Structural Characterization of 3-Dehydroquinate Synthase II

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

Majoring in Biochemistry

At the Institute of Fundamental Sciences, Massey University, Turitea, Palmerston North, New Zealand

Jack Somerton

2016

Acknowledgements

I would like to thank my co-supervisors, Associate Professor Gill Norris and Doctor Mark Patchett, and the Institute of Fundamental Sciences, Massey University, for the opportunity to undertake such a challenging, thought-provoking and rewarding experience.

To Associate Professor Gill Norris for her knowledge, encouragement and direction throughout my two years of study. For generously going out of your way when I needed it the most and for hosting the many X-lab functions.

To Doctor Mark Patchett for his abundance of enthusiasm, knowledge and the many interesting discussions. For always having a solution to dead-end problems and for being a calming influence.

To Trevor Loo for teaching me the basics of life in the lab and having the patients for dealing with my many queries.

To all my X-lab colleagues of whom have shared this experience with me. Everlasting laughs, confusion, excitement and friendships.

Thank you to my friends and family for supporting me during all my years of study. I cannot put into words how grateful I am.

Abstract

Aromatic amino acids tryptophan, tyrosine and phenylalanine are derived from a common precursor, chorismate, which is produced by the seven-step shikimate pathway in plants, fungi, *Bacteria* and *Archaea*. In *Archaea* the shikimate pathway typically begins with the alternative substrates, L-aspartate semialdehyde and 6-deoxy-5-ketofructose-1-phosphate, and so requires different enzymes to catalyse the first two steps in the pathway compared to those used in the (common) bacterial pathway. The archaeal enzyme for the second step, 3-dehydroquinate synthase type 2 (DHQS2), catalyses the oxidative deamination of 2-amino-3, 7-dideoxy-D-threo-heptoulsonic acid (formed in step 1) followed by cyclisation to produce 3-dehydroquinate, at which point the alternative and common shikimate pathways converge. No DHQS2 structures have yet been determined, and because DHQS2 enzymes have little sequence homology with their DHSQ1 analogues, they may have a novel fold. Bioinformatic methods were used to predict the solubility, stability and likelihood of sequenced DHQS2s to form crystals, and the five highest ranked were chosen for structural studies. *Methanococcus maripaludis, Desulfatibacillum alkenivorans*, *Methanospirillum hungatei*, and *Archaeoglobus veneficus* DHQS2 open reading frames were amplified by PCR and cloned into a modified pETite32a(+) vector in order to produce recombinant protein with an N-terminal, a C-terminal or no His_8-tag. Soluble recombinant DHQS2 proteins were produced in *Escherichia coli* DL41 (DE3), then purified by immobilized metal-ion affinity chromatography followed by size exclusion chromatography. C-terminally-tagged *M. maripaludis* DHQS2 with bound cofactor NAD^+ crystallised in two conditions: (i) 1.0 M ammonium sulfate, 0.1 M Bis-Tris pH 5.5 with 1% (w/v) PEG 3350; and (ii) 0.1 M CAPS at pH 10.5 with 40% (v/v) 2-methyl-2,4-pentanediol, but unfortunately the crystals were not of diffraction quality. Structure prediction using bioinformatic tools and/or far and near Circular Dichroism spectroscopy indicated that recombinant *M. maripaludis* DHQS2 was likely to have a secondary structure dominated by α-helices and had tertiary structure; recombinant *A. veneficus* was likely to have a secondary structure dominated by β-strands and had tertiary structure, while recombinant *D. alkenivorans*, and *M. hungatei* were more likely to assume a molten globule structure dominated by β-strands.

Table of Contents

Abbreviations

ADH	2-amino-3,7-dideoxy-D-threo-hept-6-ulsonic acid
ADHS	2-amino-3,7-dideoxy-D-threo-hept-6-ulsonic acid synthase
AmBic	Ammonium bicarbonate
AS	Anomalous scattering
ASA	L-Aspartate semi-aldehyde
BPB	Bromophenol blue
BSA	Bovine serum albumin
C-*Ave*DHQS2	C-terminally His_8-tagged *Archaeoglobus veneficus* 3-Dehydroquinate synthase type 2
CD	Circular dichroism
C-*Dal*DHQS2	C-terminally His_8-tagged *Desulfatibacillum alkenivorans* 3-Dehydroquinate synthase type 2
C-*Mhu*DHQS2	C-terminally His_8-tagged *Methanospirillum hungatei* 3-Dehydroquinate synthase type 2
C-*Mma*DHQS2	C-terminally His_8-tagged *Methanococcus maripaludis* 3-Dehydroquinate synthase type 2
C-*Mst*DHQS2	C-terminally His_8-tagged *Methanosphaera stadtmanae* 3-Dehydroquinate synthase type 2
Cryo-EM	Cryo-Electron Microscopy
C-*Saz*DHQS2	C-terminally His_8-tagged *Sulfurihydrogenibium azorense* 3-Dehydroquinate synthase type 2
C-score	Confidence score
DAH7P	3-Deoxy-D-arabino-heptulsonic acid 7-phosphate
DAH7PS	3-Deoxy-D-arabino-heptulsonic acid 7-phosphate synthase
DDH	3,7-Dideoxy-D-threo-hepto-2,6-diulosonic acid
DHQ	3-Dehydroquinate
DHQS	3-Dehydroquinate synthase
DHQS2	3-Dehydroquinate synthase type 2
DHS	Dehydroshikimate
DKFP	6-Deoxy-5-ketofructose 1-phosphate
DMSO	Dimethyl sulfoxide

DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotides
DTT	Dithiothreitol
E4P	Erythrose 4-phosphate
EP	Expert pool
EDTA	Ethylenediaminetraacetic acid
ESI	Electrospray ionization
F-1,6-P	Fructose-1, 6-phosphate
G6P	Glucose-6-phosphate
gDNA	Genomic DNA
HCl	Hydrochloric acid
His_8-tag	Eight histidine tag
HMMs	Hidden Markov models
HPAP	Hydroxypyruvaldehyde phosphate
IMAC	Immobilisation metal ion affinity chromatography
IOR	Indolpyruvate oxioreductase
IPTG	Isopropyl-β-D-thiogalactopyranoside
I-TASSER	Iterative threading assembly refinement
LB	Luria broth
MAD	Multi-wavelength anomalous dispersion
MCS	Multiple cloning site
MGS	Massey genome service
MIR	Multiple isomorphous replacement
ModpETite	Modified pETite vector
NAD^+	Nicotinamide adenine dinucleotide
N-*Ave*DHQS2	N-terminally His_8-tagged *Archaeoglobus veneficus* 3-Dehydroquinate synthase type 2
N-*Dal*DHQS2	N-terminally His_8-tagged *Desulfatibacillum alkenivorans* 3-Dehydroquinate synthase type 2
N-*Mhu*DHQS2	N-terminally His_8-tagged *Methanospirillum hungatei* 3-Dehydroquinate synthase type 2
N-*Mma*DHQS2	N-terminally His_8-tagged *Methanococcus maripaludis* 3-Dehydroquinate synthase type 2

NMR	Nuclear magnetic resonance
N-*Mst*DHQS2	N-terminally His_8-tagged *Methanosphaera stadtmanae* 3-Dehydroquinate synthase type 2
N-*Saz*DHQS2	N-terminally His_8-tagged *Sulfurihydrogenibium azorense* 3-Dehydroquinate synthase type 2
OD_{600}	Optical density at 600 nm
PABA	*p*-aminobenzoic acid
PCR	Polymerase chain reaction
PDB	Protein data bank
PEP	Phosphoenol pyruvate
PONDR	Predictor of natural disordered regions
SCRs	Structurally conserved regions
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
Se-Met	Seleno-methionine
SOB	Super optimal broth
TCEP	Tris-(2-carboxyethyl)phosphine
TEV	Tobacco etch virus
T_m	Melting temperature
VRs	Variable regions

List of Figures

List of Tables

Chapter 1: Introduction

1.1 The Shikimate Pathway

All organisms require aromatic compounds such as aromatic amino acids, folate, ubiquinone, and vitamins K and B. Plants, fungi and microorganisms synthesize these aromatic compounds *via* the shikimate pathway (1). Mammals lack the shikimate pathway even though they are essential to sustain life, and so must acquire these aromatic compounds from their diet. This pathway is essential for the viability of bacteria since mutation or deletion of any shikimate pathway genes results in strains that are unable to survive (2). The shikimate pathway comprises of seven enzymatic steps for the conversion of the primary metabolites phosphoenol pyruvate (PEP) and erythrose 4-phosphate (E4P) to chorismate, which is the last common metabolite in the biosynthetic pathways of various aromatic compounds. The absence of the shikimate pathway in mammals has meant that it has become a target for herbicides and antibiotics. For example, glyphosate a common ingredient in herbicides inhibits plant (3) and microorganism (4) growth but is non-toxic towards mammals.

1.2 An Alternate Shikimate Pathway for 3-Dehydroquniate Biosynthesis

The alternative Shikimate pathway for *Archaea* was first discovered when researchers began to analyse the shikimate gene cluster of the archaeon *Methanococcus jannaschii.* Four out of the seven enzymes of the canonical pathway were identified by sequence similarity to their bacterial and eukaryotic counterparts. As more genomes were sequenced, gene cluster comparisons with bacterial and eukaryotic shikimate pathways allowed identification of the missing genes (5), (6). This approach was used to identify an open reading frame (ORF) RMJ07785 in *M. jannaschii* that encodes a GHMP-kinase family related shikimate kinase which has no sequence similarity to either bacterial or eukaryotic shikimate kinases (7). Similarly, the ORFs encoding the enzymes responsible for the first two enzymatic steps of the shikimate pathway had very low sequence similarity to those of their prokaryotic counterparts. As the organism maintained the ability to synthesize aromatic amino acids it became obvious that for this organism at least, the pathway must be initiated from a different metabolite (8). Homology searches comparing the shikimate pathway gene clusters of several archaea*; M. jannaschii* (9), *Archaeoglobus fulgidus* (10) , *Methanococcus maripaludis* (11),

Aquifex aeolicus (12), *Pyrococcus abyssi* and *Pyrococcus horikoshii* (5), showed that the organisms used some different steps (enzymes) compared to those of the standard shikimate pathway. For example, *P. horikoshii* appeared to lack an aromatic amino acid biosynthetic pathway while most of the enzymes in the pathway were found in the genome of *P. abyssi* (8). The first two genes coding the enzymes of the shikimate pathway in *M. jannaschii* showed that the *aro*G/F/H and *aro*B1 genes were "missing" from the genome. In their place were two non-homologous genes since named, *aroA'* and *aroB'*. These encode the first two steps of a non-canonical archaeal shikimate pathway where the first common product, 3-dehydroquintae (DHQ), continues through the pathway in a canonical fashion (Fig. 1.1). *AroA'* encodes the enzyme 2-amino-3,7-dideoxy-D-threo-hept-6-ulosonic acid synthase (ADHS) in *M. jannaschii*, which catalyses the initial step of the pathway, a transaldol condensation between 6-deoxy-5-ketofructose 1-phosphate (DKFP) and L-aspartate semi-aldehyde (ASA) resulting in the loss of hydroxypyruvaldehyde phosphate (HPAP) and the formation of 2-amino-3, 7-dideoxy-D-threo-hept-6-ulsonic acid (ADH).

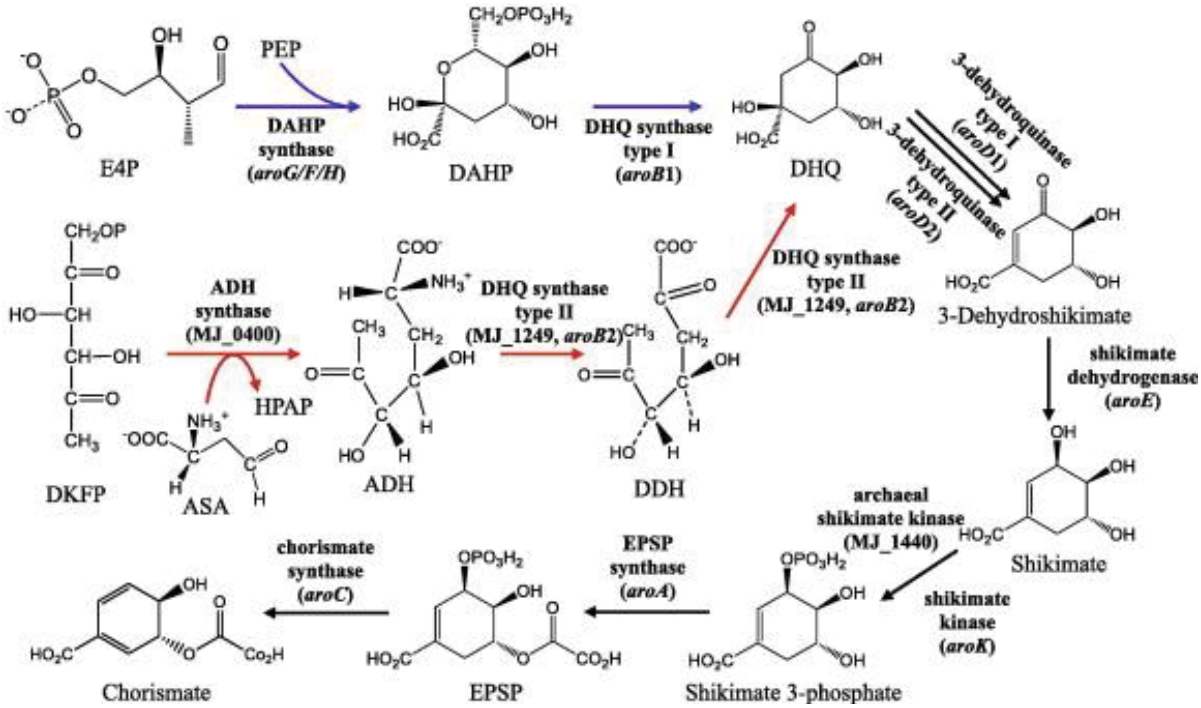

Fig. 1. 1. An alternate shikimate pathway towards 3-Dehydroquinate biosynthesis. Reprinted with permission from *Molecular Phylogenetics and Evolution*, **2014**, 75, 154–164. Copyright © 2014 Elsevier. Schematic representation of the shikimate pathway. The blue arrows define the standard shikimate pathway, the red arrows the alternative pathway.

In comparison, the canonical pathway uses 3-deoxy-D-arabino-heptulsonic acid 7-phosphate synthase (DAH7PS) to convert E4P and PEP into 3-deoxy-D-arabino-hept-6-ulosonate 7-phosphate (DAH7P) (Fig. 1.1).

The second canonical step is carried out by 3-dehydroquinate synthase (DHQS), whereas in *M. jannaschii*, it is carried out by 3-dehydroquinate synthase type 2 (DHQS2) using a two-step reaction. The reaction undergoes an oxidative deamination to produce 3,7-dideoxy-D-threo-hepto-2,6-diulosonic acid (DDH) followed by a cyclisation reaction to form DHQ (Fig. 1.1).

It is not clear why the standard pathway and the non-canonical pathway utilize different precursors as initial substrates for their respective pathways. At the extreme temperatures in which these archaea live, E4P, the substrate for DAH7PS which catalyses the first committing step into the shikimate pathway, is chemically unstable (13). Although archaea have the ability to synthesize E4P and PEP, [^{13}C] labelling of ribose in *M. maripaludis* was inconsistent with the formation of E4P by the carboxylation of a triose showing that E4P could not be a precursor of aromatic amino acids (14). An alternative first step was however not established although some suggestions were made.

Biosynthesis of amino acids in the halophilic archaeon *Haloarcula hispanica* was explored using [^{13}C] labelled glycerol and analysed by two dimensional [^{13}C, ^{1}H] nuclear magnetic resonance (NMR). A lack of [^{13}C] labelling of the aromatic amino acid tyrosine indicated the presence of a novel biosynthetic pathway that had so far not been characterised. In the standard shikimate pathway, E4P and PEP substrates combine to form the aromatic ring of tyrosine (15). In addition, research into aromatic acid biosynthesis in cell-free extracts of *Methanohalophilus mahii*, found no evidence of DAH7PS activity (16). These results coincided with a report by White *et al* (2004) who suggested that DKFP and ASA are alternative precursors to DHQ in the archaeon *M. jannaschii* (17). Incubation of *M. jannaschii AroA'* gene *mj0400* and *AroB'* gene *mj1249* products with E4P and PEP (the normal precursors) failed to produce DHQ or dehydroshikimate (DHS), showing that the normal shikimate pathway was not operating. To test if the next gene in the pathway was functional, DHQ was incubated with cell extract and found to be quantitatively converted into DHS, showing that the first steps in the biosynthesis produced DHQ (17). However, no detectable [^{13}C] labelled pyruvate as part of PEP was incorporated into DHS, showing that PEP and E4P could not be precursors. Incubation of cell extracts with ASA, DKFP, and

nicotinamide adenine dinucleotide (NAD^+), a common co-factor for many enzymes, on the other hand resulted in the production of DHQ. DKFP is formed by condensation of methyglyoxal with dihydroxyacetone phosphate, a compound produced from glucose-6-phosphate (G6P). Six-carbon chain sugars other than DKFP were then tested to see if they would condense with ASA, but DKFP was found to be the only substrate. When DKFP was incubated with cell extracts containing [^{13}C] G6P, label incorporation was slightly decreased indicating that DKFP may not be the preferred precursor (17). A mixture of the *mj0400* and *mj1249* gene products did however catalyse the formation of DHQ from ASA and DKFP suggesting that both genes are part of an aromatic amino acid biosynthetic gene cluster in *M. jannaschii.* Therefore, it was concluded that, DKFP undergoes an aldol condensation with ASA to form ADH *via* a transaldolase reaction catalysed by the protein product of MJ0400 in *M. jannaschii i.e.* ADHS. ADH then undergoes a NAD^+ dependant oxidative deamination to produce 3,7-dideoxy-D-threo-hept-2,6-di-ulosonic acid which is cyclized to form DHQ by the protein encoded by MJ1249 *i.e.* DHQS2.

Deletion of the DHQ dehydratase, an enzyme leading to chorismate biosynthesis in the shikimate pathway in *M. maripaludis,* (Fig. 1.1), results in an aromatic amino acid auxotroph. In this mutant, the aryl acids phenylacetate, hydroxyphenylacetate and indoleacetate can be converted to aromatic amino acids *via* indolepyruvate oxioreductase (IOR) thus regulating the *de novo* shikimate pathway (18) (Fig. 1.2). A deletion mutant of 'aroA', which encodes ADHS, the first step in the DKFP pathway, required aromatic amino acids for growth. In contrast, deletion of the 'aroB' gene, encoding the enzyme catalysing the second step of the DKFP pathway, DHQS2, did not require aromatic amino acids for growth (19). This result suggests that MMP0006, a homolog of *M. jannaschii* MJ1249, which encodes DHQS2, was not required for aromatic amino acid biosynthesis in *M. maripaludis*.

Genomic analysis of *Haloarchaea salinarum* (20) showed that its aromatic amino acid biosynthetic pathway was similar to that in *M. jannaschii*. The first two steps of the shikimate pathway are carried out by OE1472F and OE1475F, homologs of MJ0400 and MJ1249 respectively. OE1472F was classified as a member of the archaeal aldolase family based on sequence alignment and close structural similarity to the transaldolase in *M.*

jannaschii. All the active site residues identified in MJ1249 are completely conserved in OE1475F. Both *M. jannaschii* ADHS and the OE1472F protein product can bind fructose-1, 6-phosphate (F-1,6-P) in the active site (21).

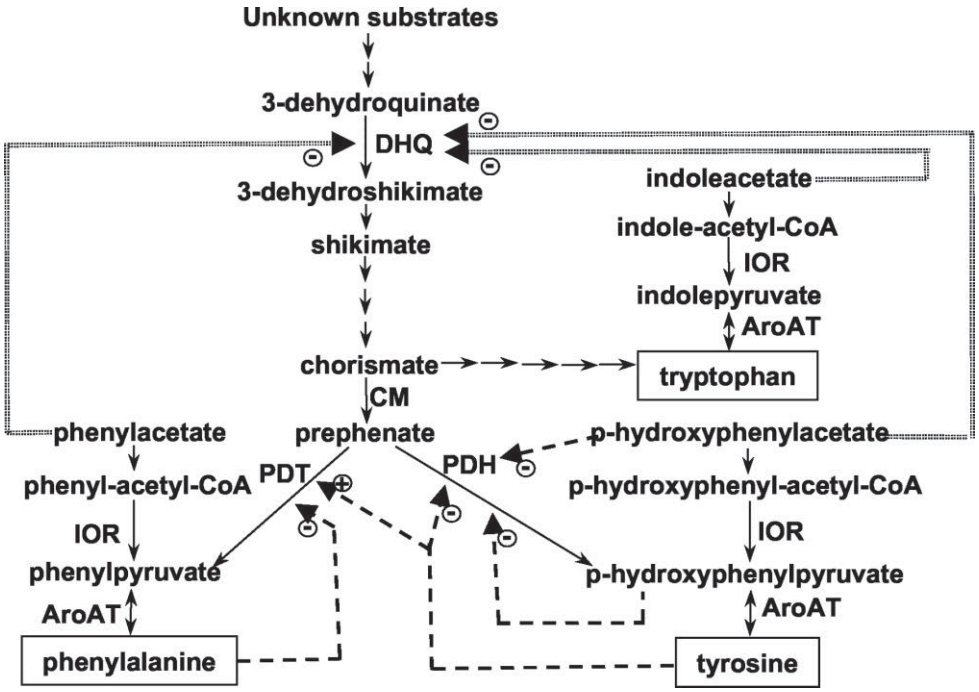

Fig. 1. 2. Pathways for the biosynthesis of aromatic amino acids in *M. maripaludis.* Reprinted with permission from *J. Bacteriol,* **2004**, 186, 4940-4950. Copyright © 2014, American Society for Microbiology. Phenylacetate, *p*-hydroxyphenylacetate, and indoleacetate, precursors for phenylalanine, tyrosine, and tryptophan, respectively. The enzymes shown are DHQ, chorismate mutase (CM), prephenate dehydratase (PDT), prephenate dehydrogenase (PDH)_, aromatic aminotransferases (AroA), and IOR. Possible transcriptional regulation of DHQ is indicated by double dashed lines. Inhibition (−) or activation (+) of the PDT and PDH enzyme activities are indicated by single dashed lines.

Interestingly, the gene product of MJ0400 does not show aldolase activity with F-1,6-P in contrast to the gene product of OE1472F which does have aldolase activity with this substrate, (21) suggesting that either ASA or F-1,6-P are possible precursors to aromatic amino acid biosynthesis in *H. salinarum*, or that the protein product of OE1472F has an additional role in the production of DKFP (22). This suggests that *Archaea* have developed an alternative strategy to produce DHQ.

As in *M. jannaschii*, the *H. salinarum* (20) MJ1249 ORF homolog OE1475F was also found to play a role in the biosynthesis of aromatic amino acids. The growth impairment of an organism containing the mutant OE1475F was restored when it was cultured in media supplemented with all three aromatic amino acids. The addition of either DHQ or shikimate in addition to aromatic amino acid supplementation stimulated the uptake of phenylalanine from the media suggesting that OE1475F might regulate specific transporters of aromatic amino acids. The positive effect of DHQ and shikimate on growth suggests that either they, or downstream metabolites, are involved in metabolic pathways other than those of aromatic amino acid synthesis (23). Interestingly, the deletion of the OE1472F gene in *H. salinarum* showed only attenuation of growth in the absence of aromatic amino acids, not cessation of growth suggesting that another gene could replace OE1472F. Alternatively a second pathway exists for aromatic amino acid biosynthesis in *H. salinarum*.

A secondary pathway has been found in *M. maripaludis* where arylamine, (a metabolite derived from *p*-aminobenzoic acid (PABA), another product that can originate from chorismate), acts as a substrate for secondary aromatic amino acid biosynthesis (18) (Fig. 1.3). PABA (derived from DHQ in the non-canonical shikimate pathway) cell extract levels are reduced when *M. maripaludis* is grown in the presence arylamine and G6P. When grown in the presence of [^{13}C] labelled G6P, the lack of labelled shikimate and DHQ showed that at least one step between the production of G6P and the production of DHQ must be as strongly regulated as DAH7PS in the equivalent standard pathway (19).

DAH7PS controls the rate-limiting step for the production of chorismate, which is then used as a substrate for the synthesis of various products. In the standard pathway, three DAH7PS isozymes are allosterically inhibited by phenylalanine, tryptophan and tyrosine respectively using a feedback mechanism (24). In the non-canonical pathway, the production of arylamine is also subject to feedback regulation in a manner reminiscent to that seen for DAH7PS, providing further evidence for its secondary role in aromatic amino acid biosynthesis

1.3 Enzymatic Reactions leading to 3-Dehydroquniate Biosynthesis

Purification, overexpression and x-ray crystallography characterization of the MJ0400 gene

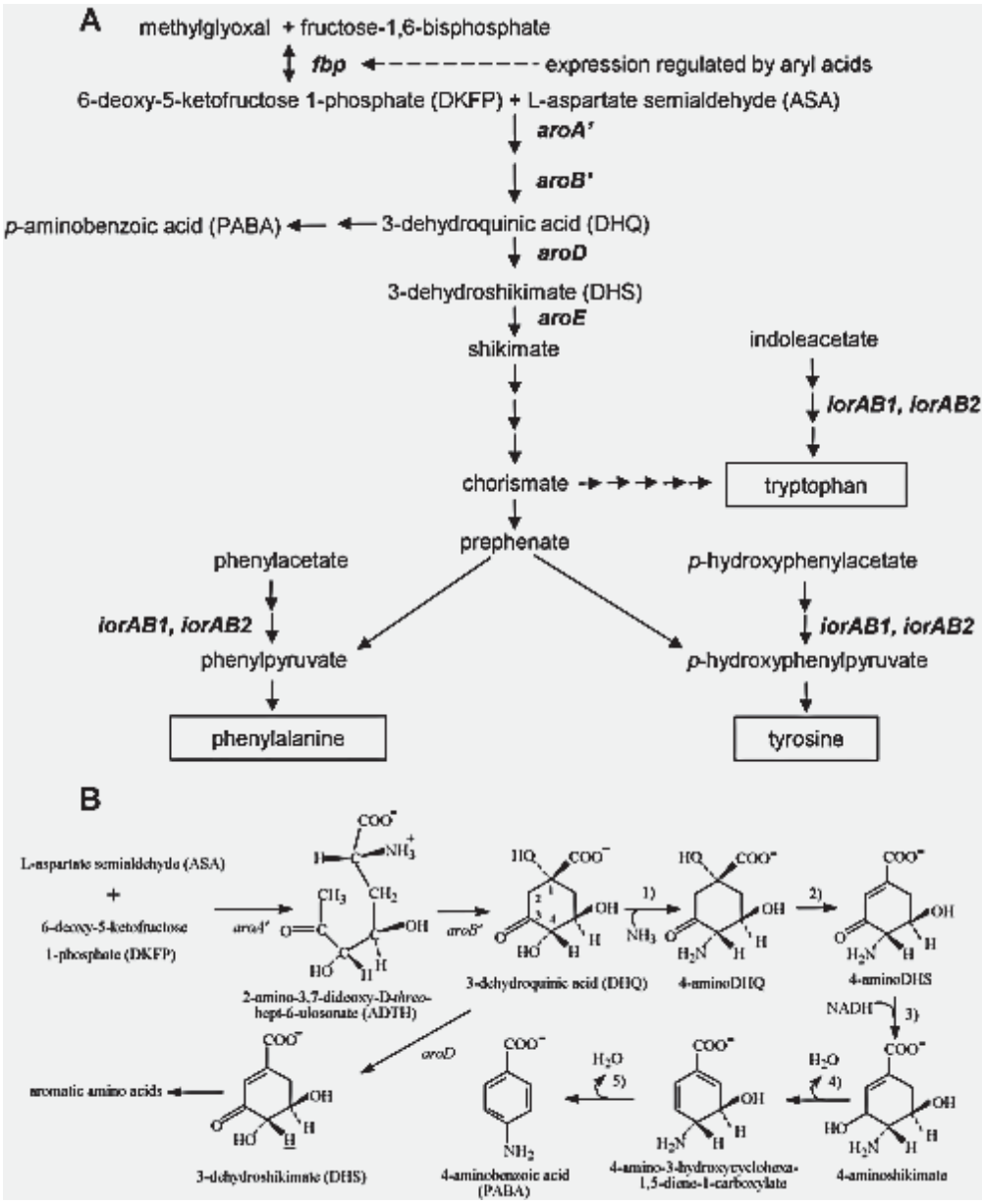

Fig. 1. 3. Pathways for the biosynthesis of aromatic amino acids and for the biosynthesis of p-aminobenzoic acid in *Methanococcus Maripaludis*. Reprinted with permission from *Molecular Microbiology*, **2006**, 62, 1117-1131. Copyright © 2006 John Wiley and Sons. Fbp, iorAB1 and iorAB2 represent fructose-1,6-bisphosphate aldolase, indolepyruvate oxioreductase I and II, respectively. Five reactions numbered: 1, 3-dehydroquinate aminotransferase; 2. 4-aminodehydroshikimate synthase; 3, 4-aminoshikimate dehydrogenase; 4, 4-aminoshikimate dehydratase and 5, 4-aminobenzoic acid synthase.

product, ADHS, from *M. jannaschii* produced a crystal structure, which gave some insight into the catalytic mechanism of the enzyme. ADHS is a member of the class 1 aldolase

superfamily and catalyses an aldol condensation reaction between DKFP and ASA resulting in the loss of HPAP. Each monomer of ADHS has a $(\beta\alpha)_8$ –barrel fold with two additional β-strands and four extra α-helices (Fig. 1.4A). Two homopentamers form the overall decameric structure, which resembles a doughnut (Fig. 1.4B). The homodecamer contains the active site on the top of the barrel and forms a covalent adduct with the substrate utilizing a strictly conserved lysine residue located on strand β6 of the barrel. In the active site, a Schiff base intermediate is formed between the εNH_2 of lysine 184 and DKFP (Fig. 1.4C). In the forward reaction, Tyr153 acts as a general acid protonating the C5-hydroxyl group of the carbinolamine intermediate and activating it as a leaving group to form the Schiff base. Asp33 acts as a general base to deprotonate the C3-hydroxyl moiety facilitating the cleavage of the C3-C4 bond. This loss of HPAP allows the resulting Schiff base intermediate to covalently bind to the aldehyde moiety of ASA to produce the product ADH (25) (Fig. 1.5). DKFP contains two carbonyl moieties capable of Schiff base formation. The C5-carbonyl forms a covalent adduct, resulting in the formation of an eneamine capable of attacking ASA. Using the ADHS−F-1,6-P structure as a model, DKFP was docked into in the active site *via* a C5-Schiff base showing that the proposed key catalytic residues are within the distance required for catalysis (Fig. 1.4C).

The crystal structure of DAH7PS from in *Thermotoga maritima*, which catalyses the first step in the standard shikimate pathway, has been solved (26). DAH7PS is a tetramer with each monomer being made up of two domains, a ferredoxin-like domain and a $(\beta\alpha)_8$ –barrel domain (Fig. 1.6A). The protein was crystallized in the presence of Cd^{2+} PEP and E4P and contains two subunits of the enzyme that associate as a dimer in the asymmetric unit (Fig. 1.6B). The two dimers form a tetramer through extensive interactions between their $(\beta/\alpha)_8$-barrel domains (Fig. 1.6C and 1.6D). As is typical for $(\beta/\alpha)_8$-barrel enzymes, the active site is located at the C-terminal end of the barrel and is formed by the residues linking the βα-loops connecting the β-strands. The active site of each barrel domain contains a bound metal cofactor, Cd^{2+} that is involved in the condensation reaction between PEP and E4P.

The reaction follows an ordered sequential kinetic mechanism where PEP binds to the enzyme first, followed by E4P. This is followed by water attack on the oxocarbenium ion,

phosphate loss, and finally DAH7P release. The loss of phosphate from PEP occurs through the cleavage of the C-O bond rather than the P-O bond of PEP. The oxygen at C2 of DAH7P is therefore obtained from water, which attacks at C2 of PEP. Glu164, Ala185, Gln182, Lys207, Arg237 and His272 position the negatively charged PEP through hydrogen-bonding and salt bridge interactions. The residues involved in E4P binding are Arg133, Thr134, Lys131, Pro132, Arg186, Asp309 and Cys102 (Fig. 1.7A).

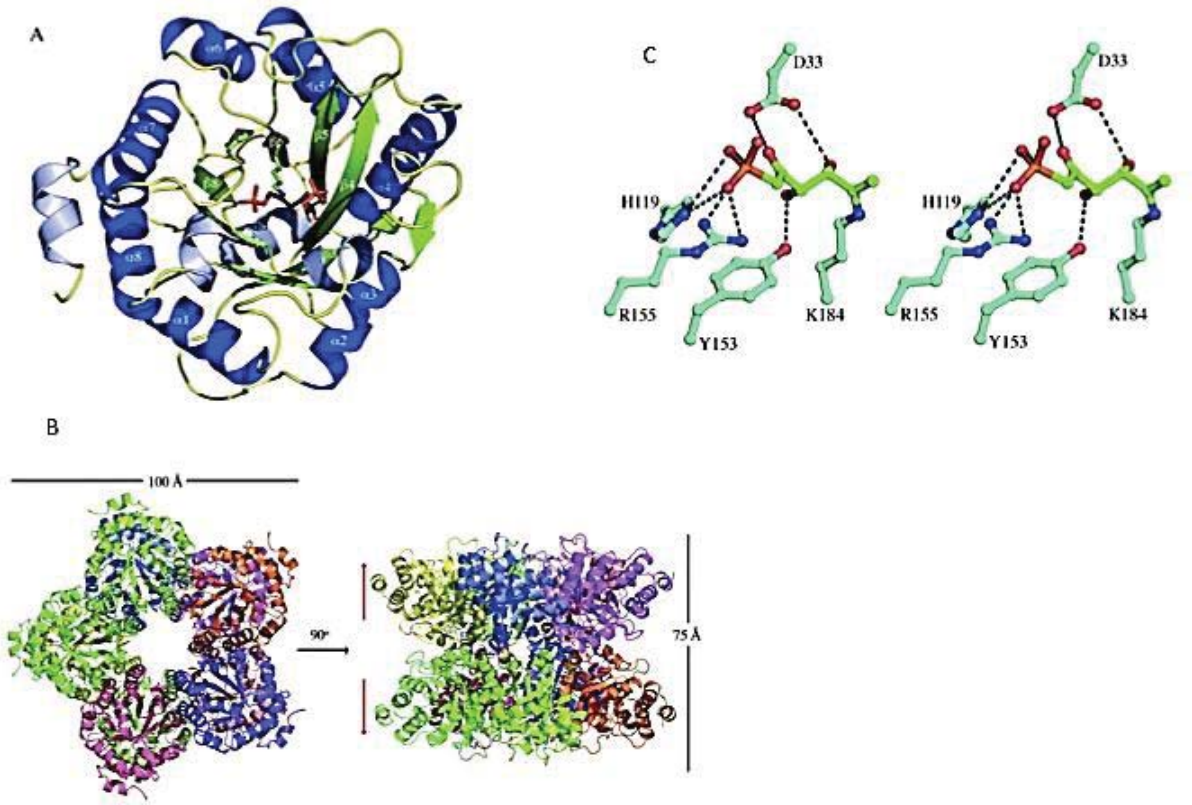

Fig. 1. 4. The Structure of ADHS in *Methanococcus jannaschii*. Adapted with permission from *Biochemistry* **2007,** 46, 10562-10571 DOI: 10.1021/bi700934v. Copyright © 2007 American Chemical Society. (A) A monomer shown as a ribbon diagram. Coloured dark green are the strands that are part of the $(\beta\alpha)_8$-barrel, and coloured dark blue are the helices that are part of the $(\beta\alpha)_8$-barrel. Coloured light green is the pair of antiparallel strands, β3a and β3b. The additional helices that are not part of the $(\beta\alpha)_8$-barrel fold are coloured light blue. The ligand, F-1,6-P is shown as ball-and-stick. (B) Decamer of ADHS. Each monomer is a different colour. Two views are shown; looking down the 5-fold axis and after a 90° rotation perpendicular to the 5-fold axis. The longest dimensions of the decamer are indicated. The red arrows represent the axis directionality of the $(\beta\alpha)_8$-barrels in the respective pentamers. (C) Diasteric views of DKFP Schiff base intermediate in the active site of *M. jannaschii* ADHS. Residues within hydrogen bonding distance of the ligand are shown.

