Biochemical Characterisation of Reconstructed Ancestral CM-DAH7PS Enzymes

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

Master of Science in Biological Sciences

at

The University of Waikato

by

Joel Patrick McMillan

The University of Waikato

2012

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Abstract

Thermophily has been proposed to be a primitive trait (Stetter, 2006, Pace, 1991, Woese, 1987) which has led to suggestions that all contemporary thermophilic species are the direct descendants of ancient thermophilic organisms. Ancestral sequence reconstruction (ASR) is a modern molecular technique which has been used to study the evolution of thermophily, however these studies have produced conflicting results. Studies utilising ASR to investigate the evolution of thermophily with two different proteins, elongation factor Tu and thioredoxin, have suggested a primitive origin for thermophily. However a recent study by Hobbs et al. (2012), in which ancestral IPMDH enzymes were reconstructed and biochemically characterised in order to investigate the evolution of thermophily in the *Bacillus* genus, suggested that thermophily may have evolved at least twice within the *Bacillus* genus, indicating that thermophily may not be strictly a primitive trait. This finding needs to be supported by data from other reconstructed enzymes.

Four ancestral *Bacillus* 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase enzymes with chorismate mutase fused to their N-termini (CM-DAH7PS) were reconstructed to investigate whether a similar trend would be observed with a different enzyme. The ages of these ancestral Precambrian enzymes were estimated to be between 570 and 1,079 million years old. The three most recent ancestral enzymes were characterised and deemed biologically feasible as evidenced by the kinetic properties they display, which are similar to those displayed by two extant CM-DAH7PS enzymes from *Bacillus subtilis* and *Bacillus stearothermophilus*. The X-ray crystal structure of the most recent CM-DAH7PS ancestor has been solved and found to be very similar to previously solved DAH7PS structures. The optimal temperatures for activity of the ancestral DAH7PS enzymes were used to infer how thermophily may have evolved in the *Bacillus* genus. Interestingly, the results conflict with the findings of Hobbs et al. (2012). This suggests that the biochemical characteristics of ancestral enzymes may be highly susceptible to the biochemical properties of the contemporary species used in the inference or that different enzymes may result in different inferences being drawn regarding the evolution of thermophily.

Acknowledgements

First and foremost, I would like to thank my supervisors Associate Professor Vic Arcus and Dr Jo Hobbs for their guidance and support throughout my project.

I would also like to thank everyone in the lab for all the advice and support you've given me over the past two years and for making the lab such a great environment to work in.

I would also like to acknowledge Associate Professor Emily Parker for her initial advice regarding the DAH7PS enzyme and for providing me with pGroESL. I would also like to thank Sebastian Reichau for coming to the rescue and supplying me with E4P, which had been a constant source of concern for me throughout the project. I would also like to thank the University of Waikato for awarding me the University of Waikato Masters Fees and Masters Research scholarships.

Thanks also go to all my friends for pretending to show interest in what I was doing and at times providing very welcome distractions from my research. Finally, I would like to thank my parents and the rest of my family for all their support and encouragement throughout the years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
AIC	Akaike information criterion
Anc2-CM-DAH7PS	CM-DAH7PS reconstructed from ANC2 node
Anc3-CM-DAH7PS	CM-DAH7PS reconstructed from ANC3 node
Anc4-CM-DAH7PS	CM-DAH7PS reconstructed from ANC4 node
Anc5-CM-DAH7PS	CM-DAH7PS reconstructed from ANC5 node
Ap-DAH7PS	DAH7PS from *Aeropyrum pernix*
APS	ammonium persulphate
ASR	ancestral sequence reconstruction
ATP	adenosine-5'-triphosphate
bp	base pair(s)
Bsel-CM-DAH7PS	CM-DAH7PS from *Bacillus selenitireducens*
Bstr-CM-DAH7PS	CM-DAH7PS from *Bacillus stearothermophilus*
Bsub-CM-DAH7PS	CM-DAH7PS from *Bacillus subtilis*
BTP	Bis-Tris propane
C-terminal	carboxyl terminus of peptide chain
cm	centimetre(s)
CM	chorismate mutase
CM-DAH7PS	CM fused to the N-terminus of DAH7PS
°C	degrees Celcius
DAH7P	3-deoxy-D-*arabino*-heptulosonate 7-phosphate
DAH7PS	3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DPA	dipicolinic acid
DSC	differential scanning calorimetry
E4P	erythrose 4-phosphate
EDTA	ethylene diamine tetraacetic acid
EF-Tu	elongation factor Tu
FL	ferredoxin-like

GDP	guanosine-5’-diphosphate
GITC	guanidinium thiocyanate
Gya	billion years ago
Gyr	billion years
h	hour(s)
His-tag	poly histidine tag
IMAC	immobilised metal affinity chromatography
ICPMS	inductively coupled plasma mass spectrometry
IPMDH	3-isopropylmalate dehydrogenase
IPTG	isopropylthio-β-D-galactosidase
kb	kilobase
kDa	kilo Daltons
KDO8PS	3-deoxy-D-*manno*-octulosonate 8-phosphate synthase
kV	kilovolts
L	litre
LB	Luria Bertani
Lm-CM-DAH7PS	CM-DAH7PS from *Listeria monocytogenes*
mA	milliamps
mAU	milli absorbance units
mM	millimoles
M	moles
MES	2(*N*-morpholino)ethanesulfonic acid
MOPS	2(*N*-morpholino)propanesulfonic acid
mg	milligrams
ML	maximum likelihood
MP	maximum parsimony
µF	microfarads
µg	micrograms
µL	microlitres
µm	micrometres
µM	micromoles
mm	millimetres
mL	millilitres

min	minutes
MME	monomethyl ether
MRCA	most recent common ancestor
MQ	milli Q
Mt-DAH7PS	DAH7PS from *Mycobacterium tuberculosis*
Mya	million years ago
Myr	million years
NB	nutrient broth
ng	nanograms
nL	nanolitres
nm	nanometres
NMR	nuclear magnetic resonance
N-terminal	amino terminus of peptide chain
OD_{600}	optical density at 600 nm
OGT	optimal growth temperature
PEG	polyethylene glycol
PCR	polymerase chain reaction
PDB	protein data bank
PEP	phosphoenolpyruvate
Pf-DAH7PS	DAH7PS from *Pyrococcus furiosus*
Pg-DAH7PS-CM	DAH7PS-CM from *Porphyromonas gingivalis*
Phe	phenylalanine
pmol	picomoles
P-protein	CM-prephenate dehydratase
RMSD	root mean square deviation
rpm	revolutions per minute
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
RO	reverse osmosis
s	second(s)
SSM	secondary structure matching
SAP	shrimp alkaline phosphatase
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis

TB	terrific broth
Temed	N, N, N’, N’-tetramethylethylenediamine
T_m	melting temperature
Tm-DAH7PS	DAH7PS from *Thermotoga maritima*
T_{opt}	optimal temperature for activity
T-protein	CM-prephenate dehydrogenase
tRNA	transfer ribonucleic acid
Trp	tryptophan
Tyr	tyrosine
UV	ultraviolet
V	volts
v/v	volume per volume
w/v	weight per volume

1 Introduction

1.1 Evolution of Thermophily

Thermophily is a property exhibited by a number of different species across a wide range of phyla from every domain of life. Numerous studies utilising a variety of different approaches have investigated the origin of thermophily and how it has evolved. These studies, however, have produced conflicting results and thus far the question of how thermophily originated and how it has evolved has yet to be resolved. Currently, the most popular hypothesis is that thermophily is a primitive trait that was exhibited by the most recent common ancestor (MRCA) of all extant organisms (Pace, 1991, Stetter, 2006, Woese, 1987). Thermophily has been proposed to be a primitive trait as the most deeply branching prokaryotes in the universal tree of life are thermophilic and the branch lengths of extant thermophilic species are generally short (Stetter, 2006). As such, this has led to suggestions that all thermophilic species are the direct descendants of ancient thermophilic organisms. The thermophilic (or hyperthermophilic) origin of life is also supported by oxygen isotope data for early diagenetic cherts, which suggest that oceanic temperatures throughout the Archaean eon (2.5-3.9 Gya) were between 55-85 °C (Knauth, 2005). Other researchers, however, disagree with the thermophilic MRCA hypothesis. Becerra et al. (2007) performed a phylogenetic analysis with protein disulphide reductases, a family of proteins known to play a crucial role in stabilising intracellular proteins in a number of thermophilic and hyperthermophilic species. Their analysis suggested that the MRCA of all extant organisms did not contain a protein disulphide oxidoreductase which argues against a thermophilic MRCA (Becerra et al., 2007). It has also been argued that the most deeply branching bacterial phyla are not thermophilic. Brochier and Philippe (2002) used a different approach to those used previously to analyse the bacterial phylogeny based on 16S rRNA and found that rather than thermophilic phyla being the most deeply branching, the mesophilic phylum Planctomycetales is the most deeply branching, suggesting a non-thermophilic ancestor for Bacteria. Galtier et al. (1999) also favour a mesophilic origin hypothesis and have proposed that thermophilic species generally have short branch lengths due to increased selective pressure acting on the rRNA of thermophiles rather than thermophiles

actually being primitive organisms. Galtier et al. (1999) inferred the rRNA sequence of the MRCA of extant organisms, analysed the G+C content of the ancestral rRNA, and from this inferred the optimal growth temperature (OGT) of the MRCA. The inferred G+C content of the MRCA rRNA sequence is incompatible with growth at high temperature, suggesting that extant thermophiles evolved from mesophilic organisms. The high temperature of the Precambrian oceans has also been challenged, with some proposing a cold, snow-covered early Earth (Runnegar, 2000). The climate of Precambrian Earth has also been proposed to have been erratic, with multiple snowball Earth's forming throughout this period (Evans et al., 1997). Ancestral sequence reconstruction (ASR), a modern molecular technique, has been used in a number of studies to investigate how thermophily has evolved. ASR can be used to study how thermophily has evolved as protein thermostability is typically correlated with the OGT of the host organism. Therefore, by reconstructing ancestral proteins and measuring their thermostabilities, it is possible to infer how thermophily may have evolved.

1.2 Ancestral Sequence Reconstruction

The properties of ancient biomolecules such as genes and proteins cannot typically be studied directly as they very rarely remain unchanged over long periods of time (Hanson-Smith et al., 2010). In 1963, Pauling and Zuckerkandl proposed that it would one day be possible to reconstruct ancient proteins by first inferring their amino acid sequences by comparing the sequences of related proteins from contemporary organisms, then synthesising the inferred ancestral proteins in the lab. Pauling and Zuckerkandl (1963) inferred the sequence of an ancient mammalian haemoglobin but due to technological limitations were unable to physically synthesise it. It was not until 27 years later that the first ancestral proteins were reconstructed (Malcolm et al., 1990, Stackhouse et al., 1990). With the exponential increase in available biological sequence data, improvements in computational power and the development of increasingly sophisticated statistical frameworks, ASR is becoming an increasingly common technique in evolutionary biology. ASR is a useful technique as it enables evolutionary biologists to investigate the properties and evolution of ancestral biomolecules. A schematic demonstrating the main steps in experimental ancestral inference studies is provided in Figure 1.1.

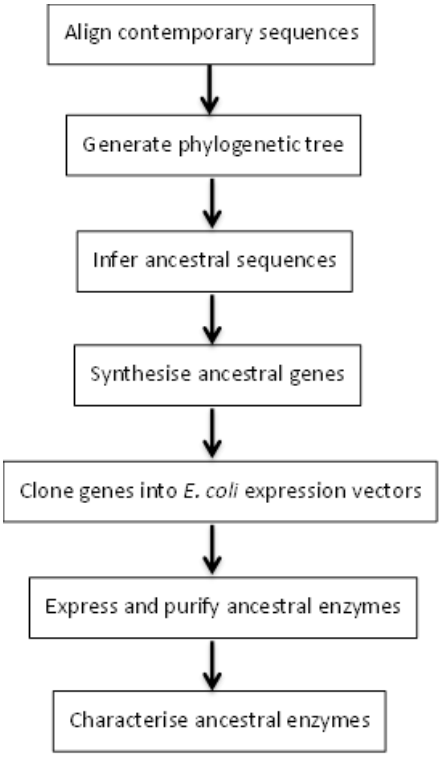

Figure 1.1 Flow diagram showing the general procedure used in ASR studies.

1.2.1 Ancestral Sequence Inference Methods

Three different methods have been used to infer ancestral biological sequences: maximum parsimony (MP), maximum likelihood (ML), and hierarchical Bayesian inference. All of these inference methods utilise an alignment of related extant biological sequences, a phylogenetic tree describing how the sequences are related and a suitable statistical model of evolution for those sequences to infer ancestral biological sequences (Hanson-Smith et al., 2010). The evolutionary model characterises the evolutionary processes responsible for trait evolution and the phylogenetic tree specifies the route through which the ancestral sequences evolved into the extant sequences. MP ancestral sequence inference was the method which was first implemented in ASR studies but has been superseded by ML ancestral sequence inference. Hierarchical Bayesian sequence inference has recently been proposed as a new method for inferring ancestral biological sequences.

1.2.1.1 Maximum Parsimony Ancestral Sequence Inference

The majority of the early ASR studies implemented MP methods to infer ancestral sequences; most notably to reconstruct ancestral RNases (Jermann et al., 1995, Stackhouse et al., 1990). Despite the extensive use of MP in early ASR studies and the many interesting results obtained, MP is no longer commonly used to infer ancestral biological sequences. The reason for this is that MP is a rather simplistic inference method which suffers from a number of inherent problems that are addressed by the more recently developed ML and hierarchical Bayesian inference methods (Cai et al., 2004). In the case of amino acid sequence inference, the MP method assigns amino acids to internal nodes (branch points) in a phylogenetic tree to minimise the number of amino acid changes along all branches of the tree, which is an extreme oversimplification of the evolutionary process (Yang et al., 1995). MP amino acid sequence inference does not take into account differing substitution patterns between amino acids or different phylogenetic branch lengths. It is unable to distinguish between equally parsimonious reconstructions, and it does not provide statistically robust measures of confidence (Cai et al., 2004).

1.2.1.2 Maximum Likelihood Ancestral Sequence Inference

One of the major disadvantages of the MP inference method is its inability to quantify the uncertainty of reconstructed ancestral sequences (Yang, 2006). Having a reliable measurement of the accuracy of reconstructed ancestral sequences is very important as, in order to make reliable conclusions about the evolutionary or functional characteristics of ancestral molecules, you must be confident that the inferred ancestral molecule is likely to be an accurate estimate of the true ancestral molecule.

The ML ancestral sequence inference method was the first method developed that provides statistically robust measures of confidence (Yang et al., 1995). Unlike the MP approach, ML inference takes into account the lengths of the branches in the phylogenetic tree and the different substitution rates between amino acids or nucleotides when reconstructing ancestral amino acid or nucleotide sequences (Yang, 2006). In ML ancestral amino acid sequence inference, the likelihood of the twenty common amino acids occupying a particular site in an ancestral

sequence is calculated for each position in the sequence at each internal node of the phylogenetic tree (Hanson-Smith et al., 2010). The likelihood is the probability of observing all the extant states given the ancestral state, the phylogenetic tree, and the evolutionary model (Hanson-Smith et al., 2010). Ancestral amino acid sequences inferred by ML are composites of the most probable amino acids at each position in the sequences. The ML method has also been referred to as the empirical Bayesian method because it uses parameter estimates to calculate the posterior probabilities of the ancestral sequences (Yang, 2006).

In order to obtain accurate ancestral biomolecules it is essential that the sequence alignment, phylogenetic tree, and evolutionary model parameters are as accurate as possible, as an error in any of these parameters is likely to generate inaccurate ancestral biomolecules. One problem with the ML inference method is the fact that it does not accommodate sampling errors in parameter estimates which can be problematic, particularly with small data sets where there may be insufficient information to estimate the parameters reliably (Yang, 2006). A hierarchical Bayesian approach to ASR has been developed by Huelsenbeck and Bollback (2001) which assigns priors to parameters and averages over their uncertainties through Markov chain Monte Carlo algorithms, potentially making the ancestral biomolecules reconstructed using hierarchical Bayesian inference more accurate than those reconstructed by ML inference.

1.2.1.3 Hierarchical Bayesian Ancestral Sequence Inference

Hierarchical Bayesian inference differs from ML inference in that it does not assume the most likely estimate of the phylogenetic tree and other parameters to be true; it integrates uncertainty in parameters such as branch lengths and tree topology by summing the likelihoods over a distribution of potential trees or parameters (Yang, 2006). Integrating this uncertainty may improve the accuracy of the reconstructions if the most likely tree and model parameters are not an accurate estimate of the true tree and model parameters.

1.2.2 Accuracy of Ancestral Sequence Reconstruction

It is rather difficult to assess the accuracy of ASR as in most cases the actual ancestral sequence is unknown. There are, however, a couple of exceptions. One such exception is a known bacteriophage T7 phylogeny that was generated by serial propogation of bacteriophage T7 in the presence of a mutagen by Hillis et al. (1992). Ancestral sequences were inferred from the terminal nodes of the phylogeny and the restriction-site maps of the inferred ancestors were compared with those of the actual viral ancestors to provide an estimate of the accuracy of MP reconstruction. It was found that MP generated restriction-site maps were greater than 98% accurate. A similar study by Oakley & Cunningham (2000) produced conflicting results which showed that MP methods did not accurately reconstruct ancient viral phenotypic characteristics. The usefulness of these experimental phylogenetic approaches is limited as they have mainly tested the accuracy of MP inference rather than the more recently developed ML and hierarchical Bayesian reconstruction methods, which are generally considered to provide more accurate estimates of ancestral character states. The relevance of results from studies on artificially accelerated bacteriophage phylogenies to the natural evolution of biomolecules over millions of years is also questionable.

Over the last ten years. the majority of experimental studies utilising ASR have used the ML inference method to reconstruct ancestral sequences. The accuracy of ML inference, however, has recently been questioned (Krishnan et al., 2004, Williams et al., 2006, Hall, 2006). It has been suggested that the hierarchical Bayesian method provides a more accurate estimation of the true ancestral character states and should be the method employed for all future ancestral reconstructions, and that results from previous studies using MP and ML inference methods may need to be re-examined (Hall, 2006, Krishnan et al., 2004, Williams et al., 2006). Krishnan et al. (2004) found that ancestral primate mitochondrial tRNA sequences inferred using the hierarchical Bayesian method were theoretically more functional than those inferred by ML because of an inherent deterministic bias in ML reconstruction. They also warned that similar reconstruction biases could be seen with reconstructed ancestral proteins. They specifically questioned the validity of the results from an ASR study by Gaucher et al. (2003). Based on the observation that the closer the ancestral elongation

factor Tu (EF-Tu) proteins were to the root of the phylogenetic tree the more thermostable they were, Gaucher et al. (2003) proposed that extant EF-Tu proteins have descended from a highly thermostable ancestor. Krishnan et al. (2004) suggested that these results may have been caused by a potential bias in ML reconstructions that favours amino acids that increase the thermostabilities of the reconstructed ancestral proteins, rather than representing a true biological phenomenon. This suggestion is supported by a more recent study by Williams et al. (2006) which found that the theoretical thermostabilities of ancestral proteins inferred by ML were overestimated when compared with proteins reconstructed using hierarchical Bayesian inference. In this study, Williams et al. (2006) performed 100 evolutionary simulations of a population of proteins undergoing nearly neutral evolution with purifying selection. Ancestral amino acid sequences were then inferred from the generated terminal amino acid sequences. The amino acid sequences of the inferred ancestral proteins and their predicted thermostabilities were then compared with the actual ancestral proteins. It was observed that, although on average the ML inferred ancestral sequences were more accurate than the ancestral sequences inferred by hierarchical Bayesian reconstruction, the ML reconstructions tended to overestimate the thermostability of ancestral proteins compared to the hierarchical Bayesian reconstructions. The results of this study must be interpreted with caution, however, as the proteins are theoretical and the thermostabilities are only predicted values.

More recent studies have produced results which conflict with the suggestion that hierarchical Bayesian ancestral inference provides a more accurate estimation of true ancestral character states. The major proposed advantage of the hierarchical Bayesian inference method is the integration of uncertainties in the parameter estimates used to infer ancestral sequences. In a recent experimental ASR study, Gaucher et al. (2008) reconstructed ancestral EF-Tu proteins to investigate the robustness of ML ASR for thermostability when uncertainties exist in the ancestral sequences and the amino acid frequencies used in the model. Results from this study demonstrated that ML ASR can be phenotypically robust to inaccuracies in *a priori* parameters. Hanson-Smith et al. (2010) assessed the accuracy of ML and hierarchical Bayesian amino acid sequence reconstructions from simulations of four-taxon ultrametric trees, non-ultrametric trees, and larger

trees derived from empirical datasets of four gene families previously used in ASR studies that used ML inference. They observed that integrating over topological uncertainty did not increase the accuracy of the reconstructed ancestral amino acid sequences.

All of the studies mentioned thus far that have directly compared the accuracies of the ML and hierarchical Bayesian inference methods have been theoretical studies, and until recently no ancestral biomolecules inferred using the hierarchical Bayesian method had even been reconstructed. Hobbs et al. (2012) were the first group to reconstruct and biochemically characterise ancestral proteins reconstructed using hierarchical Bayesian inference. They reconstructed ancestral *Bacillus* 3-isopropylmalate dehydrogenase (IPMDH) enzymes inferred by both ML and hierarchical Bayesian inference. The thermodynamic and kinetic properties of these enzymes were compared with each other and with IPMDH enzymes from contemporary *Bacillus* species to assess the biological feasibility of the reconstructed ancestral enzymes. It was found that although the ML and hierarchical Bayesian inferred ancestral IPMDH enzymes displayed similar levels of thermophily, the hierarchical Bayesian inferred ancestral IPMDH enzymes were biologically unrealistic as they exhibited aberrantly high K_M values and were kinetically unstable.

Further investigation is required to definitively determine which inference method generates biomolecules that most accurately estimate the nature of the actual ancestral biomolecules. Thus far, the theoretical simulation studies are inconclusive and suffer from a number of inherent weaknesses. More ancestral proteins from different protein families inferred by the hierarchical Bayesian method need to be reconstructed to determine whether hierarchical Bayesian inference consistently produces biologically unrealistic proteins, or if the reconstructed IPMDH enzymes are simply an anomaly. However, based on the evidence currently available, the method employed to infer ancestral proteins should be ML.

1.2.3 ASR Studies

ASR has so far been used to investigate the evolution of elongation factors (Gaucher et al., 2008, Gaucher et al., 2003), steroid hormone receptors (Bridgham

et al., 2010, Bridgham et al., 2006, Bridgham et al., 2009, Ortlund et al., 2007, Carroll et al., 2011, Li et al., 2005, Thornton et al., 2003), visual pigments (Chang et al., 2002, Chinen et al., 2005, Yokoyama et al., 2008, Shi and Yokoyama, 2003), fluorescent proteins (Field and Matz, 2010, Ugalde et al., 2004), dehydrogenases (Hobbs et al., 2012, Thomson et al., 2005), glycoside hydrolases (Malcolm et al., 1990), carbohydrate-binding proteins (Konno et al., 2007), ribonucleases (Stackhouse et al., 1990, Jermann et al., 1995, Zhang and Rosenberg, 2002), serine proteases (Chandrasekharan et al., 1996), and thioredoxins (Perez-Jimenez et al., 2011). These studies have provided important insights into the evolution of these biomolecules.

