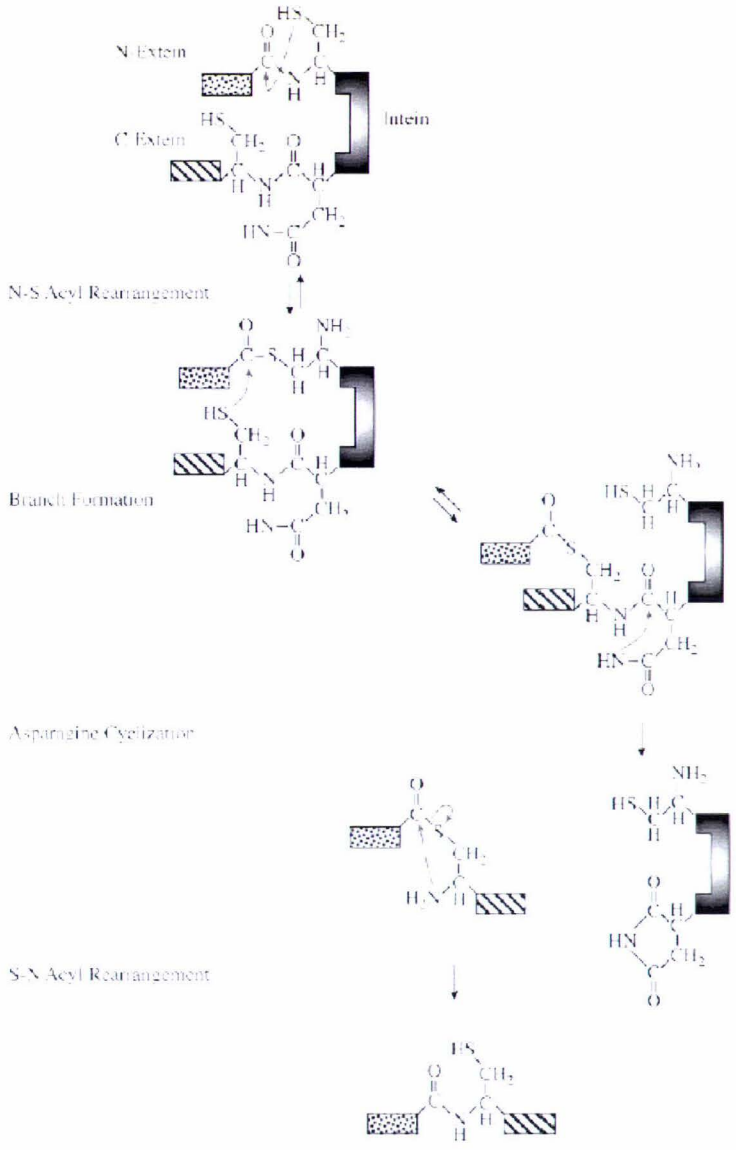

Figure 1.12 Mechanism of the Intein *Sce* VMA1 Splicing Reaction. Attack of the intein N-terminal cysteine sulfhydryl group initiates a N-S acyl rearrangement, which generates a thioester bond between the intein N-terminal cysteine residue and the N-extein C-terminal residue. Trans(thio)esterification of the N-extein acyl group to the C-extein cysteine forms the branched intermediate. Attack of an intein C-terminal asparagine excises the intein , and the reaction is completed by a spontaneous S-N acyl rearrangement, regenerating a native peptide bond between the N- and C-extein segments.

(Figure adapted from Evans and Xu, 2001)

The intein *Sce* VMA1 is from the yeast *Saccharomyces cerevisiae* and was discovered in 1990 (Hirata *et al.*, 1990; Kane *et al.*, 1990). Researchers at New England Biolabs then removed endonuclease domain (figure 1.13) at the gene level and mutated key catalytic residues to prevent the exteins religating. Although inteins lacking the

endonuclease domain (mini inteins) are now known to occur naturally, the *Sce* VMA1 intein without the endonuclease domain was artificially created, before mini inteins were discovered in nature.