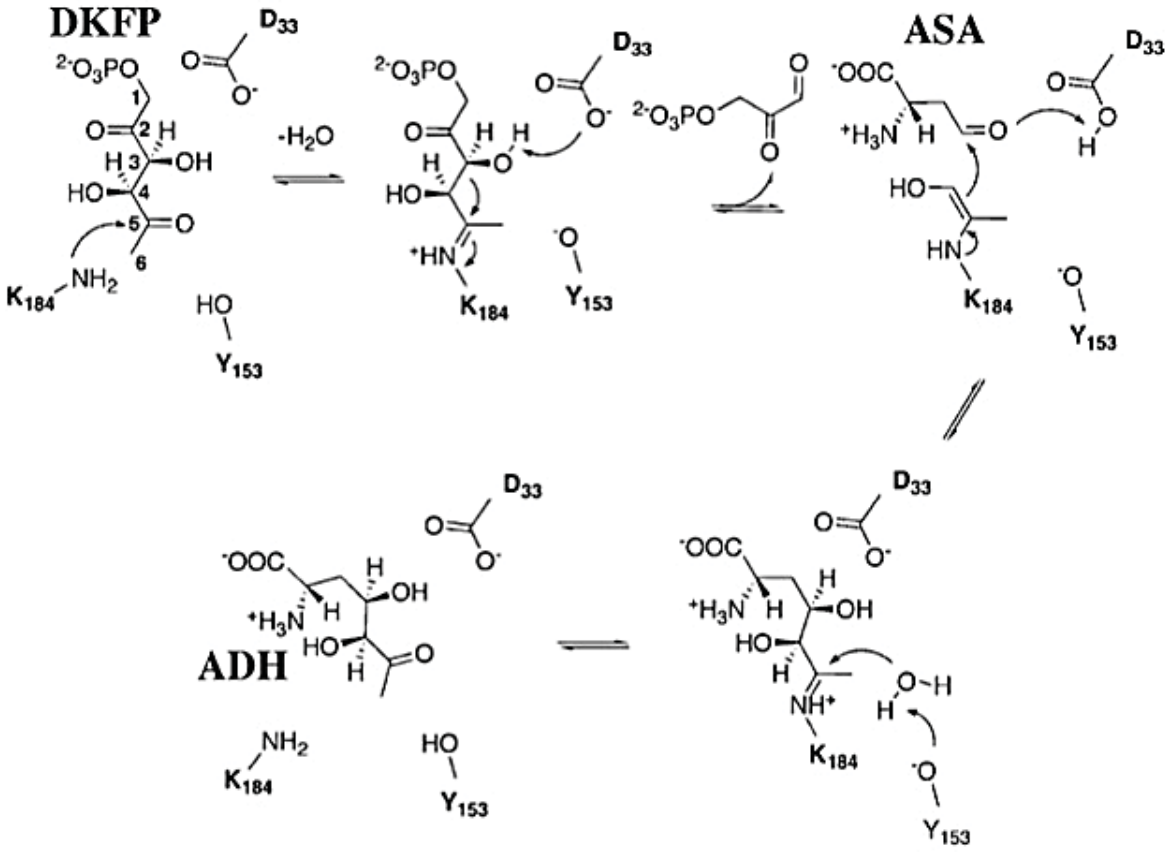

Fig. 1. 5. Catalytic mechanism of ADHS. Adapted with permission from *Biochemistry* **2007,** 46, 10562-10571. Copyright © 2007 American Chemical Society. ADHS catalyses the cleavage of DKFP in the active site utilising a class 1 aldolase mechanism. Tyr153 acts as the general acid protonating the C5-hydroxyl of the carbinolamine intermediate and activating it as a leaving group in the formation of the Schiff base in the forward reaction. Then, Asp33 acts as a general base to deprotonate the C3-hydroxyl moiety facilitating cleavage of the C3-C4 bond. HPAP dissociates from the active site allowing the formation of a carbon-carbon covalent bond between the Schiff base intermediate and the aldehyde moiety of ASA.

In the active site, Cd^{2+} coordinates with E4P, PEP, the catalytic waters Wa and Wb, and residues Glu164 and Lys131 (Fig. 1.7B). The crystal structure of DHQS shows that it assembles as a dimer of dimers. Each monomer is made up of an N-terminal α/β domain and a C-terminal α-helical domain. The N-terminal domain has a Rossman fold but one that is in an inverted orientation unlike the standard Rossman fold (Fig. 1.8). The C-terminal domain has no detectable structural relationship to other known structures. As usual, the active site is located in a cleft between the two domains. The C-terminal domain contains most of the residues involved in catalysis as well as substrate and metal ion binding sites. The metal ion,

in this case Zn^{2+}, interacts with residues Glu194, His271 and His287.

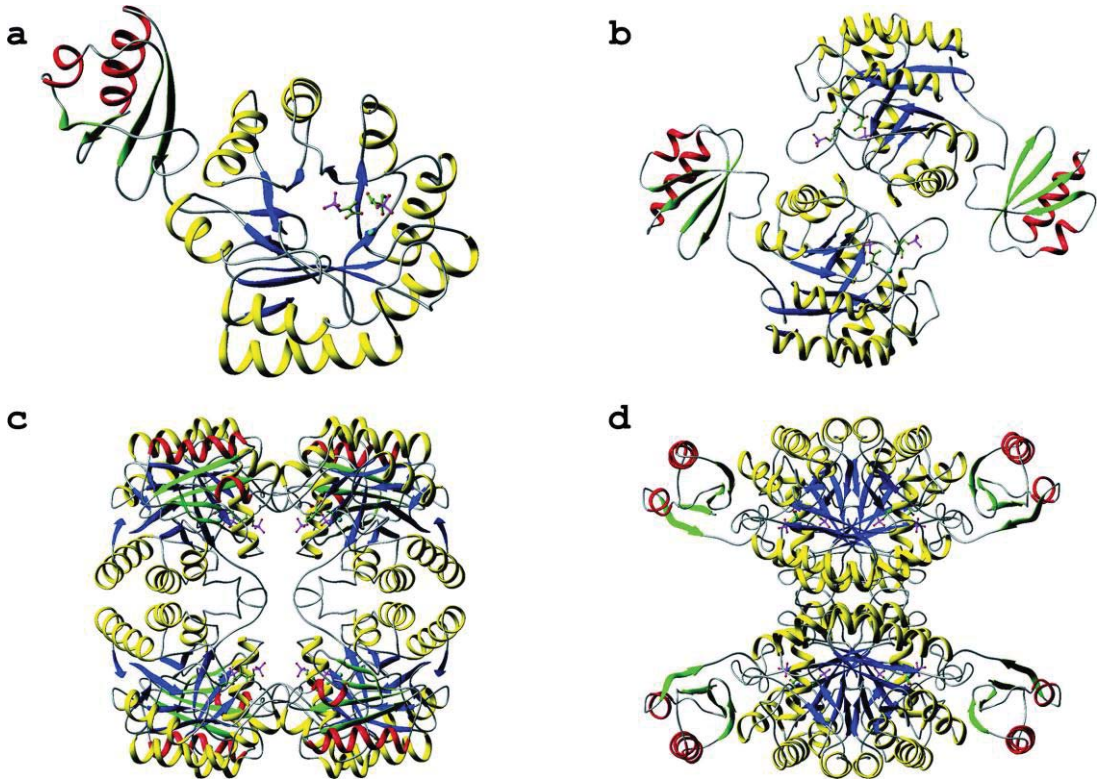

Fig. 1. 6. The structure of DAH7PS in *Thermotoga maritima*. Reprinted with permission from the *Journal of Molecular Biology*, **2004**, 341, 455–466. Copyright © 2004 Elsevier. The Structure of *T. maritima* DAH7PS–Cd^{2+}–PEP–E4P complex. The α-helices and β-sheets are yellow and blue in the $(\beta/\alpha)_8$-barrel domain and red and green in the FL domain. The active site located on the $(\beta/\alpha)_8$-barrel contains Cd^{2+} (cyan), PEP, and E4P (both green/purple). (A) Monomer. (B) Dimer, present in the asymmetric unit. (C) and (D) Two views of a tetramer that differ by a 90° rotation around the vertical axis.

The mechanism for conversion of DAH7P to DHQ by DHQS involves a number of distinct reactions. Firstly, the cofactor NAD^+ accepts a proton from DAH7P, which is then transferred to a water molecule before being passed on to the adjacent His275. The enzyme then catalyses phosphate elimination, by binding phosphate in such a way that the scissile bond is susceptible to nucleophilic attack by Lys356. The final steps involve opening the substrate ring structure followed by an aldol condensation of the C2 and C7 carbons to from DHQ (Fig. 1.9) (27).

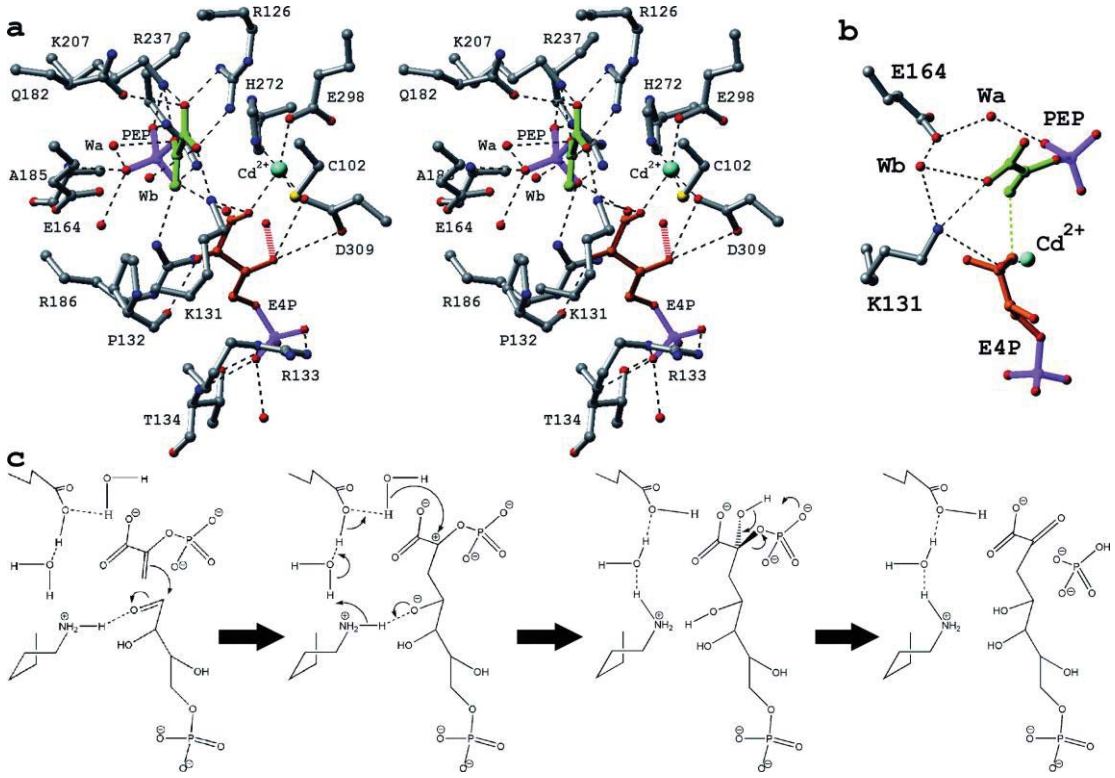

Fig. 1. 7. Catalytic Mechanism of DAH7PS in *Thermotoga maritima*. Reprinted with permission from the *Journal of Molecular Biology*, **2004**, 341, 455–466. Copyright © 2004 Elsevier. (A) A stereo diagram showing the conformation of the active site of *T. maritima* DAH7PS–Cd^{2+}–PEP–E4P is modified by rotation around its σ-bonds. Steric conflict between the hydroxyl of E4P and a water molecule is shown as a red line. (b) The relative positions of participants of the DAH7PS reaction. Hydrogen bonds and the (E4P)–Cd^{2+} bond are shown as black dotted lines. The green dotted line connects the substrate atoms forming the covalent bond in the condensation reaction. (c) The catalytic mechanism of DAH7PS. Hydrogen bonds are shown as black dotted lines.

DHQS2, an enzyme that produces the same product as DHQS in the alternative shikimate pathway, catalyses the deamination and cyclization of ADH to produce DHQ. While a number of enzymes have been bioinformatically identified, no structure of DHQS2 has yet been solved and it is not completely understood how it functions. DHQS2 has little sequence homology to the standard DHQS enzymes, yet both enzymes produce the same product DHQ from different substrates. The mechanism used by DHQS2 to accomplish this is yet to be characterised but may be reflected in the structure of the enzyme. Whether this is similar to the standard enzymes or completely different is yet to be reported and knowledge of the three dimensional structure of the enzyme may provide insight into possible mechanisms.

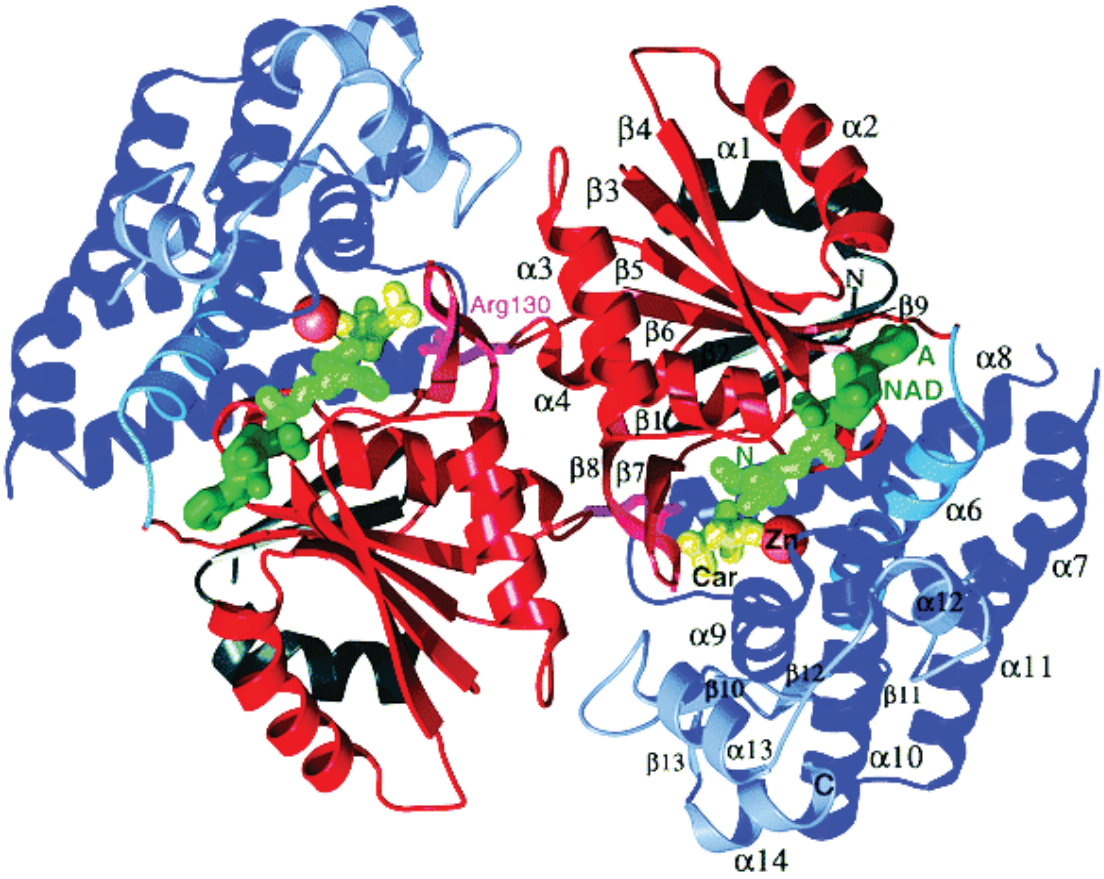

Fig. 1. 8. The Structure of DHQS. Reprinted with permission from; *Nature*, **1998**, 394, 299-302. Copyright © 1998 Nature. The N-terminal β-hairpin and connecting α-helix are shown in grey, the Rossman fold domain is in red, the β-hairpin insertion is shown in light pink, the C-terminal domain central α-helix is shown in mid-blue, the surrounding five α-helices in dark blue and the C-terminal α-helices and a distorted β-hairpin in grey-blue. The nicotinamide mononucleotide (N) and adenosine (A) moieties of NAD^+ are shown in green. The inhibitor carbaphosphonate is labelled Car and resembles DAH7P. Arg130 is shown in pink.

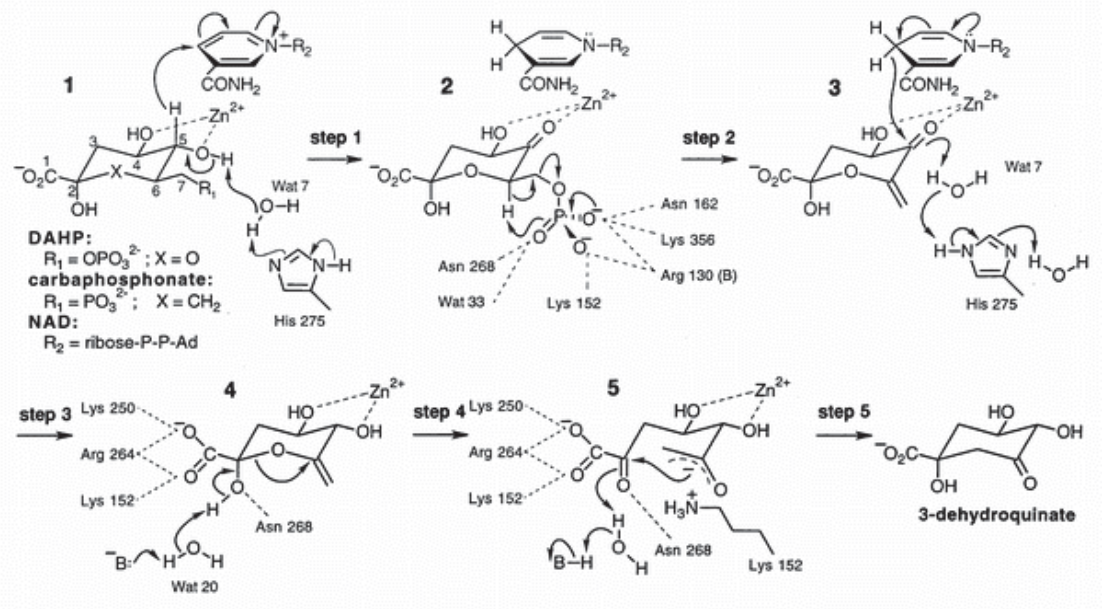

Fig. 1. 9. Catalytic Mechanism of DHQS. Reprinted with permission from *Nature*, **1998**, 394, 299-302. Copyright © 1998 Nature. Step 1 is oxidation of the hydroxyl group attached to C5. Step 2 is a β-elimination of inorganic phosphate. Step 3 is a reduction of the ketone at C5. Step 4 is a ring opening, and step 5 is an intramolecular aldol condensation. Ad, adenosine; P, phosphate; B, A general base Lys152 is shown fulfilling two possible roles.

1.4 Evolution and the Advantages of an Alternate Pathway

Genomic analyses have provided clues about the potential metabolic strategies that allow several of the many archaeal species to survive (28). Archaeal species, in which DAHP synthase and DHQS are "missing", have *aroA'* and *aroB*' genes that code for the enzymes that replace the normal shikimate pathway genes allowing them to use different starting materials for chorismate synthesis. Most of these archaea live in highly-stressed environments such as high temperatures (> 80° C), high salt and high pressure (29). Some bacteria that live in extreme environments also use the non-canonical shikimate pathway. *A. aeolicus*, a chemolithotroph, grows optimally at 85° C and lacks the normal shikimate pathway genes (12). Growth studies suggest that it does not import its aromatic amino acids from the environment, and therefore must synthesize them (30). *A. fulgidus, M. thermoautotropicum and M. jannaschii* are all hyperthermophiles that also lack the standard genes of the shikimate pathway. While there are many examples of bacteria that use the

alternative pathway to synthesise aromatic amino acids, it seems to be more common in archaea that live in extreme habitats (31).

The structure and evolution of the standard shikimate pathway in eukaryotes has been comprehensively investigated. There are two possibilities that could explain its origin; it was either of ancient origin or was acquired by prokaryote-to-eukaryote horizontal gene transfer (32). Little is known, however, about the evolution of the non-canonical shikimate pathway in prokaryotes. Phylogenetic analysis of ADHS and DHQS2 clusters in *Aquificae* and *Deltaproteobacteria* provided data in which the evolutionary distance between archaeal and bacterial horizontal gene transfers could be calculated (Fig. 1.10). The results showed that ancient horizontal gene transfer occurred between archaea and bacteria and that the non-canonical shikimate pathway was spread more widely in selected bacterial lineages, (33) corresponding with the divergence of the domains *Bacteria* and *Archaea* through the recruitment of primitive enzymes with different catalytic functions. This explains why enzyme pairs are employed to reach a common point in a pathway, use alternative mechanisms and are likely to have different structures and architectures. While some enzyme pairs share similar structural characteristics, the structures and phylogenetic architectures of the enzymes constituting parts of the shikimate pathway are different indicating that the evolution of the shikimate pathway cluster between species occurred separately before the divergence of domains *Bacteria* and *Archaea*.

One hypothesis is that the non-canonical pathway evolved because of a lack of ability of the organisms concerned to synthesizes or scavenge E4P and/or PEP from their habitats. It may be that archaea 'prefer' to use ASA and DKFP even though E4P and/ or PEP are available. As E4P is chemically unstable, at room temperature (13), it may not exist at the elevated temperatures at which hyperthermophiles grow. As PEP is an important energy-rich metabolite, minimizing its consumption may provide an evolutionary advantage. While PEP is essential for the common sixth step in the shikimate pathway, it is not used in the first step of the alternative pathway, which would result in half its consumption.

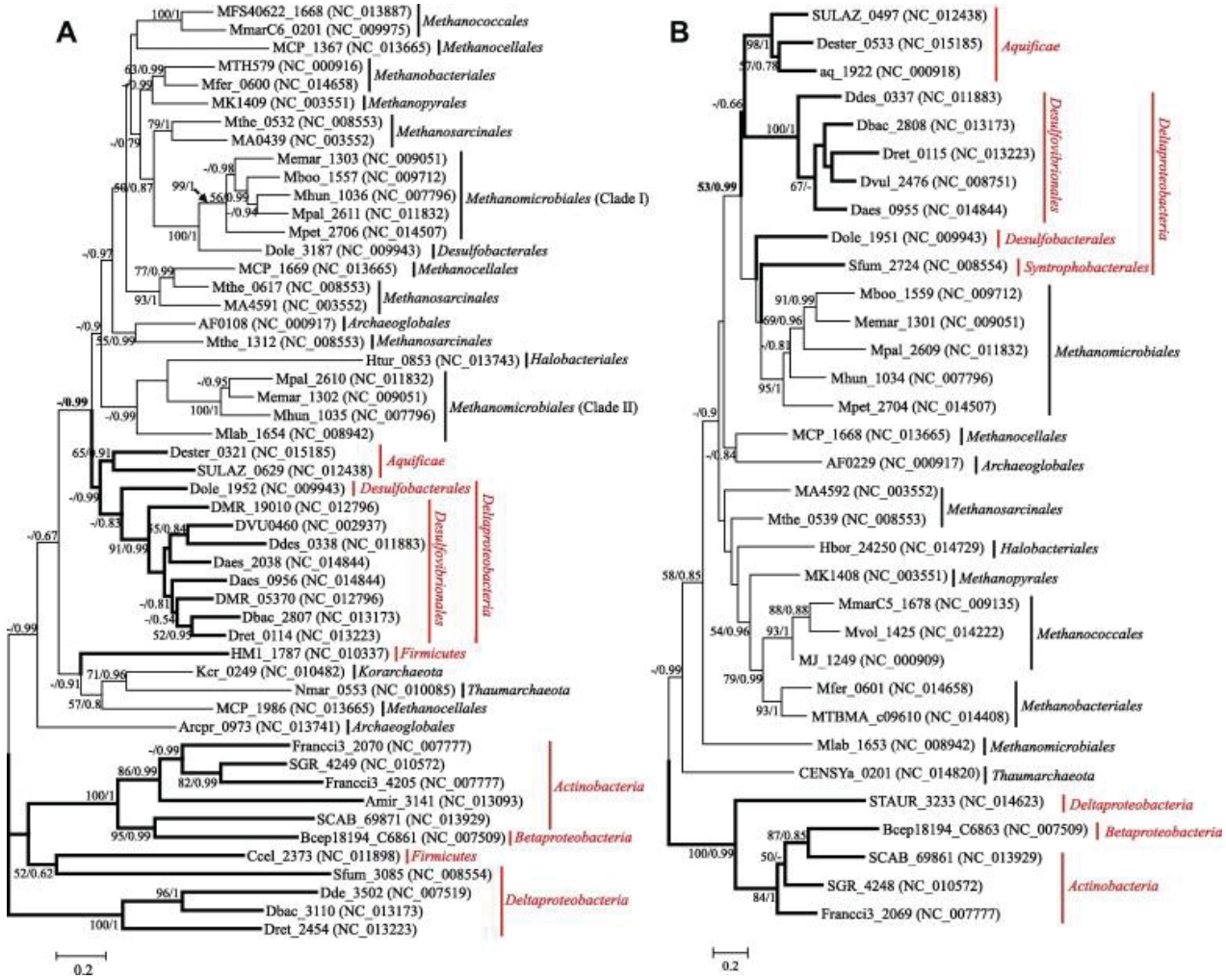

Figure 1. 10. The evolutionary distance between *Archaea* and *Bacteria*. Reprinted with permission from *Molecular Phylogenetics and Evolution*, **2014**, 75, 154–164. Copyright © 2014 Elsevier. The phylogenetic trees of ADHS (A) and DHQS2 (B). (A) This phylogeny was calculated from an amino acid alignment of 51 sequences and 233 sites (without ambiguous alignments) and based on LG model with a gamma-shaped distribution of rates across sites (LG + G). (B) This phylogeny was calculated from an amino acid alignment of 34 sequences and 190 sites (without ambiguous alignments) and based on LG + I + G model. Numbers on the branches indicate bootstrap support under maximum likelihood (left) and posterior probabilities (right). Only bootstrap values above 49% and posterior probabilities above 0.49 are shown. The scale bar denotes the number of accepted substitutions per site.

1.5 Protein Crystallization

There are two phases to crystallisation, nucleation and growth. During nucleation, molecules must overcome an energy barrier to from a periodically ordered aggregate of critical size. Growth is achieved by making the solid state more attractive to individual molecules than the free, solution state. To promote both nucleation and growth, a condition of supersaturation

must be created in the crystallisation medium (mother liquor). Supersaturation is a non-equilibrium state characterised by a tendency to form a solid phase, and to increase the proportion of material in that phase until a balance is restored. This is accomplished by altering the characteristics of the mother liquor in a way that discomforts the macromolecules, disrupts the solvent, and induces separation. A problem however, is that a protein has more than one solid state, therefore, not only protein crystals may develop but also precipitates, oils and salt crystals.

Protein homogeneity is an important principle in protein crystallography. Crystallisation presumes that identical units are available for incorporation into a lattice. In forming crystals, molecules organize by orderly self-association to produce a periodically repeating, three-dimensional array. It is necessary to facilitate molecular interactions while avoiding the formation of precipitate, non-specific aggregation or phase separation. Impurities may contribute to a failure to nucleate and are frequently the crucial obstacle to diffraction quality crystals. Most protein crystals degrade and lose diffraction quality upon aging but can be stabilized and preserved by temperature change or addition of precipitating reagent. Stability is important because no homogeneous protein lattice can remain so if its molecules alter their form, folding and conformational changes. Proteins may crystallise under a wide variety of conditions and form numerous polymorphs with a diversity of physical and diffraction properties.

1.5.1 Crystallisation Methods

The vapour-diffusion method consists of an aqueous drop where the protein and the crystallization agents are mixed in an amount lower than that required for the formation of crystals (34). The crystallization mixture is placed in the vicinity of a reservoir that contains a high concentration of salt or other non- volatile precipitating agent where it is equilibrated against the reservoir. Slow diffusion of water from the crystallization solution into the concentrated solution of salt is due to the difference in osmotic pressures of the two solutions. This diffusion leads to a decrease in volume of the crystallization mixture, and the gradual concentration of the crystallization solution. This leads to the supersaturation of the crystallization solution for the nucleation of the protein crystal to occur (35). An individual

crystallization trial proceeds through a range of conditions, thereby conducting a self-screening process.

Sitting and hanging drop vapour diffusion methods are easy to perform and allow flexible screening with minimal sample volume. The sitting-drop technique has benefits over hanging-drop plating, such as cost and time efficiencies, but crystals often adhere to the hardware surface. The hanging-drop technique reduces the occurrence of hardware crystal adherence and improves the crystal shape and size because of the inverted position of the drop, but this method has the disadvantage of requiring silicone grease and a siliconised cover slip. The hanging-drop method permits the transference of a cover slip containing the crystallization drop from one reservoir to another without disturbing the drop. This provides more flexibility for changing the conditions than the batch method where any change other than temperature involves disturbance of the crystallization drop itself. Higher quality crystals have been obtained by transferring cover slips from nucleation to growth conditions (36).

Seeding permits the use of poorly diffracting crystals to seed into similar but non-identical conditions. Interestingly, such strategy resulted in the formation of better quality crystals with a 10% reduction of the unit cell (37).

The microbatch technique is a batch method in which the molecule to be crystallized is mixed with the crystallising agents at the start of the experiment. The concentration of the ingredients is such that supersaturation is achieved immediately upon mixing, thus the composition and the volume of a trial remain constant and crystals will only form if the precise conditions have been correctly chosen (38).

1.5.2 Crystallization Conditions

The function of precipitants in the crystallization drop is to alter the protein-solvent or protein-protein contacts so that the protein molecules precipitate out of solution as ordered crystals and not as disordered aggregates (39). pH is the single most determinant of protein solubility. A protein is least soluble, hence likely to crystallise at a pH equivalent to its pI

(34). The solubility of most proteins varies as a function of temperature. Protein solubility usually increases with increasing temperature in low ionic strength conditions such as in the presence of the precipitant polyethylene glycol (PEG). Crystallization has been reported to occur for proteins over a whole range of temperatures, from 0 °C to 60 °C, (40) although crystallisation experiments are usually conducted at either 4 °C or at room temperature.

1.6 Protein Characterisation

A protein has a number of properties that leads to its overall characterization. Once the gene encoding the protein of interest has been fully sequenced it can be bioinformatically analysed using a number of tools freely available on the web. Examples of such tools are PFAM, which searches for known domains, TOPCONs, which can be used to predict membrane topology and signal peptides, PHYRE2 a protein fold recognition server as well as the myriad of algorithms in EXPASY TOOLs that can be used to calculate the theoretical pI, monoisotopic mass and extinction coefficient

Measuring the exact monoisotopic mass of the protein using mass spectrometry can indicate the presence of post translational modifications, and the secondary and tertiary structures can be probed using far and near UV circular dichroism (CD), which may give an indication the protein is folded. Of course the ultimate test of whether a protein is correctly folded is to test for its biological activity. Where a suitable assay is not available such a test is not possible.

1.6.1 Circular Dichroism (CD)

CD refers to the differential absorption of the left and right circularly polarized components of plane-polarized radiation. The plane-polarized radiation is spit into its two circularly components by passage through a modulator subjected to an alternating electric field. Proteins possess a number of chromophores, which can give rise to CD signals (41). In the far UV region (180-260 nm), peptide bond absorption gives rise to CD spectra and can be analysed to give the proportions of secondary structural elements such as α-helices and β-sheets. The near UV region (260-320 nm) reflects the environments of the aromatic amino acid side chains and thus gives very basic information about the tertiary structure of the

protein i.e. whether the protein is likely to be folded or in a molten globule conformation. CD is used extensively to study the folding and denaturation of proteins (42).

1.6.2 X-ray Crystallography

Knowledge of protein structure is vital for elucidating protein function and understanding the mechanisms in metabolic pathways. Crystallography enables the three-dimensional structures of large proteins to be determined. However, unlike crystals formed by inorganic compounds, protein crystals can contain up to 60% solvent making structure determination challenging. The success of diffraction experiments depends on crystalline order, which usually deteriorates if the crystals become dehydrated (43). Before crystallisation can be attempted, sufficient quantities of protein must be purified, usually as a recombinant protein from bacterial, yeast or mammalian cells (44). An important development has resulted from the use of autotrophic strains for the incorporation of seleno-methionine (Se-Met) into recombinant protein, because selenium allows the structure to be experimentally phased using multi-wavelength anomalous dispersion (MAD) (45). The Australian Synchrotron is the closest x-ray source capable of MAD methods using the MX2 portal beamline. To maximize the chances that crystals will grow, the protein must be as homogenous as possible, and while techniques for assessing protein aggregation have advanced considerably with the use of dynamic light scattering, it is not possible to predict the conditions under which a particular protein will crystallise. The approach is therefore to screen a wide range of crystallisation conditions using the smallest number of experiments. This is done using a sparse matrix approach (46). Once crystals have been obtained, they are manually harvested from their growth drop, dipped in cryoprotectant and flash-cooled before being irradiated with x-rays (47). Very high synchrotron source flux densities allow weakly diffracting and smaller crystals to be used for structure determination (48). However, radiation damage still remains an issue and can result in failed structure solution due to increased signal to noise and the onset of structural damage (49). Crystallographic software, such as PHENIX (50), is used to analyse diffraction data and is capable of phasing, calculating electron density maps, interpreting maps, building structures and then refining the structures.

1.6.3 Solving the phase problem

Molecular replacement methods are able to predict the 3-D structure of a protein sequence by using phasing information derived from a homologous protein of known structure (51). When the sequence identity between two proteins falls to 30% or less, the alignment process becomes increasingly unreliable (52). Fragment based molecular replacement identifies structurally conserved regions (SCRs) that tend to have well defined secondary structure. The protein backbone of the query protein can be substituted by regions of suitably homologous structures. Variable regions (VRs) of an alignment are most often loop regions where deletions, mutations and insertions are common. VRs are modelled by searching for regions of similar length between SCRs that meet steric requirements. The restraint based homology method alignment aims to derive geometrical restraints of protein backbone and side chain angles (53).

In multiple isomorphous replacement (MIR), the idea is to make a change to the crystal that will perturb the structure factors and, by the way that they are perturbed, make deductions about possible phases. The introduction of a heavy atom will change the scattered intensity significantly. One reason for this is that "heavy" atoms contribute disproportionately to the overall intensity. The contributions from the lighter atoms will tend to cancel out, because they will scatter with different phase angles. In comparison, all of the electrons in a heavy atom will essentially scatter in phase with one another. One crystal containing just the protein (native crystal) and one containing bound heavy atoms (derivative crystal), can be used to measure diffraction data. The differences in scattered intensities will largely reflect the scattering contribution of the heavy atoms, and these differences can be used solve the phase problem. Once the heavy atoms are located in the crystal, approximate structure factors can be computed and the structure solved.

Anomalous scattering (AS) refers to a small change in the phase angle of a diffracting X-ray that is due to strong X-ray absorbance. MAD methods use multiple, often three, wavelengths that result in small differences in the intensities of pairs of diffracted x-rays that can be measured and used to phase the structure.

1.6.4 Other methods for determining the 3D structures of proteins.

Nuclear magnetic resonance can also be used to analyse protein structure and dynamics. 2D ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra measure the distances between a nitroxide spin-label fixed at a specific protein site and the backbone amide protons of the same sample protein, generating distances from a single pair of NMR spectra. The distances can be used as a source of restraints for structure determination (54). As The molecular weight of a protein increases, its molecular tumbling slows, leading to transverse relaxation rates and, as a result, reduced and signal intensity and broad peaks (55).

Cryo-electron microscopy (cryo-EM) is another technique used to obtain 3D structural information on proteins. It generally results in lower resolution structures compared to those generated by x-ray diffraction and NMR, and is based upon the principle of imaging radiation-sensitive specimens in a transmission electron microscope under cryogenic conditions. Resolution is directly influenced by the wavelength of the imaging radiation source: the shorter the wavelength, the higher the attainable resolution (56). Recent advances in electron detectors and image processing have seen examples of cryo-EM resolving protein structures to 3.5 Å with electron density maps that are equivalent to that from lower resolution x-ray crystallography (57).

1.7 Conclusion

The shikimate pathway evolved before the divergence of domains *Archaea* and *Bacteria*. Horizontal gene transfers between archaea and bacteria transferred only distinct catalytic functions, resulting in many non-orthologous genes being recruited then tailored to adapt various enzyme functions to their environments. The fact there is variance in genomic architectures of closely related phylogenies is not surprising. It is likely that the first two genes of the standard shikimate pathway were not simultaneously transferred resulting in non-canonical pathways that have adapted to extreme habitats. This was an advantage to the species involved as it enabled them to utilize different precursors that were more stable in

extreme habitats, to sustain functional enzyme activity and to decrease the consumption of energy-rich molecules.

Comparisons of genomic sequences of bacterial and archaeal genomes provided the first indication that a different pathway existed for the synthesis of similar compounds. Labelling studies showed that the precursors of the shikimate pathway in *M. jannaschii* are DKFP and ASA rather than PEP and E4P and studies on *H. salinarum* provided evidence that more than one non-canonical pathway may exist to achieve the same endpoint. The first enzyme of the alternative pathway ADHS exists as a homodecamer. Each monomer forms a β-barrel with the active site at the top of the barrel. The active site Lys184 located on strand β6 of the barrel forms a covalent adduct with DKFP through the formation of a Schiff base intermediate. Tyr153 then acts as the general acid and Asp33 as a general base in a cleavage and condensation reaction to produce the product ADH (25).