The majority of ASR studies have reconstructed binding proteins, while very few have reconstructed enzymes. Apart from the studies of Hobbs et al. (2012) and Perez-Jimenez et al. (2011), all of the enzymes reconstructed have been evolutionarily young (30-80 Myr old) and their amino acid sequences differed from their extant descendants at only a small number of sites. As the age of the ancestral sequences targeted for reconstruction increases, and the level of sequence divergence increases, accurately inferring ancestral proteins becomes increasingly difficult. Reconstructing ancestral biomolecules accurately, becomes particularly difficult when extant sequences are highly divergent as a large number of changes have occurred since sequence divergence, making it more likely that erroneous amino acids may be inferred in ancestral sequences. This may result in an ancestral protein displaying different phenotypic characteristics to the true ancestral protein. Despite the difficulty associated with the reconstruction of ancient Precambrian enzymes, there are many interesting questions about the evolution of Precambrian life that can potentially be investigated by using ASR to reconstruct ancestral biomolecules and then biochemically characterising them. Precambrian proteins from three different protein families have been reconstructed to date, all of which have been reconstructed primarily to study the origin and evolution of thermophily. One of these is the EF-Tu protein family, a small, highly conserved family which present charged aminoacyl-tRNAs to the ribosome during translation (Gaucher et al., 2008). Another is the small oxidoreductase enzyme thioredoxin, which has a relatively simple catalytic mechanism involving very few active site residues

(Holmgren, 1995). The most recent protein reconstructed is the large, structurally complex, dimeric enzyme IPMDH which has a more complex mode of action than thioredoxin (Dean and Dvorak, 1995, Holmgren, 1995). Reconstructing enzymes as opposed to binding proteins is desirable as their activity can be easily measured, and their thermodynamic and kinetic properties compared with extant enzymes to not only provide insights into the evolution of biophysical properties but also to act as an internal control for accurate ancestral reconstruction. Measurement of enzymatic activity can act as an effective control because of the strict structural requirements for enzymatic activity. Any errors in the ancestral inference are therefore likely to affect enzymatic activity, resulting in either inactive or biologically unrealistic enzymes. Reconstructing structurally complex multimeric enzymes provides an additional internal control because errors in ancestral inference could prevent the formation of the appropriate quaternary structure, thereby eliminating activity. If biologically realistic, large, structurally complex, multimeric ancestral enzymes are able to be reconstructed, it is likely that their biochemical properties will be an accurate estimation of the properties of the true ancestral enzymes.

1.2.4 Using ASR to Study the Evolution of Thermophily

To date, the question of how thermophily has evolved is far from resolved. The two major hypotheses are: (1) thermophily is a primitive trait and all extant thermophilic organisms are directly descended from ancient thermophilic organisms; and (2) thermophily is a more recent adaptation and extant thermophiles have evolved from mesophilic organisms. Several ASR studies have attempted to determine how thermophily has evolved but, like other approaches, have produced differing results.

Protein thermostability has been shown to be strongly correlated with the OGT of the host organism (Gromiha et al., 1999). Therefore, if it is possible to reconstruct Precambrian proteins accurately using ASR, it should be possible to infer the OGT of the ancient host organism by determining the melting temperatures (T_m) or optimal temperatures for activity (T_{opt}) of inferred ancestral proteins. This information can then be used to provide clues about the temperatures of the

environments particular ancestral species inhabited and how different ancestral species may have evolved to adapt to different environmental temperatures.

The first ASR study performed for the purpose of investigating the evolution of thermophily was a study by Gaucher et al. (2003) in which they reconstructed ancient bacterial EF-Tu proteins to study the evolution of thermophily in Bacteria. Elongation factors were chosen because of their low level of sequence divergence and the close correlation between the thermostabilities of EFs and the OGT of the host organism (Gaucher et al., 2003). Two different phylogenetic trees were used in their reconstruction to ensure any interpretations of their data were robust to differences in the evolutionary models employed. EF-Tu proteins from the most recent bacterial common ancestors were reconstructed for both trees as well as the MRCA for mesophilic bacteria. The thermostabilities of the three ancestral EF-Tu and the EF-Tu from three contemporary organisms were determined by measuring the proteins' ability to bind GDP at different temperatures. It was found that the most recent bacterial common ancestors constructed from the different phylogenetic trees displayed the same optimal binding temperature as the extant thermophilic EF-Tu from *Thermus aquaticus*. The EF-Tu from the MRCA of mesophilic bacteria was found to be more thermostable than EF-Tu from extant mesophilic descendants. From these observations Gaucher et al. (2003) suggested that the palaeoenvironment of the most recent bacterial common ancestor was ~65 °C and that the MRCA of mesophilic bacteria lived at a higher temperature than its descendants, which suggests a primitive origin for thermophily.

The work with EFs was continued by Gaucher et al. (2008) in order to not only resolve some of the issues raised by Williams et al. (2006), regarding the initial study of Gaucher et al. (2003), but also to investigate how thermophily had evolved over time in greater detail. Two phylogenies from the literature were used to reconstruct ancestral EFs. The two phylogenies represent the two main competing views of bacterial evolution. One phylogeny places hyperthermophilic lineages at the basal branches of the bacterial tree, whereas the other phylogeny places them in a more derived position in the tree. For the ancestral EFs reconstructed from both phylogenies, a progressive decrease in the thermostabilities of ancestral EFs is observed as the age of the ancestral elongation factors decreases from ~3.5 to ~0.5 Gyr old (Figure 1.2).

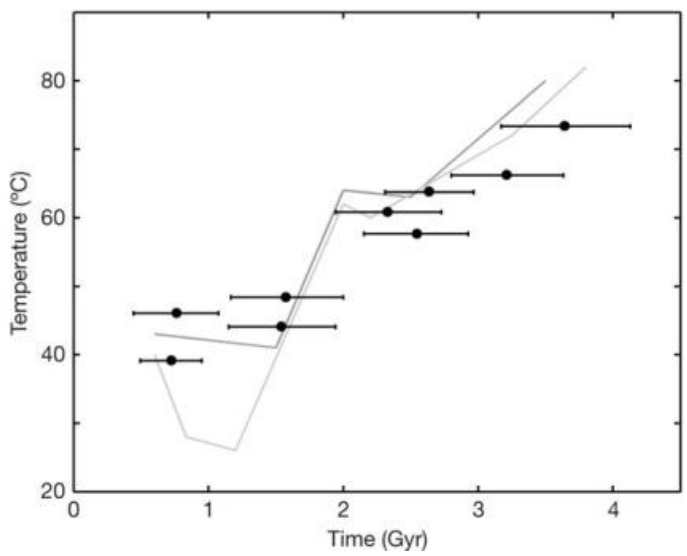

Figure 1.2 Plot of ancestral EF melting temperatures against time.
Black dots represent the melting temperatures of the reconstructed EF proteins. The horizontal bars are the confidence intervals of the estimated ages of the ancestral EFs. The grey lines represent the temperatures of the ancient ocean inferred from $\delta^{18}O$ isotope ratios. Figure is from Gaucher et al (2008).

A correlation was also noted between the progressive decrease in the thermostabilities of the ancestral EFs and a decrease in the temperature of the ancient oceans as determined by $\delta^{18}O$ and $\delta^{30}Si$ isotope ratios (Knauth, 2005, Robert and Chaussidon, 2006). However, the use of stable isotope data to infer ancient oceanic temperatures is controversial (Jaffres et al., 2007) as the changes in the isotope ratios may simply be the result of varying seawater composition, which would suggest a more temperate ancient ocean than the one proposed. Inferring the temperature of the entire ancient ocean from the inferred OGTs of ancient bacteria is also questionable. While the inferred OGTs of ancient bacteria may be used to suggest the temperature of the environment that particular ancestral species occupied, it should not be used to infer ancient oceanic temperatures. The results of Gaucher et al. (2008) suggest that thermophily is a primitive trait that was exhibited by the earliest forms of bacteria, that modern thermophilic bacteria are the direct descendants of these primitive thermophiles, and that ancient bacteria inhabited thermophilic environments. This, however, does not mean that the temperature of the ancient oceans was high; it may simply indicate that ancient bacteria inhabited thermal environments, such as hot pools or deep-sea hydrothermal vents. In this context it should be noted that extant species inhabit a wide range of temperatures and that a prediction of current oceanic temperatures based on one of these species could be extremely inaccurate.

The Gaucher et al. (2003) and Gaucher et al. (2008) studies are not the only ASR studies to suggest that thermophily is a primitive trait. A more recent ASR study conducted by Perez-Jimenez et al. (2011) also supported the idea of thermophily being a primitive trait. Seven Precambrian thioredoxin enzymes from the Bacteria, Archaea or Eukarya domains, aged between ~1.4 and ~4 Gyr old, were reconstructed and their thermostabilities measured. It was found that as the age of the ancestral thioredoxin enzymes decreased, the thermostabilities of the enzymes also decreased, as was seen with ancient bacterial EF-Tu proteins (Gaucher et al., 2008, Gaucher et al., 2003).

A very recent study conducted by Hobbs et al. (2012) used ASR as a tool to investigate the evolution of thermophily in a single genus. Four ancestral IPMDH enzymes from the *Bacillus* genus were reconstructed and biochemically characterised to investigate the evolution of thermophily over a much finer time scale (~670 to ~950 Mya) than previous ASR studies that have reconstructed Precambrian proteins. T_m, T_{opt}, and the $\Delta G^{\ddagger}$ for unfolding were measured for the four reconstructed ancestral IPMDH enzymes and compared with IPMDHs from contemporary psychrophilic, mesophilic, and thermophilic *Bacillus* species. Their results suggested that the MRCA of *Bacillus* was a thermophile, which would appear to support the hypothesis that thermophily is a primitive trait. However, contrary to the studies of Gaucher et al. (2008) and Perez-Jimenez et al. (2011), a temporal decrease in the thermophilic nature of the ancestral enzymes was observed before the reappearance of thermophily in the most recent ancestral enzyme reconstructed. It was also noted that the IPMDHs from the MRCA of *Bacillus* and the most recent reconstructed ancestral enzyme have different mechanisms of thermophily. This led Hobbs et al. (2012) to propose that thermophily may have evolved at least twice within the *Bacillus* genus, and that thermophily is not strictly a primitive trait. The fluctuating trend in thermophily suggests the general decrease in thermophily observed in previous studies (Perez-Jimenez et al., 2011, Gaucher et al., 2008, Gaucher et al., 2003) may be due to under-sampling and that fluctuations in thermophily may be observable if sampled over finer time scales. However, this suggestion is based on data from only one enzyme. More enzymes need to be reconstructed to strengthen this finding.

1.3 3-Deoxy-D-*arabino*-heptulosonate 7-Phosphate Synthase

A suitable candidate for ASR is the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase (DAH7PS; EC 2.5.1.54) protein family. DAH7PS is a suitable candidate for ASR because it is a core metabolic enzyme making it less likely that horizontal transfer of *dah7ps* has occurred (Didelot et al., 2010). In *Bacillus*, *dah7ps* occupies a region of the genome that displays little evidence of recombination (Didelot et al., 2010), decreasing the chance of spurious ASR (Arenas and Posada, 2010). DAH7PS is an interesting candidate for ASR because it is a large, structurally and mechanistically complex, multimeric enzyme. In *Bacillus* species, DAH7PS also has a chorismate mutase (CM) domain fused to the N-terminus. CM-DAH7PS would be the most complex Precambrian enzyme reconstructed thus far, potentially making reconstruction of active ancestral enzymes very difficult. Reconstructing Precambrian CM-DAH7PS enzymes would be a good test of the accuracy of ML ancestral sequence inference as the structural and mechanistic constraints are so great that any errors in the inference are likely to result in inactive enzymes. DAH7PS is a metal-dependent enzyme that catalyses the first committed step in the shikimate pathway. This pathway is responsible for the biosynthesis of chorismate, a precursor for many aromatic compounds such as the aromatic amino acids phenylalanine (Phe), tyrosine (Tyr), and tryptophan (Trp), as well as isoprenoid quinones, folates, and other secondary metabolites. The shikimate pathway has been identified in microorganisms, plants, and apicomplexan parasites (Campbell et al., 2004, Dewick, 1995, Roberts et al., 1998). The majority of bacterial species appear to contain complete shikimate pathways.

The reaction catalysed by DAH7PS is a stereospecific aldol-type condensation between phosphoenolpyruvate (PEP) and erythrose 4-phosphate (E4P) to form 3-deoxy-D-*arabino*-heptulsonate 7-phosphate (Figure 1.3). This reaction is highly stereospecific; the *si* face of C3 of PEP attacks the *re* face of the aldehyde of E4P resulting in the cleavage of the C-O bond and the release of phosphate, instead of the typical P-O bond cleavage (Deleo and Sprinson, 1968, Onderka and Floss, 1969). 3-Deoxy-D-*arabino*-heptulosonate 7-phosphate is then converted via six further enzyme catalysed reactions into the aromatic compound chorismate. The first step in the biosynthesis of aromatic amino acids from chorismate is either a

rearrangement of chorismate to prephenate catalysed by chorismate mutase (CM; EC 5.4.99.5) in the Phe/Tyr biosynthetic branch, or a conversion to anthralinate by anthralinate synthase in the Trp biosynthetic branch (Knaggs, 2001).

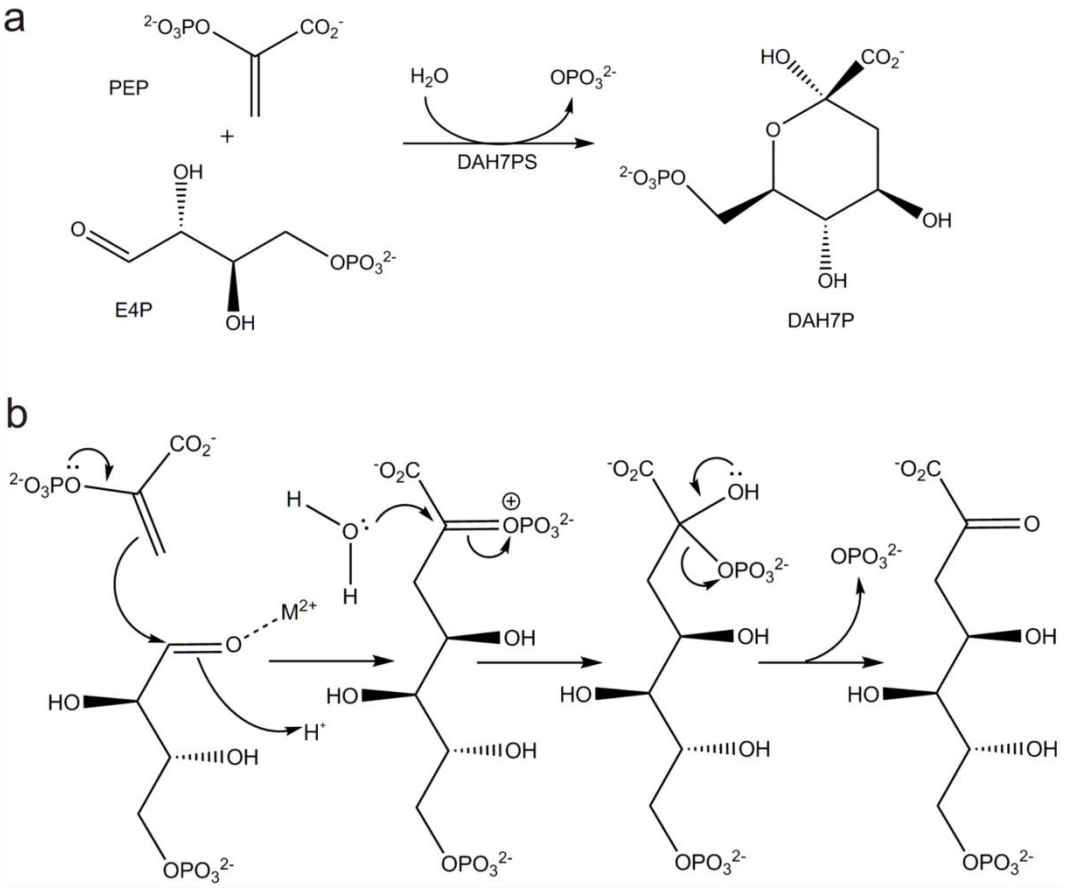

Figure 1.3 DAH7PS catalysis
(**a**) Overall reaction catalysed by DAH7PS. (**b**) Proposed acyclic reaction mechanism for the aldol-type condensation of phosphoenolpyruvate (PEP) and erythrose 4-phosphate (E4P) to form 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAH7P). Modified from Reichau et al. (2011).

In addition to DAH7PS enzymes, the DAH7PS protein family also contains the related enzyme 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase (KDO8PS). Phylogenetic analysis of over 100 homologous DAH7PS and KDO8PS amino acid sequences from microbial and plant sources divides the enzymes into two distinct subfamilies: type I and type II (Jensen et al., 2002). The type II subfamily consists of DAH7PS enzymes from either plant or microbial sources (Gosset et al., 2001). The type I subfamily can be further subdivided into two subfamilies: type Iα and type Iβ. Subfamily Iα consists solely of DAH7PS enzymes, whereas subfamily Iβ contains not only DAH7PS but also KDO8PS enzymes, meaning type Iβ enzymes can be further subdivided into subfamilies $I\beta_D$ (type Iβ DAH7PS

enzymes) and $I\beta_K$ (type Iβ KDO8PS enzymes). Representatives from all of the different subfamilies have been structurally and biochemically characterised. A summary of the DAH7PS enzymes is provided below in Table 1.1.

Table 1.1 Summary of some of the biochemically and structurally characterised DAH7PS enzymes.

Classification	Species	PDB Codes	References
Type II	*Mycobacterium tuberculosis*	2B7O	Webby et al. (2005a)
Type Iα	*Escherichia coli*	1QR7	Shumilin et al. (1999)
	Saccharomyces cerevisiae	1HFB	Hartmann et al. (2003)
	Salmonella typhimurium		DeLeo et al. (1973)
	Neurospora crassa		Hoffman et al. (1972)
Type $I\beta_D$	*Thermotoga maritima*	1RZM	Shumilin et al. (2004)
	Pyrococcus furiosus	1ZCO	Schofield et al. (2005)
	Aeropyrum pernix	1VS1	Zhou et al. (2012)
	Listeria monocytogenes	3TFC	Light et al. (2012)
	Bacillus subtilis		Huang et al. (1974)
	Porphyromonas gingivalis		Wu & Woodard (2006)

Type II DAH7PS enzymes (50 kDa) are typically larger than type I enzymes (typically less than 40 kDa), and share less than 10% sequence similarity with type I enzymes. Despite the very low sequence similarity between the subfamilies, a structural comparison between the type II DAH7PS from *Mycobacterium tuberculosis* and X-ray crystal structures for type I DAH7PS enzymes revealed a common catalytic scaffold and ancestry for the two enzyme subfamilies (Webby et al., 2005a). All structurally characterised DAH7PS enzymes are $(\beta/\alpha)_8$-barrel proteins (members of the TIM-barrel superfamily) with active sites located at the C-terminal end of the barrel, as is typical for $(\beta/\alpha)_8$-barrel enzymes (Figure 1.4). Many DAH7PS enzymes have decorations to their core catalytic $(\beta/\alpha)_8$-barrel, the nature of which can differ greatly (Webby et al., 2005a, Shumilin et al., 2004). These decorations have often been implicated in allosteric regulation of DAH7PS activity, with a surprising number of different regulatory mechanisms having evolved (Webby et al., 2010, Cross et al., 2011).

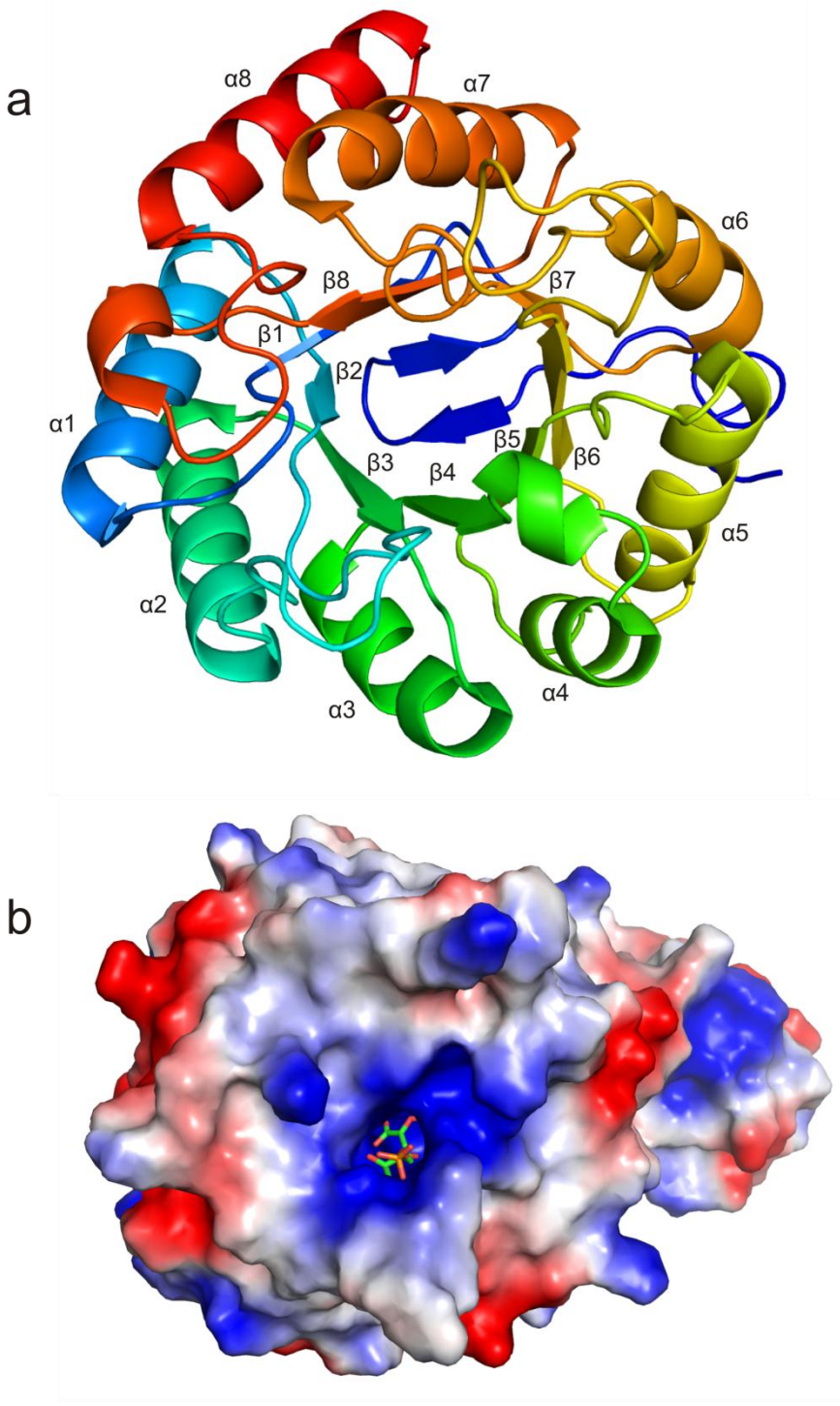

Figure 1.4 Cartoon representation and electrostatic potential surface model of DAH7PS monomer.
(**a**) Rainbow-coloured view of DAH7PS monomer from *Pyrococcus furiosus* (PDB code 1ZCO) looking in from the C-terminal end of the $(\beta/\alpha)_8$-barrel. Rainbow colouration starts with the N-terminal extension coloured dark blue and ends with α8 coloured red. (**b**) Electrostatic potential surface model of a DAH7PS monomer from *Thermotoga maritima* (PDB code 1RZM) with substrates E4P and PEP bound in the active site at the C-terminal end of the $(\beta/\alpha)_8$-barrel. Red, white and blue colouration represents surface areas with negative, neutral and positive electrostatic potential, respectively.

Allostery is an important mechanism for controlling the activity of metabolic enzymes. Allosteric control is exerted when the binding of a molecule to a particular site causes a functional response at a remote area of the protein (Jiao et al., 2012). It has been demonstrated by ^{13}C NMR that carbon flow into the shikimate pathway is controlled primarily through feedback inhibition of DAH7PS (Ogino et al., 1982). Apart from the DAH7PSs from *Pyrococcus furiosus* and *Aeropyrum pernix* all biochemically characterised DAH7PSs exhibit some form of allosteric regulation by downstream products (Cross et al., 2011, Jiao et al., 2012, Light et al., 2012). The different methods of allosteric regulation of DAH7PS enzyme activity are discussed below.

1.3.1.1 Synergistic Allosteric Regulation

The determination of the X-ray crystal structure of the DAH7PS from *M. tuberculosis* (*Mt*-DAH7PS; PDB code 2B7O; Webby et al., 2005) revealed the presence of two major additions to the core $(\beta/\alpha)_8$-barrel. The first addition is at the N-terminus of the protein and plays a role in dimerization and the closing of the N-terminal end of the barrel. The second addition is a pair of helices that result in an extended α2-β3 loop. *Mt*-DAH7PS has a dimer-of-dimers quaternary structure and displays unusual synergistic allostery (Webby et al., 2010). Inhibition studies revealed that *Mt*-DAH7PS activity is not inhibited by Phe, Tyr, or Trp separately but enzymatic activity is inhibited by combinations of Phe and Trp or Tyr and Trp (Webby et al., 2010). An X-ray crystal structure with both Phe and Trp bound revealed that Trp binds at the tetramer interface and Phe binds at the dimer interface causing the β2-α2 loop, which is involved in binding E4P, to become more flexible, resulting in a 10-fold decrease in activity (Webby et al., 2010).