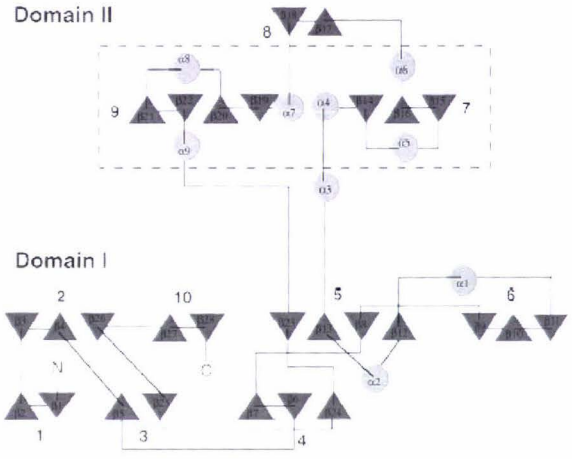

Figure 1.13 The Domain Structure of the Native *Sce* VMA1 Intein. The circles and triangles represent α helices and β strands, respectively. Numbers identify the ten different β sheets. Domain I is the nuclease domain identified by the DNA binding motifs, and domain II is thought to contain the protease activity.

(Figure adapted from Duan *et al.* 1997)

The changes to this intein allowed researchers at New England Biolabs to construct a series of vectors that contain a CBD (Watanabe *et al.*, 1994) as a purification tag, which can be removed by the *Sce* VMA1 intein that is located between the tag and the protein of interest. The CBD is inserted into a loop of the intein, and does not affect its activity. The vector pTYB4 encodes a splicing defective mutant *Sce* VMA1 intein (Asn454Ala) that only performs the N-terminal junction cleavage. The C-terminal splicing activity has been lost (Chong *et al.* 1996). Another vector in the series, pTYB11, encodes another splicing defective mutant *Sce* VMA1 intein (His453Glu) mutation at the penultimate amino acid resulting in defective C-terminal splicing activity (Chong *et al.* 1996). This allows the solubility tag to be removed from the N termini of the target protein, (figure 1.14).