Further investigation into the structure and function of this DHQS2 is needed to fully understand the mechanism of non-canonical chorismate biosynthesis. It is not known how the enzyme produces DHQ from the substrate ADH. Knowledge of the three dimensional structure of DHQS2 enzyme will allow structural homology studies, and may throw light on possible mechanisms. X-ray diffraction methods are the best methods to carry out such structural characterisation because DHQS2, at 40Kda, is too large for NMR, and cryo-EM, is likely to produce a low resolution structure. As low sequence homology of DHQS2 to known proteins rules out structure solution using molecular replacement methods, MAD methods are most likely to result in successful structure determination. The way different enzymes have adapted to use different or available substrates is of interest to those investigating direct evolution. As structure is intrinsically linked to function, understanding changes in structure may provide information on the mechanism used.

1.8 Thesis Synopsis

The objective of this project was to structurally characterize DHQS2. Doing so will complete the structural characterisation of an important enzyme in the alternative shikimate pathway and provide further insight into its catalytic mechanism and evolution. It would be

interesting to functionally characterise DHQS2 as DHQS2 was reported as being involved in aromatic amino acid biosynthesis in *M. jannaschii* and *H. salinarum* but not in *M. maripaludis.* However, the assay to characterise its function requires cloning and production of recombinant ADHS along with the synthesis of DKFP and ASA precursors. This was not the primary objective of this work, and so was not attempted due to time constraints.

Project objectives:

- Selection of target DHQS2 sequences using bioinformatics. The criteria for selection were:
    1. A high likelihood that the protein can be crystallised based on Xtalpred Expert Pool (EP) classification
    2. Minimum predicted disorder from Predictor of Natural Disordered Regions (PONDR)
    3. A high methionine content required for seleno-methionine labelling of recombinant protein.
    4. DNA commercially available

- Develop a vector system for cloning that allowed the introduction of a cleavable eight-histidine residue tag at the N-terminus, the C-terminus or with no tag that allowed the expression of the protein with the fewest extra residues possible. This plasmid should place the DHQS2 gene under the control of T7-RNA polymerase promoter to ensure production of mg quantities of recombinant DHQS2.

- Determine the optimum conditions for the production of soluble recombinant seleno-methionine derivatives of DHQS2 in *E. coli*.

- Develop a purification scheme for the recombinant protein.

- Analyse the recombinant protein to see if it has the expected mass (mass spectrometry) and that it is folded (circular dichroism spectrometry).

- Determine the conditions for the production of diffraction quality crystals.

- Solve the structure of recombinant protein using x-ray crystallography and MAD methods.

Chapter Two: Methods

2.1 General Methods

Purified Water

Water was purified by passage through a Sybron/Barnstead NANOpure II filtration system (Barnstead Lab Water Products, Lake Balboa, California United States), containing two ion exchange and two organic filters. This water is referred to as Milli-Q water. All Milli-Q water used in the cloning of genes was autoclaved before use.

Centrifugation

Centrifugation was carried in three centrifuges: a SORVALL RC6+ (Thermo Scientific, Waltham, Massachusetts, United States), a SORVALL Heraeus multifuge® (Thermo Scientific) X1R or a MiniSpin® centrifuge (Eppendorf, Hamburg, Germany).

pH Measurement

The pH of buffers used in this project was measured using a Model 20 pH/Conductivity Meter (Denver Instrument Company, Bohemia, New York, United States) and a Schott pH probe (Hattenbergstr, Germany). The pH of the solution was adjusted using 10 M NaOH (Ajax), 10 M acetic acid (Ajax) and 10 M HCl (Fisher Chemicals) where appropriate.

Media

All *E. coli* cultures were grown in Luria Broth (LB) unless otherwise stated. LB (Invitrogen, United States) was made up (25 g/L) with Milli-Q water system and sterilized by autoclaving at 121 °C and 15 psi for 20 minutes. Media for agar plates were prepared by adding 1.5 % agar (Oxoid, United Kingdom) to the liquid media prior to autoclaving.

SOC media was used to incubate freshly transformed *E. coli* cells after heat shock. Super optimal broth (SOB) consists of 20 g/L bacto-tryptone (Merck), 5g/L bacto-yeast extract (Invitrogen), 0.5 g/L NaCl (Univar), 2.5 mM KCl (Merck), 10 mM $MgCl_2$ (BDH) and 10mM glucose (Invitrogen). All components were added to Milli-Q water, except for glucose, and autoclaved. Solutions of glucose (100 mM) and $MgCl_2$ (1M) were filter-sterilized (0.2 μM) and added prior to use to convert SOB into SOC.

Antibiotic Stocks

Stock solutions of kanamycin (30 mg/mL) and ampicillin (100 mg/mL) in Milli-Q water was filter sterilised (0.2 μm) and stored at -20 °C.

Polymerase Chain Reaction

Polymerase chain reaction (PCR) is used to amplify DNA. A PCR reaction set-up using 0.02 U/μL Q5® High-Fidelity DNA Polymerase (New England BioLabs, Ipswich, Massachusetts, United States) contained Q5 reaction buffer (New England BioLabs), 200 μM deoxynucleotides (dNTPs) (New England BioLabs), 0.5 μM forward and reverse oligonucleotide primer (Integrated DNA Technologies, Coralville, Iowa, United States), ~ 1 ng genomic DNA (gDNA) or plasmid, and nuclease-free water.

An Analytik jena thermocycler (Biometra, Goettingen, Germany) was used to cycle through 30 rounds of denaturation, annealing and extension of template DNA resulting in an amplified product. An initial denaturation of 30 seconds at 98 °C was used while 20 seconds was used in each cycle. An annealing temperature of 58 °C, for 30 seconds, was used because it was 3 °C below the lowest melting temperature (T_m) of the primers used. Each cycle had an extension time of 2 minutes based on the size of the 2.4kb vector or 1kb DHQS2 template were 25 seconds of extension time assembles approximately 1kb of template. A final extension time of 5 minutes was used after the 30^{th} cycle.

The presence of a Gibson assembled product was detected using colony PCR where multiple colonies, that had been transformed with the product, were picked as template DNA. The modpETite seq primers, shown in Table 2. 1, anneal to the vector and produce a product with an approximate size of 1.5Kb. If the plasmid does not contain the insert the PCR product would be ~500bp. The resulting PCR product was visualized by 1 % (w/v) agarose gel electrophoresis.

Agarose Gel Electrophoresis

DNA fragments were separated on the basis of size using agarose gel electrophoresis. Agarose gels (1.0 % (w/v)) were prepared by adding 0.3 g agarose to 30 mL 1X TAE buffer (40 mM Tris-HCl, 20 mM acetic acid and 2 mM EDTA at pH 8.0) and heated until dissolved. All gels were run using a Sub-Cell® GT Agarose Gel electrophoresis system (Bio-Rad, Hercules, California, United States) in 1X TAE buffer. 5 μL of DNA sample was premixed with 2 μL of loading buffer (0.2 % (w/v) bromophenol blue in 50 % (v/v) glycerol) and then loaded into the wells. Electrophoresis was carried out at 80 V generated using a PowerPac™ Basic (Bio-Rad, Hercules, California, United States) power supply until the dye front had migrated to the other end of the gel. The DNA on the gels was visualized by staining the gel in ethidium bromide (0.5 μg/mL), followed by exposure to ultraviolet light (302 nm). Images of the gels were recorded using a Gel Doc™ Gel Documentation System (Bio-Rad, Hercules, California, United States).

Quantification and Size Determination of DNA Fragments

The concentration of DNA samples was estimated by comparison with DNA samples of known concentration run on an agarose gel. The approximate size of a DNA band was estimated by comparing its migration through the gel against that of DNA standards with known size (1 Kb Plus DNA ladder, Invitrogen).

All purified DNA concentrations were estimated using a Nanodrop® ND-1000 Spectrophotometer (Thermo Scientific, Waltham, Massachusetts, United States).

DNA Gel Purification

During PCR, it is possible for primers to anneal to other sites as well as that intended. In this case multiple products are formed by the PCR. A Wizard® SV Gel and PCR Clean-Up System kit (Promega, Madison, Wisconsin, United States) was used to purify the desired DNA product. Agarose gel electrophoresis separates DNA fragments based on their size, allowing the desired gene product to be cut from the gel and processed using the Wizard kit. 10 μL of membrane binding solution was added to 10 mg of gel and incubated at 65 °C for 10 minutes or until the gel was completely dissolved. The DNA in the solution containing the dissolved gel slice was bound to a membrane, which was repeatedly washed in an ethanol solution, before the DNA was eluted with DNAse- and RNase- free water.

Restriction Digest

A restriction digest was used as an alternative to colony PCR to detect the presence of a DNA product. The online tool double digest finder (New England BioLabs, United States) was used to determine an appropriate buffer for the simultaneous activity of two restriction enzymes. Both N and C-terminal modpETite vectors containing insert require *Bam*HI plus *Kpn*I and *Nde*I plus *Kpn*I restriction enzymes respectively. Both combinations were compatible with NEBuffer 1.1 (New England BioLabs). Restriction enzymes were added at 0.05 U/μL to NEBuffer 1.1. The digest reaction was left to incubate at 37 °C overnight.

DNA Sequencing

DNA sequencing services were provided by the Massey Genome Service (MGS). DNA sequencing was carried out on an ABI3730 DNA Analyzer capillary sequencer (Thermo Scientific) using BigDye™ (Thermo Scientific) labelled dideoxy chain termination chemistries.

Preparation of Chemically Competent *E. coli* DH5α and DL41 (DE3) Cells

The procedure used is that previously described by Nojima (58). The desired *E. coli* strain was streaked on an LB plate and incubated overnight at 37°C. A single colony was inoculated into 50 mL SOB medium in a 500 mL flask. The culture was grown at 25°C with shaking at 200 rpm until the optical density at 600 nm (OD_{600}) was approximately 0.4. The flask was chilled on ice for ten minutes, and then centrifuged at 1068x g for 15 minutes at 4°C. The cell pellet was resuspended in 17 mL ice-cold transformation buffer (TB) (10 mM PIPES (BDH), 15 mM $CaCl_2.2H_2O$ (Ajax) and 250 mM KCl (Ajax)), adjusted to pH 6.8 with 5 M KOH (Merck). 55 mM $MnCl_2.4H_2O$ (BDH) was added after pH adjustment and filtered through a 0.22 μm sterile filter (ReliaPrep™, AHLSTROM, Helsinki, Finland). Cells were re-pelleted by centrifugation (1690x g) then resuspended in 4 mL TB. 300 μL of dimethyl sulfoxide (DMSO) (Sigma®) was added before 0.1 mL aliquots of cells were flash frozen in liquid nitrogen to be stored at -80°C.

Transformation of pETite / modified pETite into competent *E. coli* DH5α and DL41(DE3) cells

Competent DH5α and DL41(DE3) cells were transformed with a modified pETite vector (modpETite) using the method described in the NEBuilder® HiFi DNA Assembly Chemical Transformation Protocol (E2621). 50 μL aliquots of cells were thawed on ice and then incubated with modpETite (2μL of ~30 ng/μL) on ice for 30 minutes. Cells were then heat shocked at 42 °C for 30 seconds before being returned to the ice for two minutes. They were then reconstituted in 950 μL of SOC media and shaken vigorously (250 rpm) at 37 °C for 60 minutes before being plated onto pre-warmed (37 °C) LB agar plates containing kanamycin (30 μg/mL) and incubated at 37 °C for ~16 hours.

Induction of Protein Expression

The expression of protein from genes inserted into the multiple cloning site of a modpETite are under the control of the lac promoter and is therefore induced by the addition of lactose or

isopropyl-β-D-thiogalactopyranoside (IPTG) (Applichem, United States). IPTG is a non-hydrolysable analogue of lactose that activates gene expression but is not metabolized by the cell like lactose. IPTG was added to the culture medium to a final concentration of 0.5 mM, unless otherwise stated, to induce protein expression

Polyacrylamide Gel Electrophoresis

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was carried out using the method of Laemmli (59) and a 4 % (w/v) stacking gel, a 12 % (w/v) separating gel, and a Mini Protean III cell (Bio-Rad, United States). SDS loading buffer (0.1 M Tris-HCl (Pure Science Ltd) pH 6.8, 20 % (v/v) glycerol (Vickers Laboratories Limited), 2 % (w/v) SDS (Affymetrix), 0.02 % (w/v) bromophenol blue (BPB) (Sigma) and 0.1 M dithiothreitol (DTT) (Sigma)), was added to an equal volume of the sample and the tubes boiled for 2 minutes. Low range SDS-PAGE molecular weight standards (Bio-Rad) were used.

Alternatively, native gels were prepared using the same components except for the addition of SDS or reducing agent, DTT, to either the 4 % (w/v) stacking or separating gels, loading buffer or running buffer, and samples were not denatured.

A constant voltage of 200 V was applied to the gels until the dye front reached the bottom of the separating gel, after which the gels were stained using Coomassie Blue. Gels were stained using a solution consisting of 1 g/L Coomassie Blue R 250 (Park Scientific) in 50 % (v/v) methanol (MeOH) (Labserv) and 10 % (v/v) acetic acid (Panreac). To remove excess dye, the gel was destained in 40 % (v/v) MeOH (Labserv) and 10 % (v/v) acetic acid (Panreac).

Cell Lysis

(a) Sonication

Cells were lysed, on ice, using a Qsonica sonicator (Qsonica, Newtown Connecticut, United States), large probe for large cultures and a mini-probe for small cultures, at 10 second bursts of 15 Watts.

(b) French Press

Large scale culture pellets, resuspended in 40 mL of IMAC lysis buffer were subjected to a pressure of 4000 psi using a French Press (WABASH, Indiana, USA).

Storage of Cultures

All strains of *E. coli* were stored at -80 °C as glycerol stocks. An aliquot of each culture was taken when the OD_{600} reached ~0.5. Glycerol was added to the culture to give a final concentration of 10 % (v/v) and the aliquots snap-frozen in liquid nitrogen and stored at -80 °C.

Storage of Enzymes

All enzymes preparations were stored as aliquots of volumes no greater than 200 μL, snap-frozen in liquid nitrogen and stored at -80 °C.

2.2 Methods for Chapter Four

Primer Design for Gibson Assembly

PCR primers for use in HiFi DNA assembly must have two sequence components; an overlap sequence, required for the assembly of adjacent fragments, and an insert (gBlock or DHQS2) specific sequence, required for template priming during PCR. The non-priming overlap sequence is added at the 5′ end of the primer (Fig, 2.1, fragment B, 5' red bar). This sequence is homologous to the 3′-terminal sequence of the adjacent fragment to be assembled (Fig. 2.1, fragment A, red bar). The length of the overlap sequence is dependent on the GC content of the sequences. The priming insert-specific sequence is at the 3′ end of the primer after the overlap sequence. The priming sequence should meet the criteria required for template annealing during PCR. To achieve efficient assembly of PCR fragments into a vector, a 15 – 25 nucleotide overlap with a $T_m \geq 48$ °C ($T_m = 2(A-T) + 4(G-C)$) is required. All primers used in throughout this project were designed using the NEBuilder® Assembly Tool (New England Biolabs) and are shown in Table 2. 1.

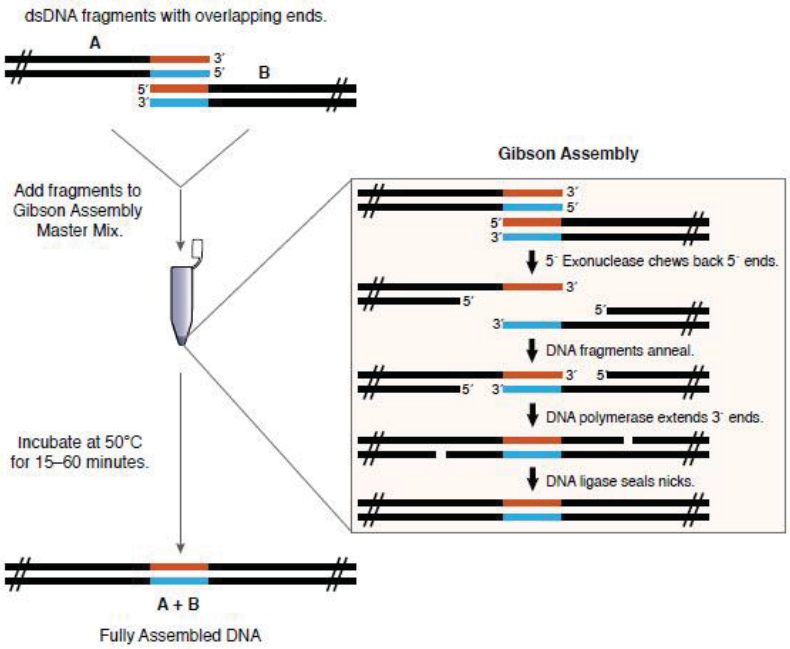

Fig. 2. 1. Gibson Assembly.

Modification of the pETite Vector by Gibson Assembly of a gBlock

The pETite C-terminal His-6 vector is an inducible T7 promoter-based expression vector chosen for its high level expression of an inserted gene (Fig 2.2, see Appendix two for the full sequence). The pETite vector confers kanamycin resistance and includes a T7 promoter site (highlighted in green) and a T7 terminator site (highlighted in red). The vector contains an *Nde*I restriction site (highlighted blue), including an ATG start codon (in bold), followed by a cloning site and a His-6 tag (highlighted yellow). The vector was modified by the addition of a gBlock (Fig. 2.3) *via* Gibson assembly to allow cloning of DHQS2 with either a N- or C-terminal His-8 tag. The resulting assembly is referred to as the modified pETite vector (modpETite, see Appendix three).

The gBlock contains overlaps that are required for assembly (underlined in Fig. 2.3), an *Nde*I restriction site (blue) with start codon (bold) followed by an N-terminal His-8 tag (yellow), a spacer arm (green), a Tobacco Etch Virus (TEV) protease site (red), a multiple cloning site (MCS) (blue) including; *Bam*HI (GGATCC), *Msc*I (TGGCCA), *Sac*I (GAGCTC), *Sal*I (GTCGAC), *Sma*I/*Xma*I (CCCGGG), *Pst*I (CTGCAG), *Kpn*I (GGTACC), and two additional histidine codons (yellow) to produce a C-terminal His-8 tag. The gBlock was designed to contain two pETite vector sequences, one before the *Nde*I start codon restriction site, before the cloning site, and another after the His-6 tag. This resulted in the correct positioning of the gBlock to allow cloning of DHQS2 with either a N- or C- terminal His-8 tag. The forward and reverse pETite primers (Table. 2.1) were then used to anneal to the same sites for both the pETite vector and gBlock (Figures 2.2 and 2.3) in order to produce overlapping ends required for Gibson assembly (Fig 2.1). The PCR products were then assembled *via* Gibson assembly.

Gibson Assembly

Gibson Assembly® was achieved using the method previously described by Gibson *et al* (2009) (60) (Fig. 2. 1) 50 – 100 ng of vector with a two-fold excess of inset, at a total of 0.03-0.2 pmols of DNA for a 2 fragment assembly, was added to 2x NEBuilder® HiFi DNA Assembly Master Mix and made up to 10 μL with Milli-Q water. The reaction mix was incubated for 15 minutes at 50 °C using an Analytik Jena thermocycler (Biometra,

Goettingen, Germany). 5 µL of assembled product was transformed by heatshock into chemically competent *E. coli* DH5α cells and plated on LB agar supplemented with kanamycin.

```
5'GGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCTAATACGACT
3'CCGTCACTCGCGTTGCGTTAATTACACTCAATCGAGTGAGTAATCCGTGGATTATGCTGA

CACTATAGGGTGTGAGCGGATAACAATTTCACGTGGAACAGCTAGAAATAATTTTGTTTAAC
GTGATATCCCACACTCGCCTATTGTTAAAGTGCACCTTGTCGATCTTTATTAAAACAAATTG
                                          REV: CTTTATTAAAACAAATTG

                                   FWD: CATCATCACCACCATCACTAATAGA
TATAAGTTGGAGATATACATATG cloned gene CATCATCACCACCATCACTAATAGA
ATATTCAACCTCTATATGTATAC cloned gene GTAGTAGTGGTGGTAGTGATTATCT
ATATTCAACCTCTATATGTATAC

G
GCGGCCGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAG
CGCCGGCGGTGGCGACTCGTTATTGATCGTATTGGGGAACCCCGGAGATTTGCCCAGAACTC

CCCAAAAAACGACTTTCCTCCTTGAATATCCGGTA5'
GGGTTTTTTGCTGAAAGGAGGAACTTATAGGCCAT3'
```

Fig. 2. 2. Nucleic Acid Sequence of the pETite Vector T7 System.

```
5'CACGTGGAACAGCTAGAAATAATTTTGTTTAACTATAAGAAGGAGATATACATATGCATC
3'GTGCACCTTGTCGATCTTTATTAAAACAAATTGATATTCTTCCTCTATATGTATACGATC
            REV: CTTTATTAAAACAAATTGATATTCAACCTCTATATGTATAC

ACCATCACCATCACCATCACGATTACGATATCCCAACGACCGAAAACCTGTATTTTCAGAGC
TGGTAGTGGTAGTGGTAGTGCTAATGCTATAGGGTTGCTGGCTTTTGGACATAAAAGTCTCG

GGATCCTGGCCAAGCAATGAGCTCAACGGCGTCGACCACAACTGGCTTCACCAGAGCGCAGA
CCTAGGACCGGTTCGTTACTCGAGTTGCCGCAGCTGGTGTTGACCGAAGTGGTCTCGCGTCT

                                                FWD: CATCATCACCACC
TGGTTATAACACACCCGGGTCTCTGCTGCAGGGTACCGGCAGCCATCACCATCATCACCACC
ACCAATATTGTGTGGGCCCAGAGAGGACGTCCCATGGCCGTCGGTACTGGTAGTAGTGGTGG

ATCACTAATAGAG
ATCACTAATAGAGCGGC3'
TAGTGATTATCTCGCCG5'
```

Figure 2. 3. Nucleic Acid Sequence of the gBlock.

DNA Purification

Once the expected insert size in a kanamycin-resistant colony had been detected by colony PCR using the forward and reverse pETite sequencing primers (pETite seq, Table 2.1), the resulting colony was grown with shaking (200 rpm), in 5 mL of LB supplemented with kanamycin at 37 °C and 200 rpm overnight. Plasmids were isolated from cells using the High Pure Plasmid Isolation Kit (Roche, Basel, Switzerland). Cells were pelleted from the 5 mL culture by centrifugation (4600x g for 20 minutes at 4°C) then resuspended in 200 µL of 50 mM Tris-HCl, 10 mM ethylenediaminetetraacetic acid (EDTA) buffer at pH 8.0 containing 0.1 mg/mL RNase A. Cells were lysed by adding 250 µL of lysis buffer (0.2 M NaOH and 1 % SDS), followed by 350 µL of binding buffer (4 M guanidine hydrochloride and 0.5 M potassium acetate pH 4.2) to precipitate unwanted cellular material. The precipitate was seperated by centrifugation (10,000 x g for 2 minutes), and the supernatant transferred by pipette to a tube containing a High Pure Filter. The filter was washed with 500 µL of wash buffer I (38 % (v/v) ethanol, 5 M guanidine hydrochloride and 20 mM Tris-HCl pH 6.6) before a final wash step using 700 µL of wash buffer II (80 % (v/v) ethanol, 20 mM NaCl, 2 mM Tris-HCl at pH 7.5). Plasmid DNA was eluted from the filter with 10 mM Tris-HCl elution buffer into a clean 1.7 mL Eppendorf tube and stored at -20 °C. The concentration and purity of plasmid and other DNA samples was determined using a Nanodrop® ND-1000 Spectrophotometer (Thermo Scientific, Waltham, Massachusetts, United States)

Assembly of DHQS2 into the C- and N- terminal His-8 ModpETite Vector

The modpETite vector (Appendix three) was primed for insertion of target DHQS2 genes by PCR to produce a linear DNA molecule. The N-terminal His-8 modpETite vector was linearised using the forward and reverse N-modpETite primers (Table 2.1), which anneal to *Kpn*I and *Bam*HI restriction sites, respectively (Fig. 2.4A). The C-terminal His-8 modpETite vector was linearised using the forward and reverse C-modpETite primers (Table 2.1), which anneal to *Kpn*I and *Nde*I restriction sites, respectively (Fig. 2.4B). The C-modpETite forward primer anneals at the *Nde*I site therefore eliminating the TEV, spacer arm and N-terminal

His8 tag from the C-terminal DHQS2 primed vector. The N-terminal DHQS2 reverse primer (Table 2.1) contains a stop codon that terminates translation before the addition of a C-terminal His8 tag. C-terminal DHQS2 primers do not include a stop codon and but use another stop codon after translation of a C-terminal His8 tag that is encoded in the original pETite vector.

Just like insertion of the gBlock into the pETite vector, the DHQS2 annealing primers (Table 2.1) contain overlapping regions, between DHQS2 and the modpETite vector, allowing insertion of DHQS2 into the MCS *via* Gibson assembly. Once the DHQS2 insert was identified (again by colony PCR), plasmid DNA from that colony was purified and sequenced. Once sequencing results had confirmed the correct DHQS2 insert and His8 tag positioning, expression plasmids were transformed into *E. coli* DL41 (DE3) cells and stored as glycerol stocks.

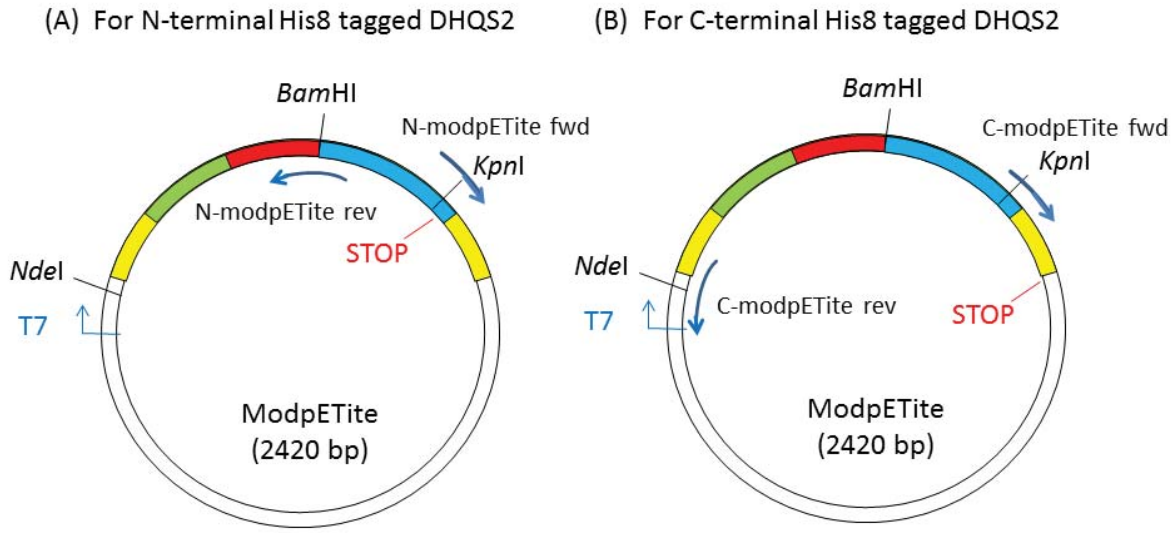

Fig. 2. 4. The ModpETite Vector. The modpETite vector is primed to encode (A) N-terminal DHQS2 and (B) C-terminal DHQS2. The gBlock sites contain His8 tags (yellow), a MCS (blue), a TEV site (red) and a spacer arm (green).

Expression Trials

Transformed *E. coli* DL41 (DE3) cells were supplemented with kanamycin, and cultured in 10 mL LB in a 50 mL flask with shaking (200 rpm) at 25 °C until an OD_{600} of ~0.8 was reached. One mL of uninduced sample was transferred to an Eppendorf tube, as an uninduced control, and the remainder of the culture split into three separate culture tubes for induction with either 0.1 mM, 0.2 mM or 0.5 mM IPTG. Samples were taken from each culture two, three and six hours after induction. Sample volumes were normalized to contain roughly the same number of cells based on their absorbance at 600 nm. Samples were centrifuged and the cell pellets resuspended SDS loading buffer, lysed by sonication and analysed by SDS-PAGE.

Solubility Trials

Constructs that showed sufficient expression were subjected to solubility trials. 4 mL cultures were grown in the conditions that optimally expressed DHQS2 as determined by the expression trials. A 1 mL sample was removed from the culture, the OD_{600} measured and a normalized volume of culture pelleted by centrifugation for whole-cell lysate. The remaining 3 mL was lysed by sonication and the soluble and insoluble material separated by centrifugation (13,000 x g for 10 minutes), normalised on the basis of absorbance at 600 nm and analysed alongside the whole-cell lysate samples by SDS-PAGE.

Table 2.1. Primers for Gibson Assembly

Function	Forward Primer	Reverse Primer
pETite	5'CATCATCACCACCATCACTAATAGAG3'	5'CATATGTATATCTCCAACTTATAGTTAAACAAAATTATTTC3'
pETite seq	5'CAACGCAATTAATGTGAGTTAGCTC3'	5'GAGGCTCGTCCTGAATGATATC3'
N-modpETite	5'GGTACCGGCAGCCATC3'	5'GCTCTGAAAATACAGGTTTTCG3'
C-modpETite	5'GGTACCGGCAGCCATC3'	5'CATATGTATATCTCCTTATAGTTAAACAAAATTATTTC3'
N-*A.veneficus*	5'TTTTCAGAGCGGATCCATGAAAGAAATTTGGTTACTAACAG3'	5'ATGGTGATGGCTGCCTTACCTCTCGATGATGAAC3'
C-*A.veneficus*	5'ACTATATGAAGGAGATATACATATGAAAGAAATTTGGTTACTAACAGAG3'	5'ATGGTGATGGCTGCCCCTCTCGATGATGAACTC3'
N-*D.alkenivorans*	5'AAACCTGTATTTTCAGAGCGGATCCAAAACCATTTGGGTAAAAGT3'	5'GATGGTGATGATGGTGATGGCTGCCTCATTTTTCCGTTATGGTCTC3'
C-*D.alkenivorans*	5'ACTATAAGAAGGAGATATACAATAGAAAACCATTTGGGTAAAAGTG3'	5'GATGGTGGTGATGATGGTGATGGCTGCCTTTTTCCGTTATGGTCTCCTTG3'
N-*M.hungatei*	5'TTTTCAGAGCGGATCCAAACAGGTATTTGTAGACCTTC3'	5'ATGGTGATGGCTGCCTTATTTCTCAATGATGGTCTCATC3'
C-*M.hungatei*	5'ACTATAAGAAGGAGATATACATATGAAACAGGTATTTGTAGACCTTC3'	5'-ATGGTGATGGCTGCCTTTCTCAATGATGGTCTCATC-3'
N-*M.maripaludis*	5'CCAACGACCGAAAACCTGTATTTTCAGAGCAAATTTGGATGGATAAAAACAAC3'	5'GTCGACGCCGTTGAGCTCATTGCTTGGCCATTATTTTTCTATAATTGTTTCTTTAATCGC3'
C-*M.maripaludis*	5'TAATTTTGTTTAACTATAAGAAGGAGATATACATATGAAATTTGGATGGATAAAAACAAC3'	5'GTGATGGTGATGATGGTGATGGCTGCCTTTTTCTATAATTGTTCCTTTAATCGC3'
N-*M.stadtmanae*	5'GACCGAAAACCTGTATTTTCAGAGCAAATTTGCATGGATAAGACC3'	5'GGTGATGATGGTGATGGCTGCCGGTACCTTATTTTTCAATAATTTCTTCTTCAATAGC3'
C-*M.stadtmanae*	5'ACTATAAGAAGGAGATATACATATGAAATTTGCATGGATAAGACC3'	5'GGTGATGATGGTGATGGCTGCCGGTACCTTTTTCAATAATTTGTTCTTCAATAGC3'
N-*S.azorense*	5'GACCGAAAACCTGTATTTTCAGAGCAAAGAGTTTATTCTTGATGCAAG3'	5'GGTGATGATGGTGATGGCTGCCGGTACCTTACTTCTCCATAATTGTTTCTTC3'
C-*S.azorense*	5'ACTATAAGAAGGAGATATACATATGAAAGAGTTTATTCTTGATGCAAG3'	5'GGTGATGATGGTGATGGCTGCCGGTACCCTTCTCCATAATTGTTTCTTCTAC3'

2. 3 Methods for Chapter Five

Minimal media

1 L cultures of *E.coli* DL41 (DE3) (a methionine auxotroph) were grown in M9 minimal media that was supplemented with seleno-methionine (Se-Met) (Sigma). As the only source of methionine, the DL41 (DE3) cells will incorporate Se-Met into proteins produced by the cell. The presence of Se-Met in place of methionine allows the use of MAD methods for structure solution. M9 minimal media consists of two separately prepared solutions; LeMaster solution I and LeMaster solution II.

LeMaster solution I

Chemical	For 10 L (g)
Alanine	5
Arginine HCl	5.8
Aspartic acid	4
Cysteine	0.3
Glutamic acid	6.7
Glutamine	3.3
Glycine	5.4
Histidine	0.6
Isoleucine	2.3
Leucine	2.3
Lysine HCl	4.2
Phenylalanine	1.3
Proline	1
Serine	20.8
Threonine	2.3
Tyrosine	1.7
Valine	2.3

Adenine	5
Guanosine	6.7
Thymine	1.7
Uracil	5
Sodium acetate	15
Succininc acid	15
Ammonium chloride	7.5
K_2HPO_4	105

23.02 g of LeMaster I was added to 1.8 g of NaOH, made up to 1 L and autoclaved.

LeMaster Solution II

Chemical	For 1 L (g)
Glucose	100
$MgSO_4.7H_2O$	2.5
$FeSO_4.7H_2O$	0.042
cH_2SO_4	83 μL
Thiamine HCl	0.01

LeMaster II was filter sterilized (0.2 μM, ReliaPrep™, Ahlstrom) and added to LeMaster I at one volume per every 10 LeMaster I volumes.

M9 minimal media was supplemented with kanamycin and 25 mg of Se-Met was added per litre.

Harvesting and Lysis of Cells

The cells were harvested approximately 6 hours after induction by centrifugation at 4600x g for 20 minutes at 4°C. The cell pellets were then stored at -80 °C until lysis.

The cell pellets were re-suspended in lysis buffer (10 mM Tris-HCl (Pure Science), 50mM NaCl (Pure Science), 1 mM tris-(2-carboxyethyl)phosphine (TCEP) (Sigma), 0.5 mM NAD^+, 10 mM imidazole (Apollo Scientific Limited) and 10% (v/v) glycerol (Vickers Laboratory Limited). Volumes smaller than 5 mL were lysed by sonication on ice and larger volumes were lysed using a French Press (WABASH, Indiana, USA) at 4000 psi. The DNA was broken up by sonication on ice and the cell debris removed by centrifugation at 13,000 x g for 20 minutes at 4°C.

Determination of Protein Concentration

Protein concentrations were determined by the method of Bradford (61), using bovine serum albumin (BSA) as a standard. Bradford reagent (Bio-Rad) was diluted 5-fold with Milli-Q water and filtered (0.2 μm). The BSA protein standards were made from dilution of a 1 mg/mL stock. 100 μL of protein standard solution or sample was mixed with 900 μL of Bradford reagent then incubated at room temperature for 20 minutes. Absorbance readings at a wavelength of 595 nm were then taken for all standards and samples using a SmartSpec™ Plus spectrometer (Bio-Rad, Hercules, California, United States). The concentration of unknown samples were read directly from the standard curve.

Immobilisation Metal Ion Affinity Chromatography

Immobilised Metal Ion Affinity Chromatography (IMAC) is the most commonly used procedure for purifying recombinant protein with either a C- or an N- terminal poly-His tag. It allows a one-step purification of the protein of interest to be achieved. In this project the metal-chelating ligand was nitrilotriacetic acid immobilized onto a polysaccharide matrix through a spacer arm. The resin was charged with Ni^{2+} ions which allowed the retention of recombinant protein through coordination of the poly-His tag to the metal ions. Proteins were eluted from the column using an imidazole gradient which competes with the His residues for coordination to the Ni^{2+}. Proteins with only few surface histidines are eluted with low concentrations of imidazole, while those with a poly-His tag are retained until high concentrations of imidazole are reached.

In this study approximately 8 mL of Chelating Sepharose resin (GE Healthcare, Little Chalfont, UK) was packed into a low pressure, 10 cm in length and 0.9 cm in diameter, Econo-column (Bio-Rad). The resin was charged with Ni^{2+} ions using a 0.1 M $NiCl_2$ (Ajax) and excess unbound metal ions were removed from the column by extensive washing with Milli-Q water. The loading buffer consisted of 50 mM Tris-HCl (Pure Science) pH 8.5, 0.5 M NaCl (Pure Science), 1 mM TCEP, 10 mM imidazole (Apollo Scientific Limited) and 0.5 mM NAD^+. The column was equilibrated with loading buffer, then crude lysate that had been filtered through a 0.8 μM filter (Ahlstrom) was loaded onto the column at a flow rate of 1 mL/min. The column was washed with 5 column volumes of loading buffer before bound proteins were eluted using a step-wise gradient of 50 mM then 100 mM imidazole, both for 33 mLs, followed by a linear gradient of 100 – 300 mM, for 70 mLs, imidazole and finally another linear gradient from 300 mM – 500 mM, for 20 mLs, of loading buffer. The chromatography was carried out at 4 °C using an AKTA Explorer 10S (GE Healthcare, Little Chalfont, UK). The flow rate was set at 1 mL/min, and 2 mL fractions were collected then analysed using SDS-PAGE. The resin was regenerated by stripping the resin using 0.5 M EDTA (Tyco) at pH 8.0 then washing extensively with Milli-Q water.