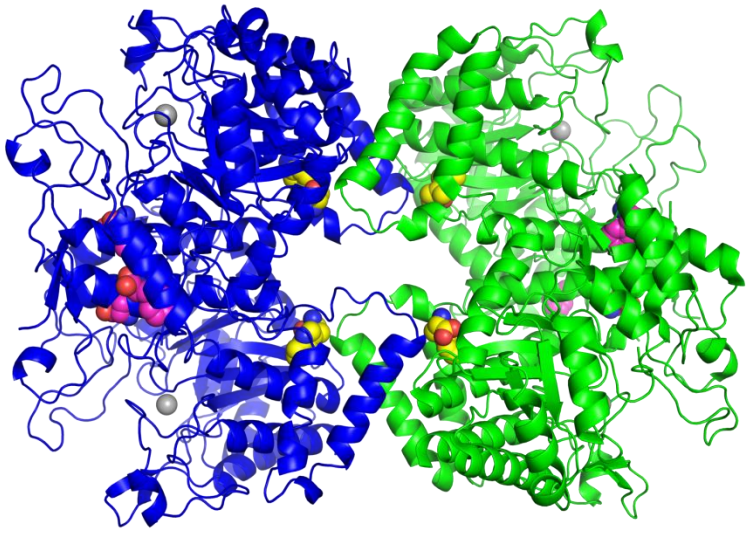

Figure 1.5 Cartoon representation of a type II DAH7PS tetramer from *Mycobacterium tuberculosis* (PDB code 3KGF).
The two DAH7PS dimers are coloured blue and green. The grey spheres represent Mn^{2+} ions bound at the active sites, bound phenylalanine is shown with yellow carbons, and bound tryptophan is shown by magenta carbons.

1.3.1.2 Allosteric Regulation of Multiple Isozymes

Type Iα DAH7PS enzymes have evolved a mode of allosteric regulation distinct from that employed by *Mt*-DAH7PS and type Iβ DAH7PS enzymes. Like *Mt*-DAH7PS, type Iα DAH7PSs are allosterically regulated by the aromatic amino acids Phe, Tyr, and Trp. However, a common mechanism for regulation of type Iα enzymes is the production of DAH7PS isozymes, each of which is sensitive to a different aromatic amino acid. *Escherichia coli*, *Neurospora crassa*, and *Salmonella typhimurium* express three type Iα DAH7PS isozymes, each sensitive to one of the aromatic amino acids (Tribe et al., 1976, Schoner and Herrmann, 1976, Hoffmann et al., 1972, Deleo et al., 1973), while *Saccharomyces cerevisiae* expresses one Phe-sensitive isozyme and one Tyr-sensitive isozyme (Schnappauf et al., 1998). X-ray crystal structures of the Phe-sensitive DAH7PS from *E. coli* and the Tyr-sensitive DAH7PS from *S. cerevisiae* reveals that both enzymes have extended N-termini that close off the N-terminal end of the $(\beta/\alpha)_8$-barrel. A β-sheet from the extended N-termini forms a three-stranded β-sheet with two additional β-sheets present in type Iα DAH7PSs between α5 and β6 to form a binding site for either Phe or Tyr which, when bound, inhibit DAH7PS activity.

1.3.1.3 Regulation of DAH7PS Activity by Fused Domains

Two further modes of allosteric regulation of DAH7PS activity by downstream products have been discovered in the type Iβ DAH7PS subfamily (Cross et al., 2011, Wu et al., 2003). These mechanisms involve large domains fused at the N- or C-termini of the DAH7PS core catalytic $(\beta/\alpha)_8$-barrel.

The $(\beta/\alpha)_8$-barrel of the DAH7PS from *Thermotoga maritima* (*Tm*-DAH7PS) is structurally very similar to the unregulated DAH7PS enzymes from *P. furiosus* (*Pf*-DAH7PS) and *A. pernix* (*Ap*-DAH7PS). In addition to the core $(\beta/\alpha)_8$-barrel, all three enzymes have two-stranded β-hairpin N-terminal extensions which seal off the N-terminal end of the $(\beta/\alpha)_8$-barrel. The major way that *Tm*-DAH7PS differ from these two enzymes is the presence of a 64-residue ferredoxin-like (FL) domain fused to the N-terminal two-stranded β-hairpin. Unlike the two archaeal enzymes, *Tm*-DAH7PS is feedback-regulated by the aromatic amino acids Phe and Tyr (Wu et al., 2003), with Tyr being the more potent inhibitor (Cross et al., 2011). Truncated *Tm*-DAH7PS without the FL domain has been shown not to be inhibited by these two aromatic amino acids, demonstrating that the FL domain is responsible for the observed feedback inhibition (Cross et al., 2011). X-ray crystal structures of *Tm*-DAH7PS with and without Tyr bound reveals the nature of this inhibition (PDB codes 3PG8 and 3PG9; Cross, et al., 2011). In the structure with no Tyr bound, the FL domains from different subunits do not associate with each other. The binding of Tyr to the FL domains results in a massive domain shift and FL dimer formation on opposite sides of the DAH7PS tetramer. The positions of the FL dimers caps the C-terminal ends of the $(\beta/\alpha)_8$-barrels, thereby inhibiting enzymatic activity by blocking substrate access to the active sites. FL domains have been identified in more than 20 microbial type Iβ DAH7PSs, suggesting this mode of allosteric regulation is reasonably widespread in the Iβ subfamily.

The last mode of regulation involves a ~80-residue CM domain fused at the N- or C-terminus of DAH7PS enzymes. This unusual fusion of two enzymes that catalyse non-sequential reactions was first identified in *B. subtilis* (Huang et al., 1974) and has subsequently been identified in a number of genera (Wu and Woodard, 2006, Light et al., 2012). The enzyme displays both CM and DAH7PS enzymatic activities, indicating that the enzyme is bifunctional (Huang et al.,

1974), however, unlike the fusion of enzymes catalysing sequential reactions which have an obvious biosynthetic benefit to the host organism, the benefit of expressing a bifunctional enzyme that catalyses non-sequential reactions is not clear. A DAH7PS enzyme from *Porphyromonas gingivalis* with a CM fused at the C-terminus of the DAH7PS (*Pg*-DAH7PS-CM) has also been characterised and shown to exhibit both DAH7PS and CM activity (Wu and Woodard, 2006). The CM enzymatic activity of the CM-DAH7PS from *B. subtilis* (*Bsub*-CM-DAH7PS) and *Pg*-DAH7PS-CM is approximately 50-fold lower than the CM activities of the bifunctional *E. coli* P-protein (CM-prephenate dehydratase) and T-protein (CM-prephenate dehydrogenase) which have evolved for true biosynthetic purposes as they catalyse sequential reactions in the Phe and Tyr biosynthetic pathways (Wu and Woodard, 2006). Wu et al. (2005) proposed that the fused CM domains provide a mechanism for feedback inhibition of DAH7PS enzymatic activity by downstream products. Wu & Woodard (2006) demonstrated that the DAH7PS catalytic activity of *Bsub*-CM-DAH7PS was inhibited by chorismate and the product of the CM catalysed reaction, prephenate, but insensitive to any of the aromatic amino acids. The DAH7PS catalytic activity of *Pg*-DAH7PS-CM was also found to be inhibited by chorismate and prephenate (Wu and Woodard, 2006). Prephenate is a far more potent inhibitor of the DAH7PS activity of both *Bsub*-CM-DAH7PS and *Pg*-DAH7PS-CM than chorismate (Wu and Woodard, 2006). Wu and Woodard (2006) separated the CM and DAH7PS domains of *Bsub*-CM-DAH7PS and *Pg*-DAH7PS-CM by domain truncation and observed that the truncated DAH7PS enzymes were not inhibited by chorismate or prephenate. This revealed that in both these enzymes the CM domains are responsible for the observed feedback inhibition. The presence of an additional, highly active CM (AroH) in organisms with fused CM-DAH7PS enzymes appears to have enabled the fused CM to evolve into a prephenate-binding domain for the sole purpose of feedback regulation of DAH7PS activity. The observed moderate sigmoidal inhibition of DAH7PS activity by chorismate *in vitro* is likely due to the binding of prephenate produced from chorismate by the moderate activity of the fused CM domain (Wu and Woodard, 2006). The low level of CM activity of these fused enzymes is likely just a relic of its evolutionary history. How binding of prephenate to the fused CM domains inhibits DAH7PS catalytic activity is not known. X-ray crystallographic data of DAH7PS enzymes with fused CM domains

at their N- or C-termini, with and without prephenate bound, should assist in determining the mechanism of feedback inhibition. Light et al. (2012) have attempted to do so and managed to solve the structure of CM-DAH7PS from *Listeria monocytogenes* (*Lm*-CM-DAH7PS) without any inhibitor bound (PDB codes 3NVT and 3TFC), but have thus far been unsuccessful in their attempts to solve the structure of this enzyme with prephenate or chorismate bound.

1.3.2 *Bacillus* DAH7PS

All DAH7PSs from *Bacillus* appear to have CM domains fused to their N-termini and be metal-dependent as they all contain the four absolutely conserved metal-chelating residues (Shumilin et al., 2004). *Bsub*-CM-DAH7PS is the only biochemically characterised enzyme from the *Bacillus* genus. *Bsub*-CM-DAH7PS was initially suggested to be a metal-independent enzyme, as it was reported that *Bsub*-CM-DAH7PS is insensitive to EDTA treatment (Jensen and Nester, 1966). However, it was recently demonstrated by using a stronger metal chelator, dipicolinic acid (DPA), that *Bsub*-CM-DAH7PS is indeed a metal-dependent enzyme like all other biochemically characterised DAH7PS enzymes, but the metals are simply bound too tightly in the active site of the enzyme to be removed by EDTA treatment (Wu et al., 2005). Recombinant *Bsub*-CM-DAH7PS expressed in *E. coli* was found to contain approximately 0.15 M of iron and 0.3 M of zinc per subunit (Wu et al., 2005). When these metals were chelated away and replaced with a range of different metal ions, only enzymes with Zn^{2+} and Cd^{2+} bound showed significant levels of activity (Wu et al., 2005). *Bsub*-CM-DAH7PS displays significantly higher K_M values for PEP and an especially high K_M for E4P compared to other bacterial DAH7PS enzymes (Wu et al., 2003, Nimmo and Coggins, 1981, Schoner and Herrmann, 1976). The molecular weight of *Bsub*-CM-DAH7PS as determined by analytical gel filtration chromatography is 156 kDa, approximately four times greater than the molecular weight of the monomer (40 kDa) as determined by SDS-PAGE (Wu et al., 2005). Therefore, *Bsub*-CM-DAH7PS likely exists as a homotetramer in solution. This is consistent with the two structurally characterised bacterial type Iβ DAH7PSs from *T. maritima* and *L. monocytogenes* which are homotetrameric (Shumilin et al., 2004, Light et al., 2012).

Of all the structurally characterised DAH7PS enzymes, *Lm*-CM-DAH7PS shares the highest sequence identity with *Bsub*-CM-DAH7PS. The DAH7PS domains from *Lm*-CM-DAH7PS form a core catalytic tetrameric structure, and N-terminal CM domains from diagonally opposite DAH7PS domains emerge on the same side of the DAH7PS catalytic core and associate to form symmetrical six-helix dimers (Figure 1.6). The CM dimer is structurally very similar to the fused CM domain from the *E. coli* P-protein (PDB code 1ECM), indicating that very little change in overall structure is required to convert a catalytic CM domain to a prephenate-binding domain. The CM dimers are not positioned symmetrically in relation to the core tetrameric $(\beta/\alpha)_8$-barrels. One of the CM domains from each of the dimers forms a small number of interactions with the catalytic DAH7PS domain to which they are fused, while the other members of the regulatory dimers do not interact with the catalytic tetramer. It is not known whether the asymmetric position of the CM dimers is biologically relevant due to there being so few interactions between the CM domains and the DAH7PS domains. The observed orientation of the CM domains may represent one of many low occupancy conformations, with this particular orientation being selected for by crystal packing forces. Currently, how binding of prephenate to CM causes inhibition of DAH7PS activity is unknown, although a similar allosteric regulatory mechanism to that of *Tm*-DAH7PS has been proposed (Light et al., 2012).

Figure 1.6 Cartoon representation of a CM-DAH7PS tetramer from *Listeria monocytogenes.*
CM domains from diagonally related DAH7PS monomers associate to form symmetrical six-helix dimers which are asymmetrically positioned in relation to the DAH7PS tetrameric core (PDB code 3TFC). Subunits are coloured to clearly illustrate which subunits interact to form the CM dimers. Most of the loops connecting the CM α-helices are absent and the CM-DAH7PS linker domains of two of the subunits are also absent.

1.4 Research Objectives

The main aim of this research was to investigate the origin and evolution of thermophily. This will be achieved by reconstructing ancestral *Bacillus* CM-DAH7PS enzymes, measuring their thermal and kinetic properties. The kinetic properties of the ancestral CM-DAH7PS enzymes will be compared to those of CM-DAH7PS from contemporary *Bacillus* species to ensure that the ancestral enzymes are biologically realistic. The thermal properties of the reconstructed CM-DAH7PS enzymes will be used to try to infer how thermophily may have evolved within the *Bacillus* genus. The results from this study will be compared with those from the ASR study of Hobbs et al. (2012) to determine

whether a similar fluctuating trend in thermophily is observed within the *Bacillus* genus when a different enzyme family is reconstructed.

A further aim of this research was to determine the 3D X-ray crystal structures of the ancestral CM-DAH7PS enzymes to assess the accuracy of the ancestral inferences and rationalise any changes in the thermal properties of the enzymes.

2 Materials and Methods

2.1 Phylogenetics and Ancestral Sequence Inference

2.1.1 Sequence Alignment

The *Bacillus stearothermophilus cm-dah7ps* nucleotide sequence was obtained from the *Bacillus* (*Geobacillus*) *stearothermophilus* Genome Sequencing Project (http://www.genome.ou.edu/bstearo.html). All other sequences were obtained from GenBank. Gene accession numbers, and species and strain details, are provided in Appendix A.

The amino acid sequences and nucleotide sequences were aligned using Clustal W2 (Larkin et al., 2007). Sequence alignment files were exported in PHYLIP format.

2.1.2 Phylogenetic Tree Construction

2.1.2.1 Determination of the Most Appropriate Models of Evolution

The most appropriate model of amino acid sequence evolution for *Bacillus* CM-DAH7PS amino acid sequences was determined using ProtTest version 2.4 (Abascal et al., 2005). The most appropriate model of nucleotide sequence evolution for *Bacillus cm-dah7ps* sequences was determined using jModelTest version 0.1.1 (Posada, 2008).

2.1.2.2 Chronogram Construction

The most appropriate model of amino acid evolution, and a multiple sequence alignment of all available *Bacillus* CM-DAH7PS amino acid sequences and two *Clostridium* DAH7PS amino acid sequences were implemented in GARLI version 1.0 (Zwickl, 2006) to generate ten ML phylogenetic trees. These trees were re-rooted in Geneious version 5.1.7 (Drummond et al., 2011) using the two *Clostridium* sequences as the outgroup. The best ML phylogenetic tree was chosen from the ten trees based on their log likelihood values, the consistency in the organisation of *Bacillus* species between the 10 ML trees, and how similar the trees were to a robust ML phylogenetic tree of 20 *Bacillus* species constructed

using the core genomes of these species (Alcaraz et al., 2010) and a ML phylogenetic tree constructed using IPMDH sequences (Hobbs et al., 2012).

The sequence alignment of *Bacillus* and two *Clostridium* DAH7PS amino acid sequences generated as in section 2.1.1, the best ML phylogenetic tree, and the most appropriate model of amino acid sequence evolution as determined in section 2.1.2.1, were implemented in GARLI version 1.0 (Zwickl, 2006) to generate 1,024 pseudoreplicate datasets. These datasets were opened in Geneious version 5.1.7 (Drummond et al., 2011) and a consensus tree was generated with a 20% support threshold and 0% burn-in.

r8s version 1.71 (Sanderson, 2003) was used to convert the ML phylogram to a chronogram. The data input was the best ML phylogenetic tree and two fixed calibration points: the point of *Bacillus* and *Clostridium* divergence (2.65 Gya) and the point of *Bacillus subtilis* and *Bacillus halodurans* divergence (950 Mya) as determined by Battistuzzi et al. (2004). FigTree version 1.3.1 (Rambaut, 2009) was used to visualise the chronogram.

2.1.3 Ancestral Sequence Reconstruction

2.1.3.1 Maximum Likelihood Sequence Inference

The extensions of the alignment files of *Bacillus* and *Clostridium* DAH7PS nucleotide and amino acid sequences were converted from .phy to .nuc and .aa, respectively. The best ML phylogenetic tree was exported from Geneious version 5.1.7 (Drummond et al., 2011) in Nexus format, manually reformatted into a Newick file, and given the extension .tre.

Three different methods of ancestral sequence inference were performed using PAML version 4.3 (Yang, 2007): nucleotide, codon, and amino acid inference. Nucleotide inference was performed in the programme BASEML using the most appropriate model of nucleotide evolution as determined by jModelTest version 0.1.1 (Posada, 2008) and the nucleotide alignment. Amino acid and codon inferences were performed in the programme CODEML using either the amino acid or nucleotide alignment and the Jones amino acid evolution model (Jones et al., 1992).

PAML version 4.3 (Yang, 2007) assigns node labels to every node within the phylogenetic tree, including the terminal nodes. To determine which label corresponds to which node, the node labelling data from the amino acid inference rst output file was extracted and transferred to a new file with a .tre file extension. This file was then opened in TreeView version 1.6.6 (Page, 1996) in order to display the labelling of the nodes within the phylogenetic tree. The sequences from the appropriate internal nodes could then be extracted from the three different inference rst output files. Ancestral nucleotide sequences were translated into amino acid sequences using Geneious version 5.1.7 (Drummond et al., 2011). The three inferred ancestral amino acid sequences were then aligned using Clustal W2 (Larkin et al., 2007). A consensus sequence was compiled and any ambiguous sites were resolved using the following criteria: if two of the inference methods predicted the same amino acid at a particular position this amino acid was chosen; if one of the amino acids at a particular position in the inferred sequences was present in the sequences of a number of extant species that grow optimally at different temperatures, this amino acid was generally chosen; at ambiguous sites the physicochemical properties of the different inferred ancestral amino acids were examined using the JTT model of amino acid classification (Taylor and Jones, 1993) and then compared to the physicochemical properties of the amino acids in contemporary *Bacillus* species at the sites of ambiguity, with the inferred ancestral amino acid with the most similar physicochemical properties to those from the contemporary species being chosen.

2.2 Cloning

2.2.1 Gene Synthesis

Genes encoding the ancestral, *B. stearothermophilus*, and *Bacillus selenitireducens* CM-DAH7PS enzymes were codon optimised for expression in *E. coli* and synthesised with *Nco*I and *Xho*I restriction sites incorporated at the 5' and 3' ends of the genes, respectively, by GENEART (Regensburg, Germany). These genes were supplied as inserts in pMA-T vectors.

2.2.2 Growth of Contemporary *Bacillus* for Genomic DNA Extraction

Bacillus subtilis Marburg 168 and *Bacillus caldovelox* (specimen #65 from Thermophile Research Unit culture collection, University of Waikato) were streaked onto nutrient agar plates and incubated overnight at 30 °C and 55 °C, respectively. A single colony of *B. subtilis* Marburg was used to inoculate 5 mL of nutrient broth in a 50 mL Falcon tube and incubated overnight at 30 °C with shaking. The same procedure was performed with a single colony of *B. caldovelox*, except incubation was performed at 70 °C for ~24 h.

2.2.3 Preparation of Glycerol Stocks

For long-term storage of bacterial strains, glycerol stocks were prepared by mixing 500 µL of overnight cultures with 500 µL of sterilised 80% glycerol in 2 mL sterile screw tubes. Cultures were frozen and stored at -80 °C.

2.2.4 Genomic DNA Extraction

B. subtilis and *B. caldovelox* cells were pelleted by centrifuging the 5 mL cultures at 4,600 rpm for 20 min. Supernatants were discarded and cells resuspended in 1 mL of nutrient broth. Aliquots (0.5 mL) from these resuspended cultures were transferred to 2 mL sterile screw tubes containing a small volume of sterile glass beads. A volume (50 µL) of 5 M guanidinium thiocyanate (GITC) was added to the cells. The cells were then shaken in a FastPrep® FP120 Cell Disrupter (Thermo Savant, USA) for three 30 s bursts at setting six, allowing at least 30 s between each burst of shaking to allow the samples to cool. To the lysed cells 50 µL each of 2 M sodium acetate pH 4.0 and a 1:1 phenol:chloroform mixture were added and then mixed on a rotatory wheel for 10 min. Lysed cells were then centrifuged at 13,000 rpm for 1 min. The top layers containing the genomic DNA were transferred to sterile 1.5 mL microcentrifuge tubes, mixed with equal volumes of isopropanol and left at room temperature for 20 min. Genomic DNA was pelleted by centrifugation at 13,000 rpm for 20 min. The supernatants were discarded and 1 mL of 70% ethanol was added to the genomic DNA. Solutions were centrifuged at 13,000 rpm for 1 min and all possible traces of ethanol removed by pipetting. Genomic DNA was left to air dry for several minutes to

remove all traces of ethanol, then resuspended in 50 µL of TE buffer and 1 µL of 1 mg/mL RNase A.

2.2.5 DNA Quantification

DNA concentration was determined by measuring the DNA absorbance at 260 nm using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, USA).

2.2.6 Primer Design

PCR primers were manually designed to amplify *cm-dah7ps* from *B. subtilis* Marburg and *B. caldovelox* genomic DNA, and to incorporate *Nco*I and *Xho*I cut sites at the 5' and 3' ends of the genes, respectively. *B. caldovelox* PCR primers were designed based on the genome sequence of the closely related species *B. stearothermophilus*. Oligo Analyzer version 1.5 (Gene Link™, USA) was used to check primers for evidence of secondary structure and potential formation of primer dimers. All primers were ordered from Sigma-Aldrich (USA) and resuspended in MQ water to a final concentration of 100 pmol/µL. Primer sequences are provided in Table 2.1.

Table 2.1 Primer sequences.

Name	**Sequence (5'-3')**
BSUBfwd	TGAAAACCATGGGCAACAC
BSUBrev	CAACTCGAGGTTGACTTTCACC
BCVXfwd	TGAGAAAACCATGGGCAATG
BCVXrev	TCTCTTACTCGAGAACGAACTGAC
T7fwd	TAATACGACTCACTATAGGG
T7rev	TAGTTATTGCTCAGCGGTGG

BSUBfwd, BSUBrev, BCVXfwd, and BCVXrev primers were manually designed as part of the present study. Underlined nucleotides indicate the position of *Nco*I and *Xho*I cut sites. The T7fwd and T7rev primers were ordered from Invitrogen, USA.

2.2.7 PCR Amplification from Genomic DNA

The optimal PCR reaction mixture for amplifying *cm-dah7ps* from *B.subtilis* and *B. caldovelox* genomic DNA was determined to consist of 5 µL 10 x PCR buffer minus Mg, 3 µL 50 mM $MgCl_2$, 1 µL 10 mM dNTP mixture, 1.5 µL forward primer, 1.5 µL reverse primer, 1 µL 100 ng genomic DNA, 0.25 µL *Taq* polymerase (Invitrogen, USA), and 36.75 µL RO water.

PCR reaction conditions used were:

95 °C	1 min	
95 °C	30 s	
45 °C*	90 s	x 29
72 °C	80 s	
72 °C	10 min	

* Annealing temperature determined using a temperature gradient around the calculated T_m of the primers.

PCR reactions were analysed by running samples on a 1% (w/v) agarose gel as described in section 2.2.8.

2.2.8 Agarose Gel Electrophoresis

Agarose gels were made up in 1 x TAE buffer containing 2-5 µL SYBR® safe DNA gel stain (Invitrogen, USA) to a concentration of 1% (w/v). DNA samples were mixed with 10 x loading dye (Invitrogen, USA) prior to loading the samples onto the gel. Fragments of DNA were separated by electrophoresis at 100 V for 45 min. DNA was visualised using a blue light box and the size of the fragments determined by comparison with the 1 kb plus DNA ladder (Invitrogen, USA).