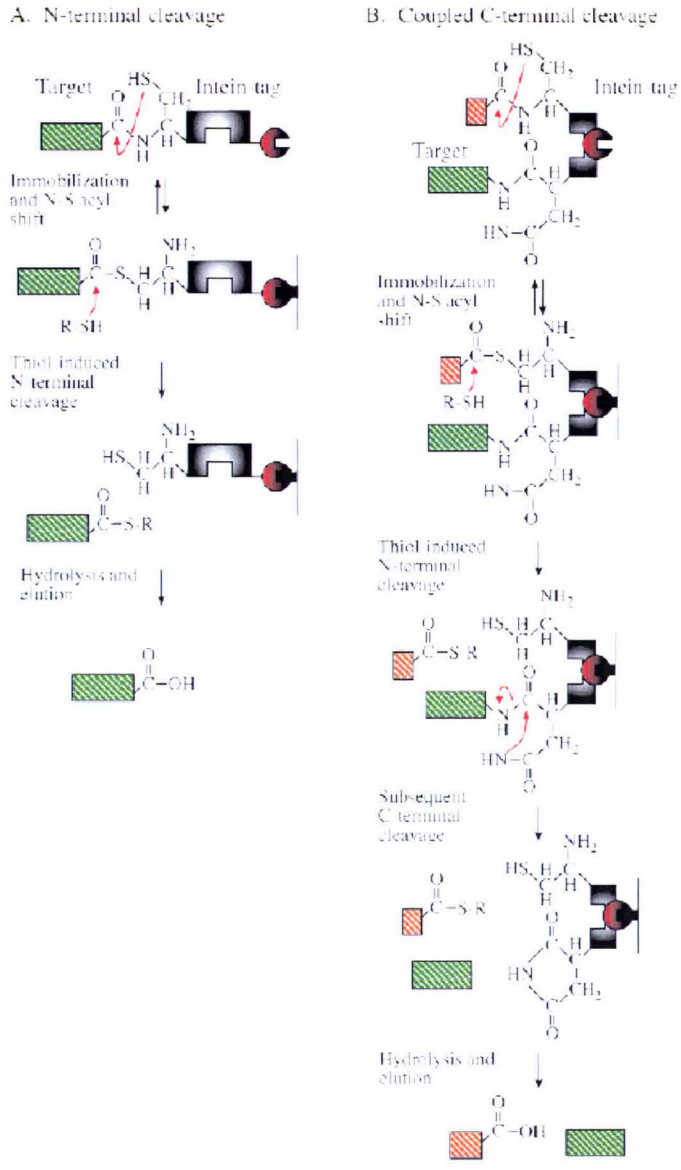

Figure 1.14 Mechanisms of Intein-Based Protein Purification.

(A) N-terminal cleavage, by the modified intein. The intein-tag is fused to the C-terminus of the target protein at the gene level. The target is released by cleavage of the bond between the target protein and the intein. This is the modified intein present in the pTYB4 vector.

(B) Coupled C-terminal cleavage by a different modified intein. This type of intein-tag allows expression and purification of the target protein with an N-terminal affinity or solubility tag, This is the modified intein present in the vector pTYB11.

(Figure adapted from Evens and Xu, 2002)

The vectors pTYB4 and pTYB11 allow the separation of the target protein under reducing conditions (figure 1.14). The target protein, intein and any solubility or purification tags are bound to a purification column or beads though an affinity tag. The

other proteins from the expression system cell lysate are washed off in a mild buffer. The target protein is then removed with a reducing agent (e.g. DTT or β-mercaptoethanol) that induces the intein to self cleave. The target protein is eluted from the column leaving the purification/solubility tags bound (Perler, 2000). Dialysis is necessary to remove a small peptide generated during the intein cleavage of proteins expressed from the vector pTYB11.

The intein leaves no amino acids on the target protein, however sometimes 1-2 amino acids must be changed at the cleavage site on the target protein, to ensure that cleavage occurs when desired, and that *in vivo* cleavage is avoided. The flexibility of the amino acids at the intein-target protein boundary means it is often possible to use amino acids with similar properties to replace the original native amino acids. Amino acids at the protease sites are very specific sequences, with little variability allowed. These amino acid changes, necessary to generate the protease site, can result in undesirable changes to the amino acids at the termini of the target protein. These amino acids may change the native sequence of the target protein and this may affect the structure and/or function of the target protein.

1.8 Chapter Summary

Studies of human methylmalonyl CoA mutase (hMCM) are required to better understand the unusual carbon skeletal rearrangement this enzyme performs, the catalysis involving a B_{12} derived cofactor, and the disease methylmalonic acidemia. Homology models based on the crystal structure of the α-subunit of a bacterial orthologue have provided a structural model. However, the decrease in homology between the bacterial and human proteins at the N-terminus results in inaccurate models of this putative dimerisation domain. Structural studies are required, and to provide a source of protein for these studies, an *E. coli* expression system is proposed.

Expression of native hMCM in *E. coli* results in insoluble inclusion bodies. A Trx solubility tag increased the yield of soluble protein but did not allow separation from the tag with the enterokinase site located between the Trx and hMCM. An alternative to protease separation of solubility tags is an intein cleavage system. Inteins may circumvent the problems associated with conventional *in trans* protease removal of tags.

1.9 Aims of this Project

This work aimed to develop a novel vector to enable expression of soluble active hMCM for structural studies. This overall goal can be subdivided into three aims:

A. Design and construct four novel expression vectors containing a solubility tag and an intein mediated removal of that tag.

B. Clone hMCM into each of these novel vectors.

C. Test the solubility of protein expressed from each of these vectors with the insoluble protein hMCM, and determine the success of the intein mediated cleavage.