Size Exclusion Chromatography

Size Exclusion Chromatography (SEC) separates proteins on the basis of molecular size. The matrix is porous, acting like a molecular sieve that excludes large molecules which are eluted in the void volume and retains small molecules that are eluted from the column between the void volume and total column volume in order of decreasing molecular weight.

SEC using a Superdex S200 HR 10/300 column (GE Healthcare, Little Chalfont, UK) was used as a final purification step in the purification of all proteins discussed in this thesis and to exchange DHQS2 into more suitable buffers for crystallisation trials. It also removes not only other protein molecules, but other non-proteinaceous material from the sample, as well as protein aggregates, rendering the sample more homogeneous. Homogeneity of the sample is extremely important for successful crystallisation. 400 μL aliquots of protein were injected onto the column at a flow rate of 0.6 mL/min and eluted over 60 minutes. The SEC buffer consisted of 10 mM K_2HPO_4, 50 mM NaCl, 2 mM DTT and 0.5 mM NAD^+. The elution was

monitored by UV absorbance at 280 and 260 nm. Peak fractions were analysed by SDS-PAGE before the most pure were pooled for crystallisation trials.

Proteomic Methods for Protein Identification.

SDS-PAGE gels were placed on a glass plate and the protein band thought to contain DHQS2 was excised and chopped into fine pieces with a new, sterile scalpel blade, then placed in an LoBind 0.5 mL tube (Eppendorf, Germany). The gel pieces were destained by repeated washings in 300 μL of 50 % acetonitrile at 45 °C until colourless, then dehydrated in 80 % acetonitrile until they shrunk and became opaque. They were then reduced by the addition of 30 μL of freshly made reducing solution (10 mM DTT and 50 mM ammonium bicarbonate (AmBic)), at 42 °C for one hour. The supernatant was carefully removed from the gel pieces and washed in 50 mM AmBic for 5 minutes before the supernatant was removed, the gel pieces dehydrated in 80 % acetonitrile and dried in a DNA SpeedVac DNA110 (Savant, Holbrook, New York, United States). 30 μL of fresh alkylating solution (20 mM iodacetamide in 50 mM AmBic) was added to the dry gel pieces and incubated in the dark for 20 minutes at room temperature. After this time, excess alkylation solution was removed by washing in 50 mM AmBic. The washing solution was removed and the gel pieces dehydrated by adding 80 % acetonitrile. Dehydration steps were repeated until the gel pieces were bone dry. The gel pieces were then swelled in 25 μL of digestion buffer (20 ng/μL T6567 trypsin (Sigma) in 50 mM AmBic) and incubated on ice for 10 minutes. 30 μL of 50 mM AmBic was added and the gel pieces incubated at 37 °C from four hours to overnight. Gel pieces were pelleted by centrifugation (10,000x g for 2 minutes) and sonicated in a sonication bath (Elmasonic S 15, Elma, Wetzikon, Switzerland). The supernatant was collected in a new tube and pooled with the supernatant from the following washes; 60 μL of 5 % formic acid in 50 % acetonitrile, then 60 μL of 0.1 % formic acid in 80 % acetonitrile. The pooled volume was reduced to a volume of 20 μL before being subjected to centrifugation at 10,000x g for 15 minutes. The top half was carefully transferred into sample vials by pipetting for mass spectrometry analysis or stored at -80 °C.

Tandem Mass Spectrometry (MS-MS)

The trypsin-digested sample was subjected to electrospray ionization (ESI) (62) at a capillary voltage of 1, 875 V and analysed using an Aligent 6520 Q-TOF mass spectrometer (Agilent Technologies, Germany). MS scans (350-1600 *m/z*) were obtained with fragmentor and skimmer voltages at 175 and 65 V, respectively. The five most abundant ions were subjected to MS-MS fragmentation at a collision energy slope of 3.6 V / 100Da with an offset of -3 V at a scanning range of 50 – 1600 *m/z*. Searches were performed using MASCOT (63) (version 2.4.1) with the following parameters: Database restricted taxonomy to *Archaea*, trypsin digestion with up to two miscleavages, carbamidomethylated cysteine as a fixed modification, charge states of +2, +3 and +4, MS and MS/MS tolerances of 20 and 50 ppm, respectively. Search results were filtered based on individual ion scores indicating identity or homology.

2.4 Methods for Chapter Six

Circular Dichroism Analysis

Circular dichroism (CD) spectra were measured on a Chirascan™ CD spectrometer (Applied Photophysics, United Kingdom). Protein samples (10 mg/mL in SEC buffer, pH 7.2) were diluted to a final concentration of 1 mg/mL with 20 mM phosphate buffer to a final volume of 1 mL, degassed and introduced to a clean 10 mm and 0.1 mm path length Precision Quartz SUPRSIL® cells (Hellma®, Germany for near and far UV CD measurements) respectively. CD spectra were measured at 25 °C. Wavelengths of 180 – 260 nm were scanned for far UV spectra, and 260 – 320 nm for near UV CD. The wavelength interval was 0.5 nm with a time point of 0.5 seconds and a 1 mm bandwidth. Raw data was analysed with Pro-Data-Viewer (Chirascan).

Crystallisation Trials

Screening for crystallisation conditions was achieved using sitting-drop vapour diffusion in 96-well intelliplates (Hampton Research, Aliso Viejo, California, United States) at 21 °C by

mixing 400 nL of protein solution (~10 mg/mL) with 400 nL of well solution using a Mosquito® robot (TTP Labtech, Melbourn, UK) for the transfer of mother liquor and protein solution from the appropriate wells. Screens included JCSG-plus™ (Molecular Dimensions, Suffolk, UK), MIDAS™ (Molecular Dimensions) and Morpheus® (Molecular Dimensions).

Screening Crystals for Diffraction

The growth of suitable looking crystals was always a cause for optimism. First a suitable cryoprotectant was found for each of these solutions. This involved mixing mother liquor with glycerol to produce 15, 20, 25, and 30 % glycerol solutions. These were then placed in a loop, (0.1 – 0.2 mm CryoLoop mounted on a CrystalCap™ Copper HT pin, Hampton Research, Aliso Viejo, California, United States) flash frozen at -120 ºC and if they formed a glass, exposed to x-rays to check for the absence of water rings. The loop was used to pick suitable crystals from a 96-well intelliplate (Hampton Research). The loop containing the crystal was then dipped into the appropriate cryoprotectant and the mounted crystal placed onto the Spider X-ray Diffractometer (Rigaku, Tokyo, Japan) consisting of a MM007 microfocus rotating anode generator (Cu Kalpha), osmic focusing mirrors, inverse phi crystal mounting system, Rigaku R-AXIS IV++ image plate detector, and an X-STREAM cryogenic device. Initial diffraction patterns were collected at two angles, 90 degrees apart, usually 0° and 90°. The data collection was controlled by CrystalClear-SM Expert 2.0 software (Rigaku, Tokyo, Japan).

Optimising Crystallisation Conditions

Further manual screening was performed in order to optimise a crystallization condition using hanging-drop vapour diffusion in 24-well plates (Linbro). 1 μL of protein was added to 1 μL of mother liquor suspended above 400 μL of the corresponding mother liquor.

Each variable within the original condition was slightly changed using a 6 x 4 matrix in order to determine the optimum conditions. For example pH was varied 0.5 pH increments against precipitant concentration in 5 % (w/v) or (v/v) increments.

Chapter Three: DHQS2 Target Selection

3.1 Selection Criteria

Obtaining diffraction quality crystals is notoriously difficult. In order to improve the likelihood of obtaining such crystals, five DHQS2 enzymes were targeted based on properties assessed using bioinformatic analyses. Due to the lack of DHQS2 crystal structures in the PDB, it was decided to produce only Se-met derivatives to ensure the solution of the possible crystal structures. Therefore, the number of methionine residues in the sequence was important. Ideally, methionine should make up 1.5-3% of the total amino acid residues (64). The percentage of methionine in each of the potential targets was determined using ExPASy (65).

It is also imperative that predicted entropic disorder is minimal, as this also adversely affects the chances of successful crystallisation. Because the genes were to be cloned and expressed in *E. coli* host cells, 127 fully sequenced prokaryotic DHQS2 genes in the Pfam (66) database were selected as potential targets. The Pfam database is a collection of protein families represented by multiple sequence alignments (67). 80 out of 127 of these strains had DNA available for purchase from DSMZ. The resulting 80 targets were subjected to bioinformatic servers XtalPred (68) and Predictor of natural disordered regions (PONDR) (69) .

XtalPred compares predicted biochemical and biophysical features of the submitted protein with corresponding distributions of crystallisation probability calculated from TargetDB (70), which is a protein crystallisation database. The prediction is made by combining individual crystallization probabilities calculated for eight protein features into a single crystallization score. The features used are: number of residues, isoelectric point, GRAVY index, predicted structural disorder, instability index, predicted coil secondary structure, predicted coiled-coil structure and insertion score. Based on this score, the protein is assigned to one of the four crystallization classes. The EP classification outputs numbers 1 - 4, were the predicted ability to crystallise decreases from 1 - 4.

Given that the amino acid sequence determines 3-D structure, it is logical that the location of certain residues in the sequence can influence local structure and hence structural stability. Disordered regions are not structural elements that have been removed from their folding context. They are defined as regions of proteins that lack a fixed tertiary structure. The definition of "disorder" does not apply to the residue side-chains but rather the protein backbone. Ordered regions have the same Ramachandran angles among the different copies of the protein, while disordered protein regions have different, often dynamic, Ramachandran angles among the set members. PONDR predicts natural regions of disorder by sequencing attributes taken over windows of 9 to 21 amino acids. The attributes, fractional composition of particular amino acids, hydropathy, and sequence complexity, are averaged over these windows. The same values are used as inputs to make predictions. PONDR outputs are between 0 and 1, where a value that exceeds or matches a threshold of 0.5 is considered disordered. Any protein sequence greater than 50% disordered was not considered as a target (Fig. 3.1).

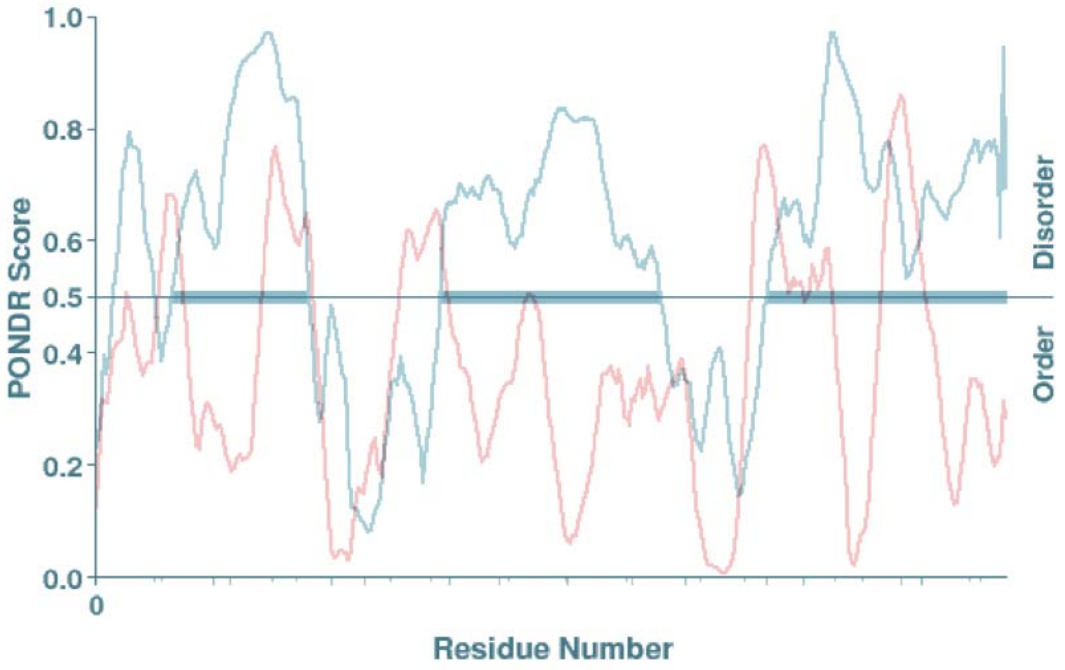

Fig. 3. 1 PONDR disorder score comparison of DHQS2 from *Methanocorpusculum labreanum* (70%, blue) and *Haloarcula marismortui* (28%, red). Disorder percentage score is calculated from the number of residues above the 0.5 threshold.

There is a conserved RPFRVNAG motif in all 80 targets generated by Logo3 (71) (Fig. 3.2), which may be significant and contain the conserved catalytic residues in the DHQS2 active site. A T-Coffee (72) sequence alignment generation-based phylogenetic tree was calculated to portray the variety in the sequences of DHQS2 (Fig 3. 3) within different prokaryotic species and was the final criteria used to select enzymes to be cloned. Species that best fitted the overall criteria: methionine content greater than 1.5%; an Xtalpred EP classification of 1 or 2; a PONDR disorder percentage below 50% were selected as targets. A sixth target, *M. maripaludis* (2.2% methionine, EP classification 3, 34% disordered), that did not meet any of the criteria, was chosen in order to test the selection strategy used.

The following target DHQS2 genes that best fit the selection criteria were; *Sulfurihydrogenibium azorense* (2.7% methionine, EP classification 1, 22% disordered), *Desulfatibacillum alkenivorans* (2.7% methionine, EP classification 1, 28% disordered), *Methanospirillum hungatei* (2.1% methionine, EP classification 1, 26% disordered), *Methanosphaera stadtmanae* (2.2% methionine, EP classification 2, 30% disordered) and *Archaeoglobus veneficus* (1.8% methionine, EP classification 1, 30% disordered) as shown in Table 3.1. *Methanococcus maripaludis* (2.2% methionine, EP classification 3, 34% disordered) did not met the selection criteria, with an EP classification higher than 2, however it was chosen as a target in order to test the viability of the predictive servers that the selection was based on.

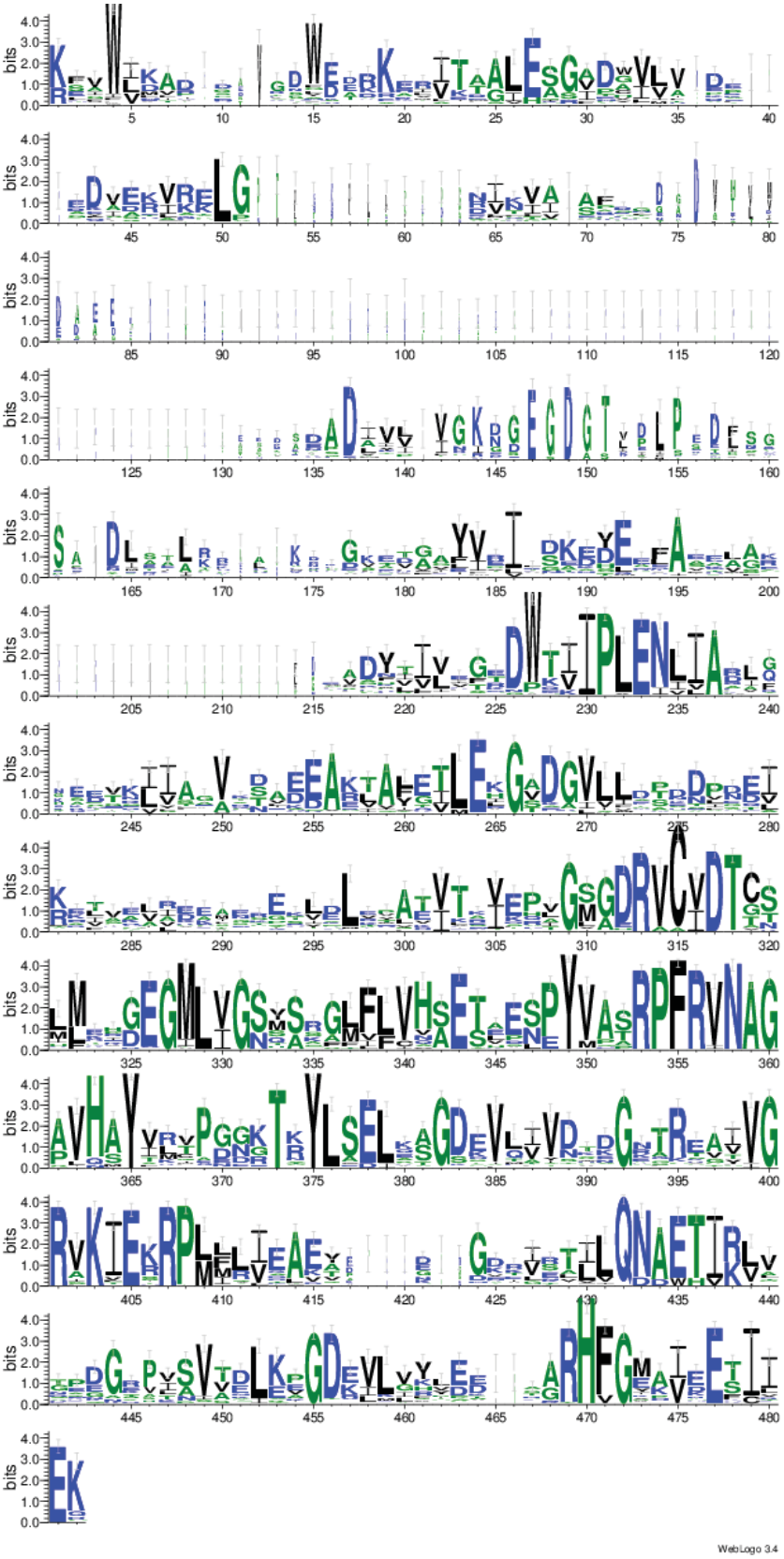

Fig. 3. 2. Common residues throughout the possible 80 target DHQS type 2 species generated by Logo3. The height of residue in bits conveys the level of conservation of each residue.

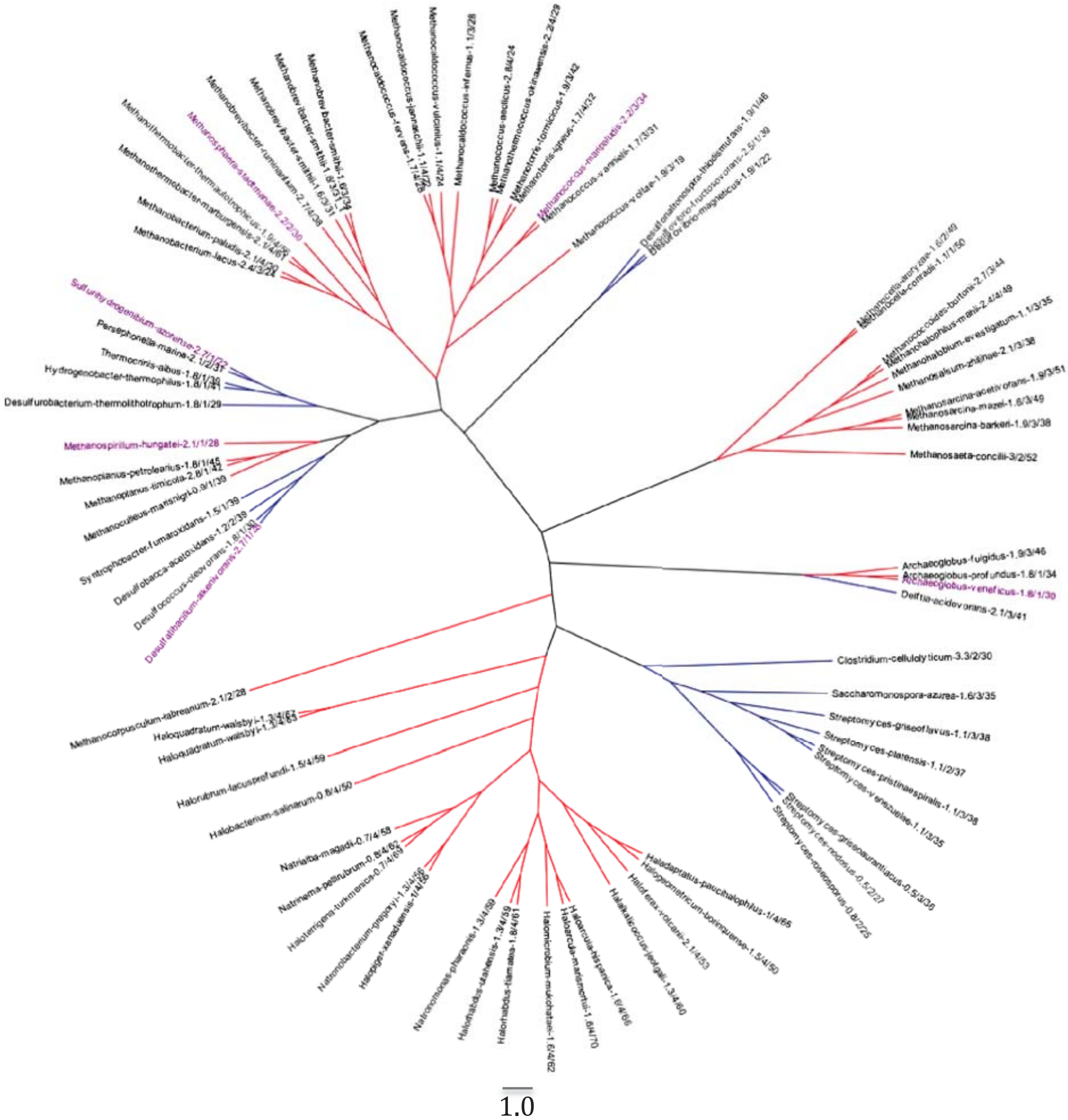

Fig. 3. 3. Phylogenetic tree of selected 3-dehydroquinate synthase type 2 (DHQS2) target species based on amino acid alignment. Alignment of FASTA sequences of target species by T-coffee software was uploaded into FigTree v1.4.2 to produce phylogenetic tree. Reference bar indicates the average distance for 1.0 amino acid changes in the aligned FASTA sequences. Red branches indicate archaeal species, while blue branches indicate bacterial species. Labels show the species name followed by methionine content, Xtalpred server EP classification, and PONDR disorder prediction server percentage. Purple labels indicate the chosen target species based on overall criteria.

Table 3. 1. Target Species selected on the basis of bioinformatics in order to increase the likelihood of characterizing DHQS2.

Target Species	Xtalpred EP classification	PONDR disorder (%)	Methionine (%)
S. azorense	1	22	2.7
M. stadtmanae	2	30	2.2
A. veneficus	1	30	1.8
M. hungatei	1	28	2.1
D. alkenivorans	1	28	2.7
M. maripaludis	3	34	2.2

3.2 Structure Predictions

Each selected DHQS2 target FASTA sequence (Appendix one) was input into a protein structure prediction iterative threading assembly refinement (I-TASSER) server (73), (74). Protein structure prediction refers to the generation of 3-dimensional models from amino acid sequences using computer algorithms. I-TASSER is a hierarchical protein structure modelling approach based on the secondary-structure enhanced Profile-Profile threading Alignment (75) and the iterative implementation of the TASSER program (76). For each submitted sequence I-TASSER predicts up to five models ranked on the basis of free energy as calculated using SPICKER (77). SPIKER identifies near-native models from a pool of protein structure decoys. Each model given a confidence score (C-score), which is based on the significance of threading template alignments and parameters of the structure assembly. The C-score is typically in the range of -5 to 2, where a higher value signifies a model with high confidence and vice-versa.

The best I-TASSER result for the *M. maripaludis* sequence predicted that it would have a secondary structure of alternating α-helices and β-strands between residues 1 – 180 and a domain rich in β-strands between domain from residues 181 – 361 (C-score of -2.04) (Fig. 3.4). The model shows an α-helical rich domain and a β-strand rich domain. I-TASSER uses the TM-align structural alignment program (78) to match the first I-TASSER model to all structures in the PDB library. In this case the top ranked analogue was alanine racemase

(1BD0) that has a homo-dimer quaternary structure.

The threading results for *S. azorense* predicted DHQS2 sequence predicted a secondary structure similar to that for M. maripaludis, with alternating α-helices and β-strands from residue 1 – 180 and a β-strand rich domain from residues 181 – 330. The C-score for the predicted model was, however, very poor, -4.11, indicating the model to be unreliable. In contrast to the secondary structure prediction, the model had a β-domain as well as a disordered domain with few secondary structural elements (Fig. 3.5). The top ranked analogue was the $\alpha_5\beta_1$ integrin headpiece (3VI3) that has a hetero-tetrameric quaternary structure.

Fig. 3. 4. Predicted Structure of DHQS2 from *M. maripaludis*. α-Helices are shown in pink, β-strands in yellow and random coils and loops in white and blue respectively.

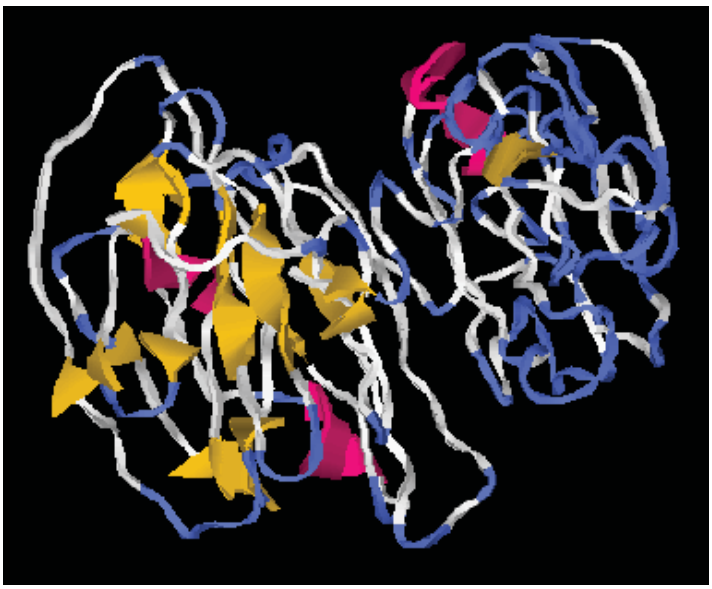

Fig. 3. 5. Predicted Structure of DHQS2 from *S. azorense*. α-Helices are shown in pink, β-strands in yellow and random coils and loops in white and blue respectively

When the DHQS2 sequence of *D. alkenivorans* was subjected to I-TASSER, the model predicted that residues 1-180 would take the form of alternating α-helices and β-strands, while 181-329 would take the form of a β-strand rich domain. The best model, shown in Fig. 3.6, had little α-helical structure in either the N- or C-terminal domains, and the C-score 0f -4.22, meant the model was unreliable. The top ranked analogue was $\alpha_5\beta_3$ integrin ectodomain (3IJE) that is a hetero-dimer.

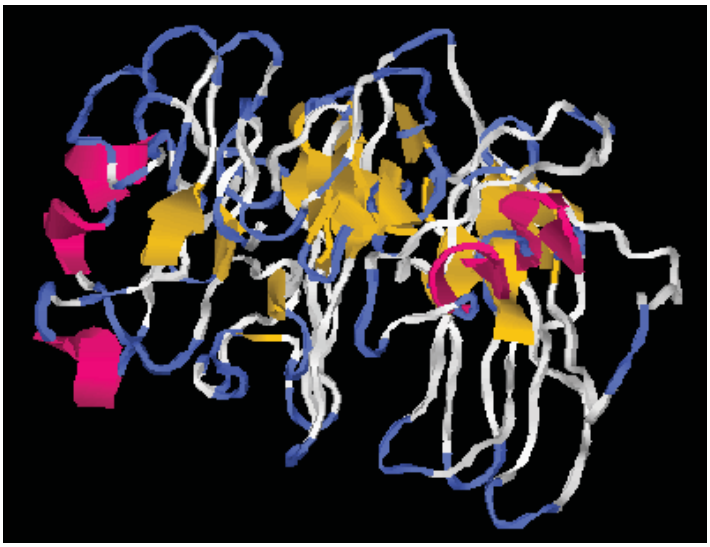

Fig. 3. 6. Predicted Structure of DHQS2 from *D. alkenivorans*. α-Helices are shown in pink, β-strands in yellow and random coils and loops in white and blue respectively

The predicted structure for the *M. hungatei* DHQS2 sequence suggested that residues 1-180 contained alternating α-helices and β-strands, and residues 181 – 327 formed a β-strand domain. The best model predicted by I-TASSER had a C-score of -3.09 (Fig. 3.7) and contained a rich α-helical domain and a β-strand domain. The top ranked analogue was alanine racemase from the pathogenic bacterium, *Pseudomonas aeruginosa* (1RCQ), a protein that exists as a homo-dimer in solution. This was also picked as being the best analogue for the *M. maripaludis* model.

Fig. 3. 7. Predicted Structure of DHQS2 from *M. hungatei*. α-Helices are shown in pink, β-strands in yellow and random coils and loops in white and blue respectively.

When the DHQS2 sequence from *M. stadtmanae* was subjected to the I-TASSER server, a model with a two domain structure was predicted (Fig. 3.8). The N-terminal α-helical domain contained residues 1- 180 while the C-terminal domain, a β-domain, contained residues 181 – 358, with a C-score of -2.67. The top ranked analogue was photosystem II D1 protease (1FC7).

The I-TASSER result for the *A. veneficus* sequence, again predicted a two domain structure; the N-terminal domain being characterised by mainly α-structure (residues 1 – 180) and the C-terminal domain containing mainly β-structure (residues 181 – 336) (Fig. 3.9). The low C-score of -4.53 indicates that the model is very unlikely to occur. The top ranked analogue is once again alanine racemase from S*treptococcus pneumoniae* (3S46), similar to that found for the predicted *M. hungatei* DHQS2 structure.

Fig. 3. 8. Predicted Structure of DHQS2 from *M. stadtmanae*. α-Helices are shown in pink, β-strands in yellow and random coils and loops in white and blue respectively

Fig. 3. 9. Predicted Structure of DHQS2 from *A. veneficus*. α-Helices are shown in pink, β-strands in yellow and random coils and loops in white and blue respectively

Chapter Four: Cloning, Expression and Solubility of DHQS2

4.1 Assembly of gBlock into the pETite Vector

The pETite vector was amplified by PCR, using the pETite forward and reverse primers (Table 2.1), and visualised by agarose gel electrophoresis (Fig. 4.1). The pETite vector is ~2200 bp long as shown in Fig. 4.1. The PCR products synthesised using two different annealing temperatures are shown in lanes 1, (53 ºC), and 2 (58 ºC).

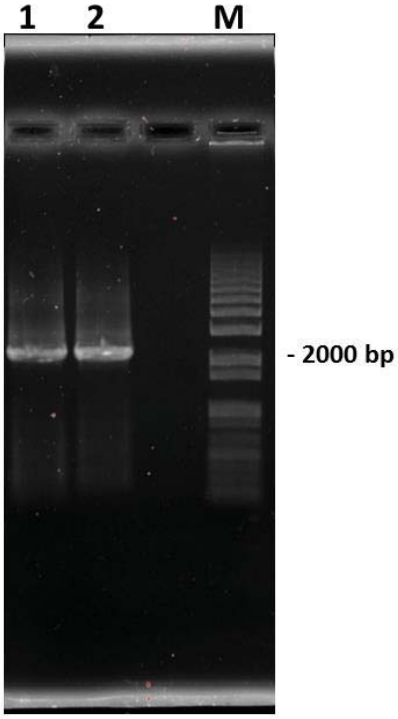

Fig. 4. 1. Amplification of the pETite vector. Lanes 1 and 2 are the PCR products synthesised using annealing temperatures of 53 ºC and 58 ºC, respectively. A 1 kb plus molecular size marker (M) was used as a size standard.

The amplified pETite vector and gBlock were assembled *via* Gibson assembly to a construct a modified pETite vector (modpETite), which enabled the production of either N- or C-terminally His_8-tagged recombinant protein. The modpETite product was transformed by heat shock into DH5α *E. coli* cells, and the incorporated gBlock was detected using colony PCR and visualised by agarose gel electrophoresis (Fig. 4.2). An expected PCR product size

of ~500 bp was seen in each of the three colonies picked, indicating the presence of a gBlock in the modpETite vector.

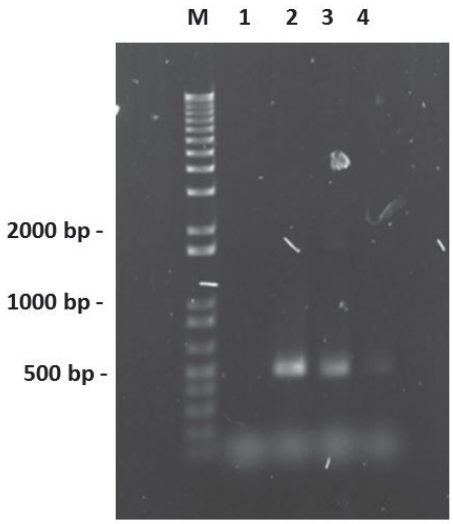

Fig. 4. 2. gBlock colony PCR. From left to right; a 1 kb plus molecular size marker (M), an empty vector negative control (1), colony one (2), colony two (3) and colony three (4).

4.2 Assembly of DHQS2 targets into a modpETite vector

The DHQS2 genes from *A. veneficus, D. alkenivorans, M. maripaludis, M. hungatei, S. azorense* and *M. stadtmanae,* were successfully amplified by PCR from the corresponding genomic DNA. For each gene, the DNA was amplified using two sets of primers; an N-terminal priming set, and a C-terminal priming set (Table 2.1). The N-terminal reverse primer encodes a stop codon to stop translation of a C- terminal His_8-tag, while the C – terminals reverse primer does not contain an in-frame stop codon. Similarly, two sets of primers were used for the modpETite plasmid, (C- modpETite forward and reverse and N-modpETite forward and reverse; Table 2.1) in order to prime for C- and N-terminally His_8-tagged modpETite vectors, respectively. Both N- and C-terminal primed modpETite vectors and DHQS2 target genomic DNA were amplified by PCR then visualized by agarose gel electrophoresis; examples are shown in Fig. 4.3. The amplified products were all of the expected size; ~2.4 kb and ~1.0 kb, for the N- and C-terminal modpETite vectors and the target DHQS2 genes respectively. Fig. 4.3 shows the PCR products for the N-terminally primed DHQS2 genes for target species *D. alkenivorans,* (N-*Dal*), *M. stadtmanae (*N-*Mst),*

*M. hungatei (*N-*Mhu) and S. azorense (*N-*Saz)*. Figure 4.4 shows the PCR products of the C-terminally primed DHQS2 genes for target species *D. alkenivorans,* (C-*Dal*), *M. stadtmanae (*C-*Mst), M. hungatei (*C-*Mhu) and S. azorense (*C-*Saz).* Fig. 4.5 shows the N- and C-terminally primed DHQS2 genes for *M. maripaludis* (N-*Mma* and C-*Mma*) and *A. veneficus* (N-*Ave* and C-*Ave*). The N- terminally primed DHQS2 genes were then assembled into the N- terminally primed modpETite vector *via* Gibson assembly, and likewise for assembly of C- terminally primed genes and vector. However before Gibson assembly, the ~2.4 kb DNA band was extracted by slicing and DNA purified, as required to remove the ~1 kb and ~5 kb secondary products and PCR reaction contaminants to ensure single plasmid assembly.

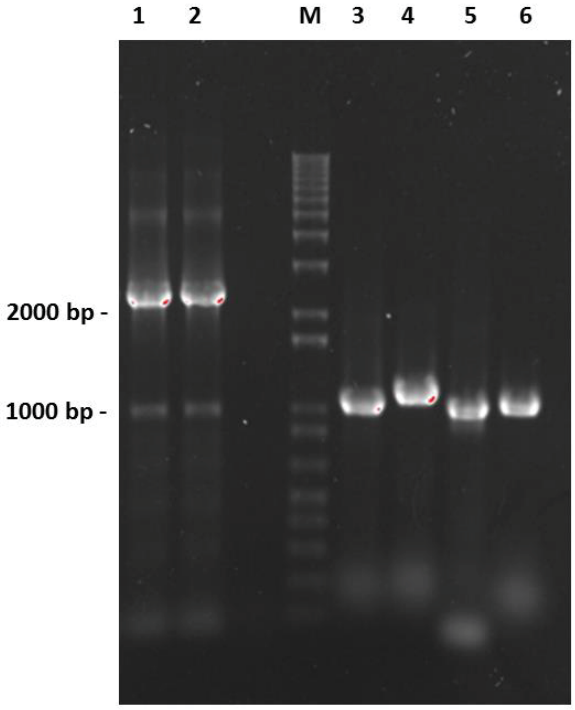

Fig. 4. 3. Amplification of C- and N- terminal pETite Vector and DHQS2 genes. The C- and N-terminally primed modpETite vectors have a molecular size of ~2.4 kb represented by the more intense bands in lanes 1 and 2 respectively. There is a secondary product 1 kb in size, which may be due to a secondary annealing site for one of the primers. A molecular size marker (M), for size determination and the DHQS2 genes *D. alkenivorans* (N-*Dal*), *M. stadtmanae (*N-*Mst), M. hungatei (*N-*Mhu) and S. azorense (*N-*Saz)*, amplified from genomic DNA, have an approximate size of 1 kb.