2.2.9 Plasmid Extraction

Single colonies or a loopful from partially-defrosted glycerol stocks containing a desired plasmid were used to inoculate 5 mL of LB in 50 mL Falcon tubes supplemented with 50 µg/mL of kanamycin and incubated overnight at 37 °C with shaking. Plasmid DNA was then extracted from cells using the QIAprep Spin Miniprep Kit (Qiagen, Netherlands) according to manufacturer's instructions.

2.2.10 Restriction Enzyme/Ligase Cloning

All reagents and enzymes used for restriction enzyme digestion and ligation reactions were purchased from Invitrogen (USA).

2.2.10.1 Restriction Enzyme Digestion

*Nco*I and *Xho*I restriction enzymes were used to digest PCR amplified fragments, pET28b (Invitrogen, USA) and the pMA-T-CM-DAH7PS plasmids supplied by GENEART (Regensburg, Germany). Restriction enzyme digestions were performed according to manufacturer's instructions and left to digest at room temperature for 3 h.

2.2.10.2 Gel Excision, DNA Extraction, and PCR Clean-up

Digested pET28b and pMA-T-CM-DAH7PS plasmids were separated by gel electrophoresis (section 2.2.8). DNA bands of the desired size were excised from agarose gels by cutting the DNA band out with a clean scalpel blade and placing the excised bands into 1.5 mL microcentrifuge tubes.

DNA was then extracted from the excised gel fragments using the QIAquick Gel Extraction Kit (Qiagen, Netherlands) according to the manufacturer's instructions. Digested PCR fragments were also cleaned-up using the QIAquick Gel Extraction Kit (Qiagen, Netherlands).

2.2.10.3 Dephosphorylation of Digested pET28b

Digested pET28b was dephosphorylated to prevent the two ends of the plasmid from ligating together. A small volume (1 µL) of shrimp alkaline phosphatase (SAP) enzyme was added to the digested vector along with 10 x SAP buffer which was added to the mixture at a 1:10 ratio. The solution was left to incubate at 37 °C for 30 min. The SAP enzyme was then inactivated by heating at 65 °C for 15 min.

2.2.10.4 *DNA Ligation*

DNA ligations were performed with dephosphorylated pET28b and purified digested PCR fragments or *cm-dah7ps* genes excised and purified from pMA-T

vectors. Ligation reactions without any insert added were also set up to act as negative controls. DNA ligations were performed using T4 DNA ligase and set up as per manufacturer's instructions. Ligation reaction mixtures were incubated overnight at 16 °C.

2.2.11 Transformation

2.2.11.1 Preparation of Electrocompetent Cells

LB (10 mL) in a 50 mL Falcon tube was inoculated with a single colony of *E. coli* DH5α or BL21 cells and incubated overnight at 37 ˚C with shaking. The 10 mL overnight starter culture was then used to inoculate 1 L of LB in a 2 L baffled flask. Cells were incubated at 37 ˚C with shaking at 200 rpm until an OD_{600} between 0.5 and 0.7 was reached. OD_{600} measurements were made periodically using a Heλios™ spectrophotometer. Once an appropriate optical density was reached, the flask was chilled on ice for 30 min then transferred to sterile centrifuge bottles and centrifuged at 4,600 rpm for 25 min at 4 °C. The supernatant was removed, the cell pellet resuspended in a total of 1 L ice-cold sterile 10% glycerol, and then centrifuged as before. The cells were pelleted and resuspended in this manner in sequentially decreasing volumes of ice-cold sterile 10% glycerol (0.5 L, 20 mL, and 2.5 mL). Small aliquots (50 µL) of the final resuspension were transferred to 1.5 mL microcentrifuge tubes and immediately frozen and stored at -80 °C.

2.2.11.2 Electroporation

A Bio-Rad Gene Pulser™ (Bio-Rad Laboratories, USA) with Pulse Controller was used to transform electrocompetent *E. coli* DH5α and BL21 cells. Initial transformations were performed with *E. coli* DH5α cells. Ligation reaction mixtures were mixed with 30 µl of sterile 10% glycerol and added to 50 µL aliquots of electrocompetent *E. coli* DH5α cells thawed on ice. The mixtures were then transferred to chilled 0.2 cm electroporation cuvettes (Bio-Rad Laboratories, USA) and electroporated with 2.5 kV at 25 µF capacitance and 200 Ω resistance. Immediately after electroporation of a sample, 925 µL of LB was added to the

cuvette. Transformed cells were then transferred to 1.5 mL microcentrifuge tubes and left for 90 min to allow for cell recovery.

2.2.11.3 Selection of Transformants by Agar Plating

Selection of transformants was performed by first spreading 200 µL of the cells onto LB agar plates supplemented with 50 µg/mL kanamycin. The remaining cells were pelleted by centrifugation at 13,000 rpm for 3 min. Supernatants were discarded, and pellets resuspended in 20 µL of LB and streaked onto LB agar plates supplemented with 50 µg/mL of kanamycin. The plates were incubated overnight at 37 °C. The number of colonies were counted and compared with the negative control plates.

2.2.12 Gene Insert Screening

Three different procedures were used to screen for *E. coli* DH5α (pET28b-CM-DAH7PS) transformants.

Colony PCR was used to screen large numbers of transformants by first suspending individual colonies in 10 µL aliquots of sterile RO water. The standard PCR mixture was made as in section 2.2.7 to a total volume of 15 µL. The DNA template in this case was 1 µL of the colony resuspensions. T7 primers were used instead of gene specific primers.

Colony PCR reaction conditions:

94 °C	3 min	
94 °C	20 s	
52.5 °C	30 s	x 29
72 °C	45 s	
72 °C	5 min	

To screen small numbers of transformants, restriction enzyme digestions (section 2.2.10.1) and standard PCR reactions (section 2.2.7), using either T7 or gene specific primers, were performed. In order to generate plasmid DNA for screening, colonies were used to inoculate 5 mL of LB containing 50 µg/mL kanamycin and grown overnight at 37 °C with shaking. Plasmid DNA was

extracted and purified as described in section 2.2.9. Purified plasmid DNA was then PCR amplified and/or digested by restriction enzymes.

PCR reaction mixtures from the two PCR methods and the restriction digest mixtures were loaded onto a 1% (w/v) agarose gel and electrophoresed to separate DNA fragments (section 2.2.8). *E. coli* DH5α (pET28b-CM-DAH7PS) transformants were identified by the presence of a DNA band of ~1,100 or 1300 bp (if T7 primers were used for PCR amplification) in size in the agarose gels. Glycerol stocks of all identified transformants were prepared and all transformants of pET28b-*Bsub-cm-dah7ps* and pET28b-*Bcvx-cm-dah7ps* were sequenced (section 2.2.12.1).

2.2.12.1 DNA Sequencing

Typically, ~20 µL of purified plasmid DNA, together with T7 forward and reverse primers diluted to 5 pmol/µL, were sent for sequencing at the Waikato DNA Sequencing Facility (University of Waikato). DNA sequencing data were compared with gene sequences from Genbank, if available, to determine whether an insert with the correct sequence had been successfully ligated into pET28b. Sequences were also examined to confirm that the C-terminal His-tag was in frame. Electropherograms were examined manually to check for any incorrect base calls.

2.2.13 Transformation of Electrocompetent *E. coli* BL21 Cells

Once *E. coli* DH5α (pET28b-CM-DAH7PS) transformants had been constructed, the plasmids were re-transformed into *E. coli* BL21 cells. Small volumes (1 µL) of purified pET28b-CM-DAH7PS plasmids were gently mixed with 50 µL of electrocompetent *E. coli* BL21 cells. Aliquots (30 µL) of the mixtures were transferred to chilled 0.2 cm electroporation cuvettes (Bio-Rad Laboratories, USA) and electroporated (section 2.2.11.2). Immediately following electroporation, 970 µL of LB was added to the cells. The cultures were then transferred to 1.5 mL microcentrifuge tubes and incubated at 37 °C for 90 min. Small volumes (10 µL) from the cultures were spread on LB agar plates containing 50 µg/mL kanamycin. The plates were incubated at 37 °C overnight. Overnight cultures were then set up by inoculating 5 mL of LB containing

50 µg/mL of kanamycin with single colonies. Glycerol stocks of the *E. coli* BL21 transformants were prepared (section 2.2.3).

2.2.13.1 Creating Double-Transformants

Double-transformant *E. coli* BL21 cells containing both pET28b-CM-DAH7PS plasmids and pGroESL (provided by Emily Parker) were prepared in an almost identical fashion as the single BL21 transformants. Double transformants were created by mixing 1 µL pGroESL plasmid and 1 µL of pET28b-CM-DAH7PS with electrocompetent BL21 cells, followed by the standard electroporation procedure (section 2.2.11.2). The selective growth media contained 34 µg/mL chloramphenicol, in addition to 50 µg/mL kanamycin, to select for both plasmids. Double transformants were generated with pET28b plasmids with ancestral *cm-dah7ps* inserts and the *B. selenitireducens cm-dah7ps* insert.

2.3 Protein Expression

2.3.1 Large Scale Expression Cultures

Two main expression systems were used to overexpress CM-DAH7PS enzymes: an IPTG induction system and an auto-induction system coupled with co-expression of the molecular chaperone complex, GroESL.

2.3.1.1 Overnight Starter Cultures

Starter cultures were prepared by adding 1 µL from a glycerol stock of a desired transformant to a 50 mL Falcon tube containing 10 mL of LB media or PA-0.5G auto-induction media supplemented with the appropriate antibiotics. Cultures were incubated overnight at 37 °C with shaking at 200 rpm.

2.3.1.2 IPTG Induced Protein Overexpression

For IPTG induced protein overexpression from single *E. coli* BL21 transformants, 10 mL of the LB starter culture was added to 1 L of terrific broth (TB) supplemented with 50 µg/mL kanamycin in a 2 L baffled conical flask. If high concentrations of protein were desired, 20 mL of LB starter culture was added to 2 L of TB media in a 5 L conical baffled flask. Expression cultures were incubated at 37 °C with shaking at 200 rpm until an OD_{600} of ~0.7 was reached.

Once an OD_{600} of ~0.7 was reached the cultures were cooled to 25 °C with shaking at 200 rpm. IPTG was added to a final concentration of 1 mM and incubation at 25 °C with shaking at 200 rpm was continued for 22 h. Cultures were then transferred to centrifuge bottles and spun at 4,600 rpm for 20 min at 4 °C to pellet cells. The supernatants were discarded and cell pellets transferred to 50 mL Falcon tubes for storage at -80 °C.

2.3.1.3 Auto-induction Coupled with GroESL Co-expression

Auto-induced protein overexpression of CM-DAH7PS, GroES, and GroEL was achieved by addition of 10 mL PA-0.5G auto-induction starter cultures to 1 L of ZYP-5052 rich medium containing 50 µg/mL kanamycin and 25 µg/mL chloramphenicol in a 2 L conical baffled flask. Expression cultures were incubated for 24 h at 25 °C with shaking at 180 rpm. Cultures were centrifuged and stored in the same manner as the IPTG expression cultures.

2.3.2 Protein Purification

2.3.2.1 Cell Lysis

Frozen cell pellets were defrosted at room temperature and resuspended in 20 mL lysis buffer (50 mM Bis-Tris propane (BTP), 500 mM NaCl, 20 mM imidazole, pH 9) containing a cOmplete Mini, EDTA-free protease inhibitor cocktail tablet (Roche, Germany). Cells were lysed on ice by sonication for six 15 s bursts with at least 1 min between each burst to ensure solutions did not overheat. Cell debris was pelleted by centrifugation at 13,000 rpm for 20 min at 4 °C.

2.3.2.2 IMAC Purification

CM-DAH7PS enzymes were purified initially by immobilised metal affinity chromatography (IMAC) using a 5 mL HisTrap™ FF column (GE Healthscience, Sweden). Following cell lysis and centrifugation, the supernatants were filtered through successive 1.2 µm, 0.45 µm, and 0.2 µm filters and then loaded onto a column pre-equilibrated with 20 mL lysis buffer. The column was connected to an ÄKTA™ Basic, Prime, or Purifier system (GE Healthcare, Sweden). Unbound protein was removed by flowing lysis buffer through the column until a base-level absorbance reading at 280 nm was reached. A gradient of 0-50% elution buffer

(50 mM BTP, 500 mM NaCl, 1 M imidazole, pH 9) at a rate of 1 mL/min over 60 mL was used to elute protein bound to the column. Protein elution was monitored by measuring the absorbance of the eluent at 280 nm and the eluent was collected as 2 mL fractions. Fractions were analysed by SDS-PAGE (section 2.7.2) to determine the size and purity of any protein within the fractions. Following purification, the column was stripped of Ni^{2+} ions with 10 mL 100 mM EDTA pH 7.5, then recharged with 100 mM $NiCl_2$.

2.3.2.3 Concentrating Protein

Proteins were concentrated by centrifugation at 3,700 rpm at 4 °C in 0.5 mL, 2 mL, or 20 mL Vivaspin concentrators (Sartorius AG, Germany) with a 10 kDa cut off until a desired concentration or volume was reached.

2.3.2.4 Size Exclusion Chromatography

Fractions from IMAC purification identified as containing CM-DAH7PS enzyme by SDS-PAGE analysis were pooled and concentrated to ~3 mL (section 2.6.2.3). A Superdex™ 200 16/60 GL column (GE Healthcare Life Science, UK) was connected to an ÄKTA Prime, Basic, or Purifier system and equilibrated with running buffer (50 mM BTP, 500 mM NaCl, pH 9). Concentrated protein was passed through a 0.2 µm filter and injected onto the column. Running buffer was flowed through the column at 0.5 mL/min for one column volume with 2 mL fractions collected from 35 mL onwards. Protein elution was monitored by measuring absorbance at 280 nm. Samples from all fractions identified as containing protein from observed increases in the absorbance at 280 nm during elution were analysed by SDS-PAGE (section 2.4.2) in order to detect the presence of CM-DAH7PS and assess its purity. The volume at which protein eluted was compared with a calibration curve to estimate the size of the proteins eluted at different stages in the purification. Fractions predicted to contain CM-DAH7PS homotetramers were pooled and concentrated to desired protein concentrations (section 2.3.2.3).

2.3.2.5 Metal Chelation and Dialysis

In order to chelate any metal ions bound to the CM-DAH7PS enzymes, 10 mM DPA was first dissolved in size exclusion running buffer (50 mM BTP, 500 mM NaCl, pH 9) by heating at 50 °C for 3-4 hours with periodic shaking. Highly concentrated protein was added to the dissolved DPA and left to chelate at room temperature for 2 ½ h with gentle shaking. The reaction mixture was added to pre wet Spectra Por® dialysis membrane (Spectrum Laboratories, USA) with 6-8 kDa cut off and the two ends sealed off with dialysis clips. The dialysis membrane was added to 1 L of size exclusion running buffer and left to dialyse for 1½ h with gentle stirring. The buffer was then replaced with fresh buffer and dialysis was continued for 1½ h. The protein solution inside the dialysis membrane was transferred to a 20 mL Vivaspin concentrator (Sartorius AG, Germany) with 10 kDa cut off and 100 µM $ZnSO_4$ was added to the solution and left for 15 min at 4 °C. The protein was then concentrated (section 2.3.2.3) to between 5-15 mg/mL.

2.4 Protein Characterisation

2.4.1 Bradford Assay

Due to the lack of any tryptophan residues within any of the CM-DAH7PS enzymes, the Bradford assay was used to measure protein concentration. Reactions were carried out in a 96 well microplate using the Bio-Rad protein assay dye (Bio-Rad Laboratories, USA). Reaction mixtures consisted of 40 µL of Bio-Rad protein assay dye, 5 µL of protein and 155 µL MQ water. Protein solutions were diluted with purification buffer so absorbance readings were within the measurable range of the assay. Reactions were mixed and left at room temperature for 5-10 min prior to measurement of the absorbance at 595 nm using a Fluostar Optima microplate reader (BMG Labtech, Germany). The absorbance measurements of the sample proteins were compared against a standard curve generated from a range of bovine serum albumin concentrations. All reactions were performed in triplicate.

2.4.2 SDS-PAGE

Protein samples at appropriate concentrations were added to 4 x SDS loading buffer (Appendix A) at a ratio of 3:1. Samples were incubated for 5 min at 95 °C before running samples on 12% SDS-PAGE gels. The recipe for making SDS-PAGE gels is provided in Appendix A. Small aliquots (15 µL) from each sample were loaded into different wells and 10 µL Precision Plus Protein™ Unstained Ladder (Bio-Rad Laboratories, USA) was added to an outside well. Gels were electrophoresed at 15 mA until the samples had run through the stacker, then at 30 mA until the dye front reached the bottom of the gel.

Gels were transferred to a microwavable box, covered with coomassie stain, heated in a microwave for 30 s and left to cool at room temperature for 5 min with gentle shaking. The coomassie stain was decanted away and the gel was covered with destain solution (10% (v/v) acetic acid). The gel was heated in a microwave for 30 s and then left to shake gently at room temperature for at least 30 min. Destaining was performed at least twice or the gel was left overnight in destain solution to fully destain. Protein sizes were estimated by comparison with the ladder.

2.4.3 Urea Unfolding Assays

Urea unfolding assays were performed to determine whether a change was observed in the intrinsic fluorescence of CM-DAH7PS enzymes when they were unfolded. A control solution was prepared by mixing 20 µL of concentrated protein with 180 µL of assay buffer (50 mM BTP, 1 mM $ZnSO_4$, pH 9). The unfolding solution was prepared by mixing 20 µL of concentrated protein with 180 µL of 8 M urea prepared in assay buffer; this solution was then left overnight at 4 °C to unfold. 2D- (with excitation at 280 nm) and 3D-wavelength scans were performed in a Hitachi F-7000 fluorescence spectrophotometer (Hitachi High Technologies Corp., Japan). The 2D- and 3D wavelength scans of the folded and unfolded samples were compared to determine whether there is any observable difference in the intrinsic fluorescence of folded and unfolded CM-DAH7PS enzymes.

2.4.4 Real-Time Protein Melts

A modified version of the real-time protein melt assay developed by Ericsson et al. (2006) was performed using a RotorGene-6000 Real-time PCR machine (Corbett Research, Australia) in order to estimate the T_m values for contemporary and ancestral CM-DAH7PS enzymes. Six different pH buffers were used to prepare real-time melt reaction mixtures: sodium acetate pH 4, sodium acetate pH 5, MES pH 6, MOPS pH 7, Tris pH 8, and glycine pH 9. The six different pH buffers were used to dilute 5000x SYPRO™ orange protein gel stain (Sigma-Aldrich, USA) to 300x stock solutions. Small volumes (7.5 µL) from each of the dilute 300x SYPRO™ orange stock solutions were added to 200 µL PCR tubes. Protein was added to the dye at final reaction concentrations ranging from 100 µg/mL to 700 µg/mL. Appropriate pH buffers were added to the tubes to a final reaction volume of 25 µL. Blanks were prepared by adding buffer alone to diluted dye solutions.

The tubes were placed in a 36-well rotor in the RotorGene-6000 Real-time PCR machine. Following equilibration at 25 °C for 90 s, the temperature was increased to 99 °C in 0.2 °C increments with 5 s delays per increment. Samples were excited at 470 nm and the emission measured at 555 nm. Data were exported in Excel format. Blank emission data were subtracted from protein emission data, which were then overlaid on a graph for comparison between proteins. T_m values were estimated from the mid-point between the onset point and the top of the emission peak.

2.4.5 DAH7PS Activity Assay

A modified version of the continuous assay developed by Schoner and Hermann (1976) was used to assess DAH7PS enzymatic activity. Enzymatic activity was assessed by monitoring the disappearance of α,β-unsaturated carbonyl absorbance of PEP (Sigma-Aldrich, USA) at 232 nm using a Heλios™ spectrophotometer (Thermo Fisher Scientific, USA). Reaction temperature was controlled with a single cell peltier unit (Thermo Fisher Scientific, USA) and the temperature accurately measured using an Omega® microcomputer thermometer model HH-72T (Omega Engineering Inc., USA).

PEP was stored long-term as a lyophilised powder at -20 °C. For biochemical enzymatic assays, PEP was dissolved at 1.5 mM concentration in assay buffer (50 mM BTP, 1 mM $ZnSO_4$, pH 9). The pH of assay buffers were determined at the assay temperature to account for changes in the pH of BTP buffer with temperature. Once dissolved in this assay buffer, PEP was stored at 4 °C for at least two weeks without any noticeable effect on the rate of enzymatic activity at the concentrations used in the present study. E4P was synthesised by Sebastian Reichau (University of Canterbury) and supplied at a concentration of 27 mM. Aliquots of E4P were stored at -80 °C to increase its lifetime. In aqueous solution E4P can exist in both monomeric and dimeric forms (Duke et al., 1981). At high E4P concentrations, and in frozen or freeze-dried preparations, a substantial percentage of the E4P exists in one of three dimeric forms, while at low E4P concentrations at room temperature E4P exists predominantly as a hydrated monomer (Duke et al., 1981). As DAH7PS enzymes can only use the monomeric form of E4P as a substrate, frozen aliquots of E4P were defrosted and left to equilibrate at room temperature for at least 2 h, or overnight at 4 °C prior to performing DAH7PS activity assays. The solutions were left to equilibrate for such long periods of time as the spontaneous conversion of E4P from the dimeric forms to the monomeric forms occurs slowly at room temperature or below. Once thawed, E4P was able to be stored for up to 2 days at 4 °C without any discernible effect on the rate of enzymatic activity at the concentrations used in the present study.

Enzyme assays were performed by pre-equilibrating the assay mixture (50 mM BTP, 1 mM $ZnSO_4$, pH 9, with varying concentrations of PEP) to the desired temperature in a 5 mm pathlength quartz cuvette (Starna® Optiglass Ltd., UK). E4P was added at varying concentrations, followed immediately by addition of CM-DAH7PS enzyme at varying concentrations. The total reaction volume was 200 µL. The reaction was mixed quickly and vigorously and the absorbance was measured at 232 nm. Absorbance data were examined within the program Vision 32 version 1.25 (Thermo Spectronic, UK).

2.4.6 Optimal Temperature for Activity

In order to determine the T_{opt} of the different CM-DAH7PSs, enzymatic activity was measured at a number of different temperatures ranging from 30-75 °C. To determine the optimum temperature for activity of *Bsub*-CM-DAH7PS and the CM-DAH7PS from *B. stearothermophilus* (*Bstr*-CM-DAH7PS), in each reaction PEP was used at a concentration of 10 x the K_M at 40 °C and 60 °C, respectively, while E4P was used at a concentration of 3 x the K_M at the respective temperatures. Due to the severe inhibition by E4P of DAH7PS enzymatic activity of the ancestral enzymes, and the variation in the degree of inhibition at different temperatures, a constant maximum concentration of E4P could not be used at each different temperature. To determine the T_{opt} of the ancestral enzymes, a large excess of PEP was used in all of the reactions (~1.4 mM). At each temperature measured for the different ancestral enzymes a number of different reactions were performed. Everything was kept constant, except the concentration of E4P which was varied in each reaction. The concentration of E4P at which maximum enzymatic activity was observed at a particular temperature was used to compare rates of activity at different temperatures. Initial rates of activity were used to determine the T_{opt} of each enzyme in order to minimise the effect any potential E4P degradation and enzyme denaturation at higher temperatures may have had on the rate of enzymatic activity. The T_{opt} of the different CM-DAH7PS enzymes were determined by comparing the change in absorbance per second at different temperatures.

2.4.7 Michaelis-Menten Kinetic Analysis

Michaelis-Menten kinetic analysis was performed at the T_{opt} for the different DAH7PS enzymes and/or 40 °C. DAH7PS enzymatic activity was measured as in section 2.7.5. To determine the K_M (E4P) for each CM-DAH7PS enzyme, PEP was added to each reaction in large excess while the concentration of E4P was varied. K_M (PEP) values for the contemporary CM-DAH7PS enzymes from *B. subtilis* and *B. stearothermophilus* were determined by adding E4P at a concentration of 3 x the K_M at the enzymes T_{opt}, while varying the concentration of PEP in the reaction. For the ancestral enzymes for which kinetic analysis was possible, the concentration of E4P at which maximum activity (at either the T_{opt} or

40 °C) was added to the reaction, while the concentration of PEP added to the reaction mixture was varied. The concentrations of E4P which could be used to determine the K_M (PEP) was limited due to the enzymatic activity of the ancestral enzymes being inhibited by high concentrations of E4P. Initial rates of activity were determined at each concentration.