After Gibson assembly the plasmids were transformed, initially into *E. coli* XL-1 cells with limited competency and then, into *E. coli* DH5α cells resulting in a higher efficiency of plasmid uptake. Colonies on each of the transformation plates were picked and checked for the presence of the expected plasmid by colony PCR amplification and agarose gel

electrophoresis (Figs. 4.6, 4.7, 4.8 and 4.9). The amplified product should be the size of the insert plus the vector at ~1.5 kb.

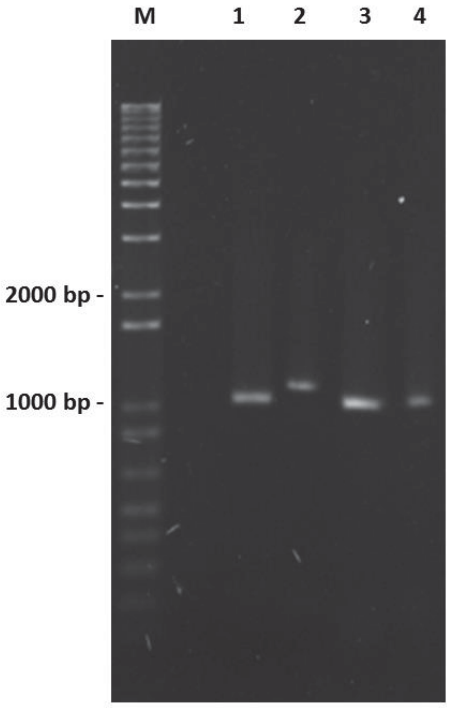

Fig. 4. 4. Amplification of DHQS2 genes. From left to right: 1 kb plus molecular size marker (M) for size determination, C-*Dal* (1), C-*Mst* (2), C-*Mhu* (3), and C-*Saz* (4).

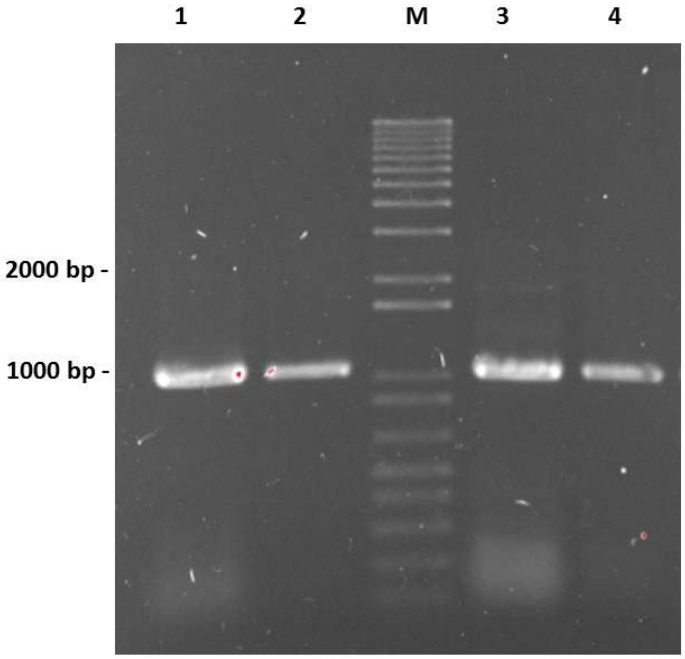

Fig. 4. 5. Amplification of DHQS2 genes continued. From left to right, C-*Mma* (1), N-*Mma* (2), 1 kb plus molecular marker (M) for size determination, C-*Ave* (3) and N-*Ave* (4).

Figure 4.6 shows a modpETite vector positive control (+), that yields a product of ~500 bp, as expected for an empty vector without the assembled DHQS2 gene. DH5α colonies were picked and screened for the presence of N- and C-*Dal* by colony PCR. Colony 2 did not have an insert as indicated by the ~500bp product resembling that of the positive control modpETite vector no insert control. C-*Dal* and N-*Dal* plasmid DNA was purified from cultures inoculated with colonies 1 and 6, respectively (Fig. 4.6).

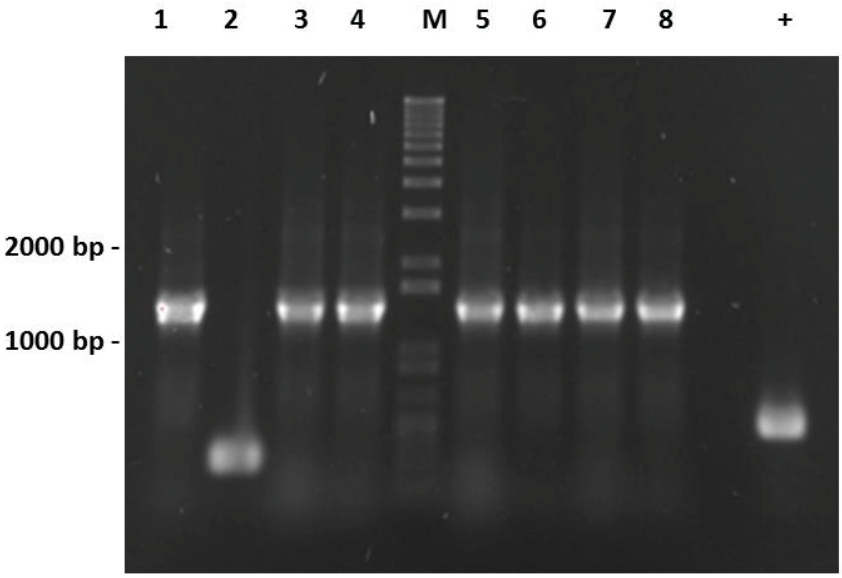

Fig. 4. 6. *E. coli* DH5α Colony PCR of *D. alkenivorans* Assembled Plasmids. Lanes 1-4 contain plasmids purified from cultures of individual C-*Dal* assembled colonies, M represents the 1 kb plus molecular size marker, lanes 5 -8 contain plasmids purified from cultures of individual N-*Dal* assembled colonies, and + is the PCR product from an empty modpETite vector template control.

Seven C-*Mma* and six N-*Mma* colonies were picked for colony PCR (Fig. 4.7A). Cultures inoculated with colonies five and ten were selected to purify plasmid from. Likewise, seven C-*Ave* and six N-*Ave* colonies were picked (Fig. 4.7B), and cultures of colonies one and 12 were subjected to plasmid DNA extraction.

None of three selected N-*Mhu* colonies picked for colony PCR contained the DHQS2 gene (Fig. 4.8). In contrast, all four C-*Mhu* colonies had a detectable DHQS2 gene, and the colony in lane five was cultured for plasmid DNA purification. No colony PCR reactions analysed in Fig. 4.9 contained an assembled DHQS2 gene, therefore it was concluded that C-*Saz*, N-

Saz, C-*Mst* and N-*Mst* were not assembled into their respective ModpETite vectors *via* Gibson assembly.

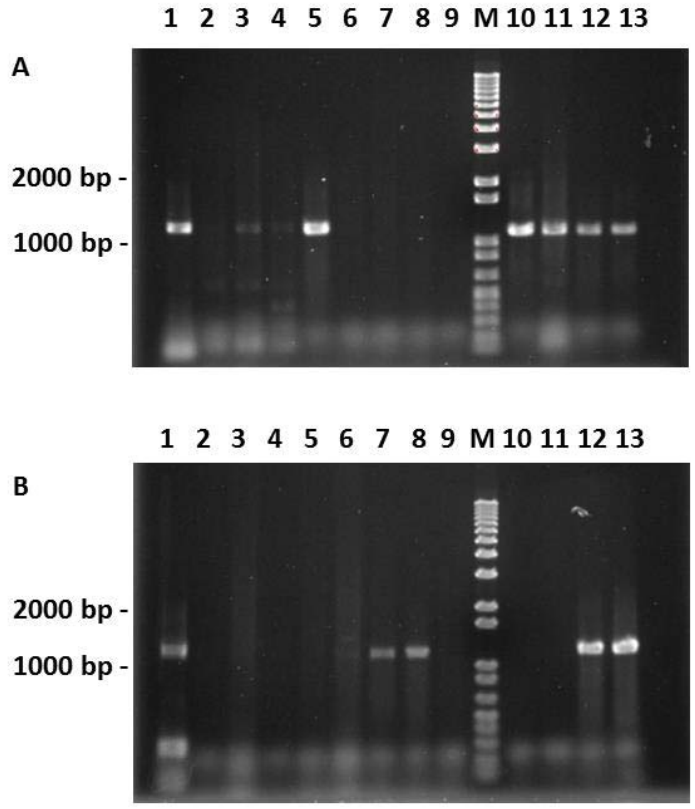

Fig. 4. 7. *E. coli* DH5α Colony PCR of (A) *M. maripaludis* and (B) *A. veneficus* Assembled Plasmids. (A) Lanes 1-7 and 8-13 are PCR products from C-*Mma* and N-*Mma* colonies respectively. (B) Lanes 1-7 and 8-13 represent C-*Ave* and N-*Ave* colonies respectively.

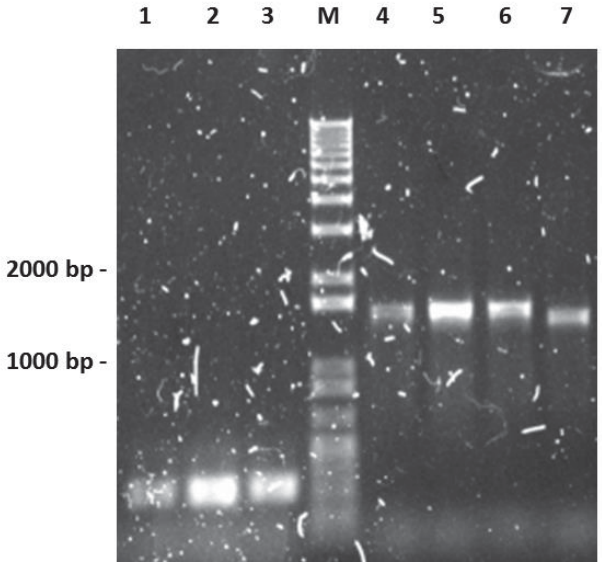

Fig. 4. 8. *E. coli* DH5α Colony PCR of *M. hungatei* Assembled Plasmids. Lanes 1-3 and 4-7 are from N-*Mhu* and C-*Mhu* colonies, respectively. M is a 1 kb plus molecular size marker.

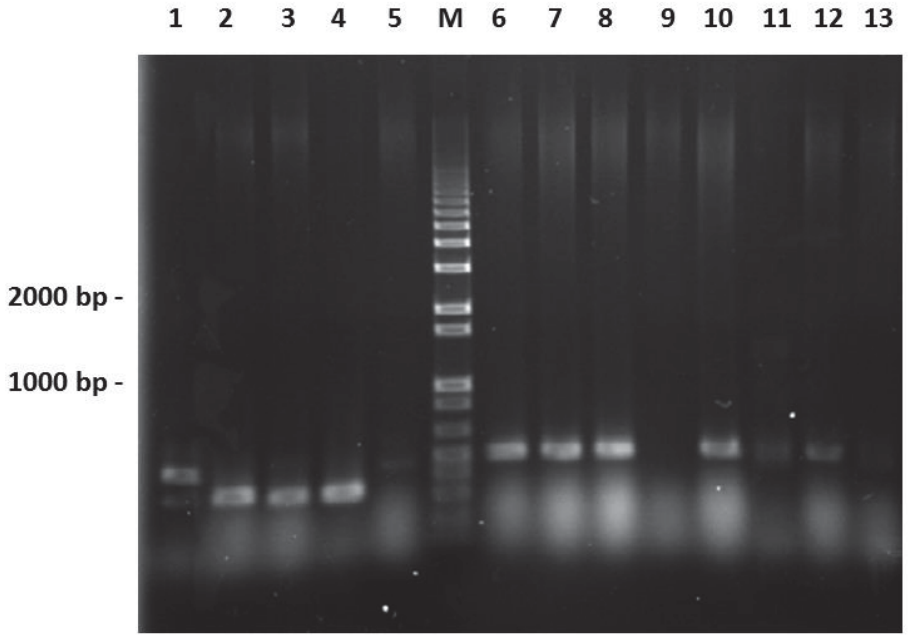

Fig. 4. 9. ***E. coli*** **DH5α Colony PCR of** ***S. azorense*** **and** ***M. stadtmanae*** **Assembled Plasmids.** Lanes 1-3, 4-6, 7-9 and 10 -13 represent C-*Saz*, N-*Saz*, C-*Mst* and N-*Mst* assembled colonies, M is a 1 kb plus molecular size marker.

Out of the 12 potential expression plasmid constructs, seven were shown to contain the inserted gene using colony PCR. These constructs were N-*Ave*DHQS2, C-*Ave*DHQS2, N-*Dal*DHQS2, C-*Da*lDHQS2, N-*Mma*DHQS2, C-*Mma*DHQS2 and C-*Mhu*DHQS2. As transformation efficiencies were acceptable for seven constructs, it appears that the remaining constructs did not assemble during Gibson assembly. A single target, C-S*az*DHQS2, was selected for an alternative cloning method to Gibson assembly, in order to determine if this cloning method was the reason for no plasmid formation, as opposed to some intrinsic property of the target DNA. A pET28b(+) vector containing a C-terminal His_8 tag was digested with restriction enzymes *Not*I and *Nco*I, and *S. azorense* genomic DNA was amplified using primers containing the same restriction enzyme sites. The vector and insert were incubated with ligase in two separate occasions using firstly Invitrogen and secondly New England Biolabs T4 DNA ligase. Both subsequent transformations were unsuccessful therefore C-S*az* could not be cloned into the pET28b(+) vector using this standard digest and ligate approach.

4.3 Expression and Solubility Trials

Sequenced constructs with the correct DHQS2 insert, start codon and and His_8 tag positioned in frame were transformed into methionine auxotroph, *E. coli* DL41 (DE3) cells. Transformation plates were supplemented with 1 % (w/v) glucose to minimize leaky expression of the *lac* operon that controls DHQS2 expression. This is because in the presence of glucose, the lac operator is not induced until all the glucose has been utilized allowing lactose uptake (79). Gene expression was induced by the non-hydrolysable lactose analog IPTG. IPTG concentrations of 0.1 mM, 0.25 mM, 0.5 mM and 1mM, were used to induce expression of N-*Mma*DHQS2 for 2, 3 and 6 hours at 25 °C and whole cell samples were analysed for protein production by SDS-PAGE as shown in Fig. 4.10. DHQS2 has a theoretical size of ~40 KDa which is the size of a band observed at high concentrations across all induced samples and is absent from the uninduced sample, strongly indicating that it represents recombinant DHQS2. A variety of IPTG concentrations and induction times were trialled to find the conditions for optimum expression of DHQS2 for each construct. From Fig. 4.10 it was established that 0.5 mM IPTG with an induction time of 6 hours led to the optimum expression of N-*Mma*DHQS2. These conditions were also found to optimally produce recombinant N-*Ave*DHQS2, C-*Ave*DHQS2, N-*Dal*DHQS2, C-*Da*lDHQS2, N-*Mma*DHQS2, C-*Mma*DHQS2 and C-*Mhu*DHQS2. When the cells were induced at 37 °C, lower cell densities were obtained which could have been due to protein toxicity. Induction at lower temperatures is likely to reduce the toxic effects of the induced protein, improving cell survival (78). The plasmid was then transformed into a number of different expression strains tailored for toxic protein expression including C41 (DE3), which has a mutation that inhibits cell death associated with protein toxicity and C41 (DE3) pLysS that includes the plasmid pLysS, which carries the gene encoding T7 lysozyme. T7 lysozyme lowers the background expression level of target genes under the control of the T7 promoter but does not interfere with the level of expression achieved following induction by IPTG. No significant increase in the level of protein expression was however observed.

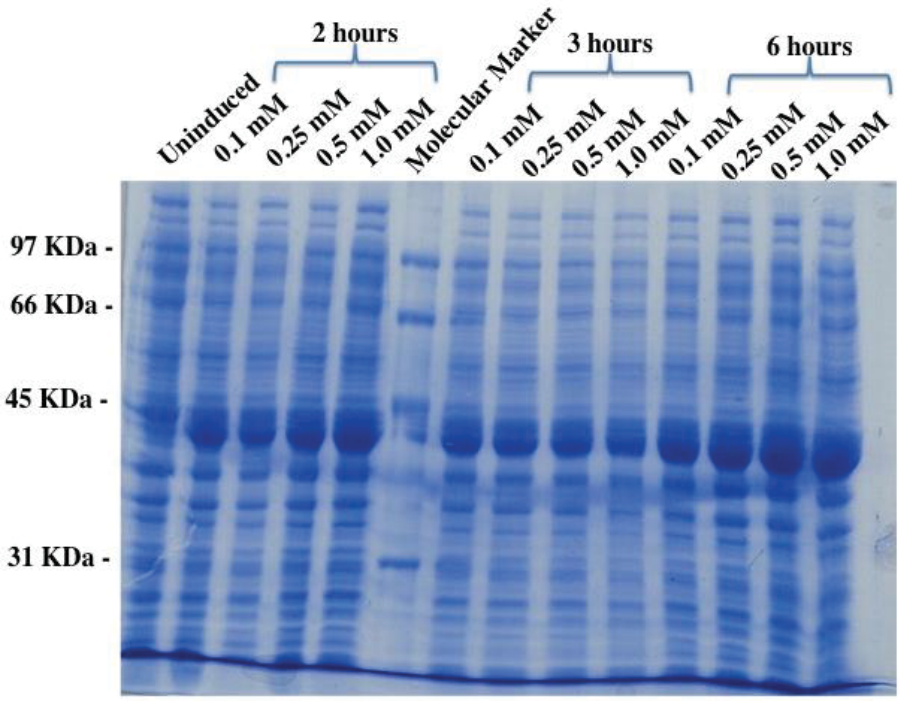

Fig. 4. 10. SDS-PAGE gel of N-*Mma*DHQS2 Expression Trials. Whole-cell samples of N-*Dal*DHQS2 taken at time intervals, 2hours, 3 hours, and 6 hours, from LB cultures that were induced with 0.1 mM, 0.25 mM, 0.5 mM and 1.0 mM concentrations of IPTG at 25 °C.

Constructs that showed recombinant protein induction were lysed using a sonicator and the whole cell lysate, supernatant and pellet fractions were analysed and compared by SDS-PAGE to determine recombinant protein solubility (Fig. 4.11, 4.12, 4.13, 4.14, 4.15 and 4.16). The over-expressed 40KDa protein band was present in both whole-cell and the soluble supernatant fractions for N-*Mma*DHQS2, C-*Mma*DHQS2 (Fig. 4.11, the difference in intensities between N-*Mma*DHQS2 and C-*Mma*DHQS2 being due to the cell density at time of induction), C-*Mhu*DHQS2 (Fig. 4.12), N-*Dal*DHQS2 (Fig. 4.13), C-*Dal*DHQS2 (Fig. 4. 14), N-*Ave*DHQS2 (Fig. 4.15) and C-*Ave*DHQS2 (Fig. 4.16). The concentration of DHQS2 was significantly less in the insoluble cell pellet fractions, indicating that the recombinant protein is mainly soluble for all constructs except C-*Mhu*DHQS2 (Fig. 4.12), which had equal amounts of DHQS2 in the soluble and insoluble fractions. C-*Mhu*DHQS2 was therefore deemed to be partially soluble. Protein solubility is crucial for further studies and the recombinant protein from all seven assembled plasmids was either soluble or partially soluble.

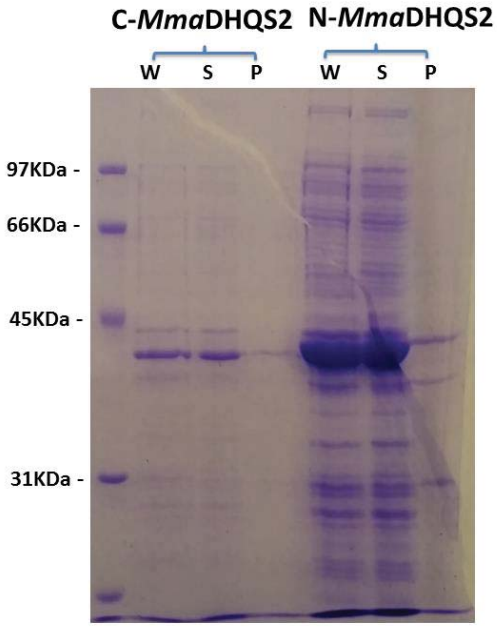

Fig. 4. 11. C-*Mma*DHQS2 and N-*Mma*DHQS2 Solubility Trials. The over-expressed DHQS2 is present at equivalent amounts in the whole cell lysate (W) and the soluble supernatant (S). DHQS2 is significantly less in the insoluble cell pellet fraction (P).

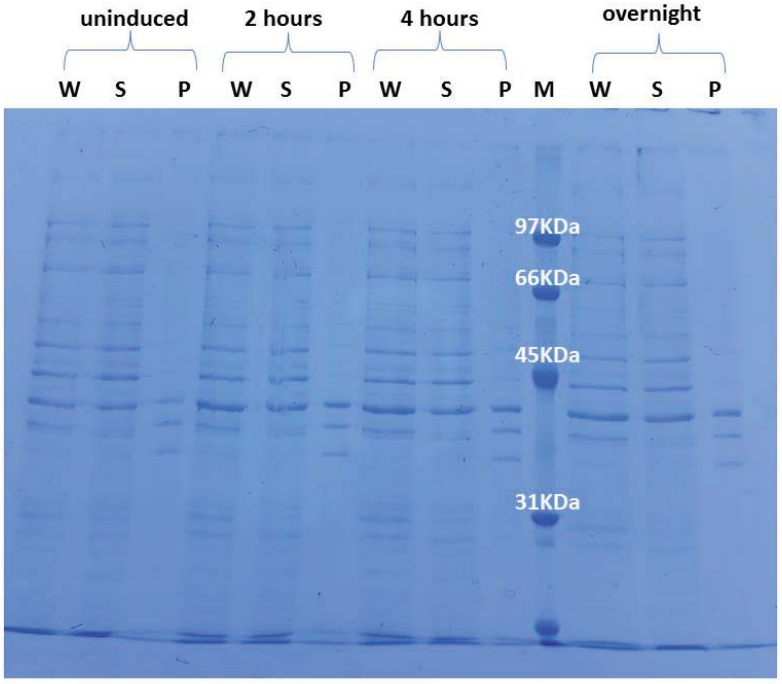

Fig. 4. 12. C-*Mhu*DHQS2 Expression and Solubility Trials. The over-expressed DHQS2 is present at equivalent amounts in the whole cell lysate (W), the soluble supernatant (S) and cell pellet fraction (P).

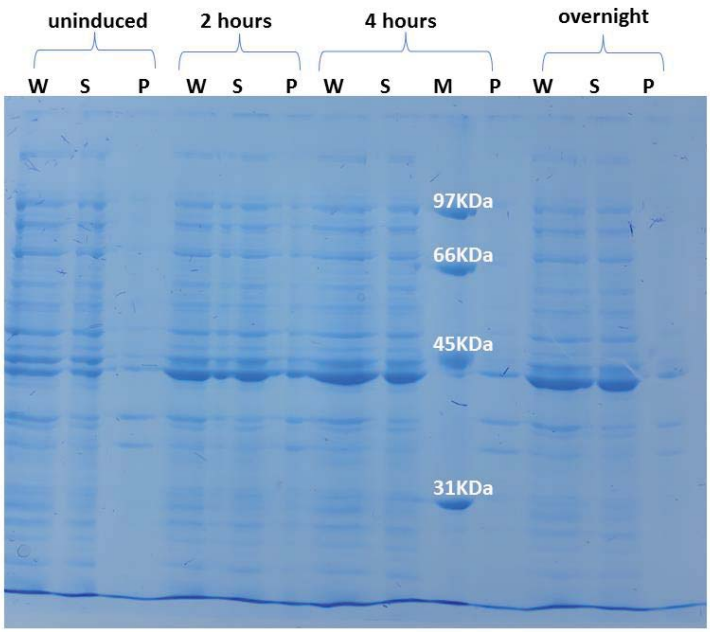

Fig. 4. 13. N-*Dal*DHQS2 Expression and Solubility Trials. The over-expressed DHQS2 is present at equivalent amounts in the whole cell lysate (W) and the soluble supernatant (S). DHQS2 is significantly less in the insoluble cell pellet fraction (P).

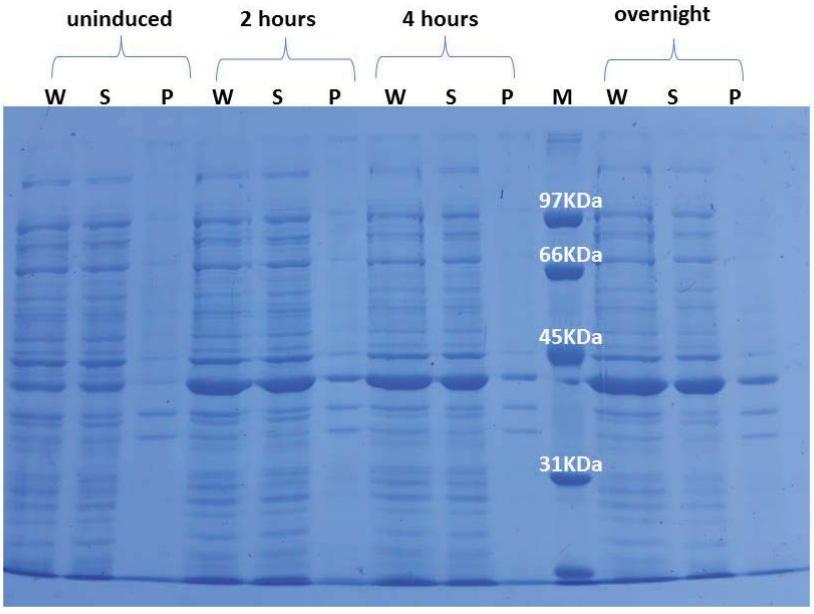

Fig. 4. 14. C-*Dal*DHQS2 Expression and Solubility Trials. The over-expressed DHQS2 is present at equivalent amounts in the whole cell lysate (W) and the soluble supernatant (S). DHQS2 is significantly less in the insoluble cell pellet fraction (P).

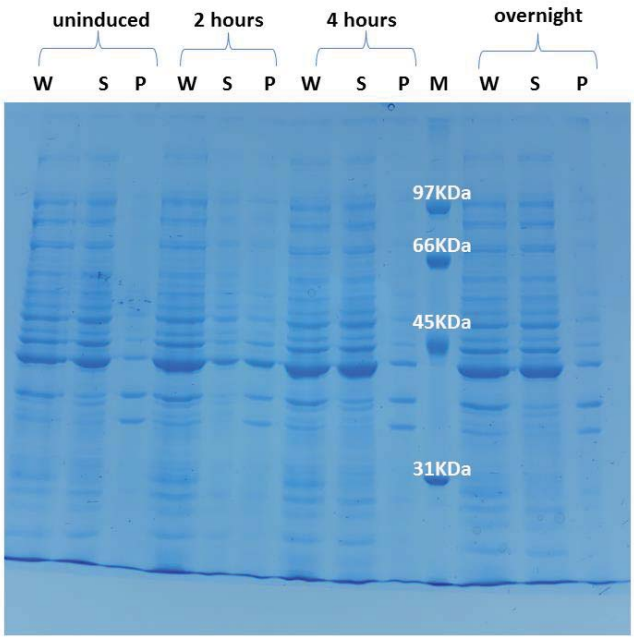

Fig. 4. 15. N-*Ave*DHQS2 Expression and Solubility Trials. The over-expressed DHQS2 is present at equivalent amounts in the whole cell lysate (W) and the soluble supernatant (S). DHQS2 is significantly less in the insoluble cell pellet fraction (P).

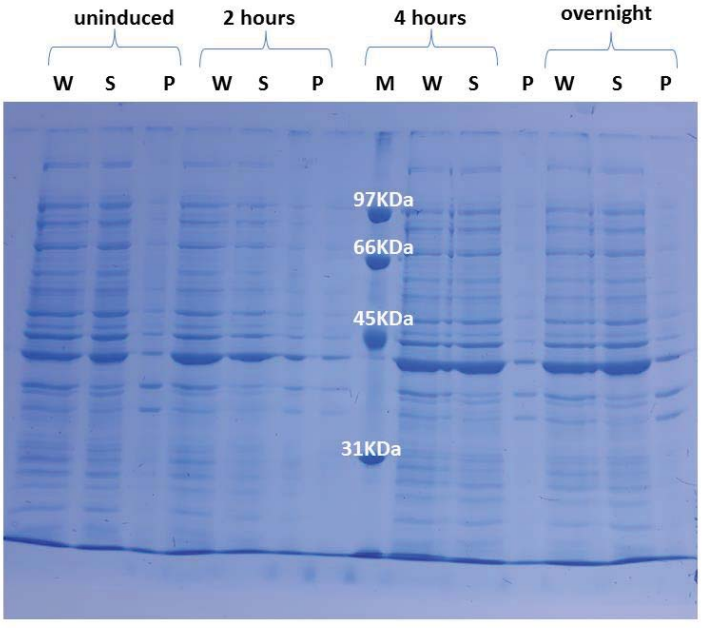

Fig. 4. 16. C-*Ave*DHQS2 Expression and Solubility Trials. The over-expressed DHQS2 is present at equivalent amounts in the whole cell lysate (W) and the soluble supernatant (S). DHQS2 is significantly less in the insoluble cell pellet fraction (P).

Chapter Five: Purification of DHQS2

5.1 Immobilised Metal Ion Affinity Chromatography (IMAC)

DL41 (DE3) cells were inoculated into 1L of M9 minimal media chosen to facilitate the incorporation of Se-Met into the recombinant DHQS2. The culture was grown to an OD_{600} of 0.5 at 37 °C with shaking at 200 rpm, then transferred to 25 °C until the OD600 reached 1. It was then induced by making the media 0.5 mM IPTG for six hours. The cells were lysed using a French press and after the insoluble material had been removed by centrifugation, the soluble fraction was subjected to immobilisation metal ion affinity chromatography (IMAC) (80). The resin, in the column, was charged with Ni^{2+} ions which promoted the retention of recombinant protein through coordination of the His_8-tag to the metal ions. After thorough washing with loading buffer, proteins were eluted from the column using increasing imidazole concentrations. As imidazole competes with the His residues to coordinate to Ni^{2+}, proteins with only few surface histidines are eluted with low concentrations of imidazole, while those with a poly-His tag are retained until high concentrations of imidazole are reached. The amount of imidazole was controlled by the ratio of loading buffer (50 mM Tris:HCl, 0.5 mM NaCl, 1 mM TCEP, 10 mM imidazole, 0.5 mM NAD^+) to elution buffer (50mM Tris:HCl, 0.5 mM NaCl, 1 mM TCEP, 0.5 M imidazole, 0.5 mM NAD+, 0.5 M imidazole). The results of a typical elution are shown in Fig. 5.1, where the concentration of imidazole is shown as a red line. Elution comprised two 33 mL steps containing 10 and 20 mM imidazole respectively, before a linear gradient from 20 to 300 mM in a volume of 70 mL was applied. The His8 tagged DHQS2 eluted at approximately 30 mM imidazole. The imidazole concentration was increased to 0.5 mM for 30 mL to ensure all His-bound proteins were removed. The elution was monitored by absorbance at 280 nm and peak fractions analysed for purity by SDS-PAGE as follows: for N-*Dal*DHQS2 (Fig. 5.2), C-*Dal*DHQS2 (Fig. 5.3), N-*Ave*DHQS2 (Fig. 5.4), C-*Ave*DHQS2 (Fig. 5.5), C-*Mhu*DHQS2 (Fig. 5.6) N-*Mma*DHQS2 (Fig. 5.7) and C-*Mma*DHQS2 (Fig. 5.8). There are double DHQS2 bands observed in Figs. 5.2, 5.4, 5.5, and 5.7, which is believed to be an artefact of highly concentrated protein in the gels.

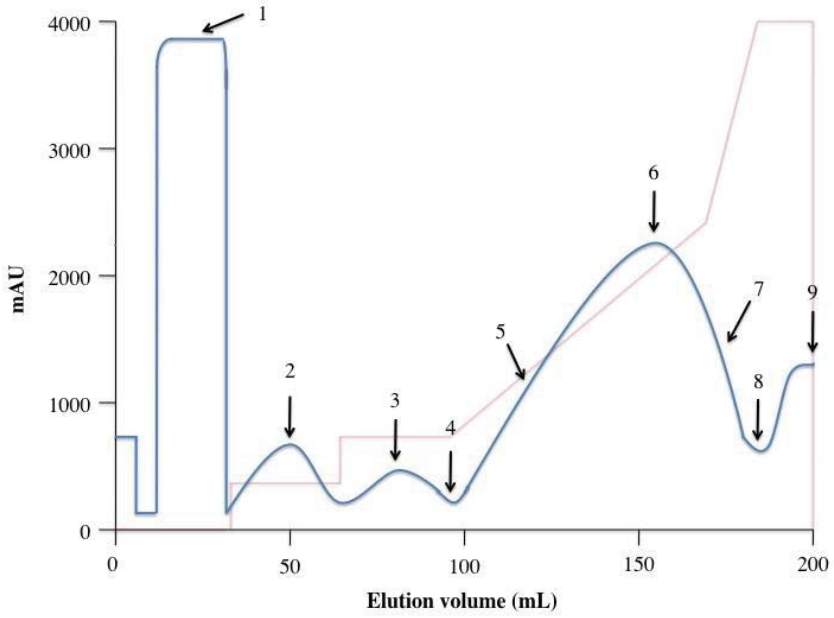

Fig. 5. 1. IMAC N-*Dal*DHQS2 elution profile. The protein absorbance at 280 nm (blue) as imidazole gradient (red) increases. Numbers 1 – 9 represent their respective lanes in figure 5. 2.

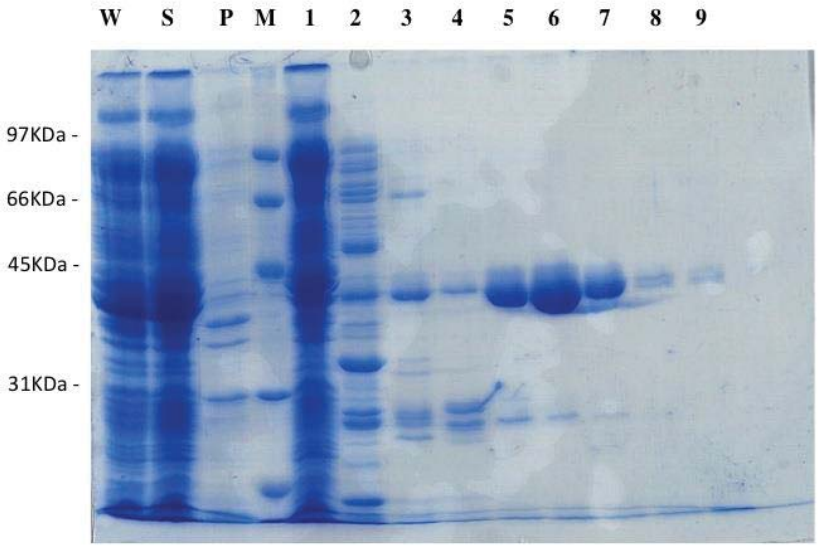

Fig. 5. 2. SDS-PAGE gel of collected fractions after IMAC of N-*Dal*DHQS2. Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Lanes 1 – 9 correspond to eluted fractions shown in Fig. 5. 1. Whole cell lysate (W), supernatant (S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lanes 1-3 show a decreasing elution of cellular protein with no affinity to the IMAC column before lanes 4-9 show the elution of the His8 tagged protein.

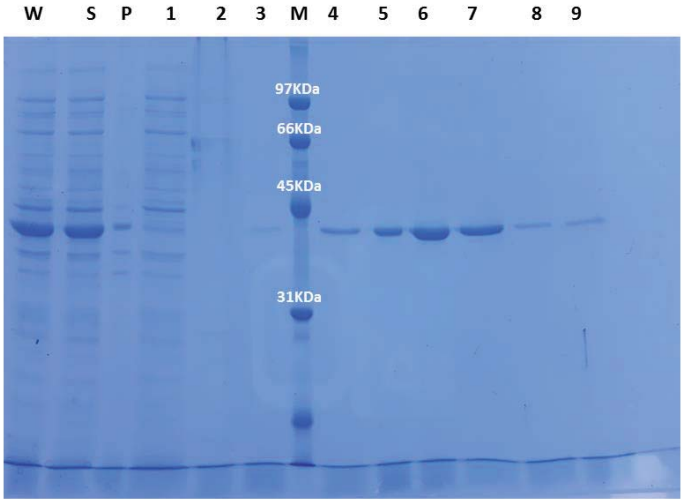

Fig. 5. 3. **SDS-PAGE gel of collected fractions after IMAC of C-*Dal*DHQS2.** Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Whole cell lysate (W), supernatant (S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lanes 1-3 show a decreasing elution of cellular protein with no affinity to the IMAC column before lanes 4-9 show the elution of the His8 tagged protein.

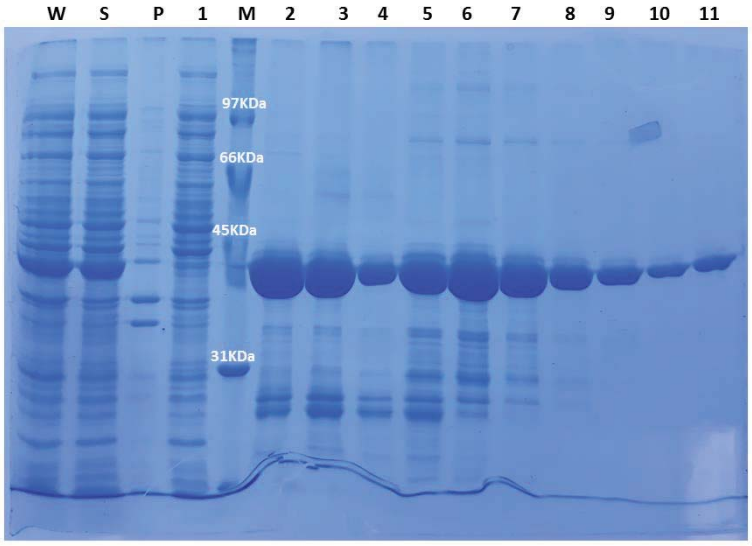

Fig. 5. 4. SDS-PAGE gel of collected fractions after IMAC of N-*Ave*DHQS2. Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Whole cell lysate (W), supernatant (S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lane

1 shows a decreasing elution of cellular protein with no affinity to the IMAC column before lanes 2-11 show the elution of the His8 tagged and contaminant protein.

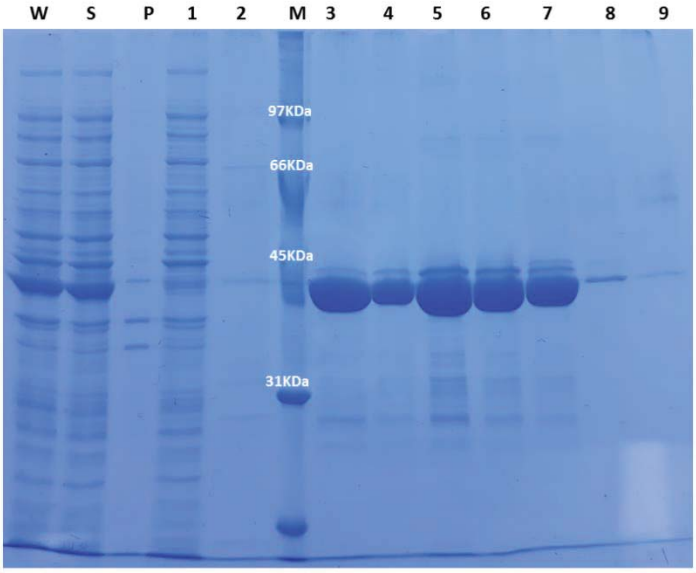

Fig. 5. 5. SDS-PAGE gel of collected fractions after IMAC of C-*Ave*DHQS2. Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Whole cell lysate (W), supernatant (S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lanes 1-2 show a decreasing elution of cellular protein with no affinity to the IMAC column before lanes 3-9 show the elution of the His8 tagged protein.

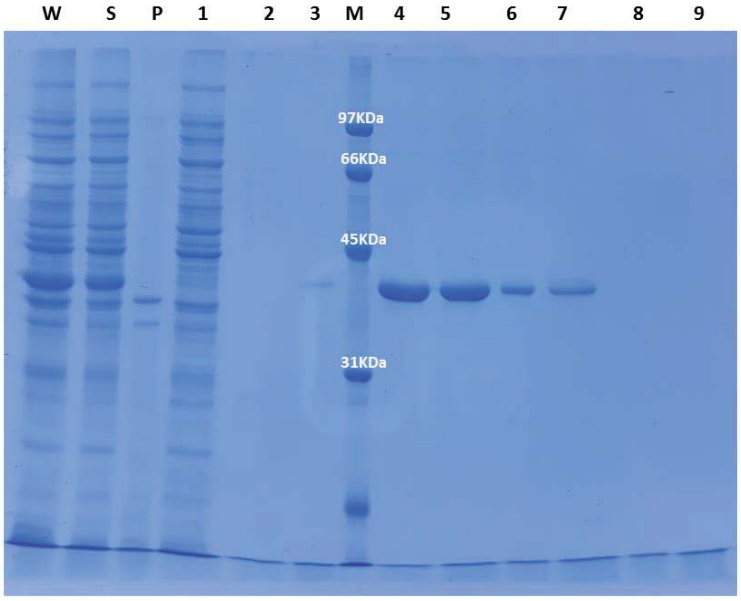

Fig. 5. 6. SDS-PAGE gel of collected fractions after IMAC of C-*Mhu*DHQS2. Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Whole cell lysate (W), supernatant

(S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lanes 1-3 show a decreasing elution of cellular protein with no affinity to the IMAC column before lanes 4-9 show the elution of the His8 tagged protein.

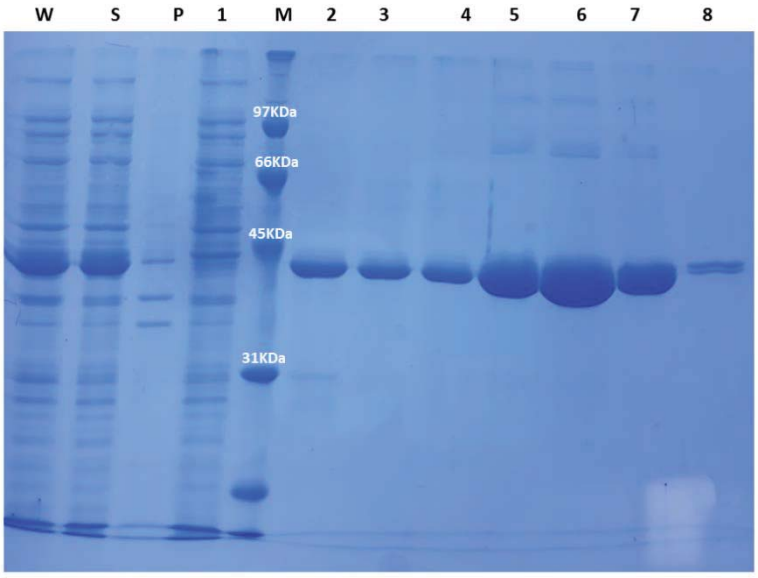

Fig. 5. 7. SDS-PAGE gel of collected fractions after IMAC of N-*Mma*DHQS2. Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Whole cell lysate (W), supernatant (S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lane 1 shows a decreasing elution of cellular protein with no affinity to the IMAC column before lanes 2-8 show the elution of the His8 tagged protein.

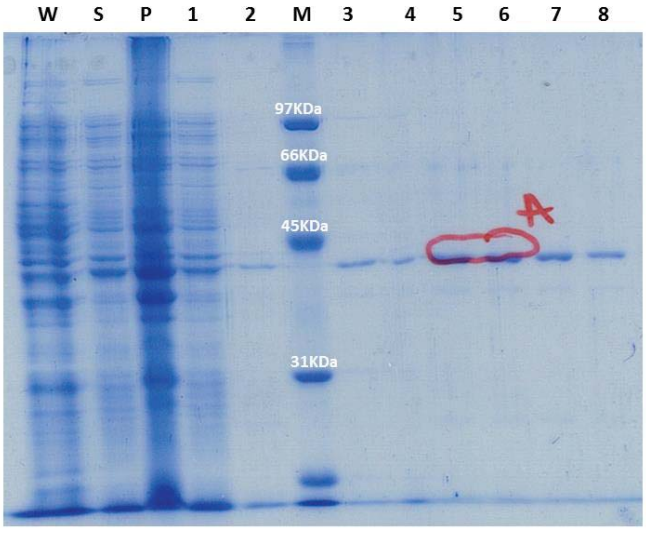

Fig. 5. 8. SDS-PAGE gel of collected fractions after IMAC of C-*Mma*DHQS2. Whole cell lysate (W), supernatant (S), cell pellet (P) and molecular marker (M). Whole cell lysate (W), supernatant (S) and pellet (P) fractions show that most of the induced protein (42 KDa) is in the supernatant. Lanes 1-2 show a decreasing elution of cellular protein with no affinity to the IMAC column before

lanes 3-8 show the elution of the His8 tagged protein. 42kDa protein bands in lanes 5 and 6, labelled A, were subjected to trypsin digest and mass spectrometry for identification.

5.2 Identification of DHQS2

After IMAC and SDS-PAGE analysis of C-*Mma*DHQS2 the protein band eluted by the imidazole gradient in lanes 5 and 6 (Fig. 5.8) was cut from the gel and subjected to a trypsin digest before peptides were extracted and analysed by tandem mass spectrometry (MS-MS). The band was identified as the C-terminal His_8-tagged DHQS2 from *M. maripaludis* with high confidence based on a peptide coverage 37 % and the presence of 10 unique peptides.

```
  1 MKFGWIKTTG NDLEERMESV KDALESSIPG IIAEKEEISS VRELGNIKIV
 51 SDNLDADVVL INKGEDLEIL KSAKLSGKET GVYVVINTKE DEVYATDVSK
101 LDFVDYVVLE GSDWTIIPLE NIIADLFSEE IKIVSVVTNV KDAEAAYEIL
151 EKGVDGVVLI PKDINEVKDF SKLIERMNSE SVKLDYATVT KIEPVGSGDR
201 VCIDTCSMME MGEGMLIGSY SRGMFLVHSE TVENPYVATR PFRVNAGPVH
251 AYILCPENKT KYLSDLKAGD KVLVVNKNGE TREAIIGRVK IEKRPLFLVE
301 AEYNGENLRT ILQNAETIRL VGEDGKPVSV VDLKVGTKVL IKPDENARHF
351 GMAIKETIIE K
```

Fig. 5. 9. Tandem MS-MS Sequenced Peptides of C-*Mma*DHQS2. Peptides unique to DHQS2, shown in red.

5.3 Size Exclusion Chromatography

IMAC fractions underneath the protein elution curve (Fig. 5.1) corresponding to lanes 5-7 (Fig. 5.2), 5-7 (Fig. 5.3), 8-11, (Fig. 5.4), 5-7 (Fig. 5.5), 4-7 (Fig. 5.6), 5-7 (Fig 5.7) and 5-7 (Fig. 5.8) were pooled and concentrated to a volume of 1-2 mL and then subjected to a size exclusion chromatography (SEC) using a Superdex S200 HR 10/300 column at a flow rate of 0.5 mL/min (Fig. 5.10). Initially, two species of C-*Mma*DHQS2 could not be completely separated, as shown by the red elution profile in (Fig. 5.10) (elution volumes were 13.17 mL and 14.85 mL, respectively). As shown in Fig. 5.11, these retention times correlate to an overall molecular weight of ~160 KDa and ~80 KDa, respectively, and equate to tetrameric and dimeric species on the basis of the average DHQS2 monomer mass being 40.0 KDa.

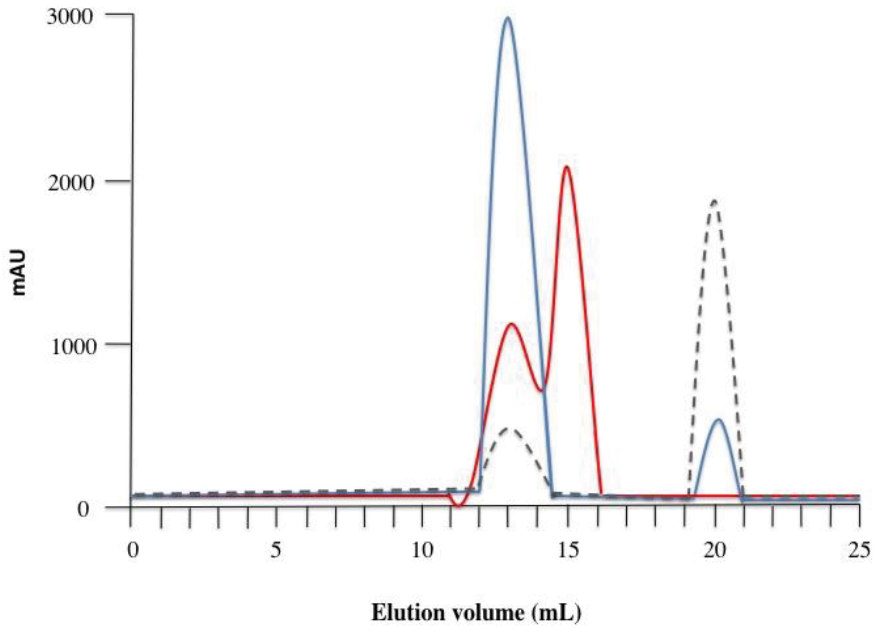

Fig. 5. 10. SEC Elution profile. Red line absorbance at 280 nm without NAD+. Blue line, absorbance at 280 nm, dashed line absorbance at 260 nm with the addition of NAD+ to 0.5 mM. Buffer, 10 mM K2HPO4, 50 mM NaCl, 2 mM DTT, 0.5 mM NAD^+, orthophosphate acid pH 7.2; flow rate 0.5 mL/min.

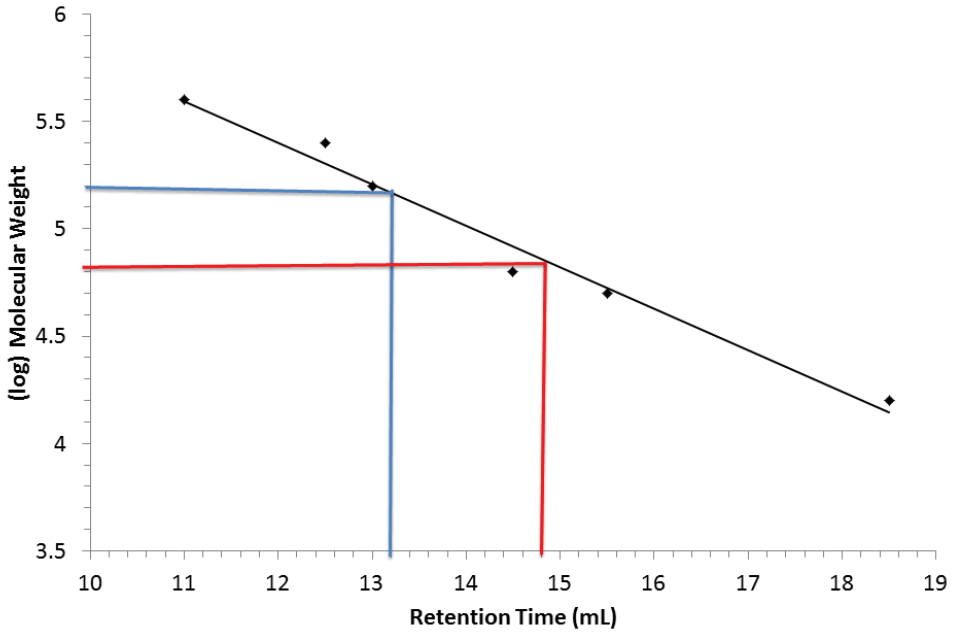

Fig. 5. 11. Superdex S200 HR 10/300 column calibration curve run in 10mM K_2HPO_4, 50mM NaCl, 2mM DTT, 0.5mM NAD^+, pH 7.2) at 0.5 mL/min. A protein species with a retention time of 14.85 mL (shown in red), relating to the major peak of the SEC elution profile without NAD^+ (shown in red, Fig. 5.10), has a molecular weight of ~80 KDa. A protein species with a retention time of 13.17 mL (shown in blue), relating the peak containing NAD+ (shown in blue, Fig. 5.10) and the minor peak of the elution profile without NAD+ (shown in red, Fig. 5.10) has a molecular weight of

~160KDa. Superdex S200 HR 10/300 column calibrationused the following markers: Ferritin (440 kDa), Catalase (250 kDa), Aldolase (158 kDa) and Ovalbumin (44 kDa).

Peak fractions run without NAD^+, were analysed using non-denaturing polyacrylamide gels (Fig. 5.12A). Lane 1 represents the tetrameric species while lane 2 represents a dimeric species. The addition of 0.5 mM NAD^+ to all buffers used in the purification process produced single homogeneous tetrameric peaks (Fig. 5.12B). NAD^+ is a cofactor required by DHQS2 to catalyze the oxidative deamination of ADH. Enzyme substrates, products and cofactors bind to specific sites on the enzyme. Cofactor binding has been shown to affect enzyme conformation (81), so it is not unreasonable to assume that when NAD^+ binds DHQS2, it changes the conformation of individual monomers to stabilise the tetrameric state as shown by SEC (Fig. 5.10) and native polyacrylamide gel electrophoresis (Fig. 5.12).
NAD^+ absorbs at a wavelength of 260 nm, shown by a dashed line in Fig. 5.10. The small 280 nm peak that coincides with the strong 260 nm peak at an elution volume of 20 mL is due to the absorbance of NAD^+. A Bradford assay and native polyacrylamide gel electrophoresis were used to further test for the presence of protein in this peak. A lack of protein as indicated by the Bradford assay indicated this peak represented unbound NAD^+.

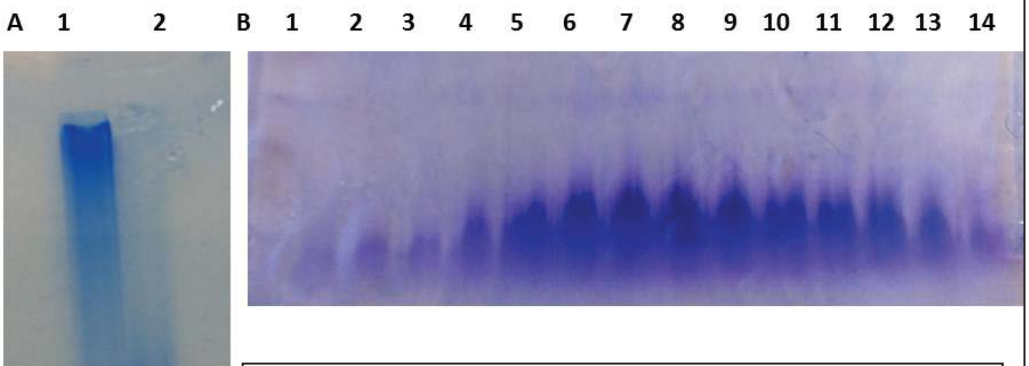

Fig. 5. 12. Native gel of C-*Mma*DHQS2 without NAD^+ (A) and with NAD^+ (B). (A) Lane 1 represents tetrameric C-*Mma*DHQS2 collected from under the minor peak, and lane 2 represents a dimeric C-*Mma*DHQS2 collected from under the major peak, shown in red, Fig. 5.10. (B) Lanes 1-14 represent homogeneous C-*Mma*DHQS2 in the tetrameric state, collected from fractions under the major blue shown in Fig. 5. 10.

5.4 Determination of Protein Concentration

Protein concentrations were measured using the Bradford assay with BSA as a standard (Fig. 5.13). The yield of protein at each stage of protein purification for the seven constructs is shown in Table 5. 1.

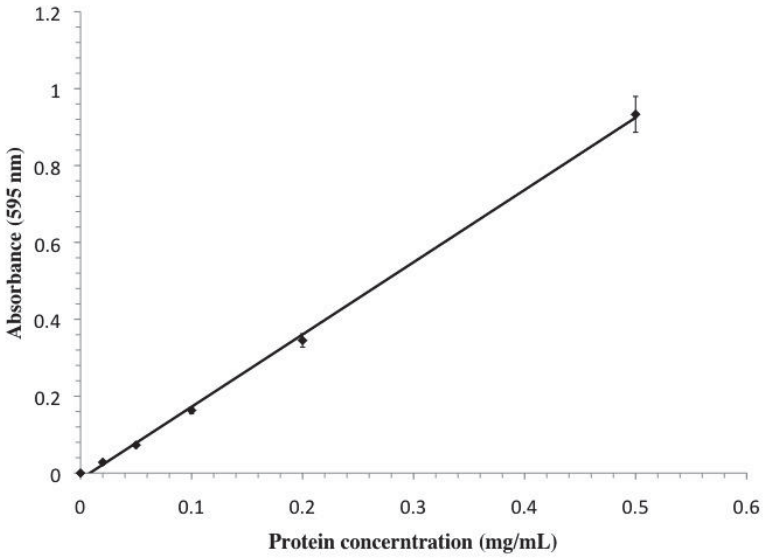

Fig. 5. 13. Bradford BSA standard curve. Line of best fit of triplicates that are within 5%. R^2 = 0.9911.

Table 5.1 DHQS2 concentrations at various stages of purification.

DHQS2 Protein purification prep	Soluble fraction after cell lysis			Pooled fractions after IMAC			Pooled fractions after SEC		
	Concentration (mg/mL)	Volume (mL)	Total (mg)	Concentration (mg/mL)	Volume (mL)	Total (mg)	Concentration (mg/mL)	Volume (mL)	Total (mg)
C-*Mma*DHQS2	8.5	35	298.9	2.9	17	49.8	1.38	4.5	6.2
N-*Mma*DHQS2	6.7	38	256.9	2.1	20	42.5	1.22	4.5	5.5
C-*Dal*DHQS2	7.3	35	250.4	0.5	28	13.4	0.13	3.9	0.5
N-*Dal*DHQS2	6.2	37	231.2	1.2	25	28.8	0.18	4.5	0.8
C-*Ave*DHQS2	6.7	36	242.2	1.4	24	34.5	0.20	5.1	1.0
N-*Ave*DHQS2	7.8	35	275.3	1.3	30	39.3	0.31	3.9	1.2
C-*Mhu*DHQS2	7.0	38	266.6	1.8	25	44.4	1.07	4.2	4.5

Chapter Six: Structural Investigations of DHQS2

6.1 Crystallisation Trials

After SEC, pooled fractions were concentrated to 10 mg/mL by ultrafiltration. Crystallisation trials were conducted using JCSG-plus™ (Molecular Dimensions, Suffolk, United Kingdom), MIDAS™ (Molecular Dimensions) and a Morpheus® (Molecular Dimensions) crystallisation screens. C-*Mma*DHQS2 was the only protein out of all the protein constructs *i.e.* N-*Ave*DHQS2, C-*Ave*DHQS2, N-*Dal*DHQS2, C-*Da*lDHQS2, N-*Mma*DHQS2 and C-*Mhu*DHQS2, that produced what could have been protein crystals (Fig. 6.1 and 6.2). Although these crystals did not diffract, neither did they produce a diffraction pattern characteristic of salt when exposed to x-rays. Interestingly, *Mma*DHQS2 did not meet the selection criteria proposed to increase the chances of crystallization; it had an XtalPred EP classification of 3 compared to the better 1 and 2 EP classifications of the other targets, yet it was the only protein that appeared to crystallise.

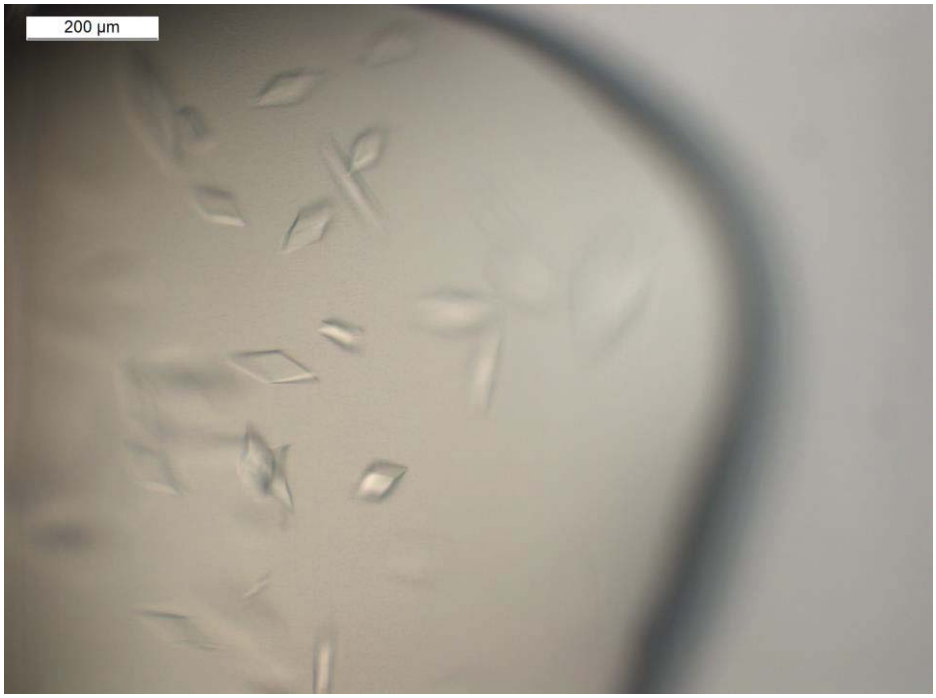

Fig. 6. 1. C-*Mma*DHQS2 crystals in JCSG$^+$ screen condition 1.0 M ammonium sulphate, 0.1 M BIS-Tris, 1% (w/v) PEG 3350 at pH 5.5.

Small protein crystals initially grew in 1.0 M ammonium sulphate, 0.1 M BIS-Tris, 1% (w/v) PEG 3350 at pH 5.5 (Fig. 6.1). In order to increase their size different conditions were trialled by varying pH, precipitant concentration, buffer concentration, and salt additive concentrations around their original concentrations. The following conditions were trialled in various combinations: 0.5, 1.0 and 1.5 M ammonium sulphate, 0.1 and 0.2 M BIS-Tris, 1%, 2% and 5% (w/v) PEG 3350 at pH 5.0, 5.25, 5.5, 5.75 and 6.0. However, despite containing the same components no crystals grew. When the original conditions were used, but made up with "in-house" reagents, no crystals grew. Repeating the experiment with the original JSCG+ solution produced crystals once more. Large crystals were grown when protein was further diluted with SEC buffer at a (v/v) ratio of 3:1 protein to SEC buffer. These large crystals were however very fragile. They were sent to the Australian synchrotron for data collection, but were found to have dissolved on arrival. The smaller crystals were sent to the synchrotron frozen, but did not diffract. Unfortunately, the premade condition from the JCSG+ screen ran out and when a new screen was purchased the same condition did not promote the formation of crystals.

Fig. 6.2. C-*Mma*DHQS2 crystals in JCSG$^+$ screen condition 0.1 M CAPS, 40 % (v/v) MPD at pH 10.5.

Small needles grew in 0.1 M CAPS, 40 % (v/v) MPD at pH 10.5 (Fig 6.2). Attempts were made to optimise these crystallisation conditions by varying the buffer concentration, pH and precipitant as follows using lab-sourced individual components: 0.05, 0.10 and 0.15 M CAPS, 30 %, 35 %, 40 % and 45% (v/v) MPD, pH 9.0, 9.5, 10.0, 10.5 and 11.0. Frustratingly, no crystals grew in these new conditions and attempts to recrystallize the protein using the same original screen condition and the same protein solution also failed to produce crystals.

Overall, the lack of diffracting crystals was very disappointing, and possible reasons are many including a limited number of screens and variety of precipitants. Precipitants in the crystallization drop alter the protein-solvent or protein-protein contacts so that the protein molecules precipitate out of solution as ordered crystals and not as disordered aggregates (39). The solubility of most proteins varies as a function of temperature where solubility usually increases with increasing temperature in low ionic strength conditions such as in the presence of the precipitant polyethylene glycol (PEG). Crystallization has been reported to occur for proteins over a whole range of temperatures, from 0 °C to 60 °C, (40) although crystallisation experiments are usually conducted at either 4 °C or at room temperature. Therefore, crystal screens and optimised conditions, that did not result in protein crystals at 25 ºC, were also grown at 4 ºC and 37 ºC in an attempt yield protein crystals, to no success. Another approach to obtain protein crystals from the initial screens was to change the protein concentration in the conditions. Initially a protein to mother liquor ratio of 1:1 was used before ratios 1:2 and 2:1, unfortunately neither change in protein concentration was successful.

The addition of cofactor, NAD^+, homogenised DHQS2, which is a crucial factor for protein to crystallise. In the same way, the product of the DHQS2 catalysed reaction, DHQ, or substrate ADH, could also potentially homogenise DHQS2 in the same or in a different conformation. These additional co-crystallisations could yield diffraction quality protein crystals. One of the main reasons for lack of success with the N-*Mma*DHQS2, N- and C-*Ave*DHQS2, N- and C-*Dal*DHQS2 and C-*Mhu*DHQS2 proteins was most likely due to a lack of protein homogeneity due to non-native folding of the protein. Solubility does not

necessarily equate to proper folding, and there is reasonable evidence that were not folded despite exhibiting the presence of secondary structure.

6.2 Circular Dichroism (CD)

CD refers to the differential absorption of the left and right circularly polarized components of plane-polarized radiation. Proteins possess a number of chromophores, which can give rise to CD signals (41). In the far UV region (180-260 nm), peptide bond absorption gives rise to CD spectra and can be analysed to give the proportions of secondary structural elements such as α-helices and β-sheets. The near UV region (260-320 nm) reflects the environments of the aromatic amino acid side chains and thus gives very basic information about the tertiary structure of the protein *i.e.* whether the protein is likely to be folded or in a molten globule conformation. The purpose of carrying out CD was to determine if the protein was properly folded or not. If not, this would explain the lack of protein crystal formation.

A buffer solution needs to have low background absorbance in these ranges in order to produce an analytical result. All DHQS2 samples were currently stored in excess NAD^+ for structural homogeny and NAD^+ absorbs highly at 260 nm, therefore the NAD^+ was removed from the protein by buffer exchange into 20 mM K_2HPO_4 and 1 mM TCEP using a Superdex S200 HR 10/300 column. The elution profile (Fig. 6.3), shows the removal of excess NAD^+ without apparently changing the conformation of the protein. This implies that NAD^+ binds reasonably tightly to the protein maintaining it in a tetrameric conformation.

The Dichroweb web server enables on-line analyses of CD spectroscopic data, providing estimated secondary structure content and graphical analyses comparing calculated structures and experimental data. The information contained in CD spectra can be treated as a sum of the characteristic individual spectra arising from each type of secondary structure present in a protein sample. Dichroweb analysis programs, including SELCON 3, CONTIN and CDSSTR, utilize a number of reference database comprised of spectra of proteins whose crystal structures, and therefore their secondary structures, are known (82). The

performances of CONTIN (83), (84) and SELCON3 (85), were examined by N. Sreerama and R. Woody (2000) (86) using the five reference protein sets of secondary structure. The performance indices, RMS differences and correlation coefficients, between the CD-predicted and X-ray values for individual secondary structures were mixed, with each method giving slightly better results for one of the secondary structures. The larger reference set performed better than the smaller reference sets, even with CD data over a smaller wavelength range, and this can be attributed to a larger representation of structural and spectral variability of β-sheets and turns. Therefore reference set 7 was chosen as the database for analysis of the CD data from Fig. 6.4, 6.5, 6.6, 6.7. 6.8, and 6.9, which were input into CONTIN (to no result) and SELCON 3 (Table 6.1), because it contains the highest number of reference proteins. CD data was also input into a CD analysis and plotting tool (CAPITO) (85) server for additional analysis (Table 6.2).

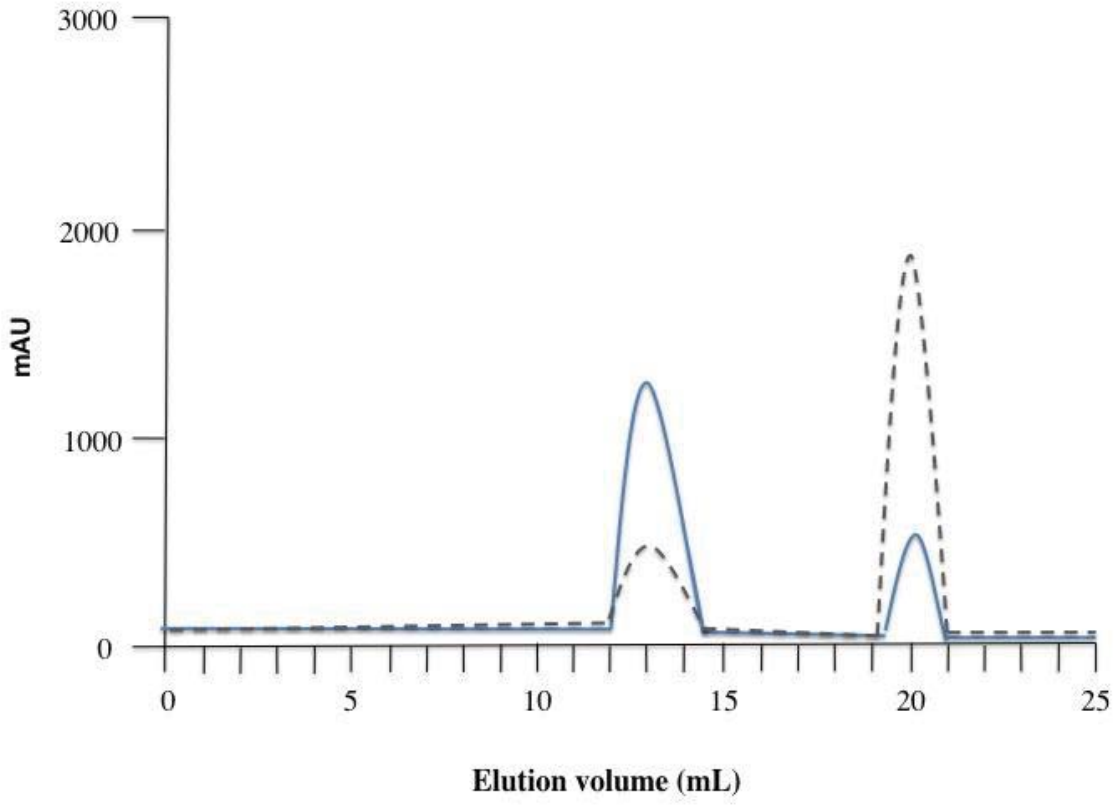

Fig. 6. 3. Elution profile of C-*Mma*DHQS2. Absorbance at 280 nm shown in blue and the absorbance at 260 nm shown by a dashed line.

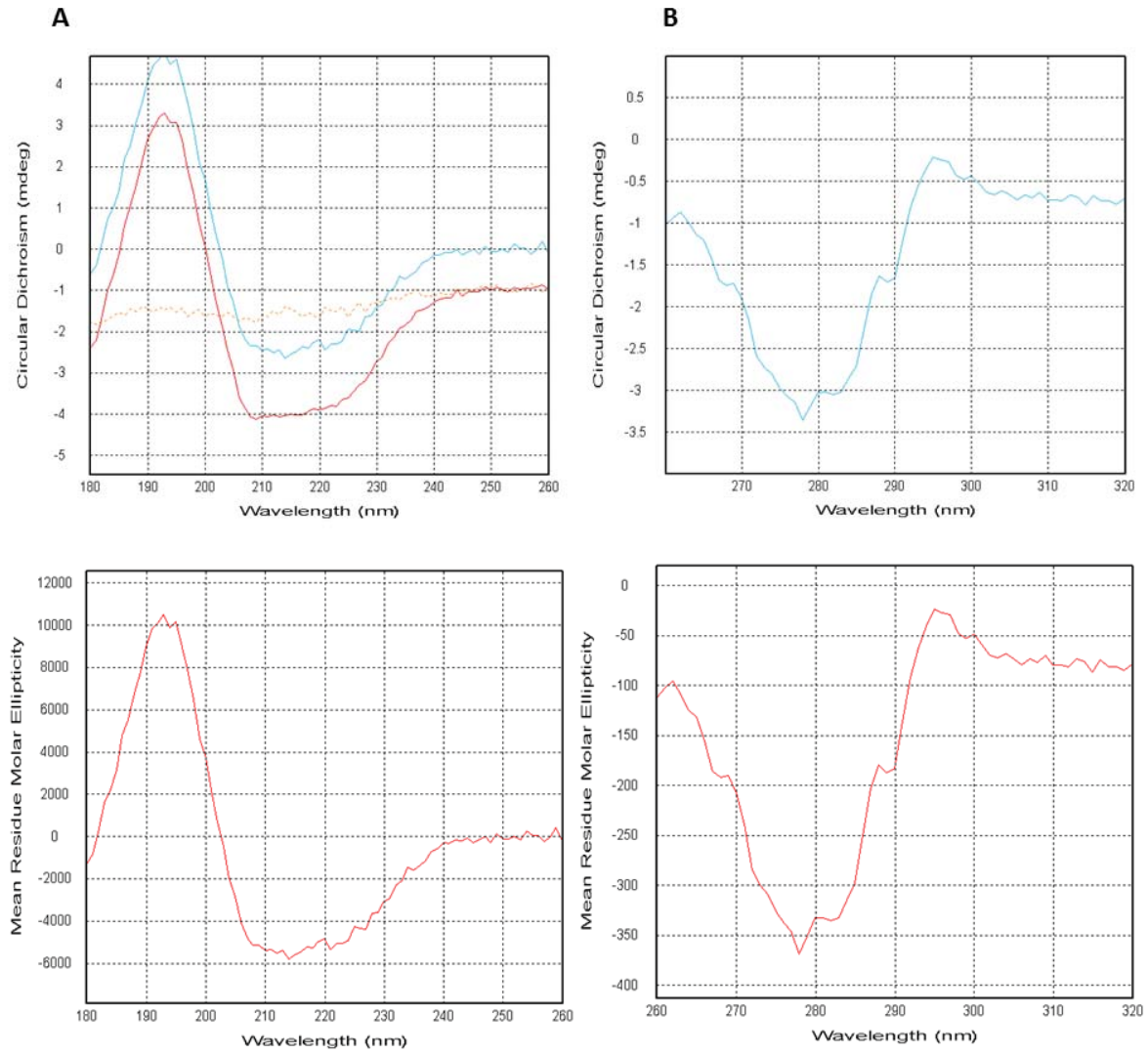

Fig. 6. 4. C-*Mma*DHQS2 far (A) and near (B) UV Circular Dichroism. The CD buffer baseline (orange) was accounted for by subtracting it from the C-*Mma*DHQS2 far UV spectra (red) to give the adjusted spectra (blue). The adjusted spectra is observed in Figs. 6.4 to 6.9.