Michaelis-Menten plots were generated from these data using Graphpad Prism version 5.01 (GraphPad Software, USA). K_M and V_{max} values were determined using a non-linear regression line of best fit. V_{max} values determined from reactions in which PEP was added in excess were used to calculate k_{cat} values for all of the CM-DAH7PS enzymes. To calculate k_{cat}, changes in absorbance per second were converted to changes in molar concentration of PEP per second by dividing the change in absorbance per second by the extinction coefficient of PEP (2840 $M^{-1}cm^{-1}$) and the path-length of the cuvette. The change in molar concentration of PEP was divided by the molar concentration of enzyme in the reaction. It was assumed that the four active sites of the CM-DAH7PS enzymes are all catalytically active, so the rate per second was divided by four to give the k_{cat} of the enzyme.

2.5 Protein Crystallography

2.5.1 General Methodology

Proteins used for crystallographic analysis were overexpressed by IPTG-induction (section 2.3.1.2), then purified by IMAC (section 2.3.2.2) and size exclusion chromatography (section 2.3.2.4). The concentration of NaCl in the size exclusion chromatography running buffer was decreased from 500 mM to 150 mM. Proteins were concentrated by centrifugation in Vivaspin concentrators (section 2.3.2.3) to concentrations typically between 20 and 30 mg/mL. Concentrations higher than this typically resulted in the formation of soluble protein aggregates. All crystallisation trials were performed at 18 °C and were observed periodically using an optical microscope.

2.5.2 Initial Crystallisation Trials

Initial crystallisation trials using the sitting drop method were prepared at 18 °C using a Mosquito® crystallisation robot (TTP LabTech Ltd., USA). A total of 384 crystallisation conditions were used in the initial crystallisation trials. From the crystallisation screens PEGRx HT™ - HR2-086, Crystal Screen HT™ - HR2-130, Index HT™ - HR2-134, and SaltRx HT™ - HR2-136 (Hampton Research, USA) 100 µL of each precipitant solution were pipetted into the large wells of four 96-2 low profile Intelli-Plate™ protein crystallisation plates (Hampton Research, USA). The Mosquito® crystallisation robot was used to mix and dispense 100 nL of protein and 100 nL of mother liquor into the small wells of the 96-well Intelli-plates. Plates were then sealed with ClearSeal film™ (Hampton Research, USA) and placed on shock-proof shelves.

2.5.3 Optimising Crystallisation Conditions

Promising crystallisation conditions identified from the initial crystallisation trials were optimised by altering the pH or concentration of the different constituents in the precipitant solutions. The hanging drop method was used for fine screening of crystallisation conditions in 24-well VDX™ plates (Hampton Research, USA). The tops of the wells were lined with glisseal®N grease (Borer Chemie, Switzerland) and 500 µL of each precipitant solution were pipetted into separate wells. A small volume of protein (1 µL) was added to 1 µL of mother liquor on a siliconised glass cover slip, which was then inverted and pressed gently on top of the pre-greased well. Plates were placed on shock-proof shelves.

2.5.4 Preparing Crystals for Data Collection

Protein crystals were transferred to a 20 µL drop of mother liquor containing 5% (v/v) glycerol using a cryo-loop (Hampton Research, USA). Crystals were left in this solution for 30 s and transferred to mother liquor solutions with increasing glycerol concentrations (10%, 15%, and 20%) and left for 30 s in each solution. Crystals were then immersed in liquid nitrogen and loaded into a SSRL automated mounting cassette (Crystal Positioning Systems, USA).

2.5.5 X-Ray Diffraction Data Collection

X-ray diffraction data were collected at the Australian Synchrotron, Melbourne, Australia using the MX1 beam-line. An ADSC Quantum 210r detector (Area Detector Systems Corp., USA) was used to measure reflections. Prior to data collection, the MOSFLM strategy function was used with two images 90° apart to assist data collection.

2.5.6 Data Processing

2.5.6.1 Indexing and Scaling

X-ray diffraction images were visualised, indexed, and integrated using the MOSFLM program (Leslie, 1992). Images were examined to determine whether any deterioration in the data was observed. Cell parameters were determined using auto spot finder and auto index functions. Images were integrated in MOSFLM with particular attention given to any changes in RMSD and mosaicity. Images were then merged using SCALA within the CCP4 program (Bailey, 1994). Output was examined and re-scaled to optimise R_{merge}, I/σ_i, and data completeness.

2.5.6.2 *Matthew's Coefficient*

The number of monomeric subunits in the asymmetric unit was determined using the MATTHEWS_COEF program within CCP4. The number of subunits in the asymmetric unit was selected based on solvent percentage.

2.5.6.3 Molecular Replacement

Molecular replacement was performed using two CM-DAH7PS 3D structures from *L. monocytogenes* as models (PDB codes 3NVT and 3TFC). Modelling of the DAH7PS and CM domains was performed separately. Molecular replacement of the DAH7PS domains was performed using PHASER within PHENIX (McCoy et al., 2007). Molecular replacement of the CM domains was performed using PHASER within CCP4 (McCoy et al., 2007).

2.5.6.4 Model Building and Refinement

Automated building was performed using the PHENIX AutoBuild wizard (Terwilliger et al., 2008). The CM domains were modelled using PHASER within CCP4 (McCoy et al., 2007). Manual building was performed within COOT (Emsley and Cowtan, 2004). The model was built into $2|F_O|-|F_C|$ and $|F_O|-|F_C|$ maps contoured to 1 σ and 3 σ, respectively. Model refinement was performed using Refmac 5.0 within CCP4 (Murshudov et al., 1997), with particular attention given to the R_{free} and R-factor values.

2.5.7 Structural Analysis

All structural images within this thesis were generated using PYMOL (DeLano, 2002). Ramachandran analysis was performed using PROCHECK (Laskowski et al., 1993) within CCP4. Average B-factor analysis was performed using the Baverage program within CCP4 (Bailey, 1994). PDBePISA (Krissinel and Henrick, 2007) was used to determine total surface area of the proteins involved in inter-subunit interactions contributing to oligomeric assembly. The closest structural homologues in the PDB were identified using PDBeFold (Krissinel and Henrick, 2004).

3 Phylogenetics and Ancestral Inference

3.1 Introduction

CM-DAH7PS from the *Bacillus* genus was identified as an appropriate enzyme for ASR as CM-DAH7PS from *B. subtilis* has previously been expressed in a soluble, active form in *E. coli*, and the enzyme has been biochemically well characterised (Wu and Woodard, 2006, Wu et al., 2005, Jensen and Nester, 1966). In addition, DAH7PS is a large, structurally complex, multimeric protein which makes it an interesting candidate for ASR as any errors in reconstruction are likely to result in inactive or biologically unrealistic ancestral enzymes. In terms of genetic integrity, DAH7PS is a core metabolic enzyme that is present in most microorganisms making it a suitable candidate for ASR, as genes which are a part of the *Bacillus* core genome are less likely to be horizontally transferred (Didelot et al., 2010). Furthermore, in *Bacillus* species, *cm-dah7ps* exists in a region of the genome for which there is little evidence of recombination (Didelot et al., 2010), decreasing the chances of inaccurate phylogenetic inference (Arenas and Posada, 2010). The *Bacillus* genus is ideal for studies utilising ASR to investigate the origin and evolution of thermophily because *Bacillus* is an ancient genus (~1 Gyr old) and contemporary species inhabit a wide range of different temperature environments.

3.2 Results and Discussion

In order to perform ancestral sequence inference, a phylogenetic tree, statistical models of evolution, and nucleotide and amino acid alignments of extant sequences must first be generated. The majority of the CM-DAH7PS nucleotide and amino acid sequences were obtained from GenBank. The *cm-dah7ps* sequence from *B. stearothermophilus* was obtained from the *Bacillus* (*Geobacillus*) *stearothermophilus* Genome Sequencing Project (http://www.genome.ou.edu/bstearo.html) located on contig 501. Accession numbers and strain details are provided in Appendix A.

Four different sequence alignments were generated as part of this study. The first two alignments generated were an amino acid alignment and a nucleotide

alignment of CM-DAH7PS sequences from *Bacillus* species only. The CM-DAH7PS amino acid sequences from *Bacillus* are very similar (77.5% overall pairwise identity) and there were no gaps in the alignment that needed to be manually resolved. The CM domains (residues 1-87) are less highly conserved (69.5% pairwise identity) than the DAH7PS domains (residues 103-366) which share 80% pairwise identity. This may indicate that there are more functional constraints acting on the DAH7PS domain than the CM domain whose main function has been proposed to be the binding of prephenate for the purpose of allosteric regulation of DAH7PS activity by downstream products. This high level of sequence conservation is clearly demonstrated in Figure 3.1 by the large black areas in the alignment, which indicate conservation of amino acids at these sites in these sequences from *Bacillus*. The high sequence similarity of the CM-DAH7PS sequences has potential implications for phylogenetic tree construction and ASR which will be discussed later. Two further alignments were generated: an amino acid alignment and a nucleotide alignment of CM-DAH7PS sequences from *Bacillus* species and DAH7PS sequences from *Clostridium butyricum* and *Clostridium acetobutylicum*. Unlike all of the *Bacillus* DAH7PS enzymes, the DAH7PS enzymes from the two *Clostridium* species do not have a CM attached to their N-termini. Instead, they have an N-terminal extension of ~80 amino acids which shares significant sequence identity with the FL domain from *T. maritima* and other putative FL domains. The first 78 amino acid residues (or 234 nucleotides) from the two *Clostridium* species sequence were therefore removed from the alignments as the inclusion of these residues would have resulted in the sequence divergence between the two *Clostridium* and the *Bacillus* species being exaggerated.

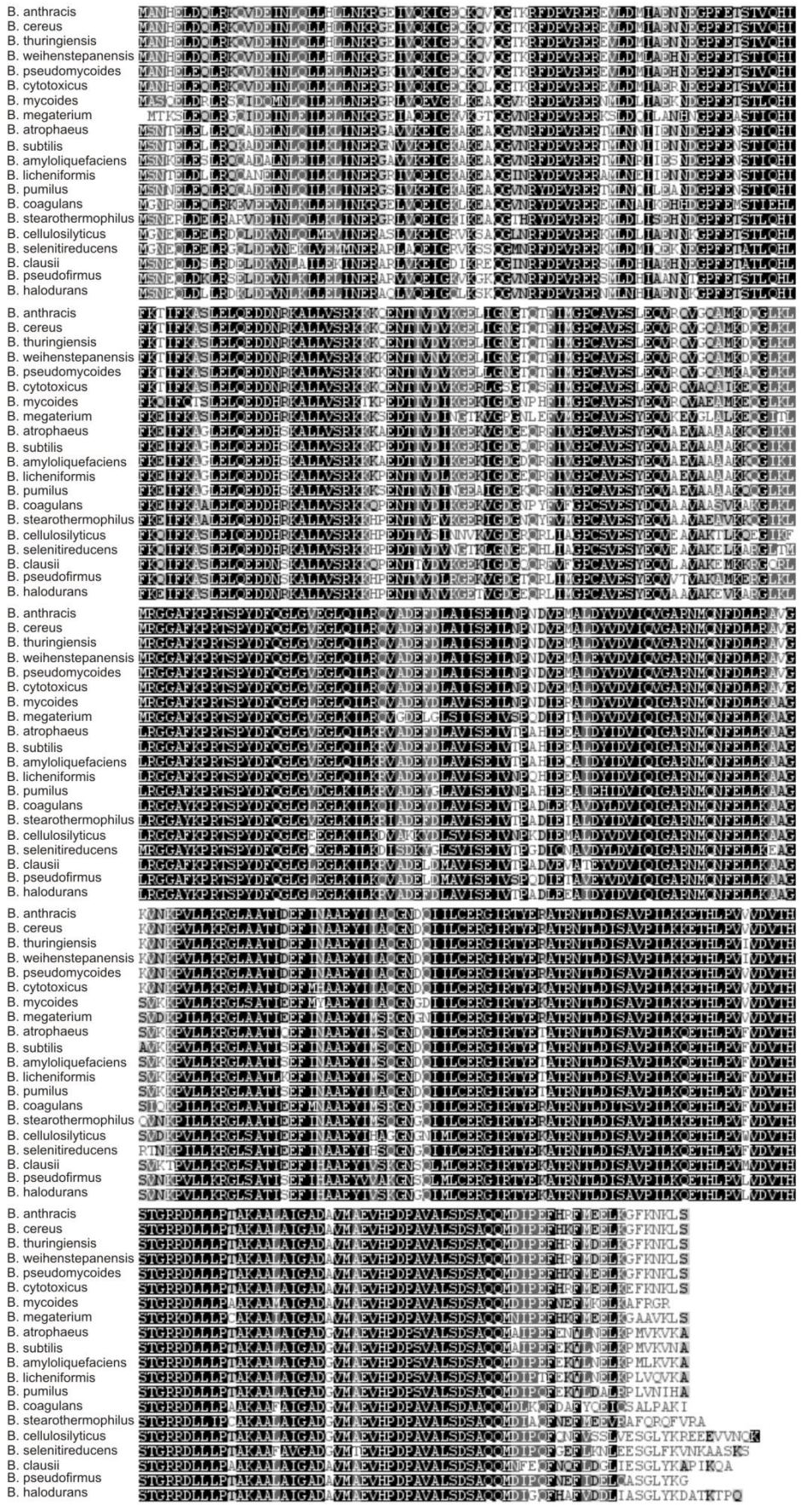

Figure 3.1 *Bacillus* CM-DAH7PS amino acid sequence alignment.
Shading indicates the percentage similarity at each position in the alignment. Black=100%, dark grey=80-100%, light grey=60-80% and white=less than 60% similarity. Alignment was made using Geneious (Drummond et al. 2011).

3.2.1 Phylogenetic Analysis

The most appropriate model of amino acid evolution for CM-DAH7PS amino acid sequences as determined by ProtTest version 2.4 (Abascal et al., 2005) was LG + *I* + *G*. This is a sophisticated statistical model that incorporates evolutionary rate variability across sites in the matrix estimation and is the best protein evolutionary model for most datasets as indicated by lower Akaike information criterion (AIC) values (Le and Gascuel, 2008). AIC scores represent the relative likelihood of a model and the model with the lowest AIC being the best model out of the models available for a particular dataset.

Ten ML phylogenetic trees were constructed using GARLI version 1.0 (Zwickl, 2006) as described in section 2.1.2.2. The phylogenetic trees were re-rooted by selecting the two *Clostridium* species as the outgroup. The log likelihood scores for all ten phylogenetic trees were -5361 indicating that no particular tree was less likely than any of the other trees constructed, so the log likelihood scores were unable to be used to help determine the best ML phylogenetic tree. The ML phylogenetic trees were compared with each other and with published *Bacillus* phylogenetic trees (Alcaraz et al., 2010, Hobbs et al., 2012). One of the ten trees was selected as the best ML phylogenetic tree based on the positions of the branches within the tree being consistently present in the majority of the constructed trees and the fact that this particular tree also had a similar branching pattern to the ML phylogenetic tree constructed by Alcaraz et al. (2010) based on the core genomes of 20 *Bacillus* species, and the ML phylogenetic tree constructed by Hobbs et al. (2012) based on the IPMDH amino acid sequences of 19 *Bacillus* species. In all three trees, *Bacillus clausii* and *B. halodurans* cluster together and diverge at basal positions in the trees. The same species were present in the *Bacillus cereus* and *B. subtilis* clades across the three trees if they were included in all the phylogenetic trees and these two clades diverge at more derived positions in the trees than the *B. clausii* and *B. halodurans* clade.

A total of 1,024 pseudoreplicate trees were generated using GARLI version 1.0 (Zwickl, 2006) and bootstrap proportions for each branch within the phylogenetic tree were generated. These numbers represent the frequency with which a given branch occurs within the pseudoreplicate datasets and allows the reliability of

each branch within the phylogenetic tree to be estimated. Bootstrap proportions for all internal nodes are displayed on the chronogram (Figure 3.2). Due to the high level of sequence similarity between CM-DAH7PS amino acid sequences, it is difficult for the phylogenetic relationships between very closely related sequences to be determined. This may explain the low bootstrap proportions observed for some of the internal nodes, especially the ANC2 and ANC3 internal nodes, 39 and 43, respectively (Figure 3.2). Despite the low bootstrap support for these nodes, their positioning was deemed to be reliable as nodes at similar positions were observed in the phylogenetic trees constructed by Alcaraz et al. (2010) and Hobbs et al. (2012).

In a phylogram, branch lengths represent the number of changes per site per unit of time and, as such, the relatedness of different sequences. In a chronogram, the unit of time is defined so the branch lengths represent the amount of time since sequence divergence. To determine which ancestral nodes should be reconstructed and allow the age of all of the internal nodes within the tree to be estimated, the CM-DAH7PS phylogram was converted to a chronogram using the program r8s version 1.71 (Sanderson, 2003). The points of *Bacillus* and *Clostridium* divergence, and *B. subtilis* and *B. halodurans* divergence were fixed based on the ages of these two nodes as determined by Battistuzzi et al. (2004): 2,650 Mya and 950 Mya, respectively. The age of the remaining nodes were then estimated from the branch lengths in the phylogenetic tree. The constructed chronogram is shown in Figure 3.2.

OGTs of contemporary *Bacillus* species were taken from numerous literature sources (Fritze and Pukall, 2001, Nogi et al., 2005, Nakamura, 1998, Gordon, 1972, Ju et al., 2009, Denizci et al., 2004). OGTs are displayed as coloured circles on the ML chronogram (Figure 3.2). The OGTs of the contemporary *Bacillus* species used in the present study are not as diverse as those used in the ASR study of Hobbs et al. (2012). Five additional species (*Bacillus atrophaeus, Bacillus cellulosilyticus, Bacillus mycoides, Bacillus pseudofirmus* and *Bacillus selenitireducens*) were included in the ML CM-DAH7PS phylogenetic tree that were not included in the ML IPMDH tree. These species are all from mesophilic organisms with OGTs between 25-30 °C or at 37 °C. *B. atrophaeus* is very closely related to *B. subtilis*, with these two species only recently being classified

as separate species (Fritze and Pukall, 2001). As such in the ML CM-DAH7PS tree, *B. atrophaeus* is part of the *B. subtilis* clade. *B. mycoides* is included as part of the *B. cereus* clade in the ML CM-DAH7PS tree, which is consistent with its taxonomic classification (Nakamura and Jackson, 1995). *B. pseudofirmus*, *B. cellulosilyticus* and *B. selenitireducens* are mesophilic, alkaliphilic and halotolerant or halophilic organisms. Based on their 16S rRNA sequences, *B. pseudofirmus* has previously been shown to be part of the *B. halodurans* and *B. clausii* clade, and *B. cellulosilyticus* and *B. selenitireducens* have been shown to diverge from basal portions of the *Bacillus* 16S rRNA phylogenetic tree (Nogi et al., 2005) which is consistent with their positioning in the ML CM-DAH7PS chronogram (Figure 3.2). The positioning of these three species at basal positions in the tree is also supported by the basal positioning of *B. halodurans* and *B. clausii*, which are also mesophilic, alkaliphilic and halotolerant or halophilic species. Unlike the study of Hobbs et al. (2012), in which a number of sequences from contemporary thermophilic and psychrophilic *Bacillus* species were used in their reconstruction, no sequences from psychrophilic species and only one sequence from a thermophilic species (*B. stearothermophilus*) were used in the reconstruction of ancestral CM-DAH7PS enzymes. Hobbs et al. (2012) were able to incorporate a number of sequences from psychrophilic and thermophilic species as they determined the sequences from these species as part of their study. The lack of sequences from psychrophilic and thermophilic species, however, should not affect the thermostabilities of the inferred CM-DAH7PS ancestral enzymes as the ASR study of Gaucher et al. (2003) demonstrated that the MRCA of mesophilic bacteria likely lived at a higher temperature than any of its descendants. The thermal properties of the ancestral IPMDH enzymes reconstructed by Hobbs et al. (2012) also demonstrate that the thermal properties of ancestral proteins are not simply averages of their extant descendants.

Once the ML chronogram had been generated, evolutionarily interesting internal nodes could be selected for ancestral sequence inference and reconstruction.

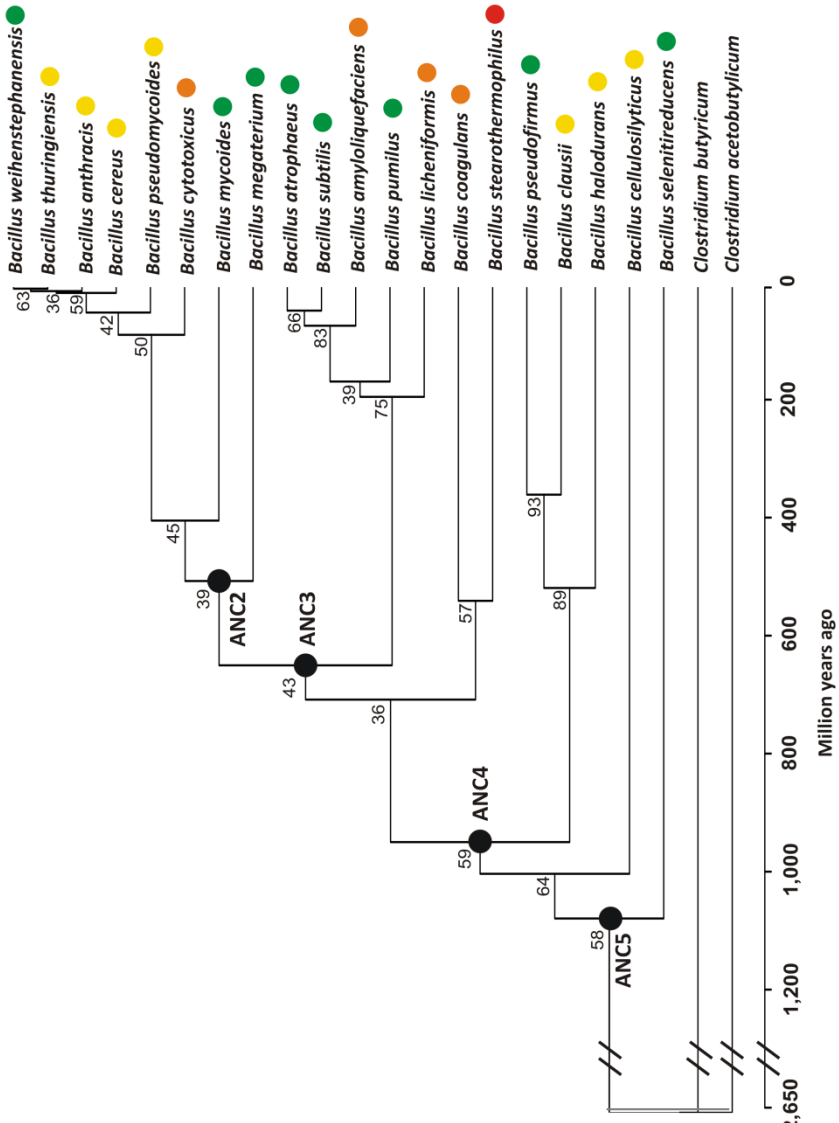

Figure 3.2 Maximum likelihood chronogram of *Bacillus* species based on CM-DAH7PS amino acid sequences.
The ancestral nodes reconstructed in the present study are indicated by black circles. The numbers at internal nodes are bootstrap proportions determined from 1,024 pseudoreplicate datasets. The two *Clostridium* species were used as an outgroup. Optimal growth temperatures were determined from the literature (Denizci, et al., 2004; Fritze & Pukall, 2001; Gordon, 1972; Ju, et al., 2009; Nakamura, 1998; Nogi, et al., 2005) and are represented by coloured circles: green (25-30 °C), yellow (37 °C), orange (45-50 °C), and red (60 °C).

3.2.2 Ancestral Inference

The most appropriate model of nucleotide evolution as determined by jModelTest version 0.1.1 (Posada, 2008) was TIM3 + *I* + *G*. The transition model (TIM) incorporates variable base frequencies, variable transition rates and two transversion rates (Posada, 2003).