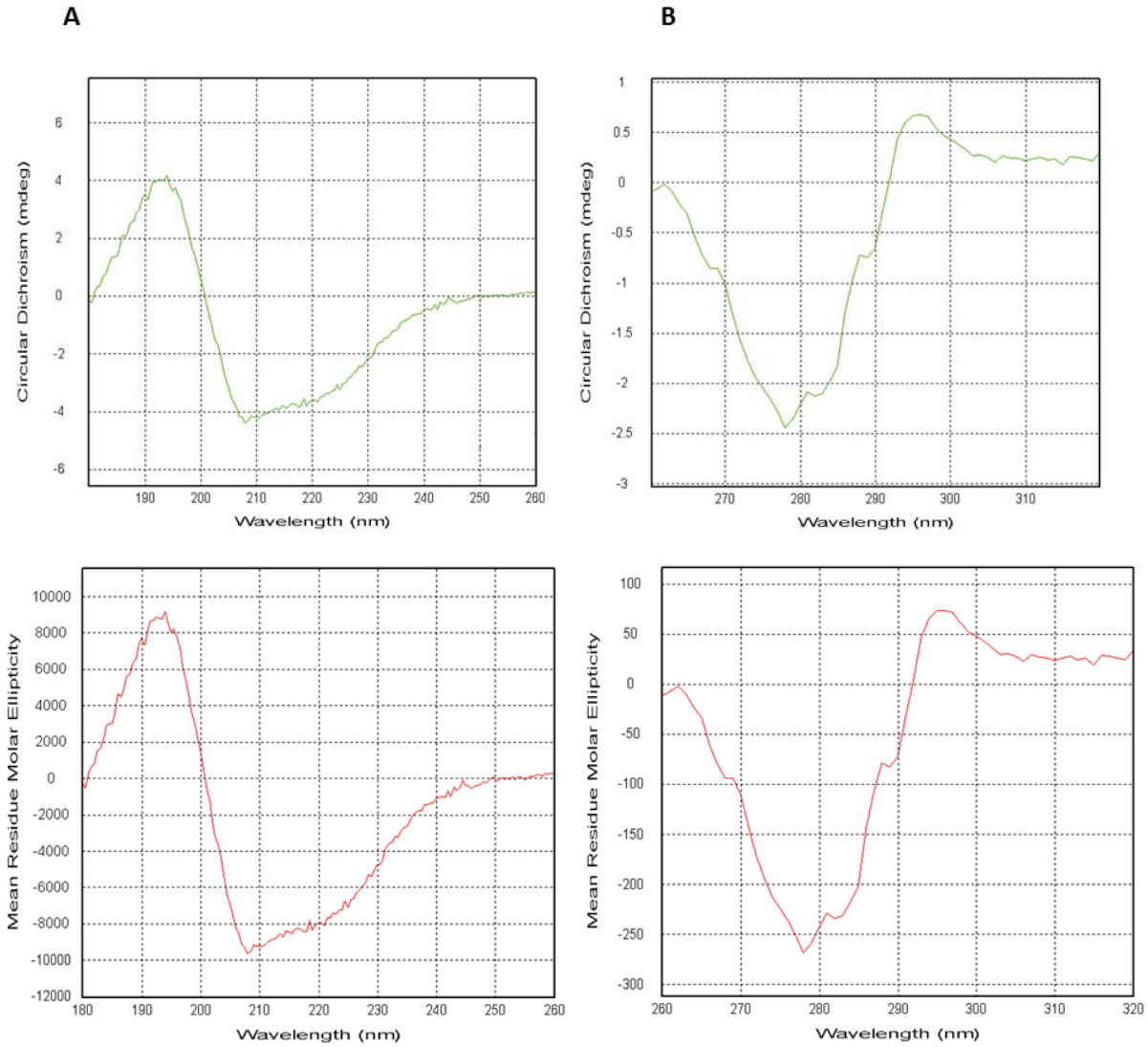

Fig. 6. 5. N-*Mma*DHQS2 far (A) and near (B) UV Circular Dichroism.

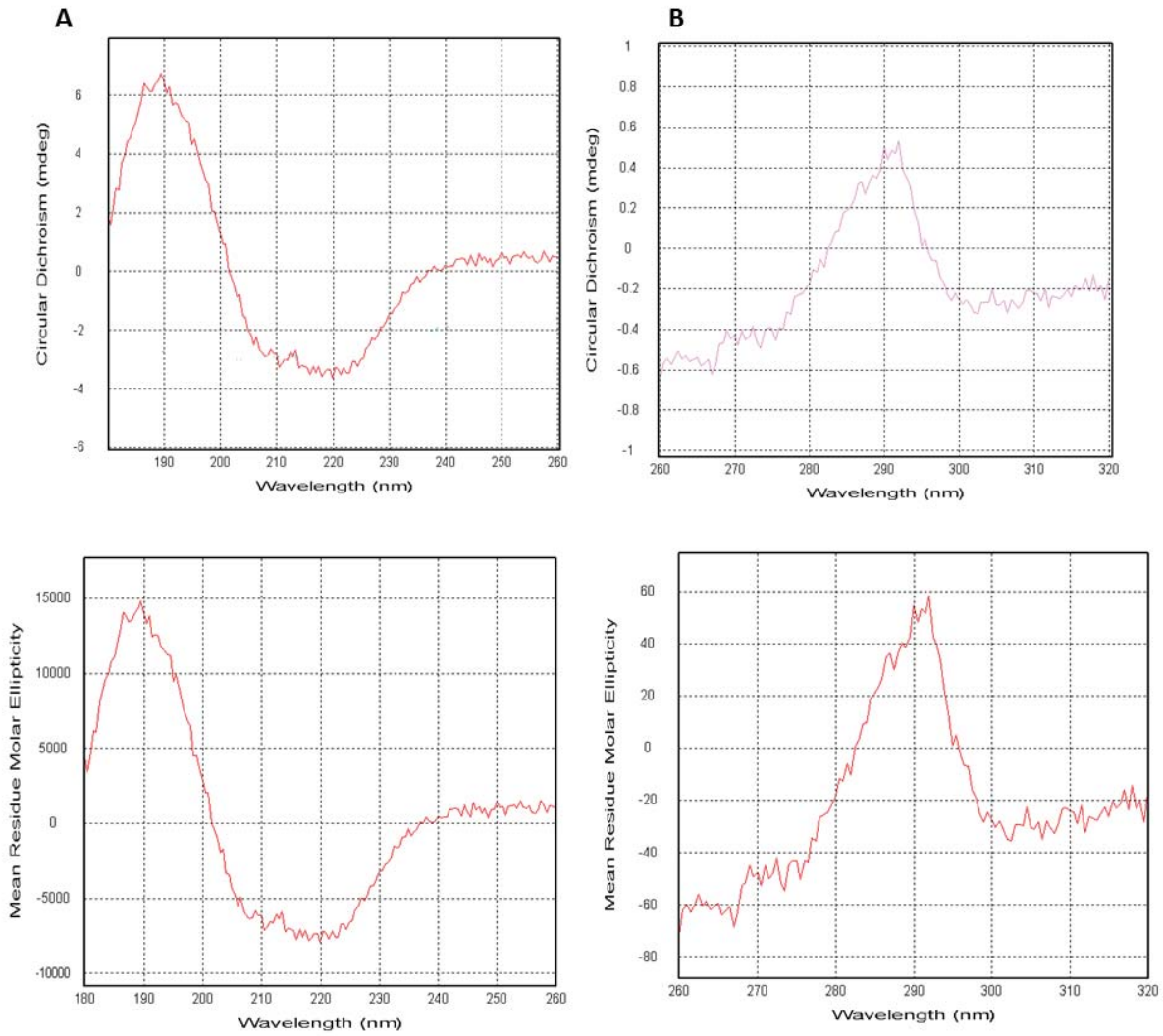

Fig. 6. 6. C-*Ave*DHQS2 far (A) and near (B) UV Circular Dichroism.

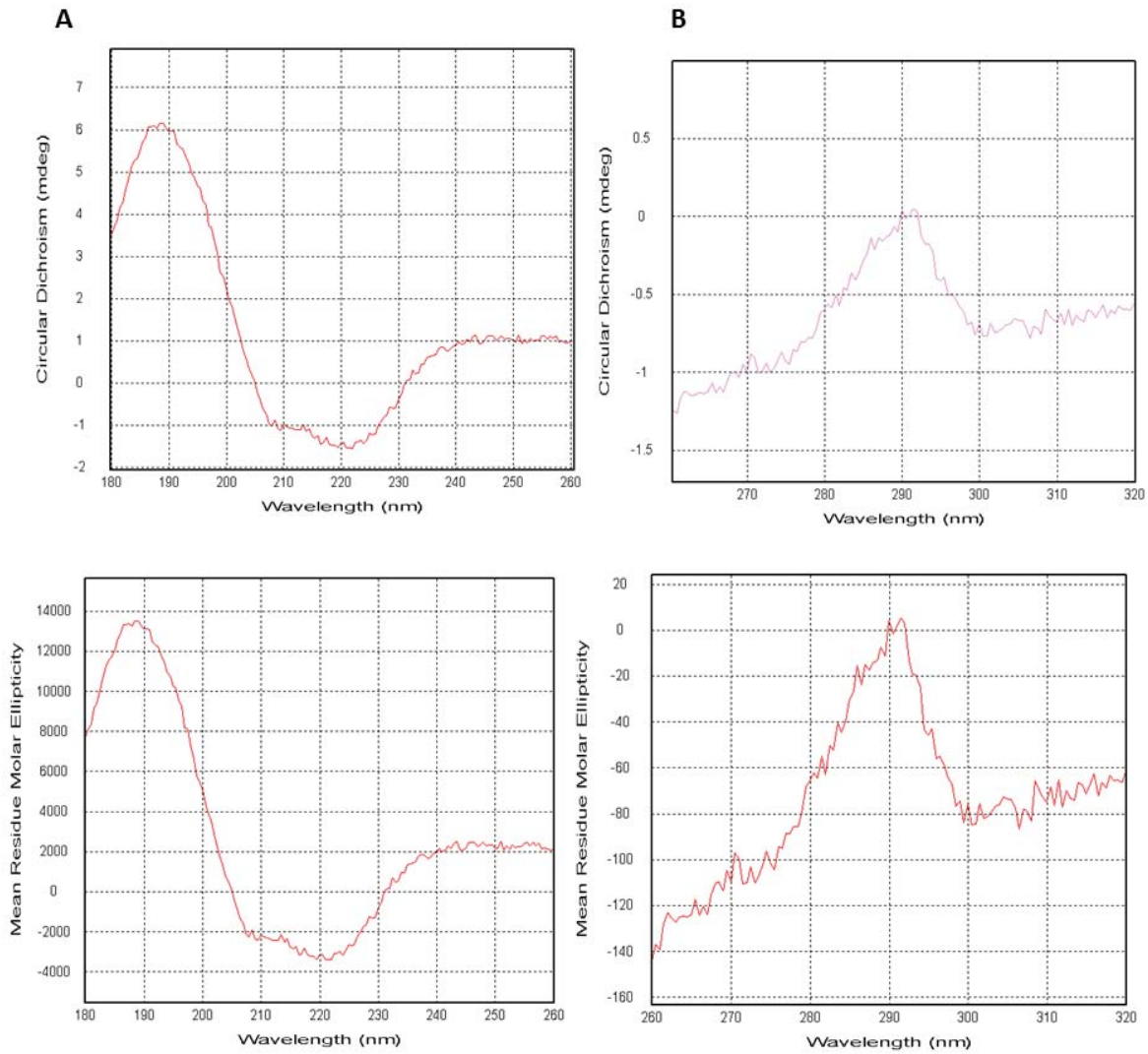

Fig. 6. 7. N-*Ave*DHQS2 far (A) and near (B) UV Circular Dichroism.

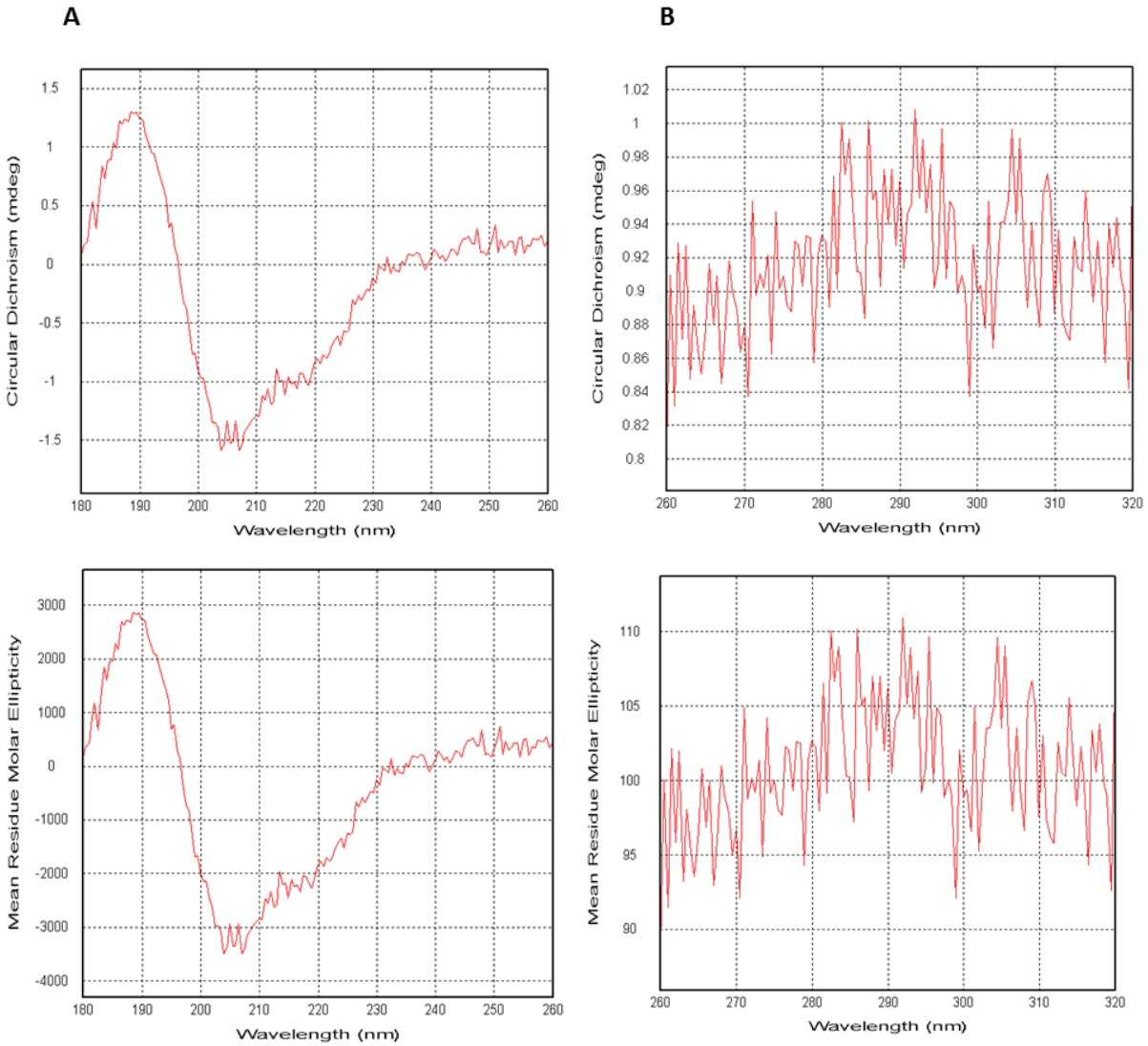

Fig. 6. 8. C-*Dal*DHQS2 far (A) and near (B) UV Circular Dichroism.

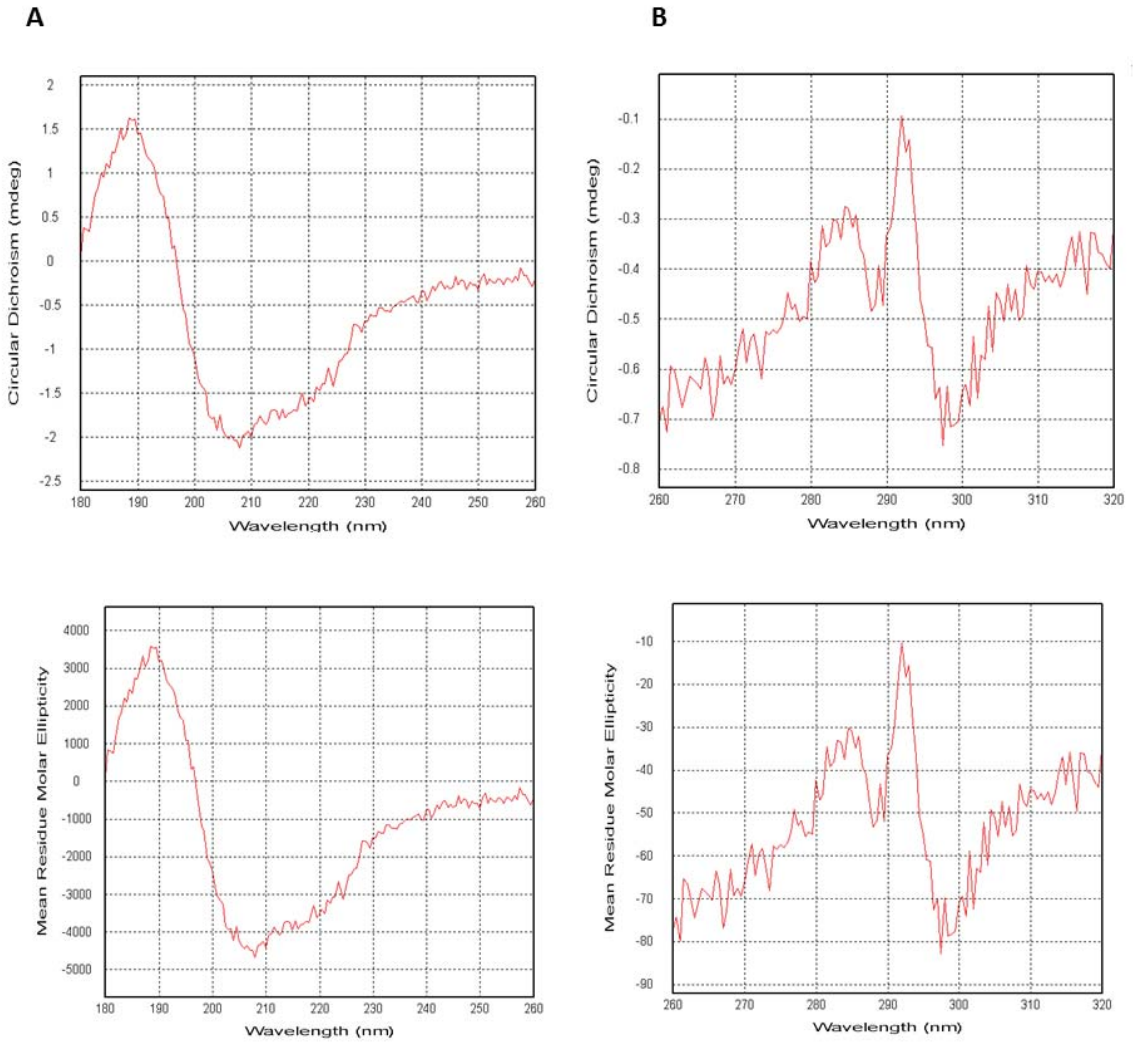

Fig. 6. 9. N-*Dal*DHQS2 far (A) and near (B) UV Circular Dichroism.

Table 6. 1. SELCON 3 Reference Set 7 Analysis of Far UV Circular Dichroism Spectra. Helix 1, Helix 2, Strand 1 and Strand 2 represent regular helices, distorted helices, regular β-sheets and distorted β-sheets, respectively.

	Helix 1	Helix2	Strand 1	Strand 2	Turns	Unordered	Total
C-*Mma*DHQS2	0.583	0.222	0.000	0.000	0.050	0.142	0.997
N-*Mma*DHQS2	0.538	0.205	0.000	0.000	0.050	0.206	0.999
C-*Ave*DHQS2	0.089	0.077	0.183	0.113	0.218	0.315	0.995
N-*Ave*DHQS2	0.013	0.043	0.261	0.137	0.192	0.346	0.996
C-*Dal*DHQS2	0.016	0.052	0.218	0.113	0.194	0.407	1.000
N-*Dal*DHQS2	0.031	0.063	0.186	0.106	0.189	0.425	1.000
C-*Mhu*DHQS2	-0.038	-0.043	0.371	0.195	0.297	0.140	0.921

Table 6.2. CAPITO analysis of Far UV Circular Dichroism Spectra.

	Helix	Strand	Unordered
C-*Mma*DHQS2	0.39	0.20	0.41
N-*Mma*DHQS2	0.31	0.12	0.57
C-*Ave*DHQS2	0.14	0.39	0.51
N-*Ave*DHQS2	0.06	0.46	0.47
C-*Dal*DHQS2	0.01	0.17	0.55
N-*Dal*DHQS2	0.08	0.41	0.57
C-*Mhu*DHQS2	0.00	0.67	0.47

The CD spectrums for the respective C- and N-terminally tagged DHQS2 proteins are similar, as expected because the only difference is the His_8-tag positioning. For all DHQS2 proteins analysed, visual inspection of the far UV spectra indicate the probability of them having mixed α-helical and β-sheet secondary structure (Figs. 6.4A 6.9A). However analyses of the results on the Dichroweb server give a different result. Both Dichroweb (Table 6.1) and CAPITO (Table 6.2) analysis of the far UV spectra show that with the exception of *Mma*DHQS2, the protein generally contains a greater proportion of β-strand. In contrast, DHQS type 1 from *Aspergillus nidulans* (87) (Fig. 6.11), has an N-terminal Rossmann-type domain and C-terminal α-helical domain architecture, suggesting that DHQS2 and DHQS1 adopt different folds. The differences observed between proteins with ~ 50% sequence homology suggests that only one of these folds is likely to be the native fold. As all protein concentrations were nominally the same, the intensities of the bands at ~193nm and may provide a clue. The intensities of the absorbances at 193 nM and 208 – 222 nm may provide a clue. Although the far UV spectra shape is consistent, the scale for *Mma*DHQS2

and *Ave*DHQS2 is six times that of *Dal*DHQS2. While this could be due to a lower protein concentration it could also be due to less defined secondary structure and more disorder. C-*Mma*DHQS2 was the only protein to crystallise, yet was predicted to be highly α-helical. I-TASSER predicted a DHQS2 structure made up of a, α-helical domain and a β-domain, although all the models had low confidence scores.

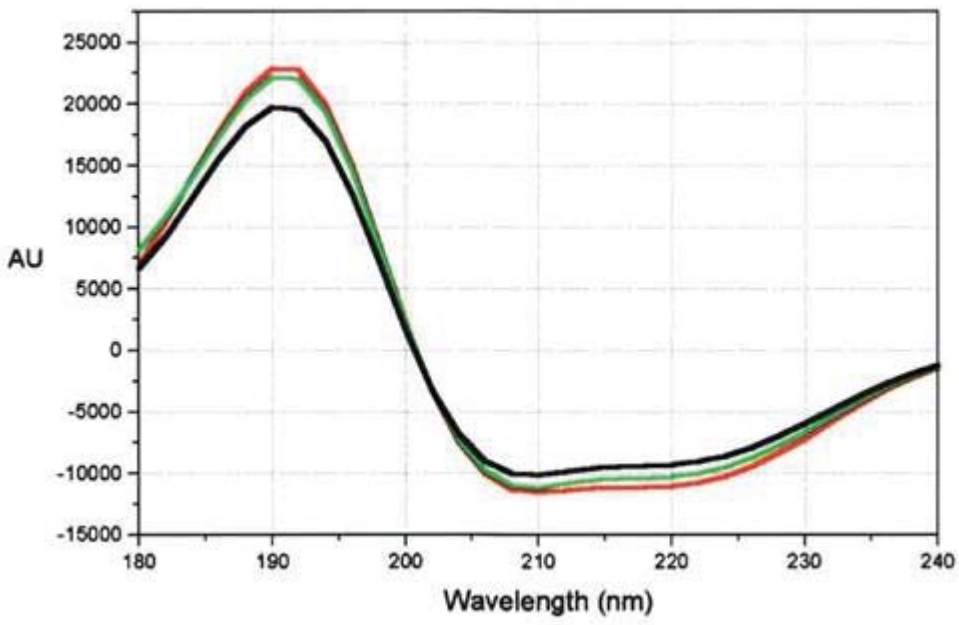

Fig. 6.10. Far-UV spectra of wild-type and mutant *Aspergillus nidulans* DHQS. Adapted with permission from the *Protein Science,* **2004**, 13, 2108–2119. Copyright © 2004 Cold Spring Harbor Laboratory Press. Wild-type shown in red, point mutations R130K and R130A shown in green and black respectively.

The spectra in the near UV region, 260 – 320 nm, arise from the aromatic amino acids. Each amino acid tends to have a characteristic wavelength profile. Tryptophan shows a peak close to 290 nm, tyrosine shows a peak between 275 and 282 nm, and phenylalanine shows weaker peaks but sharper bands with fine structure between 255 and 270 nm. The actual shape and magnitude of the near UV CD spectrum depends on the number of each type of aromatic amino acid, their mobility, their spatial disposition in the protein, and the nature of their environment (41). In the near UV spectra for *Ave*DHQS2 (Fig. 6.6B and 6.7B), there is a peak at 290 nm that is representative of tryptophan in a chiral environment, therefore it seems

as if the *Ave*DHQS2 has tertiary structure. In the *Mma*DHQS2 near UV spectra (Fig. 6.4B and 6.5B) there is a trough comprising the wavelengths corresponding aromatic amino acids. This near UV spectrum has been observed in other proteins such as the molybdate-sensing protein ModE from *E. coli* (40) (Fig. 6. 11), and as a result, *Mma*DHQS2 was deemed to have viable tertiary structure. The near UV spectra of *Dal*DHQS2 (Fig. 6.8B and 6.9B) has no real tertiary structure but does have secondary structure (Figs. 6.8A and 6.9A), suggesting that the protein is a molten globule with no stable tertiary structure. A possible reason for this lack of tertiary structure could be due to the *E. coli* machinery being unable to properly fold *Dal*DHQS2 into its quaternary structure. In future, it may be beneficial to use an expression system containing n inducible chaperone such a GroEL, trigger factor or Tig , to assist the *E. coli* machinery. Even though C-MhuDHQS2 was concentrated enough to see on an SDS-PAGE gel, its CD signal was too weak to produce a viable spectra.

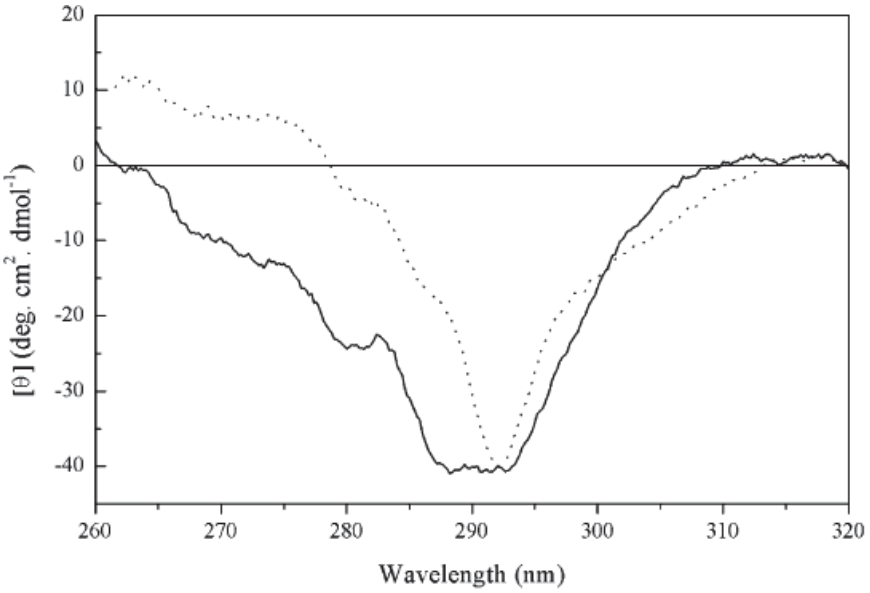

Fig. 6.11. Near UV CD trough-like spectra of molybdate-sensing protein ModE from *E. coli*. Adapted with permission from *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics,* **2005**, 1751, 119-139. Copyright © 2005 Elsevier.

Converting near UV CD data to the preferred mean residue molar elipticity units shifts the scale into the negatives. This is has been observed in a previous study (88) (Fig. 6.12), where glutamate dehydrogenase (Fig. 6.12A) and NAD^+ (Fig. 6.12B) CD spectra are relatively positive in comparison to the glutamate dehydrogenase-NAD^+ complex (Fig. 6.11C). It was

concluded that binding of NAD^+ significantly changes the observed CD spectra. The stoichiometry between DHQS2 and NAD^+ could have been determined using CD, and in hindsight, this should have been done. But the number of NAD^+ molecules bound to each DHQS2 was not known, and differences in the NAD+ to protein ratio between samples are most likely to be the reason for the differences in intensity observed.

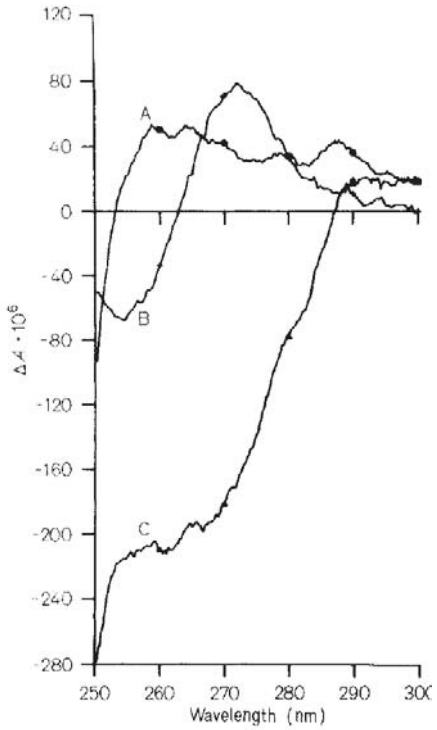

Fig. 6. 12. CD Spectra of (A) Glutamate dehydrogenase, (B) NAD+, and (C) Glutamate dehydrogenase-NAD+ complex. Taken from with permission from the *European journal of Biochemistry*, **1980**, 112, 521-531. Copyright © 1980 Blackwell Publishing ltd.

*Mma*DHQS2, *Ave*DHQS2 and N-*Dal*DHQS2 near UV spectra, all had a peak between 290 and 295 nm. This is most likely to be due to NAD^+ bound to the protein as previously described by Bayley and O'Neil (1980) (88) (Fig. 6.13).

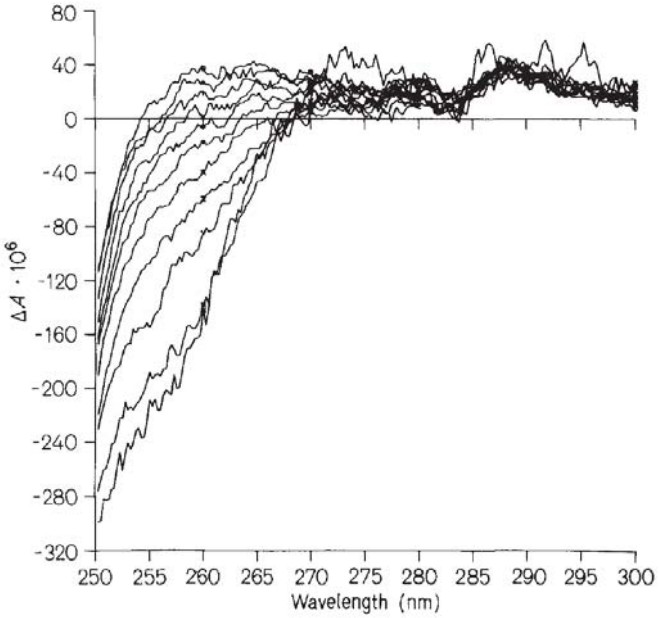

Fig. 6. 13. Circular Dichroism of Glutamate dehydrogenase in the presence of increasing NAD^+. Taken from with permission from the *European journal of Biochemistry*, **1980**, 112, 521-531.

Chapter Seven: Conclusions

The aim of this thesis was to firstly use and test, on a small scale, the viability of protein structure bioinformatic predictive servers. And secondly, to characterise the structure of DHQS2.

The difficulty of protein crystallisation drives an increased focus towards the prediction of protein structure from its amino acid sequence. Bioinformatic servers such as XtalPred, PONDR and I-TASSER predict protein properties using various algorithms based on current knowledge of folds in the PDB, amino acid charges, spatial conformation restraint laws of peptide bonds etc. The ultimate goal would be to predict structures of proteins at reasonable resolution with an acceptable level of confidence. Structure prediction is now a readily available tool used as an initial gage of possible protein structure. For this reason, five of the six selected DHQS2 targets, *S. azorense*, *D. alkenivorans, M. hungatei, M. stadtmanae* and *A. veneficus,* were selected from 80 potential targets based on bioinformatic criteria. It is peculiar that the only target selected to test this basis of selection because it scored below the selection criteria, *M. maripaludis*, was in the end the only protein to crystallise. Unfortunately however the crystals were not of diffraction quality. It would have been interesting to compare the predicted I-TASSER folds and CD data to a high-resolution structure. A larger data set of targets would be necessary to determine the viability of the selection criteria however that was outside the time constraints of a master's thesis.

The predictive bioinformatic servers used, Xtalpred and PONDR, presume that the protein sequence will have a monomeric conformation. It has been observed that DHQS2 adopts a multimeric structure, which as a result, changes the conditions on which these predictive servers base their calculations. This could explain why these bioinformatic servers predicted a high likelihood of crystallisation that did not transition into reality.

Far and near UV CD spectra analysis gave some insight into possible reasons why N- and C-*Dal*DHQS2, N- and C-*Ave*DHQS2, N-*Mma*DHQS2 and C-*Mhu*DHQS2 did not crystallise at all, and C-*Mma*DHQS2 did not produce diffraction quality crystals. Dichroweb and CAPITO

analysis of each spectrum predicted high proportions of random structure despite an attempt to account for disorder using PONDR. The analyses of the far UV spectra for those proteins that appeared to have some tertiary structure indicated one to have mainly α structure while the other contained mainly β structure. For proteins with 50% sequence homology this would be highly unlikely, indicating that either one or both had a non-native fold. DalDHQS2 clearly was not folded and because it exhibited secondary structure could have been in a molten globule state.

Protein homogeneity is crucial for the formation of an ordered state resulting in crystal formation. During the purification process it was observed that the addition of NAD^+, the cofactor required by DHQS2 to carry out the oxidative deamination of ADH, stabilised the protein into a single form believed to be tetrameric based on SEC.

7.1 Future Experiments

Possible future experiments include:

- Determination DHQS2 activity by coupled-enzyme assay. This would require the purification of another recombinant enzyme, ADHS, and the synthesis of precursor substrates DKFP and ASA. This would indicate whether or not DHQS2 is active by detecting the presence of product DHQ. If DHQ is produced then this would show that DHQS2 is properly folded, although CD experiments suggest otherwise.

- Co-crystallisation of DHQS2 with either substrate, ADH, or product, DHQ. Just as NAD^+ appears to stabilise DHQS2, a similar effect may be observed with ADH and DHQ, where substrate binding in the absence of cofactor may encourage DHQS2 to adopt a stable folded conformation.

- Small Angle X-ray Scattering (SAXS) analysis for further structural investigation of DHQS2. SAXS provides low resolution information on the shape, conformation and assembly state of proteins. This could be used to detect the conformational changes

of DHQS2 binding either cofactor, product or substrate, NAD^+, DHQ and ADH, respectively .It may also determine whether DHQS2 is intrinsically disordered.

References

1. Bentley R (1990) The Shikimate Pathway - A Metabolic Tree With Many Branches. *Critical Reviews in Biochemistry and Molecular Biology* 25(5):307-384.
2. Moral CH, del Castillo EF, Fierro **L**P**,** Cortés **V**A, Castillo **A**J, Soriano **C**A, Salazar SM, Peralta RB, Carrasco NG (1998) Molecular characterization of the *Aeromonas hydrophila* aroA gene and potential use of an auxotrophic aroA mutant as a live attenuated vaccine. *Infection and Immunity* 66(5):1813-1821.
3. Singh BK & Shaner DL (1998) Rapid determination of glyphosate injury to plants and identification of glyphosate-resistant plants. *Weed Technology* 12(3):527-530.
4. Jaworski EG (1972) Mode of Action of N-Phosphonomethylglycine - Inhibition of Aromatic Amino-Acid Biosynthesis. *Journal of Agricultural and Food Chemistry* 20(6):1195-1204.
5. Higuchi S, Kawashima T, & Suzuki M (1999) Comparison of pathways for amino acid biosynthesis in archaebacteria using their genomic DNA sequences. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences* 75(8):241-245.
6. Overbeek R, Fonstein M, D'Souza M, Pusch GD, & Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* 96(6):2896-2901.
7. Daugherty M, Vonstein V, Overbeek R, & Osterman A (2001) Archaeal shikimate kinase, a new member of the GHMP-kinase family. *Journal of Bacteriology* 183(1):292-300.
8. Selkov E, Maltsev N, Olsen GJ, Overbeek R, & Whitman WB (1997) A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* (1-2):197,11-26.
9. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA**,** FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273(5278):1058-1073.
10. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Peterson S, Reich CI, McNeil LK, Badger JH, Glodek A, Zhou L, Overbeek R, Gocayne JD, Weidman JF, McDonald L, Utterback T, Cotton MD, Spriggs T, Artiach P, Kaine BP, Sykes SM, Sadow PW, D'Andrea KP, Bowman C, Fujii C, Garland SA, Mason TM, Olsen GJ, Fraser CM, Smith HO, Woese CR, Venter JC (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390(6658):364-370.
11. Hendrickson EL, Kaul R, Zhou Y, Bovee D, Chapman P, Chung J, Conway de Macario E, Dodsworth JA, Gillett W, Graham DE, Hackett M, Haydock AK, Kang A, Land ML, Levy R, Lie TJ, Major TA, Moore BC, Porat I, Palmeiri A, Rouse G, Saenphimmachak C, Söll D, Van Dien S, Wang T, Whitman WB, Xia Q, Zhang Y,

Larimer FW, Olson MV, Leigh JA (2004) Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. *Journal of Bacteriology* 186(20):6956-6969.
12. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 392(6674):353-358.
13. Blackmore PF, Williams JF, & Macleod JK (1976) Dimerization of erythrose 4-phosphate. *Febs Letters* 64(1):222-226.
14. Tumbula DL, Teng Q, Bartlett MG, & Whitman WB (1997) Ribose biosynthesis and evidence for an alternative first step in the common aromatic amino acid pathway in *Methanococcus maripaludis*. *Journal of Bacteriology* 179(19):6010-6013.
15. Hochuli M, Patzelt H, Oesterhelt D, Wuthrich K, & Szyperski T (1999) Amino acid biosynthesis in the halophilic archaeon *Haloarcula hispanica*. *Journal of Bacteriology* 181(10):3226-3237.
16. Fischer RS, Bonner CA, Boone DR, & Jensen RA (1993) Clues From A Halophilic Methanogen About Aromatic Amino-Acid Biosynthesis In Archaebacteria. *Archives of Microbiology* 160(6):440-446.
17. White RH (2004) L-aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannaschii*. *Biochemistry* 43(23):7618-7627.
18. Porat I, Waters BW, Teng Q, & Whitman WB (2004) Two biosynthetic pathways for aromatic amino acids in the archaeon *Methanococcus maripaludis*. *Journal of Bacteriology* 186(15):4940-4950.
19. Porat I, Sieprawska-Lupa M, Teng Q, Bohanon FJ, White RH, Whitman WB (2006) Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p-aminobenzoic acid in the archaeon *Methanococcus maripaludis*. *Molecular Microbiology* 62(4):1117-1131.
20. Gonzalez O, Gronau S, Pfeiffer F, Mendoza E, Zimmer R, Oesterhelt D (2009) Systems Analysis of Bioenergetics and Growth of the Extreme Halophile *Halobacterium salinarum*. *Plos Computational Biology* 5(4).
21. Gulko MK, Dyall-Smith M, Gonzalez O, & Oesterhelt D (2014) How Do Haloarchaea Synthesize Aromatic Amino Acids? *Plos One* 9(12).
22. White RH & Xu HM (2006) Methylglyoxal is an intermediate in the biosynthesis of 6-deoxy-5-ketofructose-1-phosphate: A precursor for aromatic amino acid biosynthesis in *Methanocaldococcus jannaschii*. *Biochemistry* 45(40):12366-12379.
23. Webby CJ, Jiao W, Hutton RD, Blackmore NJ, Baker HM, Baker EN, Jameson GB, Parker EJ (2010) Synergistic Allostery, a Sophisticated Regulatory Network for the Control of Aromatic Amino Acid Biosynthesis in *Mycobacterium tuberculosis*. *Journal of Biological Chemistry* 285(40):30567-30576.
24. Morar M, White RH, & Ealick SE (2007) Structure of 2-amino-3,7-dideoxy-D-threo-hept-6-ulosonic acid synthase, a catalyst in the archaeal pathway for the biosynthesis of aromatic amino acids. *Biochemistry* 46(37):10562-10571.
25. Shumilin IA, Bauerle R, Wu J, Woodard RW, & Kretsinger RH (2004) Crystal structure of the reaction complex of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from *Thermotoga maritima* refines the catalytic mechanism and indicates a

new mechanism of allosteric regulation. *Journal of Molecular Biology* 341(2):455-466.
26. Carpenter EP, Hawkins AR, Frost JW, & Brown KA (1998) Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis. *Nature* 394(6690):299-302.
27. Schleper C, Jurgens G, & Jonuscheit M (2005) Genomic studies of uncultivated archaea. *Nature Reviews Microbiology* 3(6):479-488.
28. Chaban B, Ng SYM, & Jarrell KF (2006) Archaeal habitats - from the extreme to the ordinary. *Canadian Journal of Microbiology* 52(2):73-116.
39. Coutinho PM & Henrissat B (1999) Life with no sugars? *Journal of Molecular Microbiology and Biotechnology* 1(2):307-308.
30. Woodard RW (2004) Unique biosynthesis of dehydroquinic acid? *Bioorganic Chemistry* 32(5):309-315.
31. Richards TA, Dacks JB, Campbell SA, Blanchard JL, Foster PG, McLeod R, Roberts CW (2006) Evolutionary origins of the eukaryotic shikimate pathway: Gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryotic Cell* 5(9):1517-1531.
32. Zhi XY, Yao JC, Li HW, Huang Y, & Li WJ (2014) Genome-wide identification, domain architectures and phylogenetic analysis provide new insights into the early evolution of shikimate pathway in prokaryotes. *Molecular Phylogenetics and Evolution* 75:154-164.
33. McPherson A (1995) Increasing the size of microcrystals by fine sampling of pH limits. *Journal of Applied Crystallography* 28:362-365.
34. Zheng B, Gerdts CJ, & Ismagilov RF (2005) Using nanoliter plugs in microfluidics to facilitate and understand protein crystallization. *Current Opinion in Structural Biology* 15(5):548-555.
35. Chayen NE (2005) Methods for separating nucleation and growth in protein crystallisation. *Progress in Biophysics & Molecular Biology* 88(3):329-337.
36. Ireton GC & Stoddard BL (2004) Microseed matrix screening to improve crystals of yeast cytosine deaminase. *Acta Crystallographica Section D-Biological Crystallography* 60:601-605.
37. Chayen NE (1998) Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques. *Acta Crystallographica Section D-Biological Crystallography* 54:8-15.
38. Behlke J, Marg A, & Paeschke M (1997) Protein crystallization with and without precipitants. *Journal of Applied Crystallography* 30(1):559-560.
39. McPherson A & Rich A (1972) X-ray crystallographic analysis of swine pancreas alpha-amylase. *Biochimica Et Biophysica Acta* 285(2):493-497.
40. Kelly SM, Jess TJ, & Price NC (2005) How to study proteins by circular dichroism. *Biochimica Et Biophysica Acta-Proteins and Proteomics* 1751(2):119-139.
41. Kelly SM & Price NC (2000) The Use of Circular Dichroism in the Investigation of Protein Structure and Function. *Current Protein & Peptide Science* 1(4):349-384.
42. Garman EF (2014) Developments in X-ray Crystallographic Structure Determination of Biological Macromolecules. *Science* 343(6175):1102-1108.
43. Haffke M, Viola C, Nie Y, & Berger I (2013) Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research. *Methods in molecular biology* 1073:131-140.

44. Hendrickson WA & Ogata CM (1997) Phase determination from multiwavelength anomalous diffraction measurements. *Macromolecular Crystallography, Pt A* 276:494-523.
45. Brodalla D, Mootz D, Boese R, & Osswald W (1985) Programmed crystal-growth on a diffractometer with focused heat radiation. *Journal of Applied Crystallography* 18:316-319.
46. Jancarik J & Kim SH (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography* 24:409-411
47. Helliwell JR (2012) The evolution of synchrotron radiation and the growth of its importance in crystallography. *Crystallography Reviews* 18(1):33-93.
48. Weik M, Ravelli RB, Kryger G, McSweeney S, Raves ML, Harel M, Gros P, Silman I, Kroon J, Sussman JL (2000) Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences of the United States of America* 97(2):623-628.
49. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D-Biological Crystallography* 68:352-367.
50. Sanchez R & Sali A (1997) Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology* 7(2):206-214.
51. Venclovas C, Zemla A, Fidelis K, & Moult J (1999) Some measures of comparative performance in the three CASPs. *Proteins-Structure Function and Bioinformatics*:231-237.
52. Havel TF & Snow ME (1991) A new method for building protein conformations from sequence alignments with homologs of known structure. *Journal of Molecular Biology* 217(1):1-7.
53. Battiste JL & Wagner G (2000) Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry* 39(18):5355-5365.
54. Hennig J & Sattler M (2014) The dynamic duo: Combining NMR and small angle scattering in structural biology. *Protein Science* 23(6):669-682.
55. Milne JL, Borgnia MJ, Bartesaghi A, Tran EE, Earl LA, Schauder DM, Lengyel J, Pierson J, Patwardhan A, Subramaniam S (2013) Cryo-electron microscopy - a primer for the non-microscopist. *Febs Journal* 280(1):28-45.
56. Bai XC, McMullan G, & Scheres SHW (2015) How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences* 40(1):49-57.
57. Laemmli UK (1970) Cleavage of structural proteins during assembly of head of bacteriophage-T4. *Nature* 227(5259):680-685.
58. Inoue H, Nojima H, & Okayama H (1990) High-efficiency transformation of *Escherichia-coli* with plasmids. *Gene* 96(1):23-28.
59. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA & Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* 6(5):343-U341.
60. Bradford MM (1976) Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding. *Analytical Biochemistry* 72(1-2):248-254.

61. Fenn JB, Mann M, Meng CK, Wong SF, & Whitehouse CM (1989) Electrospray ionization for mass-spectrometry of large biomolecules. *Science* 246(4926):64-71.
62. Perkins DN, Pappin DJC, Creasy DM, & Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18):3551-3567.
63. Pike ACW, Garman EF, Krojer T, von Delft F, & Carpenter EP (2016) An overview of heavy-atom derivatization of protein crystals. *Acta Crystallographica Section D-Structural Biology* 72:303-318.
64. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research* 31(13):3784-3788.
65. Finn RD*, et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44(1):279-285.
66. Eddy SR (1996) Hidden Markov models. *Current Opinion in Structural Biology* 6(3):361-365.
67. Slabinski L ,Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A (2007) XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* 23(24):3403-3405.
68. Xue B, Dunbrack RL, Williams RW, Dunker AK, & Uversky VN (2010) PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochimica Et Biophysica Acta-Proteins and Proteomics* 1804(4):996-1010.
69. Chen L, Oughtred R, Berman HM, & Westbrook J (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* 20(16):2860-2862.
70. Crooks GE, Hon G, Chandonia JM, & Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Research* 14(6):1188-1190.
71. Notredame C, Higgins DG, & Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302(1):205-217.
72. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics* 9(40).
73. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. *Nature Methods* 12(1):7-8.
74. Wu ST & Zhang Y (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research* 35(10):3375-3382.
75. Zhang Y & Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale'. *Proceedings of the National Academy of Sciences of the United States of America* 101(20):7594-7599.
76. Zhang Y & Skolnick J (2004) SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry* 25(6):865-871.
77. Zhang Y & Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* 33(7):2302-2309.
78. Rosano GL & Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology* 5.
79. Porath J, Carlsson J, Olsson I, & Belfrage G (1975) Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature* 258(5536):598-599.

80. Lowe CR & Dean PDG (1973) Affinity chromatography of lactate-dehydrogenase on immobilized nucleotides. *Biochemical Journal* 133(3):515-520.
81. Whitmore L & Wallace BA (2004) DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Research* 32:668-673.
82. Provencher SW & Glockner J (1981) Estimation of globular protein secondary structure from circular-dichroism. *Biochemistry* 20(1):33-37.
83. Vanstokkum IHM, Spoelder HJW, Bloemendal M, Vangrondelle R, & Groen FCA (1990) Estimation of protein secondary structure and error analysis from circular-dichroism spectra. *Analytical Biochemistry* 191(1):110-118.
84. Sreerama N & Woody RW (1993) A self-consistent method for the analysis of protein secondary structure from circular-dichroism. *Analytical Biochemistry* 209(1):32-44.
85. Sreerama N & Woody RW (2000) Estimation of protein secondary structure from circular dichroism spectra: Comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Analytical Biochemistry* 287(2):252-260.
86. Wiedemann C, Bellstedt P, & Goerlach M (2013) CAPITO-a web server-based analysis and plotting tool for circular dichroism data. *Bioinformatics* 29(14):1750-1757.
87. Park A, Lamb HK, Nichols C, Moore JD, Brown KA, Cooper A, Charles IG, Stammers DK, Hawkins AR (2004) Biophysical and kinetic analysis of wild-type and site-directed mutants of the isolated and native dehydroquinate synthase domain of the AROM protein. *Protein Science* 13(8):2108-2119.
88. Bayley PM & O'Neil KTJ (1980) The bonding of oxidised coenzyme to bovine-liver glutamate-dehydrogenase studied by circular-differnce spectrometry. *European journal of Biochemistry* 112(3):521-531.

Appendix One

FASTA and Nucleotide Sequences of DHQS2 Targets

>Sulfurihydrogenibium-azorense
KEFILDARGLPKTAITTAIESGVDYILIEEEKANEIKKLGRVKLILQDKEGNLTEDFVIV
KIENKQDEEKAANLAKSGKKVIVQTTDWTIIPLENLIAQSPNIYSIVKDAEEAKTSIGIL
EKGVRGVVLKTDNLNEIKKVAKIIKEDTEKLNMVKAKVTSIKPVGMGDRVAVDTTSLLKK
GEGMLVGNSSAGMILVHAETEESEYVASRPFRVNAGAVHMYTRVPNGKTTYLSELSSGKE
VMVYDFNGNGRVVYVGRAKVERRPMLLIEAEYEGKKLSAILQNAETIRVVKPDGTPVSVV
DLKVGDEIVGYVEEAGRHFGMKVEETIMEK - 330

ATGAAAGAGTTTATTCTTGATGCAAGAGGACTACCAAAAACTGCAATAACTACTGCTATAGAGTC
CGGAGTAGATTACATACTTATTGAAGAAGAAAAGGCAAACGAAATAAAAAAGTTAGGTAGAGTA
AAACTCATACTTCAAGACAAAGAAGGAAATCTTACAGAAGACTTCGTTATAGTAAAGATAGAAAA
TAAACAAGATGAAGAAAAAGCAGCAAACTTGGCAAAAAGCGGTAAAAAAGTTATAGTTCAAACT
ACAGACTGGACTATAATTCCTTTAGAAAACCTTATAGCCCAATCACCTAACATATACTCTATAGTA
AAGGATGCAGAGGAAGCAAAAACATCTATAGGAATCCTTGAAAAAGGTGTAAGAGGAGTAGTTC
TTAAAACAGACAACTTGAACGAAATAAAAAAAGTAGCGAAAATTATAAAAGAGGATACAGAAAA
GTTAAACATGGTAAAAGCAAAGGTGACAAGTATAAAACCCGTAGGTATGGGGGATAGAGTAGCT
GTGGATACAACGTCCTTACTTAAAAAAGGAGAAGGGATGCTTGTAGGTAACTCTTCAGCTGGCAT
GATTTTAGTACATGCTGAGACAGAAGAGTCAGAGTATGTAGCATCACGACCTTTTAGAGTAAATG
CTGGAGCTGTCCATATGTATACCCGTGTACCAAACGGTAAAACTACATATCTAAGTGAGCTTTCAT
CTGGAAAAGAAGTTATGGTATATGACTTTAATGGAAACGGAAGAGTTGTGTACGTAGGAAGGGCA
AAAGTAGAAAGAAGACCTATGCTCTTAATAGAAGCTGAGTACGAAGGGAAAAAACTCTCTGCTAT
TCTACAGAATGCAGAAACTATAAGAGTTGTAAAACCAGATGGAACACCTGTATCAGTAGTAGACT
TAAAAGTTGGTGATGAAATAGTAGGGTACGTAGAAGAAGCTGGAAGACACTTTGGAATGAAAGTA
GAAGAAACAATTATGGAGAAGTAA

>Desulfatibacillum-alkenivorans
KTIWVKVDPWNKDLVTTALEGGADALMVPEGYAEKVKELGRITTIAPDGDMIPGQDLSTI
TITSMDDEDQIIQAQAMGPVIVDTPDWSIIPLENLIAKGARVVHPAKSLDEAKTAAGILE
HGVWGVLIDTPDPAELKRILAFLKGEGTTAPVVPVEIMEVKAVGMGDRVCVDTCTDMKPG
EGMLVGNSSAAMFLIHSESIENPYVAARPFRVNAGAVHAYTKVPGNKTRYLSELGAGDEV
LITDYKGKCSIGVVGRLKIEKRPLLLITAKAGDKQITTICQNAETIRLVAADGKPLSVVA
LKPGDKVLASLEEGGRHFGHKIKETITEK - 329

ATGAAAACCATTTGGGTAAAAGTGGATCCCTGGAACAAAGACCTTGTCACCACCGCCCTGGAGGG
CGGCGCGGACGCGCTTATGGTTCCCGAGGGATACGCTGAAAAAGTCAAGGAACTGGGCCGGATAA
CCACTATCGCCCCGGACGGGGACATGATTCCCGGCCAGGACCTGAGCACCATCACCATTACCAGC
ATGGACGACGAAGACCAGATTATCCAGGCACAGGCCATGGGGCCGGTCATTGTGGACACCCCGGA
CTGGAGCATCATCCCCCTGGAAAACCTCATCGCCAAGGGCGCCAGGGTGGTGCATCCGGCCAAGA
GCCTGGACGAAGCCAAGACCGCCGCCGGCATCCTGGAGCATGGAGTCTGGGGCGTGTTGATCGAC
ACGCCGGACCCGGCCGAGCTTAAACGCATCCTGGCCTTTTTAAAAGGCGAGGGCACGACCGCTCC
GGTGGTTCCGGTGGAAATCATGGAAGTCAAGGCCGTGGGCATGGGCGACCGGGTGTGCGTGGACA
CCTGCACGGACATGAAGCCCGGAGAAGGCATGCTGGTGGGCAATTCCTCGGCGGCCATGTTTCTG
ATTCACTCCGAAAGCATTGAGAATCCCTATGTGGCGGCTCGTCCTTTCAGAGTGAACGCAGGCGCG
GTCCATGCCTACACCAAGGTCCCGGGCAATAAAACCCGGTATCTTTCCGAACTCGGCGCAGGGGA
CGAAGTGCTTATCACGGATTACAAGGGCAAGTGCTCCATCGGCGTTGTGGGCAGGCTGAAAATCG
AAAAACGGCCTTTGCTGTTGATTACCGCCAAGGCAGGCGATAAACAGATAACCACCATTTGCCAA
AACGCCGAAACCATCCGCCTTGTGGCTGCCGACGGCAAACCCCTTTCCGTGGTGGCTCTCAAGCCA
GGCGACAAAGTGCTGGCCAGCCTGGAAGAAGGCGGGCGTCACTTCGGCCATAAAATCAAGGAGA
CCATAACGGAAAAATGA

>Methanospirillum-hungatei
KQVFVDLRPWDKELAIAALESGAAGVIADSAGPVRELGRILVIAPDGDLIPGQDIHEITI
GNTEDQARAMEAARTCRIIVHTPDWTIIPLENLVACGDNVIAVVSDIKEAEQALTVLEKG
VSGVLVKTDDPDLVRSICRMVQSGISGQQLHRLTVTTVKPAGMGERVCVDTCSLMVDGEG
MLVGNTSSGFFLVHAETLVNPYVAPRPFRVNAGGVHAYLQVPEGKTAYLADLKAGDRVMI
VHGNGSCREATVGRVKIERRPLFLVEAESECQKVSIILQNAETIRLVRPDNSAVSVTSLK
PGDVVLGRVESGGRHFGMAIDETIIEK – 327

ATGAAACAGGTATTTGTAGACCTTCGTCCGTGGGATAAGGAACTTGCAATAGCCGCCCTTGAATCA
GGAGCTGCAGGAGTAATTGCAGATTCAGCAGGTCCGGTCAGAGAACTGGGCCGTATTCTGGTCAT
CGCTCCTGATGGGGATCTAATTCCGGGGCAGGATATTCATGAGATTACTATTGGGAATACAGAGG
ACCAGGCTCGTGCAATGGAAGCTGCCAGGACCTGCCGCATTATCGTGCACACTCCTGACTGGACG
ATAATTCCTCTTGAAAACCTGGTGGCATGCGGGGACAATGTAATAGCCGTCGTTTCAGATATAAAA
GAGGCTGAGCAGGCTCTGACTGTACTTGAAAAAGGAGTCTCGGGTGTCCTCGTGAAAACGGATGA
TCCGGACCTTGTCAGATCCATATGTCGGATGGTTCAGTCAGGAATATCCGGACAACAACTCCACCG
TCTGACCGTAACCACTGTAAAACCTGCCGGAATGGGTGAGAGGGTCTGTGTCGATACCTGTTCTCT
TATGGTTGACGGCGAAGGGATGCTTGTTGGAAATACATCATCCGGATTTTTTCTCGTTCATGCAGA
GACGCTTGTAAATCCCTATGTAGCGCCCAGGCCGTTCAGGGTGAATGCAGGTGGCGTTCATGCATA
TCTGCAGGTTCCGGAAGGAAAGACTGCATATCTGGCTGACCTGAAGGCAGGTGACCGGGTGATGA
TAGTTCATGGAAACGGATCCTGCCGGGAAGCGACTGTTGGCAGGGTGAAAATTGAACGGCGTCCG
CTCTTTCTTGTTGAAGCAGAGAGCGAGTGTCAGAAGGTATCAATTATCCTGCAGAATGCAGAGACC
ATCAGGCTTGTTCGCCCGGATAATTCGGCAGTATCTGTAACCTCACTCAAACCCGGTGATGTTGTC
CTGGGCCGTGTAGAAAGTGGCGGTCGTCATTTTGGTATGGCAATTGATGAGACCATCATTGAGAA
ATAA

>Methanosphaera-stadtmanae
KFAWIRPNGTWNDRKEAIVDSLESGFDHIMDLDNAETIKKLGSVTIISDKEDSDITLLGL
NNKITMADIKKAQESGKEVAAYVEINNKDDELLVSKLGTVADYVILKGKNWKVIPLENII
ASLQNRTSKIIVDVPNYEEAKLALETMEHGSDGVLLSSNDGNEIRKLGALIEKVSKESYD
LKAATVTKVESVGIGDRVCVDTCSMMNVGDGMLVGSFASGLFLVHSETLESEYVASRPFR
VNAGPVHAYVMTPENKTRYLSELEAGDEVVTLNSKGEANTVIVGRVKIEKRPLLLIEAKY
KNSRIRTLVQNAETIRLVNDKGEPISVSKLKVGDKVLAYFSEAARHFGMAIEEQIIEK – 358

ATGAAATTTGCATGGATAAGACCTAATGGTACTTGGAATGATCGTAAAGAAGCTATTGTTGATTCA
TTAGAATCTGGATTCGATCACATAATGGATTTAGATAATGCTGAAACTATTAAGAAATTAGGTAGT
GTAACAATAATATCAGATAAAGAAGATTCTGACATTACACTTCTTGGATTAAATAACAAAATTACC
ATGGCAGATATTAAAAAAGCACAAGAAAGTGGTAAAGAAGTAGCAGCATATGTTGAAATTAATA
ACAAGGATGATGAGTTACTAGTATCTAAACTAGGTACAGTAGCAGACTATGTTATATTAAAAGGT
AAAAATTGGAAAGTTATCCCTTTAGAAAATATCATTGCCAGTTTACAAAATAGAACTTCTAAAATC
ATAGTTGATGTTCCAAATTATGAAGAAGCAAAATTAGCTCTTGAAACAATGGAACATGGATCTGA
TGGTGTGTTATTATCATCTAATGATGGAAATGAAATAAGAAAATTAGGTGCCCTAATAGAAAAAG
TATCAAAGGAATCATATGATCTTAAAGCTGCAACTGTTACTAAAGTAGAATCTGTTGGTATTGGTG
ATAGAGTATGTGTAGATACATGTTCAATGATGAATGTTGGAGATGGAATGCTTGTAGGTTCATTTG
CATCTGGACTATTTTTAGTTCATAGTGAAACATTAGAAAGTGAATATGTTGCATCACGTCCATTTA
GAGTAAATGCTGGTCCAGTACATGCTTATGTTATGACTCCAGAGAATAAAACAAGATATCTTTCAG
AATTAGAAGCTGGAGATGAAGTTGTTACATTAAATAGTAAAGGTGAAGCTAACACTGTTATTGTTG
GCCGTGTAAAAATAGAAAAACGTCCTTTACTTCTTATTGAAGCAAAATATAAGAATTCTAGAATAA
GAACTCTTGTACAGAATGCTGAAACAATTAGACTAGTAAATGATAAAGGTGAACCAATCTCTGTTT
CTAAACTAAAAGTTGGAGATAAAGTTCTTGCATACTTTAGTGAAGCTGCAAGACACTTTGGAATGG
CTATTGAAGAACAAATTATTGAAAAATAA

>Archaeoglobus-veneficus
KEIWLLTEGDNWSEVKEQVKDAIEIGFTGVVVKDEFAEKAAKLGRIAVVGKSDGEGLIGK
GKAYFLDITSAEAQQEAIELSEHADYLFLSFSDWRIIPLENLIAMRKRAKIIAEVSSVDD
ARVVLTTLEKGADGIAVKGNREELLGYFKAVMEDSGRLNLKPAKIVEIKPLGVGERVCID
TVTLMSVGEGMLVGNKAGFMFLVASESEESEYVASRPFRVNAGSVNAYIKVGDKTKYLAE
LKTGDEVEIVRYDGAARKSYVGRVKIERRPLMLVRAVANGEEGSVILQNAETIKLITSEG
KHKSVAELKAGDEVLVWVGQKARHFGVGVEEFIIER – 336

ATGAAAGAAATTTGGTTACTAACAGAGGGCGATAACTGGAGTGAAGTCAAGGAACAGGTGAAGG
ATGCTATAGAGATAGGCTTTACTGGCGTCGTTGTGAAGGATGAATTTGCGGAAAAAGCCGCAAAG
CTGGGTAGAATAGCTGTTGTTGGGAAGAGCGACGGCGAAGGACTTATTGGGAAAGGGAAGGCGT
ACTTCTTGGACATAACGTCTGCGGAAGCGCAGCAAGAAGCTATTGAGCTTTCGGAACATGCGGAT
TATCTCTTCCTGAGCTTTTCAGACTGGCGCATCATCCCTCTCGAAAACCTCATAGCCATGCGCAAG
CGGGCGAAGATCATTGCTGAAGTTAGCAGTGTTGATGACGCAAGGGTCGTGCTGACGACACTCGA
GAAAGGAGCGGATGGCATTGCTGTTAAAGGCAACAGGGAGGAACTGCTTGGCTACTTCAAGGCCG
TGATGGAGGACTCTGGAAGACTGAACCTCAAACCGGCAAAAATCGTGGAGATAAAACCTCTCGGT
GTCGGGGAGAGGGTTTGCATAGACACAGTTACGCTCATGAGCGTTGGAGAGGGGATGCTCGTTGG
CAACAAAGCCGGTTTTATGTTCCTCGTTGCGAGTGAGAGTGAGGAGAGCGAGTACGTCGCTTCCCG
CCCATTCAGAGTAAACGCAGGCAGTGTAAACGCGTACATAAAAGTTGGTGACAAAACGAAGTACC
TTGCAGAACTCAAGACAGGCGATGAAGTGGAAATTGTGAGGTACGATGGAGCAGCGAGGAAGAG
CTACGTTGGAAGGGTTAAGATAGAGCGTCGCCCCCTGATGCTCGTGAGAGCAGTAGCCAACGGAG
AGGAGGGCAGCGTAATACTGCAGAACGCCGAAACAATAAAGCTGATAACGTCGGAAGGGAAGCA
CAAATCAGTTGCAGAACTCAAAGCAGGCGACGAGGTACTCGTCTGGGTTGGCCAAAAGGCGAGGC
ACTTCGGCGTGGGTGTGGAGGAGTTCATCATCGAGAGGTAA

>Methanococcus-maripaludis
KFGWIKTTGNDLEERMESVKDALESSIPGIITEKEEISSVRELGNIKIVSDNLDADVVLI
NKGEDLEILKSAKLSGKETGVYVVINTKEDEVYATDVSKLDFVDYVVLEGSDWTIIPLEN
IIADLFSEEIKIVSVVTNVKDAEAAYEILEKGVDGVVLIPKDINEVKDFSKLIERMNSES
VKLDYATVTKIEPVGSGDRVCIDTCSMMEMGEGMLIGSYSRGMFLVHSETVENPYVATRP
FRVNAGPVHAYILCPENKTKYLSDLKAGDKVLVVSKNGETREAIIGRVKIEKRPLFLVEA
EYNGENLRTILQNAETIRLVGEDGKPVSVVDLKVGTKVLIKPDENARHFGMAIKETIIEK - 360

AAATTTGGATGGATAAAAACAACAGGAAATGACCTTGAAGAGAGAATGGAATCGGTAAAAGATG
CTCTTGAGAGTTCAATTCCCGGAATCATTGCAGAAAAGGAAGAAATAAGTTCTGTAAGAGAACTT
GGAAATATAAAAATTGTTTCTGATAATCTCGATGCAGATGTCGTTTTAATTAATAAAGGGGAAGAT
TTAGAAATTTTAAAATCTGCAAAACTCTCTGGAAAAGAAACCGGTGTTTACGTTGTAATAAACACT
AAAGAAGATGAAGTATACGCTACAGACGTAAGCAAACTCGATTTTGTCGATTACGTTGTTTTAGAA
GGAAGCGACTGGACAATTATTCCTCTTGAAAACATCATTGCAGATTTATTCAGCGAAGAAATAAA
AATTGTTTCAGTTGTAACAAACGTAAAAGACGCGGAAGCCGCATACGAAATTTTAGAAAAAGGAG
TTGACGGTGTAGTTTTAATTCCTAAAGATATTAACGAAGTAAAAGACTTCTCAAAATTAATTGAAA
GAATGAACTCTGAAAGCGTAAAACTCGACTACGCAACAGTTACAAAAATTGAACCTGTTGGAAGT
GGCGATCGGGTCTGTATTGACACATGTTCCATGATGGAAATGGGAGAAGGAATGTTAATTGGATC
CTACTCACGAGGGATGTTTTTAGTTCATTCAGAAACTGTTGAAAACCCGTATGTCGCTACAAGACC
ATTTAGGGTAAATGCCGGACCTGTACATGCCTATATCTTGTGCCCTGAAAACAAAACTAAATATTT
AAGCGATTTAAAAGCAGGCGATAAAGTTTTAGTTGTTAATAAAAACGGTGAAACGAGAGAAGCTA
TTATCGGACGCGTGAAAATCGAAAAAAGACCATTATTCCTTGTTGAAGCAGAATATAACGGGGAA
AATTTAAGAACAATCCTCCAAAACGCAGAAACTATCCGATTAGTTGGAGAAGATGGAAAACCAGT
TTCAGTTGTTGATTTAAAAGTTGGAACCAAAGTATTGATAAAACCAGATGAAAACGCGAGACACT
TTGGAATGGCGATTAAAGAAACAATTATAGAAAAA

Appendix Two

pETite vector T7 region

T7 promoter

5'GGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCTAATACGACTCACTA
3'CCGTCACTCGCGTTGCGTTAATTACACTCAAT

TAGGGTGTGAGCGGATAACAATTTCACGTGGAACAGCTAGAAATAATTTTGTTTAACTATAAGAA

GGAGATATACAT**ATG** CATCATCACCACCATCAC**TAA**TAGAGCGGCCGCCACCGCTGAGCAATAAC

T7 terminator

TAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATAT

CCGGAT3'

pETite Vector Full Sequence

GGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCTAATACGACTCACTATAGGGTGTGA
GCGGATAACAATTTCACGTGGAACAGCTAGAAATAATTTTGTTTAACTATAAGAAGGAGATATACAT**ATG**
[clonedgene]
CATCATCACCACCATCACTAATAGAGCGGCCGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGTCTAAACGG
GTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGGTAACGAATTCAAGCTTGATATCATTCAGGACGA
GCCTCAGACTCCAGCGTAACTGGACTGCAATCAACTCACTGGCTCACCTTCACGGGTGGGCCTTTCTTCGGTAGA
AAATCAAAGGTCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACC
AGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAGGTAACTGGCTTCAGCAGAGCGCAGATA
CCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTC
GCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGA
TAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACC
TACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGG
TATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTAT
AGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCATCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGG
AAAAACGCCAGCAACGCAGAAAGGCCCACCCGAAGGTGAGCCAGGTGATTACATTTGGGCCCTCATCAGAGGTTT
TCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATG
TCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCC
ATGTTAAGGGCGGTTTTTTCCTGTTTGGTCATTTAGAAAAACTCATCGAGCATCAAGTGAAACTGCAATTTATTC
ATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTC
CATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCC
TCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGCTTA
TGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCACCAACCAAACCG
TTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCCGTTAAAAGGACAATTACAAACAGGAATC
GAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACC
TGGAATGCTGTTTTCCCTGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATG
GTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGCCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCT
TTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCG
ACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTCGAGCAAGAC
GTTTCCCGTTGAATATGGCTCATAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTATT
TCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAAGTCAAAAGCCTCCGGTCGGAGGCTTTTGACTTTCT
GCTATGGAGGTCAGGTATGATTTAAATGGTCAGTATTGAGCGATATCTAGAGAATTCGTC

Appendix Three

Mod pEtite Vector

ACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACT
CACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAAC
CTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGA
ATGGCAAAAGCTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCG
CACCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCCGTTAAAAGGACAAT
TACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGAT
ATTCTTCTAATACCTGGAATGCTGTTTTCCCTGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGA
TAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGCCTGACCATCTCATCTGTAACATCAT
TGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCG
CACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCG
GCCTCGAGCAAGACGTTTCCCGTTGAATATGGCTCATAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCG
GTAGTGATCTTATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAAGTCAAAAGCCTCCGGTCGGAG
GCTTTTGACTTTCTGCTATGGAGGTCAGGTATGATTTAAATGGTCAGTATTGAGCGATATCTAGAGAATTCGTCG
GCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCTAATACGACTCACTATAGGGTGTGAG
CGGATAACAATTTCACGTGGAACAGCTAGAAATAATTTTGTTTAACTATAAGAAGGAGATATACATATGCATCAC
CATCACCATCACCATCACGATTACGATATCCCAACGACCGAAAACCTGTATTTTCAGAGCGGATCCTGGCCAAGC
AATGAGCTCAACGGCGTCGACCACAACTGGCTTCACCAGAGCGCAGATGGTTATAACACACCCGGGTCTCTGCTG
CAGGGTACCGGCAGCCATCACCATCATCACCACCATCACTAATAGAGCGGCCGCCACCGCTGAGCAATAACTAGC
ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGGTAACGAAT
TCAAGCTTGATATCATTCAGGACGAGCCTCAGACTCCAGCGTAACTGGACTGCAATCAACTCACTGGCTCACCTT
CACGGGTGGGCCTTTCTTCGGTAGAAAATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGC
TTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAGG
TAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGA
ACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGT
GTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCA
CACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGC
TTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTC
CAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCATCGATTTTTGTGAT
GCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCAGAAAGGCCCACCCGAAGGTGAGCCAGGTGAT
TACATTTGGGCCCTCATCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAG
CGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTA
ATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCATTTAGAAAAACTCATCGA
GCATCAAGTGAA