Ancestral nodes in similar positions to the ancestral nodes reconstructed by Hobbs et al. (2012) were chosen for reconstruction to allow the results from the present study to be more easily compared with those from Hobbs et al. (2012). The labelling of the ANC2, ANC3, and ANC4 nodes reflects their similar positioning in the chronogram to the position of the IPMDH ANC2, ANC3, and ANC4 enzymes reconstructed by Hobbs et al. (2012). In addition, *Anc5*-CM-DAH7PS was reconstructed because it is evolutionarily interesting as the ANC5 node represents the MRCA of contemporary *Bacillus* species, and the unusual contemporary species *B. selenitireducens*, which is able to respire oxyanions of selenium and arsenic (Blum et al., 1998), diverges directly from this branch point. The estimated ages of the ANC2, ANC3, ANC4, and ANC5 nodes are 570 Myr, 650 Myr, 950 Myr, and 1,079 Myr old, respectively. The ANC2 and ANC3 nodes from the CM-DAH7PS chronogram are estimated to be considerably younger than the corresponding nodes from the IPMDH chronogram constructed by Hobbs et al. (2012), which are 820 Myr and 850 Myr old, respectively. The ages of the ANC4 nodes are exactly the same as this was one of the calibration points which was fixed in order to convert the phylograms to chronograms.

The ML inference program PAML version 4.3 (Yang, 2007) was used to perform three different inference methods: nucleotide, amino acid, and codon inference (section 2.1.3.1). The ancestral nucleotide and codon inferences were translated into amino acid sequences and aligned with the inferred amino acid sequence using Clustal W2 (Larkin et al., 2007). Consensus sequences were compiled and any ambiguous sites were resolved using the criteria described in section 2.1.3.1. There were very few ambiguous sites (less than 30 amino acids) in each of the ancestral sequences. Most of these ambiguous sites were resolved by selecting the amino acid inferred by two of the inference methods. Only five or fewer amino acid residues from each of the ancestral nodes needed to be resolved using the

other criteria. The low levels of ambiguity may mean that the ancestral enzymes are more likely to be biologically realistic, despite the structural complexity of the CM-DAH7PS, as lower levels of ambiguity reduces the likelihood of amino acids being incorrectly inferred. The ancestral sequences contain all the essential metal-binding and active site residues present in extant DAH7PSs.

The ancestral amino acid sequences are highly similar to the contemporary *Bacillus* sequences with all of the ancestral sequences sharing more than 71% identity with any of the contemporary *Bacillus* sequences. This, again, increases the likelihood of accurate ancestral inference of such a structurally complex enzyme as the high levels of sequence identities between the ancestral and contemporary sequence suggests a conservation of function. The sequence similarities between all of the extant sequences and the four ancestral sequences are provided in Table 3.1 and the ancestral amino acid sequences are provided below.

***Anc2*-CM-DAH7PS**

MTNKELEQLREQVDEINLQILELLNERGRIVQEIGKVKEAQGVNRFDPVRERKML
DLIAENNDGPFETSTLQHIFKEIFKASLELQEDDHRKALLVSRKKKPENTIVDIK
GEKIGDGNQQFIMGPCAVESYEQVREVAEAMKEQGLKLMRGGAFKPRTSPYDFQG
LGVEGLQILRQVADEFDLAVISEIVTPNDIEMALDYVDVIQIGARNMQNFELLKA
AGSVNKPVLLKRGLAATIEEFINAAEYIMSQGNGQIILCERGIRTYERATRNTLD
ISAVPILKKETHLPVVVDVTHSTGRRDLLLPTAKAALAIGADAVMAEVHPDPAVA
LSDSAQQMDIPEFNKFMEELKAFGNKLS

***Anc3*-CM-DAH7PS**

MSNKELEQLREQVDEINLQILELINERGRIVQEIGKVKEAQGVNRFDPVRERKML
DLIAENNDGPFETSTLQHIFKEIFKASLELQEDDHRKALLVSRKKKPENTIVDIK
GEKIGDGNQQFIMGPCAVESYEQVAEVAEAVKEQGLKLLRGGAFKPRTSPYDFQG
LGVEGLQILKRVADEFDLAVISEIVTPADIEKALDYVDVIQIGARNMQNFELLKA
AGSVNKPVLLKRGLAATIEEFINAAEYIMSQGNGQIILCERGIRTYERATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAALAIGADGVMAEVHPDPAVA
LSDSAQQMDIPQFNKFMEELKAFGNKKA

***Anc4*-CM-DAH7PS**

MSNEQLEELRDQLDEVNLKLLELINERARLVQEIGKVKSAQGVNRFDPVRERKML
DLIAENNKGPFETSTLQHIFKQIFKASLELQEDDHRKALLVSRKKHPENTIVDVK
GEKVGDGKQRLIMGPCAVESYEQVAAVAKAVKERGLKLLRGGAFKPRTSPYDFQG
LGLEGLKILKRVADEFDLAVISEIVTPADIEEALDYVDVIQIGARNMQNFELLKA
AGSVNKPVLLKRGLSATIEEFINAAEYIVSQGNGQIMLCERGIRTYEKATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAALAIGADGVMAEVHPDPAVA
LSDSAQQMDIPQFNEFVDDLIASGLYKAATKTAQQK

***Anc5*-CM-DAH7PS**

MGNEQLEELRDQLDEVNLKLVEMMNERARLAQEIGRVKSSQGMNRFDPVRERKML
DMIAEKNEGPFETATLQHLFKQIFKASLELQEDDHRKALLVSRKKHPEDTIVDVN
GTKIGDGEQHLIAGPCSVESYEQVEAVAKELKERGLKLLRGGAFKPRTSPYDFQG
LGQEGLEILKDVADKYGLSVISEIVTPGDIENAVDYVDVIQIGARNMQNFELLKE
AGRTNKPILLKRGLSATIEEFINAAEYIHSQGNGQIILCERGIRTYEKATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAAFAVGADGVMTEVHPDPAVA
LSDSAQQMDIPQFDEFLKNLEESGLFKVKKAASKSK

The contemporary sequence which is most closely related to *Anc2*-CM-DAH7PS is the sequence from *B. mycoides* (88% sequence identity). It is also highly similar to the other species in the *B. cereus* clade and, although out of all of the contemporary sequences it shares the lowest sequence identity with the sequence from *B. selenitireducens*, the level of sequence identity observed is higher than any of the sequences from species within the *B. cereus* and *B. subtilis* clades have with the *B. selenitireducens* sequence. The most closely related contemporary sequence to *Anc3*-CM-DAH7PS is from *B. stearothermophilus* (88% sequence identity). *Anc3*-CM-DAH7PS also shares high sequence identity with sequences from species which are part of the *B. subtilis* clade, and shares higher sequence identity with the *B. selenitireducens* sequence than the contemporary species from the *B. cereus* and *B. subtilis* clades. *Anc4*-CM-DAH7PS is highly similar in sequence to the sequences from *B. halodurans* (92% sequence identity) and *B. pseudofirmus* (89% sequence identity). *Anc4*-CM-DAH7PS shares a greater sequence identity with the sequence from *B. selenitireducens* than any contemporary sequence does. However, *Anc5*-CM-DAH7PS shares the highest sequence identity with the sequence from *B. selenitireducens* of all the ancestral enzymes (96% sequence identity). In contrast, *Anc5*-CM-DAH7PS shares low sequence identity with all other contemporary sequences, except the sequence from *B. cellulosilyticus* with which it shares 83% sequence identity. These observed sequence identities are consistent with the topology of the tree and the positioning of the contemporary and ancestral sequences.

Table 3.1 Sequence identities between extant and ancestral CM-DAH7PS amino acid sequences.

	BANT	BATR	BCEL	BCER	BCLA	BCOA	BCYT	BHAL	BLIC	BMEG	BMYC	BPSF	BPSM	BPUM	BSEL	BSTR	BSUB	BTHU	BWEI	ANC2	ANC3	ANC4	ANC5
BAMY	74	96	74	74	76	76	74	78	92	75	77	75	75	89	71	79	95	74	74	83	86	81	74
BANT	-	75	71	99	70	71	95	75	76	78	81	74	97	74	69	77	75	99	98	86	82	76	72
BATR		-	75	75	75	76	75	79	93	76	77	75	75	89	71	79	97	75	74	83	86	82	74
BCEL			-	71	76	73	71	80	74	73	75	80	71	73	80	74	74	71	70	78	80	85	83
BCER				-	70	71	95	75	76	78	81	74	98	75	69	77	75	99	98	86	82	76	72
BCLA					-	73	71	82	75	68	74	85	71	74	73	75	75	70	70	76	78	83	77
BCOA						-	70	78	76	73	74	74	71	75	73	80	75	71	70	78	80	79	75
BCYT							-	75	76	76	82	75	95	75	70	77	75	95	93	86	82	77	73
BHAL								-	79	71	76	87	75	78	77	80	80	75	74	80	83	92	79
BLIC									-	74	78	77	75	90	72	81	92	76	75	83	86	81	75
BMEG										-	77	74	77	74	71	75	76	77	76	85	82	75	72
BMYC											-	78	81	76	72	79	77	81	80	88	85	81	75
BPSF												-	74	75	76	76	75	75	74	80	82	89	80
BPSM													-	75	70	77	75	97	96	86	82	77	72
BPUM														-	70	79	89	75	75	81	84	80	73
BSEL															-	74	71	69	68	75	76	81	96
BSTR																-	80	77	76	85	88	84	75
BSUB																	-	75	74	83	86	81	74
BTHU																		-	98	86	82	77	72
BWEI																			-	84	81	75	71
ANC2																				-	96	86	78
ANC3																					-	90	80
ANC4																						-	84

Abbreviations: *Bacillus amyloliquefaciens* (BAMY), *Bacillus anthracis* (BANT), *Bacillus atrophaeus* (BATR), *Bacillus cellulosilyticus* (BCEL), *Bacillus cereus* (BCER), *Bacillus clausii* (BCLA), *Bacillus coagulans* (BCOA), *Bacillus cytotoxicus* (BCYT), *Bacillus halodurans* (BHAL), *Bacillus licheniformis* (BLIC), *Bacillus megaterium* (BMEG), *Bacillus mycoides* (BMYC), *Bacillus pseudofirmus* (BPSF), *Bacillus pseudomycoides* (BPSM), *Bacillus pumilus* (BPUM), *Bacillus selenitireducens* (BSEL), *Bacillus stearothermophilus* (BSTR), *Bacillus subtilis* (BSUB), *Bacillus thuringiensis* (BTHU), and *Bacillus weihenstephanensis* (BWEI).

In addition to inferring ancestral sequences, PAML also provides posterior probabilities for every residue inferred in the ancestral sequences as well as the average posterior probabilities for all of the residues in ancestral sequences. These posterior probabilities provide an estimation of the likelihood a particular amino acid occupied a particular site in the protein during its evolution given the data and the evolutionary model. The average posterior probabilities of the four inferred ancestral sequences are shown in Table 3.2. The posterior probabilities of all the inferred sequences are high, indicating that there is a high level of confidence in the inferred ancestral sequences.

Table 3.2 Posterior probabilities of inferred ancestral sequences.

	Amino Acid Inference	**Codon Inference**	**Nucleotide Inference**
***Anc2*-CM-DAH7PS**	0.955	0.843	0.835
***Anc3*-CM-DAH7PS**	0.964	0.834	0.861
***Anc4*-CM-DAH7PS**	0.959	0.848	0.864
***Anc5*-CM-DAH7PS**	0.961	0.861	0.870

Posterior probabilities were determined by PAML version 4.3 (Yang, 2007) during ancestral inference.

There is a potential for bias to be introduced in ancestral inferences as a result of long branches in the phylogenetic tree and the fact that amino acid substitution matrices have high equilibrium frequencies for hydrophobic amino acids (Gaucher et al., 2008). This bias can lead to a spurious increase in hydrophobic amino acids in inferred ancestral protein sequences which can potentially increase the thermostability of the reconstructed enzymes (Gaucher et al., 2008). This could result in inaccurate inferences being made regarding the evolution of thermophily. However, there are no significant differences between the overall amino acid compositions of the four ancestral CM-DAH7PS sequences. The hydrophobic content of *Anc5*-CM-DAH7PS is 41% and the other three ancestral CM-DAH7PS enzymes have hydrophobic contents of 44%. The thermostabilities of the reconstructed CM-DAH7PS enzymes are therefore unlikely to be affected by any amino acid selection bias.

Following inference and sequence analysis, the four ancestral genes encoding these ancestral proteins were synthesised by GENEART (Regensburg, Germany). These gene sequences are provided in Appendix B.

4 Cloning, Protein Expression, Purification and Characterisation

4.1 Introduction

Modern molecular techniques enable gene sequences of inferred ancestral amino acid sequences to be synthesised, allowing the enzymes encoded by these synthesised genes to be biochemically characterised. This characterisation can then reveal how certain properties may have originated and evolved. Measurement of the thermal properties of reconstructed Precambrian proteins has previously been used to infer the origin and evolution of thermophily (Hobbs et al., 2012, Gaucher et al., 2008, Perez-Jimenez et al., 2011). The two earlier studies inferred a primitive origin for thermophily and a steady decrease in thermophily over time. However, the more recent study by Hobbs et al. (2012) investigating the evolution of thermophily within the *Bacillus* genus suggested a fluctuating trend in thermophily and that thermophily may have evolved at least twice within the *Bacillus* genus. This inference, however, was based on data from only a single enzyme so, in order to ensure that this inference is robust in regards to the type of enzyme being reconstructed, more enzymes from the *Bacillus* genus need to be reconstructed and biochemically characterised. In addition, the biological feasibility of the ancestral enzymes can be evaluated by comparing the kinetic properties of the ancestral enzymes with those displayed by their contemporary counterparts.

4.2 Results and Discussion

4.2.1 Cloning of *cm-dah7ps* Genes

Before the ancestral and contemporary CM-DAH7PS enzymes could be biochemically characterised, the *cm-dah7ps* genes had to be cloned into an appropriate vector for expression of recombinant CM-DAH7PS in *E. coli*. The CM-DAH7PS enzymes from *B. subtilis* and *B. caldovelox* were selected as appropriate contemporary enzymes to be cloned and characterised as *B. subtilis* is a mesophile and *B. caldovelox* is a thermophile. Expression and characterisation

of these two enzymes would enable the relationship between the thermal properties of CM-DAH7PS enzymes and the OGT of the organism to be determined. The CM-DAH7PS from *B. subtilis* was also chosen because *B. subtilis* is one of the most studied bacteria and the CM-DAH7PS enzyme from *B. subtilis* has previously been expressed and purified from *E. coli*. Furthermore, the biochemical properties of this enzyme are well known (Wu and Woodard, 2006, Wu et al., 2005, Jensen and Nester, 1966). The CM-DAH7PS from *B. caldovelox* was chosen as there was a lack of *cm-dah7ps* sequences from thermophilic *Bacillus* species available, with the only available sequence being from *B. stearothermophilus* which is a close relative of *B. caldovelox*. Cloning and sequencing of *cm-dah7ps* from *B. caldovelox* genomic DNA would allow an additional sequence from a contemporary thermophilic species to be incorporated into the ancestral inference. Before cloning of the *cm-dah7ps* genes from these organisms could be performed, the genomic DNA was first extracted from *B. subtilis* and *B. caldovelox* cultures (section 2.2.2)

The *cm-dah7ps* genes from *B. subtilis* and *B. caldovelox* were amplified from genomic DNA using the primers listed in Table 2.1. The *B. caldovelox* genome has yet to be sequenced and so the primers for *cm-dah7ps* from *B. caldovelox* were designed against the genome of the very closely related species *B. stearothermophilus*. This approach had been used successfully by Hobbs et al. (2012) with *leuB* from *B. caldovelox.* Following PCR amplification, large bands of ~1,100 bp in length, the expected length of the *cm-dah7ps* genes from these *Bacillus* species, were observed following agarose gel electrophoresis.

Once the *cm-dah7ps* genes had been successfully amplified, the genes could then be cloned into an appropriate expression vector. The plasmid pET28b was selected as an appropriate expression vector as a C-terminal His-tag was desired for IMAC purification. A N-terminal His-tag was deemed unsuitable as a His-tag fused to the N-terminal FL regulatory domain of *Tm*-DAH7PS has been shown to contribute to interactions with the opposing FL domain, thereby locking the protein in the inactive tyrosine-bound conformation (Cross et al., 2011). The *cm-dah7ps* genes were inserted into pET28b and transformed into the *E. coli* cloning strain DH5α. Potential positive transformants, as identified by growth on selective agar plates containing 50 µg/mL kanamycin, were screened by one of

three procedures: colony PCR, PCR following plasmid extraction, or restriction enzyme digestion (section 2.2.12). Successful cloning of the *cm-dah7ps* from *B. subtilis* into pET28b was confirmed by DNA sequencing (the DNA sequence is provided in Appendix B). The plasmid from the successful transformant was then retransformed into *E. coli* BL21 to allow expression of recombinant CM-DAH7PS via a T7 promoter.

Despite numerous attempts, the *B. caldovelox cm-dah7ps* gene could not be successfully cloned. The reason for this failure was revealed when digested and undigested PCR amplified *cm-dah7ps* from *B. caldovelox* were run side-by-side on an agarose gel (section 2.2.8). Analysis of the banding pattern revealed the presence of one large band (~1,100 bp) in the undigested sample, whereas two smaller bands (~850 and ~250 bp) were present in the digested sample (Figure 4.1). This banding pattern indicates that there is either an *Nco*I or *Xho*I cut site within the *cm-dah7ps* gene from *B. caldovelox*, which meant that these two restriction enzymes could not be used to clone this gene. This result highlights that using the DNA sequence of *B. stearothermophilus* to design PCR primers to amplify genes from the closely related *B. caldovelox* is not ideal as even a single nucleotide substitution can introduce a restriction site. New primers incorporating different restriction enzyme sites could have been designed to try to re-amplify and clone the *B. caldovelox cm-dah7ps* gene, however, due to time constraints, the gene for *cm-dah7ps* from *B. stearothermophilus* was chemically synthesised instead.

The synthesised contemporary and ancestral *cm-dah7ps* genes were inserted into pET28b, transformed into *E. coli* DH5α and then subsequently into *E. coli* BL21. A list of all of the transformants generated as part of this study is provided in Appendix A.

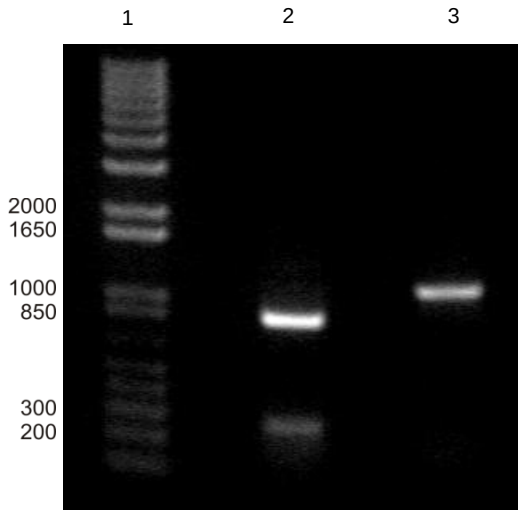

Figure 4.1 Agrarose gel of digested and undigested amplified *cm-dah7ps* from *B. caldovelox*.
Lane 1 is a 1 kb plus DNA ladder, lane 2 is a sample of *cm-dah7ps* amplified from *B. caldovelox* which has been digested by the restriction enzymes *Nco*I and *Xho*I, and lane 3 is a sample of undigested *cm-dah7ps* amplified from *B. caldovelox*. The presence of two smaller bands in the digested sample (~850 and 250 bp in size) when compared to the size of the undigested full length gene (lane 3, ~1,100 bp) demonstrates that a cut site for one of these restriction enzymes is present within the gene.

4.3 Protein Expression and Purification

Following successful cloning of the *cm-dah7ps* genes into pET28b, the CM-DAH7PS enzymes could then be expressed and purified. Two main methods were used to express recombinant CM-DAH7PS enzymes: IPTG-induction (section 2.3.1.2) and auto-induction coupled with co-expression of GroES and GroEL (section 2.3.1.3). Expression of ancestral CM-DAH7PS enzymes was initially induced by addition of IPTG and all of the contemporary enzymes were expressed in this manner. Towards the end of the study, expression and purification of stable and active ancestral CM-DAH7PS enzymes was no longer possible using this method of induction. Instead, the expression of the ancestral enzymes was coupled with expression of recombinant *E. coli* GroES and GroEL heat shock proteins from a separate pGroESL plasmid. These heat shock proteins form a GroEL-GroES chaperonin complex that assists proper protein folding in an ATP-dependent manner (Horwich et al., 2007).

4.3.1 Protein Purification

Recombinant CM-DAH7PS enzymes were purified by IMAC purification (section 2.3.2.2), which utilises high affinity binding to Ni^{2+} ions of the His-tag fused to the C-termini of the recombinant CM-DAH7PS enzymes to separate the recombinant protein from the majority of the other cellular proteins. An example UV trace of an IMAC purification of recombinant CM-DAH7PS is provided in Figure 4.2. SDS-PAGE analysis revealed that the fractions corresponding to the large UV absorbance peak contained a protein with a molecular weight of ~40 kDa, the calculated size of all of the *Bacillus* CM-DAH7PS monomers, indicating successful, soluble expression of recombinant CM-DAH7PS (Figure 4.2). From the SDS-PAGE gel it can also be seen that, although the CM-DAH7PS enzyme is quite pure after IMAC purification, there are a small number of faint protein bands of different sizes present, so further purification was necessary.

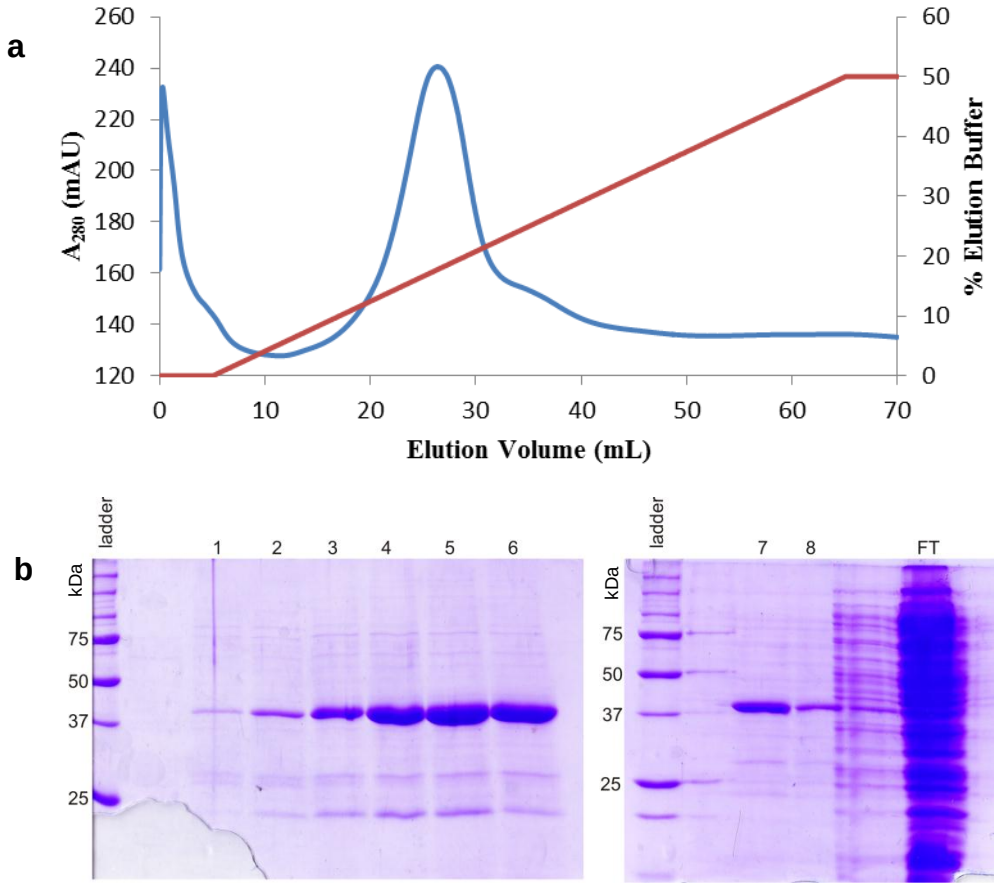

Figure 4.2 IMAC purification of a recombinant *Bacillus* CM-DAH7PS enzyme expressed in *E. coli* BL21.
(a) Elution of recombinant His-tagged CM-DAH7PS enzyme from IMAC column. Protein elution, as a result of increasing concentration of elution buffer, was monitored by measuring UV absorbance at 280 nm. (b) SDS-PAGE gels demonstrating the size and purity of the CM-DAH7PS enzyme from 2 mL fractions collected during purification. Lane 1 is a sample from 10-12 mL elution volume, lanes 2-8 are samples from 2 mL fractions collected from 18-32 mL elution volume. FT is protein from the cellular supernatant which was not bound to the IMAC column. The lane beside lane 7 is overflow from the ladder and the lane beside FT is overflow from the FT lane.

Following IMAC purification, the CM-DAH7PS enzymes were further purified by size exclusion chromatography (section 2.3.2.4) which purified the CM-DAH7PS enzymes based on their size. An example UV trace and SDS-PAGE gel are shown in Figure 4.3.

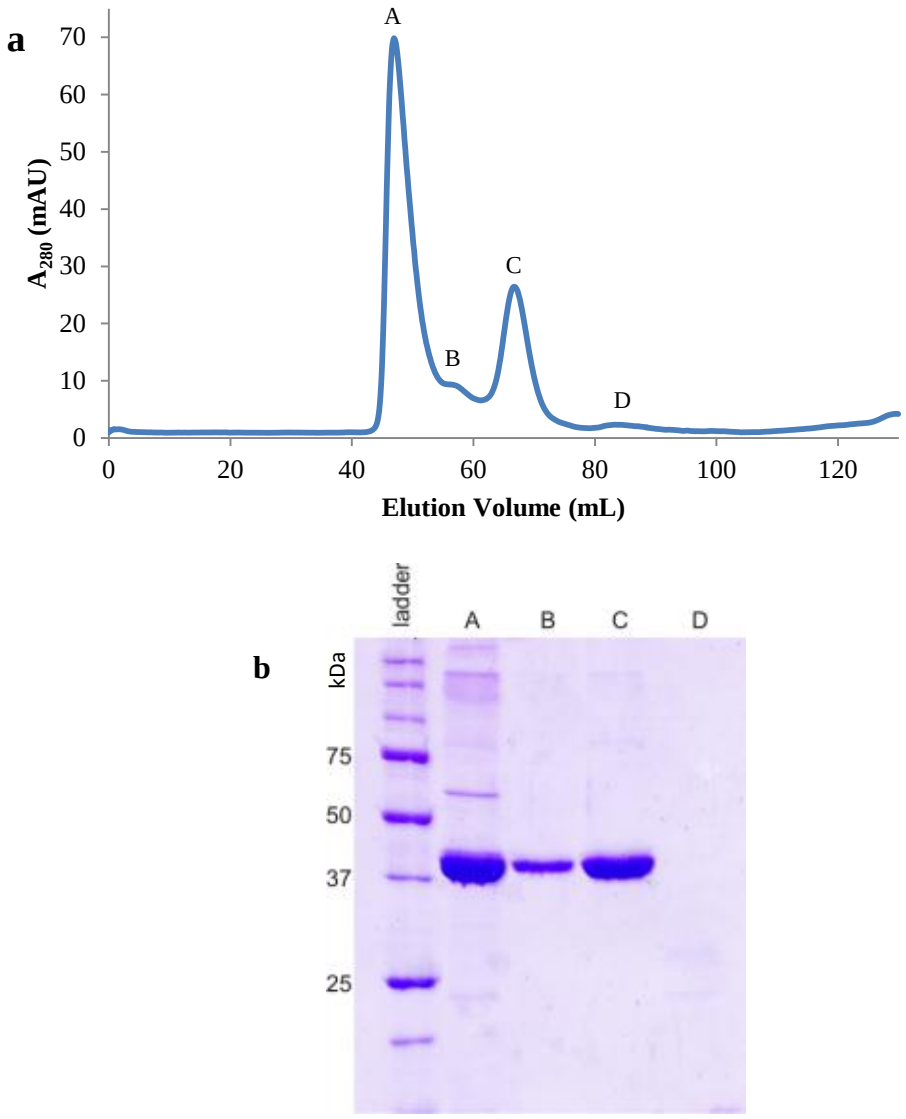

Figure 4.3 Size exclusion chromatography purification of recombinant *Bacillus* CM-DAH7PS.
(a) UV trace of size exclusion chromatography purification of a recombinant *Bacillus* CM-DAH7PS. (b) SDS-PAGE gel of samples from the four peaks in UV absorbance.

At least two peaks in UV absorbance were always observed during size exclusion chromatography purification of all of the different recombinant enzymes (peaks A and C in Figure 4.3). The molecular weights of protein eluted during size exclusion chromatography were estimated by comparing the volume of buffer eluted from the column prior to protein elution to a calibration curve (Figure 4.4) generated for the S200 16/60 size exclusion chromatography column using the high and low molecular weight Gel Filtration Calibration Kits (GE Healthcare, UK).

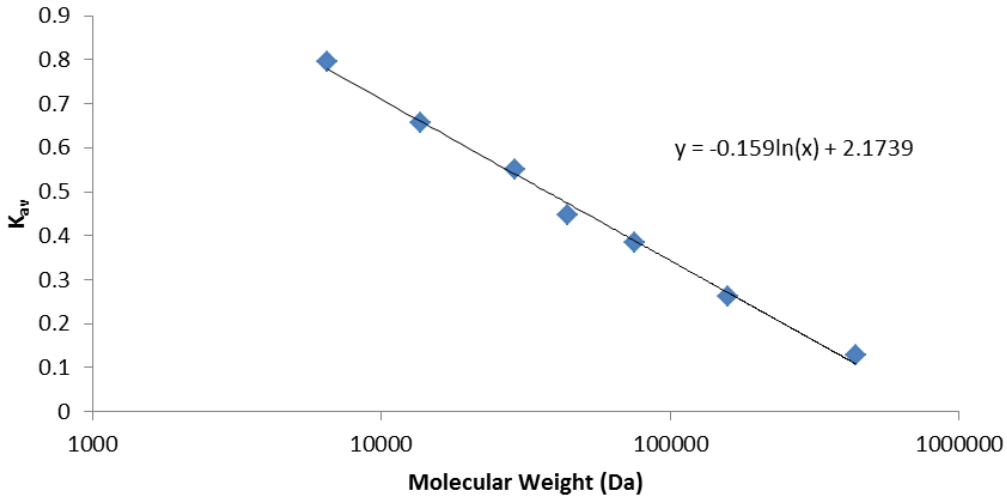

Figure 4.4 Calibration curve for the S200 16/60 size exclusion column. K_{av} is calculated from the equation below.

The elution volumes of the peaks from the size exclusion chromatography purifications were substituted into the following equation:

$K_{av} = (V_e - V_o)/(V_c - V_o)$

Where V_e = elution volume
V_o = column void volume
V_c = geometric column volume

The equation from the calibration curve (Figure 4.4) was rearranged to calculate the molecular weight of the CM-DAH7PS complexes:

Molecular weight = $e^{((K_{av}-2.1739)/-0.159)}$

All the fractions from the large peaks in absorbance (peaks A, B and C) were shown to contain high concentrations of recombinant CM-DAH7PS by measuring the protein concentration using the Bradford assay (section 2.4.1) and by running samples on SDS-PAGE gels (Figure 4.3). The peaks in absorbance are quite low relative to the concentration of protein in the samples as none of the recombinant CM-DAH7PS enzymes contain any tryptophan residues. Peak A is the void volume and the fractions corresponding to this absorbance peak contained primarily large soluble aggregates of CM-DAH7PS. Fractions from the second, much smaller absorbance peak (peak B), contained smaller pure soluble aggregates of CM-DAH7PS, ~400 kDa in size. Fractions from the third

absorbance peak (peak C) contain pure CM-DAH7PS enzyme with a molecular weight of 164 kDa, indicating that CM-DAH7PS in these fractions had likely adopted a homotetrameric quaternary structure (predicted molecular weight of CM-DAH7PS homotetramer is 163 kDa). This is the proposed quaternary structure adopted by the catalytically active *Bsub*-CM-DAH7PS purified by Wu and Woodard (2006). The very slight peak in absorbance at ~84 mL (peak D) represents a molecular weight of ~40 kDa, the predicted molecular weight of the monomeric recombinant CM-DAH7PS enzymes, however the protein concentration was too low to detect any bands on an SDS-PAGE gel. Occasionally, a fifth peak was observed immediately after elution of the homotetrameric CM-DAH7PS between peaks C and D. The fractions from this absorbance peak were also shown to consist of pure CM-DAH7PS by SDS-PAGE analysis and the estimated molecular weight of the protein was ~80 kDa, which is the predicted weight of a homodimeric CM-DAH7PS enzyme (data not shown). DAH7PS enzymatic activity assays were performed with all the different multimers of CM-DAH7PS as described in section 2.4.5. Only homotetrameric CM-DAH7PS displayed any DAH7PS catalytic activity, indicating that *Bacillus* CM-DAH7PS must adopt a homotetrameric structure in order to be catalytically active.

4.3.2 Approaches Used to Obtain Stable, Active CM-DAH7PS Enzymes

A number of different purification buffers, including variations in salt concentration, pH and buffering agent, were tested in order to obtain contemporary and ancestral CM-DAH7PS enzymes which were stable and active. The *Bsub*-CM-DAH7PS and *Bstr*-CM-DAH7PS enzymes were found to be tolerant to a wide range of conditions, however the ancestral enzymes were found to be highly sensitive to the buffer conditions used, in particular to the pH of the solutions. Apart from *Anc5*-CM-DAH7PS, all of the enzymes were stable when purified in 50 mM BTP buffer at pH 9.0 with 500 mM NaCl. The CM-DAH7PS enzymes purified in these buffers also displayed enzymatic activity in 50 mM BTP, 1 mM $ZnSO_4$, pH 9.0 assay buffer. $ZnSO_4$ was added to the assay buffer at a concentration of 1 mM to increase the percentage of CM-DAH7PS enzymes with Zn^{2+} ions bound at their active sites. Metal analysis performed by Wu et al. (2005) showed that as-isolated recombinant *Bsub*-CM-DAH7PS expressed in *E. coli*

contained mostly Zn^{2+} ions and that, when compared with other metal ions, Zn^{2+} bound with the highest affinity.

Anc5-CM-DAH7PS purified in high salt, high pH buffers was very unstable. To improve enzyme stability, protein expression was performed at 18 °C and purification was performed at 4 °C. Glycerol was also added to the purification buffers at a concentration of 5% (v/v) as glycerol can have a stabilising effect on proteins by shifting protein ensembles to more compact states, and has also been shown to prevent some proteins from aggregating (Vagenende et al., 2009). None of these alterations to the expression and purification procedures resulted in *Anc5*-CM-DAH7PS that was stable above 25 °C or displayed any enzymatic activity. The lack of enzymatic activity could indicate that the ancestral inference of this enzyme was inaccurate. The CM-DAH7PS sequences from ancestral and contemporary *Bacillus* were compared to identify differences in *Anc5*-CM-DAH7PS which might explain its lack of activity. T321 is only present in *Anc5*-CM-DAH7PS and the CM-DAH7PS from *B. selenitireducens* (*Bsel*-CM-DAH7PS), which directly branches from the ANC5 ancestral node (Figure 3.2). All of the other enzymes have hydrophobic amino acid residues at this position. This residue is adjacent to the conserved metal binding residue E322. DAH7PS from a selection of species from the Firmicutes phylum, other bacterial phyla and the Euryarchaeota and Crenarchaeota phyla also have hydrophobic residues at this position. As such, the possibility that the enzyme from *B. selenitireducens* might be inactive was explored. *B. selenitireducens* was first isolated from anoxic lake sediment (Blum et al., 1998), therefore it is unlikely that *B. selenitireducens* can obtain all of the aromatic compounds required for cell growth from its environment and a functional shikimate pathway is likely to be essential. The *B. selenitireducens* genome does not appear to contain any additional type I or type II *dah7ps* genes, so the presence of an inactive CM-DAH7PS from *B. selenitireducens* is considered to be unlikely. It was, therefore, thought that the threonine residue in the CM-DAH7PS from *B. selenitireducens* may have been mistakenly annotated at this position due to an error during genome sequencing of this organism. This could then have resulted in threonine being incorrectly inferred at this position in *Anc5*-CM-DAH7PS, resulting in the observed instability and inactivity.

In order to confirm that an error in sequencing had not occurred, and that the *Bsel*-CM-DAH7PS encoded by the gene sequence in GenBank is active, this enzyme was cloned, expressed, purified and assayed. To obtain *Bsel*-CM-DAH7PS, expression in *E. coli* BL21 was induced by the addition of IPTG and the enzyme was purified successfully in the same buffers used to purify the other enzymes. At low enzyme concentrations, *Bsel*-CM-DAH7PS displayed moderate activity but was inactive when concentrated, which was likely the result of enzyme aggregation. The enzyme was very unstable, as activity was lost above 30 °C and there was a rapid loss of activity when the enzymatic activity was assayed at 28 °C. The fact that *Bsel*-CM-DAH7PS displayed enzymatic activity suggests that an error may have occurred in the inference of *Anc5*-CM-DAH7PS at a site other than T321, although analysis of all of the ancestral and contemporary *Bacillus* amino acid sequences did not reveal any obvious candidates.

Subsequent to the analysis of *Bsel*-CM-DAH7PS activity, it was observed that purifications of *Anc2*-CM-DAH7PS, *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS expressed using IPTG-induction were now inactive when concentrated above ~3 mg/mL, when previous purifications of these enzymes had displayed no such limits up to 30 mg/mL. The enzymes were also observed to denature when heated to 40 °C or when stored overnight at room temperature or 4 °C. No such issues were observed with the *Bstr*-CM-DAH7PS and *Bsub*-CM-DAH7PS enzymes. The reason for the sudden change in the properties of the ancestral CM-DAH7PS enzymes was unclear. It was thought that this loss of activity could be due to suboptimal metal ions being bound to the active site of the ancestral enzymes, as the metal analysis of *Bsub*-CM-DAH7PS performed by Wu et al. (2005) had revealed that significant DAH7PS enzymatic activity was only observed when particular metal ions were bound. $ZnSO_4$ was added to the expression media at a concentration of 100 mM to increase the proportion of enzymes with Zn^{2+} bound, however this did not result in active enzyme. In order to ensure the Zn^{2+} ions were actually binding at the active sites of the enzymes, the metals bound to the as-isolated ancestral enzymes were removed using the chelator DPA and replaced with Zn^{2+} ions. This did not restore the activity of the ancestral enzymes. Another possible causative factor for the change in the ancestral enzyme properties that

was investigated was the formation of disulphide bonds within the enzymes. It has been shown previously with the DAH7PS from *M. tuberculosis* that if a reducing agent was not added to the purification buffers, only partially active enzyme was obtained (Webby et al., 2005a). Addition of the reducing agent β-mercaptoethanol to the purification and assay buffers at 2 mM or 5 mM did not restore stability and activity to the ancestral CM-DAH7PS enzymes. Expression of recombinant ancestral CM-DAH7PS enzymes in auto-induction medium was also attempted, however these enzymes were still very unstable and displayed only minimal enzymatic activity.

It was hypothesised that the ancestral CM-DAH7PS enzymes were potentially unstable and minimally active due to improper folding of the proteins within the *E. coli* BL21 cells during protein expression. Soluble expression of type II DAH7PS enzymes from *Helicobacter pylori* and *Actinosynnema pretiosum* has previously been shown to only be possible by co-expression with recombinant GroES and GroEL heat shock proteins (Webby et al., 2005b, Ma et al., 2012). Therefore, this approach was applied to the ancestral CM-DAH7PS enzymes. The chaperone proteins and CM-DAH7PS enzymes were co-expressed using auto-induction medium. There were no observable changes in the amount of tetrameric CM-DAH7PS enzymes purified relative to the amount of aggregated CM-DAH7PS enzymes eluted during size exclusion chromatography. However, the enzymatic activity of *Anc2*-CM-DAH7PS, *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS were restored to previous levels and the ancestral enzymes were again able to be concentrated to high levels without aggregation occurring.

Anc5-CM-DAH7PS co-expressed with GroES and GroEL showed noticeable enzymatic activity. However, the enzyme was still very unstable and noticeably lost activity during kinetic analysis, despite the enzyme being held on ice. This meant that it was not possible to measure the kinetic or thermal properties of this enzyme.

Once active and stable enzyme had been obtained for all of the enzymes, with the exceptions of *Anc5*-CM-DAH7PS and *Bsel*-CM-DAH7PS, these enzymes were biochemically characterised.

4.3.3 Kinetic Analysis

Due to a limited supply of E4P, all analysis was performed at pH 9.0, rather than first determining the respective pH optima of the enzymes and then performing the analyses at their respective pH optima. Initial kinetic analysis with *Bsub*-CM-DAH7PS, *Anc2*-CM-DAH7PS, *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS was performed at 40 °C. It was reasoned that 40 °C was a sufficiently low temperature at which the enzymatic rate of activity of all of the enzymes should be unaffected by enzyme denaturation. The kinetic properties of *Bstr*-CM-DAH7PS was measured at 60 °C as this was considered an appropriate starting temperature for an enzyme from a thermophilic organism.

Significant inhibition of DAH7PS activity of *Anc2*-CM-DAH7PS and *Anc3*-CM-DAH7PS by E4P was observed at higher concentrations of E4P. This phenomenon was not observed with the CM-DAH7PSs from the contemporary species, *B. subtilis* and *B. stearothermophilus* however, interestingly, when the activity of *Bsel*-CM-DAH7PS was tested, the enzyme was also found to be inhibited by high E4P concentrations. This inhibition of activity by E4P and the effect it had on the fitting of Michaelis-Menten curves to the data in the K_M (E4P) plots can be seen in Figure 4.5. Above the concentrations displayed in the plots the inhibition of activity by E4P became very severe, meaning these data points could not be included in the Michaelis-Menten plots.

In order to minimise the use of the limited E4P available, K_M values for (PEP) for the *Bstr*-CM-DAH7PS, *Bsub*-CM-DAH7PS, *Anc2*-CM-DAH7PS and *Anc3*-CM-DAH7PS enzymes were measured by adding E4P at the K_M (E4P) concentrations of the different enzymes and varying the concentrations of PEP added to the reaction mixtures. Michaelis-Menten plots of *Anc2*-CM-DAH7PS and *Anc3*-CM-DAH7PS are provided in Figure 4.5, and plots of *Bsub*-CM-DAH7PS and *Bstr*-CM-DAH7PS are shown later in Figure 4.6 as coincidentally their respective T_{opt} values are near 40 and 60 °C, respectively. Table 4.1 summarises the initial kinetic data collected for these four enzymes. The k_{cat} values were calculated from the data used to determine the E4P K_M as this gave a reliable V_{max} unlike the PEP K_M determination where E4P was rate-limiting. Kinetic analysis was also performed with *Anc4*-CM-DAH7PS, but the

enzyme was not at a high enough concentration to obtain accurate activity measurements at low substrate concentrations. It was evident, however, that the enzymatic activity of *Anc4*-CM-DAH7PS was also inhibited by high concentrations of E4P.

Table 4.1 Initial Michaelis-Menten kinetic properties of CM-DAH7PS enzymes.

	K_M (E4P) (mM)	K_M (PEP) (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (E4P) (s^{-1} mM^{-1})
Bsub-CM-DAH7PS	0.85	0.09	0.9	1.1
Bstr-CM-DAH7PS	0.78	0.15	5.8	7.4
Anc2-CM-DAH7PS	1.47	0.23	2.3	1.6
Anc3-CM-DAH7PS	0.41	0.13	2.1	5.1

The kinetic properties of *Bsub*-CM-DAH7PS, *Anc2*-CM-DAH7PS, and *Anc3*-CM-DAH7PS were determined at 40 °C. The kinetic properties of *Bstr*-CM-DAH7PS were determined at 60 °C.

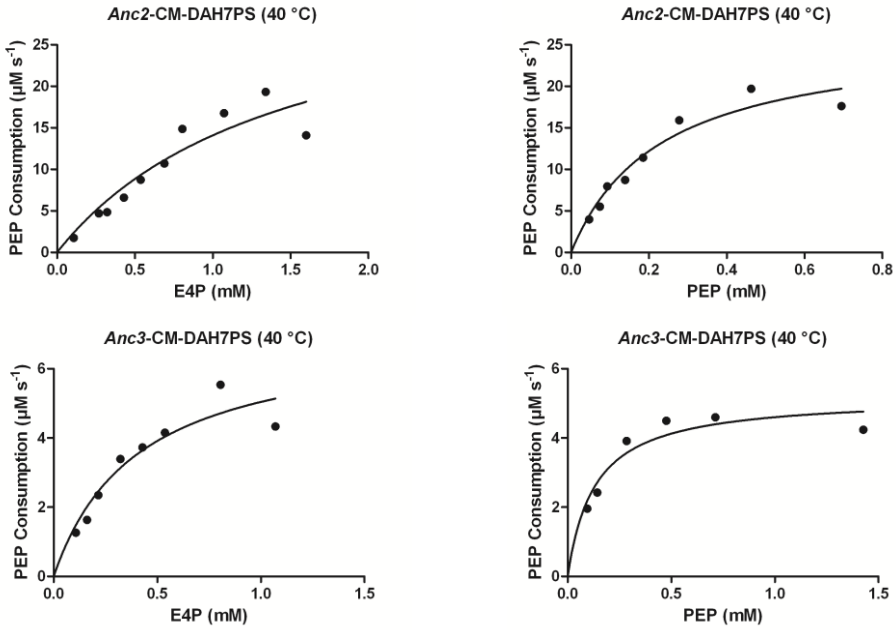

Figure 4.5 Michealis-Menten plots for *Anc2*-CM-DAH7PS and *Anc3*-CM-DAH7PS for E4P and PEP at 40 °C.

The inability to measure the reaction rate of CM-DAH7PS at high concentrations of E4P may have a number of implications for determining the enzymes kinetic properties and T_{opt}. When determining the K_M of one substrate, the other substrate is kept at a constant concentration. If possible, it is desirable for the concentration of this substrate to be added in excess of the K_M for the second substrate so that availability of this substrate is never a rate limiting factor in the reactions. This

enables more accurate V_{max} values to be determined and therefore more accurate K_M and k_{cat} values.

To minimise any potential effect of E4P availability on the rates of activity measured when determining the K_M (PEP) of the enzymes, only the initial rates of activity were measured. Due to the limitations on the concentration of E4P which could be added to the reactions used to determine K_M (PEP), the enzymes V_{max} and k_{cat} values were determined from the K_M (E4P) Michaelis-Menten plots as the maximal rate of activity is likely to be closer to the actual maximal rate of enzymatic activity. The extinction coefficient of PEP is temperature-dependent (Schofield et al., 2004), however over the range of reaction temperatures used in this study the extinction coefficient does not vary greatly so the extinction coefficient was treated as a constant.

Following determination of the T_{opt} of the different enzymes, kinetic analysis at the respective T_{opt} values of the different enzymes was performed. Kinetic analysis of the *Bstr*-CM-DAH7PS and *Bsub*-CM-DAH7PS enzymes at their respective T_{opts} had already been performed, as coincidentally, these two enzymes had already been kinetically characterised near their T_{opts} of 58 °C and 39 °C, respectively. The kinetic properties of *Anc2*-CM-DAH7PS and *Anc4*-CM-DAH7PS were determined near their respective T_{opts} of 48 °C and 39 °C as described in section 2.4.7. A summary of the kinetic data for these enzymes is provided below in Table 4.2.

Table 4.2 Summary of kinetic properties of CM-DAH7PS enzymes.

	K_M (E4P) (mM)	K_M (PEP) (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (E4P) (s^{-1} mM^{-1})
Bsub-CM-DAH7PS	0.85	0.09	0.9	1.1
Bstr-CM-DAH7PS	0.78	0.15	5.8	7.4
Anc2-CM-DAH7PS	1.72	0.55	6.2	3.6
Anc3-CM-DAH7PS	0.41	0.13	2.1	5.1
Anc4-CM-DAH7PS	1.00	0.11	0.2	0.2

All kinetic data were determined near the T_{opt} of the respective enzymes, except for *Anc3*-CM-DAH7PS which did not display Michaelis-Menten kinetics at its T_{opt}. For this enzyme, data at 40 °C are presented.

Despite no evidence of substrate inhibition with PEP at 40 °C, for the determination of the K_M (E4P) for *Anc2*-CM-DAH7PS at 48 °C, PEP could only be used at 3 x K_M rather than 10 x due to inhibition by higher concentrations of PEP. The Michaelis-Menten kinetic constants of *Anc3*-CM-DAH7PS at its T_{opt} of 60 °C were unable to be assessed as the data collected did not fit the Michaelis-Menten equation (Figure 4.6), despite the *Anc3*-CM-DAH7PS activity data determined at 40 °C being amenable to such analysis. The kinetic properties of the enzymes at their respective T_{opt} values (except for *Anc3*-CM-DAH7PS) are summarised in Table 4.3 and the Michaelis-Menten plots are displayed in Figure 4.6.

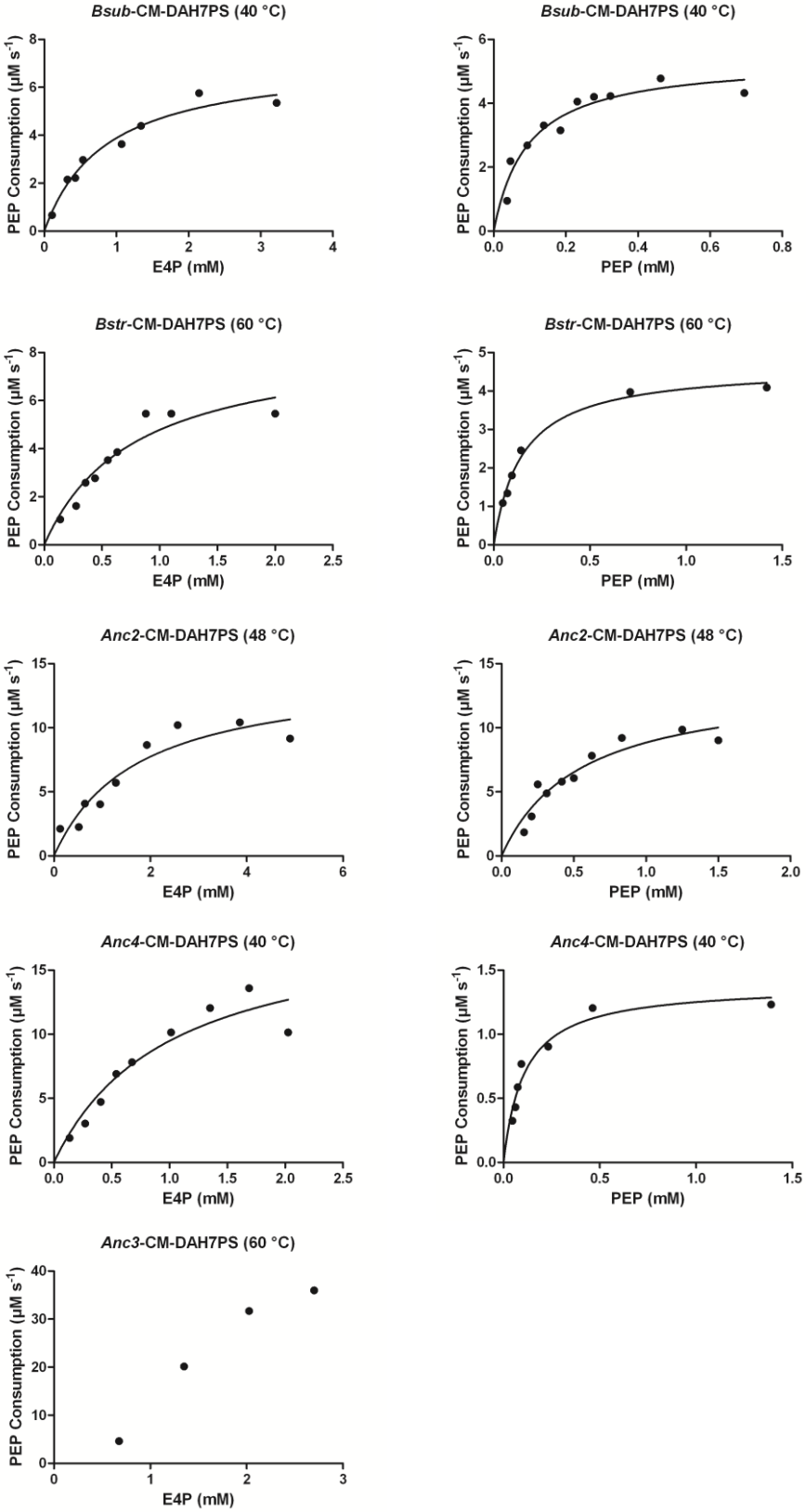

Figure 4.6 Michaelis-Menten plots of CM-DAH7PS near the enzymes' respective T_{opt} values.

4.3.4 T_{opt} Determination

Activity assays were performed to determine the optimal temperature for DAH7PS activity of the different CM-DAH7PS enzymes. The *Bstr*-CM-DAH7PS and *Bsub*-CM-DAH7PS enzymes assayed were expressed using IPTG induction, whereas the *Anc2*-CM-DAH7PS, *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS enzymes were all expressed by auto-induction coupled with co-expression of the GroES and GroEL heat shock proteins.

In the reactions performed to determine the T_{opt} of the *Bstr*-CM-DAH7PS and *Bsub*-CM-DAH7PS enzymes, PEP was used at a concentration 10 x greater than the K_M at 60 °C and 40 °C, respectively, and E4P was used at a concentration 3 x greater than the K_M at 60 °C and 40 °C, respectively. The pH of the assay buffers were adjusted at the reaction temperatures used to allow for ΔpK_a/°C (-0.016; Mohan, 2003). In order to minimise any effect that substrate availability may have on the rate of activity, particularly at higher temperatures only the initial rates of activity were used to determine T_{opt}.

A fixed concentration of E4P could not be used when determining the ancestral enzymes T_{opt} values as the severity of inhibition by E4P changed with temperature. To measure the T_{opt} of the ancestral enzymes, PEP was used at a concentration 10 x greater than the K_M at 40 °C in all of the reactions, except for determination of the T_{opt} of *Anc2*-CM-DAH7PS, for which it was only possible to use PEP at 5 x the K_M (E4P) at 40 °C because higher concentrations inhibited activity. At every temperature measured, multiple reactions were performed with different concentrations of E4P. At each temperature, the reaction which displayed the fastest rate of activity was used to compare to the rates of activity at the different temperatures. The thermoactivity profiles of the CM-DAH7PS enzymes are displayed in Figure 4.7 and the T_{opts} of the CM-DAH7PS enzymes are listed in Table 4.3.

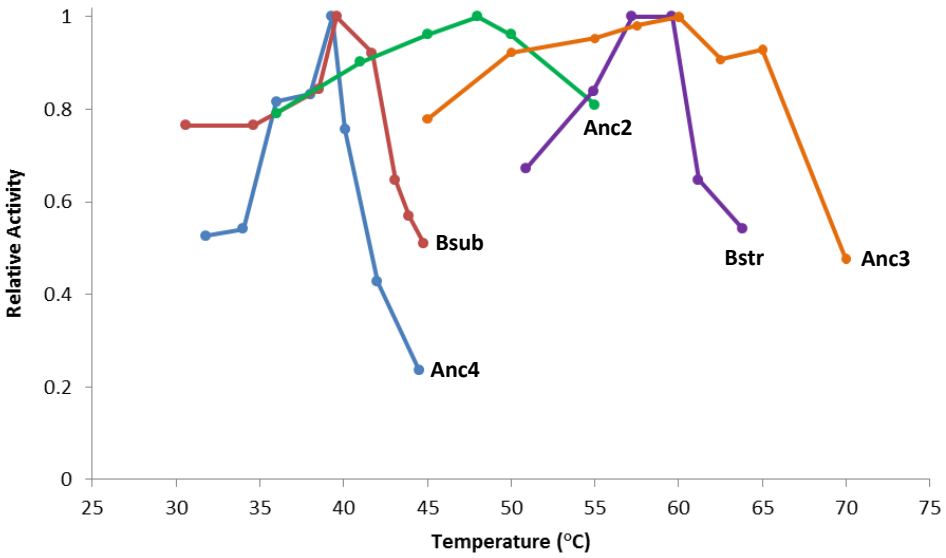

Figure 4.7 Thermoactivity profiles of contemporary and ancestral *Bacillus* CM-DAH7PS enzymes.
The graph demonstrates the relative rates of activity of two contemporary CM-DAH7PS enzymes and three ancestral CM-DAH7PS enzymes at different reaction temperatures. *Anc4*-CM-DAH7PS (blue), *Bsub*-CM-DAH7PS (red), *Anc2*-CM-DAH7PS (green), *Bstr*-CM-DAH7PS (purple) and *Anc3*-CM-DAH7PS (orange).

4.3.5 Real-Time Protein Melt Analysis

T_m data were obtained for some of the CM-DAH7PS by performing real-time protein melt assays (section 2.4.4). The concept behind this method is that temperature-induced denaturation can be monitored by an increase in fluorescence emission of an added hydrophobic fluoroprobe. The fluorescence of the hydrophobic fluoroprobe is quenched when in aqueous solution but as the enzyme begins to denature, exposing the hydrophobic core of the enzyme, the fluoroprobe binds to these exposed hydrophobic surfaces, reducing the quenching of the fluoroprobe and resulting in measurable increases in fluorescence (Ericsson et al., 2006). T_m values estimated using this method have been shown to correlate well with T_m data obtained by circular dichroism and differential scanning calorimetry (DSC) for several proteins (Ericsson et al., 2006, Easter, 2010).

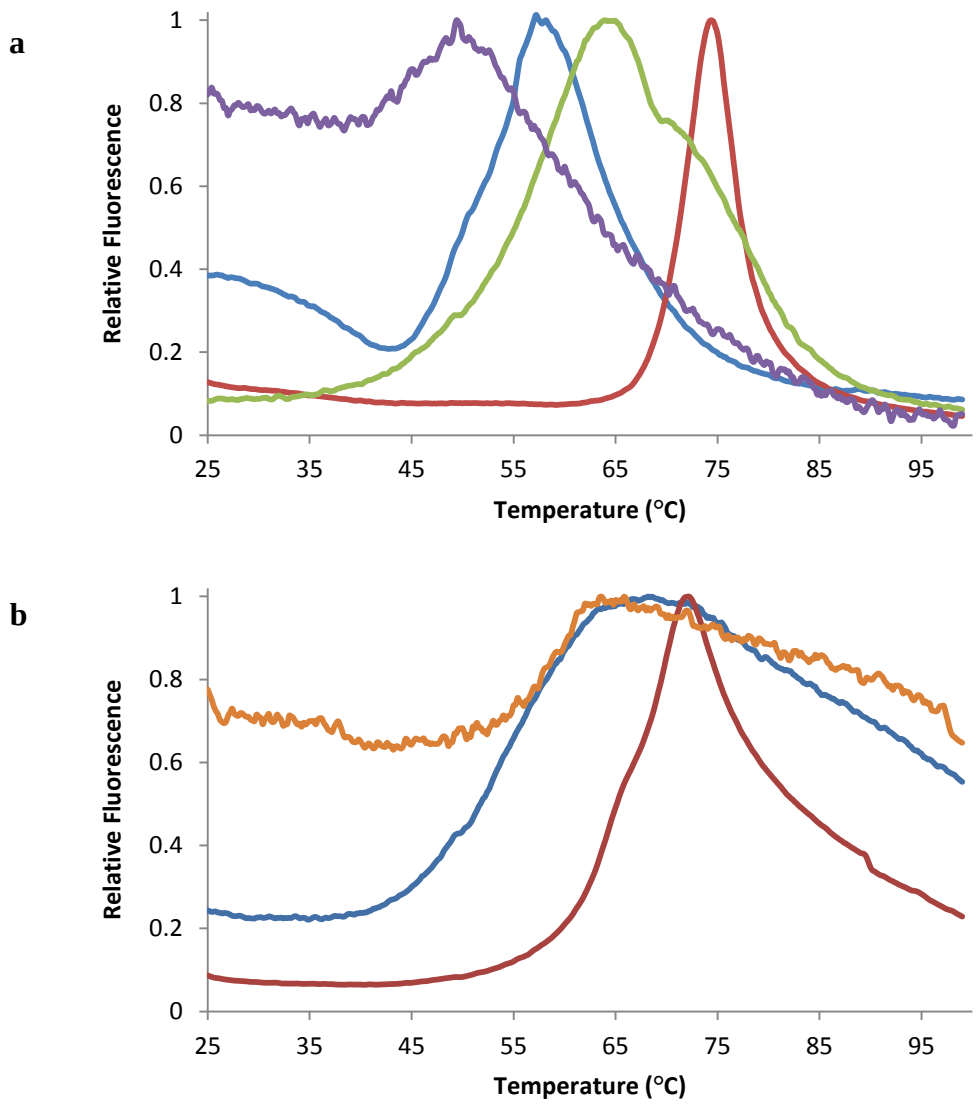

Figure 4.8 Real-time protein melts of CM-DAH7PS enzymes.
(**a**) Real-time melts of *Bacillus* CM-DAH7PS enzymes at pH 7. (**b**) Real-time melts of *Bacillus* CM-DAH7PS enzymes at pH 9. *Bsub*-CM-DAH7PS (blue), *Anc2*-CM-DAH7PS (red), *Bstr*-CM-DAH7PS (green), *Anc4*-CM-DAH7PS (purple) and *Anc3*-CM-DAH7PS (orange).

Protein melt data from which it was possible to estimate the T_m of some of the CM-DAH7PS enzymes were obtained from reactions performed at pH 7.0 and 9.0 (Figure 4.8). T_m values were estimated from the mid-point between the onset point and the top of the emission peak. The emission data resulting from the denaturation of the *Anc4*-CM-DAH7PS and *Anc3*-CM-DAH7PS enzymes were not of high enough quality to estimate a T_m. The T_m values of *Bsub*-CM-DAH7PS, *Bstr*-CM-DAH7PS and *Anc2*-CM-DAH7PS at pH 7.0 were

estimated to be 52 °C, 56 °C and 71.5 °C, respectively, but are unlikely to be the true melting temperatures of the enzymes, as the T_m of the enzymes change with pH, as can be seen by comparing the melt curves of *Anc2*-CM-DAH7PS and *Bsub*-CM-DAH7PS at pH 7.0 with the corresponding curves at pH 9.0 (Figure 4.8). The data suggest, as expected, that *Bstr*-CM-DAH7PS is more thermostable than *Bsub*-CM-DAH7PS, and that *Anc2*-CM-DAH7PS may be more thermostable than the CM-DAH7PS from the thermophilic species *B. stearothermophilus*. The emission data from the real-time protein melts of *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS suggest that these enzymes may both be less thermostable than the *Bsub*-CM-DAH7PS enzyme.

Urea unfolding assays were performed in an attempt to obtain an alternative measure of enzyme stability and ensure the biological feasibility of the ancestral enzymes. Despite the lack of tryptophans in the CM-DAH7PSs, it was hoped that the tyrosine and phenylalanine residues present would provide sufficient intrinsic fluorescence. However, it was found that urea unfolding could not be used to determine the free energy of unfolding for the CM-DAH7PS enzymes as no changes in the intrinsic fluorescence of the enzymes as observable between folded and unfolded enzymes.

4.3.6 Kinetic and Thermal Properties of Ancestral and Contemporary CM-DAH7PS

A summary of the kinetic and thermal properties of contemporary and ancestral *Bacillus* CM-DAH7PS enzymes is provided below in Table 4.3.

Table 4.3 Summary of the kinetic and thermal properties of CM-DAH7PS enzymes.

	K_M (E4P) (mM)	K_M (PEP) (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (E4P) (s^{-1} mM^{-1})	T_{opt} (°C)	T_m (°C)
Bsub	0.85	0.09	0.9	1.1	40	52
Bstr	0.78	0.15	5.8	7.4	58	56
Anc2	1.72	0.55	6.2	3.6	48	72
Anc3	0.41	0.13	2.1	5.1	60	-
Anc4	1.00	0.11	0.2	0.2	39	-

All kinetic data were determined near the T_{opt} of the respective enzymes, except for *Anc3*-CM-DAH7PS which did not display Michaelis-Menten kinetics at its T_{opt}. For this enzyme, data at 40 °C are presented.

The kinetic properties of the ancestral enzymes are similar to those of the contemporary enzymes, indicating that *Anc2*-CM-DAH7PS, *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS are biologically realistic enzymes. The ancestral enzymes do, however, display some differences in their kinetic properties. *Anc2*-CM-DAH7PS has a K_M for PEP more than three-fold higher than any of the other enzymes and the k_{cat} of *Anc4*-CM-DAH7PS is four-fold lower than any of the other enzymes.

When the T_{opt} and T_m data for each enzyme are compared, a discordance is observed. The difference of 24 °C between the T_{opt} and T_m of *Anc2*-CM-DAH7PS is very large compared to a difference of only 12 °C between these values for the *Bsub*-CM-DAH7PS enzyme. It is evident from the real-time protein melt curves of the different enzymes (Figure 4.8) that the estimated T_m from the real-time protein melt curves are unlikely to be accurate estimates of the actual T_m of the CM-DAH7PS enzymes as the melting profiles of the different enzymes are very different. There is also no clear correlation between the T_{opt} and the estimated T_m

of the enzymes, with the T_m of *Bstr*-CM-DAH7PS being 2 °C lower than the T_{opt} of the enzyme, whereas the T_{opt} of *Anc2*-CM-DAH7PS is 24 °C lower than the estimated T_m of this enzyme. It has been shown that the T_m of a number of proteins estimated using the real-time protein melt assay were similar to the T_m determined using more robust techniques (Easter, 2010, Ericsson et al., 2006). This, however, does not appear likely to be the case for CM-DAH7PS enzymes from *Bacillus*. The inability to obtain reliable data for *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS using the real-time protein melt assay also indicates that this is not a suitable method for estimating the T_m of *Bacillus* CM-DAH7PS enzymes. The real-time protein melt assay relies on the increase in fluorescence emission upon binding of the hydrophobic fluoroprobe to exposed hydrophobic surfaces of the enzymes as a proxy for protein unfolding. This may not be the case for all proteins. This method has been found to produce poor thermal unfolding data for other proteins, namely Lsr2 and a *Bacillus* α-glucosidase (Joanne Hobbs and Emma Summers, personal communication).

The T_{opt} of the CM-DAH7PSs from the contemporary species, *B. subtilis* and *B. stearothermophilus*, appear to be correlated with the OGT of the host organisms, with the enzyme from the thermophilic *B. stearothermophilus* demonstrating maximum catalytic activity at a temperature 18 °C higher than the enzyme from the mesophilic organism *B. subtilis*. This means that the T_{opt} of the ancestral CM-DAH7PS enzymes could potentially be used to infer the OGT of the ancestral host organisms. There are, however, some limitations associated with the determination of the T_{opt} values of the ancestral CM-DAH7PS enzymes that need to first be acknowledged. The main limitation which may have affected the accuracy of some of the determined T_{opt} values of the ancestral enzymes is the inability to add E4P at a concentration largely in excess of the K_M (E4P) due to the severe substrate inhibition observed. If all of the enzymes were affected to the same degree by substrate inhibition this would be less of an issue when using this data to infer how thermophily may have evolved in *Bacillus*. However, this is not the case with the T_{opt} values determined in this study, as the activity of the contemporary enzymes was not inhibited by high concentrations of E4P. The degree of substrate inhibition was also not consistent for all of the ancestral enzymes. This means that the rate of activity of some of the enzymes may have

been affected to different degrees by substrate availability, potentially affecting the magnitude of any observed differences in the T_{opt} of the enzymes. This may also mean that the T_{opt} values determined for some of the enzymes may be artificially low, as the slight increase in enzymatic activity due to an increase in reaction temperature may have been counteracted by the increase in the kinetic constants of the enzymes. An increase in the kinetic constants as the temperature increased has been observed with the DAH7PS from *T. maritima*, with the K_M (E4P) in particular being greatly affected and increasing three-fold when the reaction temperature was raised by 10 °C (Wu et al., 2003). This effect may have been greatest for *Anc2*-CM-DAH7PS because only 3 x the K_M (PEP) concentration could be used in the reactions. This may have resulted in the T_{opt} of the enzyme being lowered relative to the T_{opt} of the other enzymes, and may explain the large discrepancy between the T_{opt} and T_m of this enzyme. E4P was added at 2 x the K_M (E4P) concentration at 48 °C because of substrate inhibition, meaning that the concentration of E4P used to determine the T_{opt} of *Anc2*-CM-DAH7PS is lower than for any of the other enzymes. The higher kinetic binding constants of *Anc2*-CM-DAH7PS may have resulted in the determined T_{opt} of this enzyme being affected to a greater extent than the other enzymes, which may result in inaccurate inferences about the evolution of thermophily within the *Bacillus* genus. At higher temperatures, the instability of E4P could also have affected the rate of enzymatic activity, with E4P shown to be very unstable above 60 °C (Schofield et al., 2004). This could have the effect of artificially lowering the T_{opt} of enzymes that have T_{opt} values at or above 60 °C, which may have had the effect of lowering the T_{opt} of *Anc3*-CM-DAH7PS and *Bstr*-CM-DAH7PS slightly.

Thus far, three ancestral *Bacillus* CM-DAH7PS enzymes (*Anc2*-CM-DAH7PS, *Anc3*-CM-DAH7PS and *Anc4*-CM-DAH7PS) estimated to be 570, 650 and 950 Myr old, respectively, have been successfully reconstructed and biochemically characterised. These enzymes are the largest and most structurally complex Precambrian enzymes reconstructed to date which, based on their similar kinetic properties to the CM-DAH7PS enzymes from *B. subtilis* and *B. stearothermophilus*, appear to have been accurately reconstructed. The T_{opt} of the two contemporary enzymes *Bsub*-CM-DAH7PS and *Bstr*-CM-DAH7PS are

correlated with the OGT of the host organisms, meaning that with the aforementioned limitations associated with determining the T_{opt} of the ancestral CM-DAH7PS enzymes in mind, some preliminary inferences about the origin and evolution of thermophily within the *Bacillus* genus can be made. In contrast to the data of Hobbs et al. (2012), which suggested that a thermophilic ancestral *Bacillus* species existed ~950 Mya, a mesophilic ancestor is proposed to have existed at this time based on the T_{opt} of *Anc4*-CM-DAH7PS. The CM-DAH7PS T_{opt} data suggest that thermophily within the *Bacillus* genus evolved between 950 and 650 Mya before the thermophilic nature began to decrease again between 650 and 570 Mya to a moderately thermophilic ancestral species that existed ~570 Mya. It is difficult to infer an evolutionary trend from only three data points however, if the data are compared to the thermal data of the ancestral IPMDH from *Bacillus* determined by Hobbs et al. (2012), it is evident that different inferences about the evolution of thermophily within the *Bacillus* genus could be made. The ANC2, ANC3 and ANC4 ancestral nodes from both the IPMDH and CM-DAH7PS *Bacillus* phylogenetic trees are in similar positions within the tree in terms of the clades which have diverged from these different nodes, however the ages of the ANC2 and ANC3 ancestral nodes differ greatly between the two trees, with the IPMDH nodes estimated to be 250 Myr and 200 Myr older than the corresponding nodes from the CM-DAH7PS chronogram. When the thermal profiles of *Anc2*-CM-DAH7PS and *Anc3*-CM-DAH7PS are compared with those of *Anc2*-IPMDH and *Anc3*-IPMDH, similar decreases in the T_{opt} are observed from ANC3 to ANC2. Based on the thermal properties of the different CM-DAH7PS and IPMDH contemporary enzymes, *Anc3*-CM-DAH7PS has a thermophilic profile and *Anc2*-CM-DAH7PS has a moderately thermophilic profile, whereas *Anc3*-IPMDH has a moderately thermophilic profile and *Anc2*-IPMDH has a mesophilic/psychrophilic profile. If the thermal profiles of the CM-DAH7PS and IPMDH ancestral enzymes are compared based on the ages of the ancestral nodes (Figure 4.9), a direct comparison is more difficult. The ages of the ANC4 nodes are the same, as this node was one of the fixed calibration points used to convert the phylograms to chronograms. The age of the ANC3 node from the CM-DAH7PS chronogram is approximately the same as the ANC1 node from the IPMDH chronogram from Hobbs et al. (2012). Both of the enzymes reconstructed from these nodes display similar thermophilic profiles. The ANC2 node from the

CM-DAH7PS chronogram is the most evolutionarily recent ancestral node to be reconstructed and is ~100 Myr younger than *Anc1*-IPMDH, the most recent ancestral IPMDH reconstructed. The ANC2 CM-DAH7PS node cannot be directly compared with any of the IPMDH nodes in terms of the estimated times of divergence, however it does indicate that a decrease in the thermophilic nature of some ancestral *Bacillus* species may have occurred between ~650 Mya and 570 Mya.

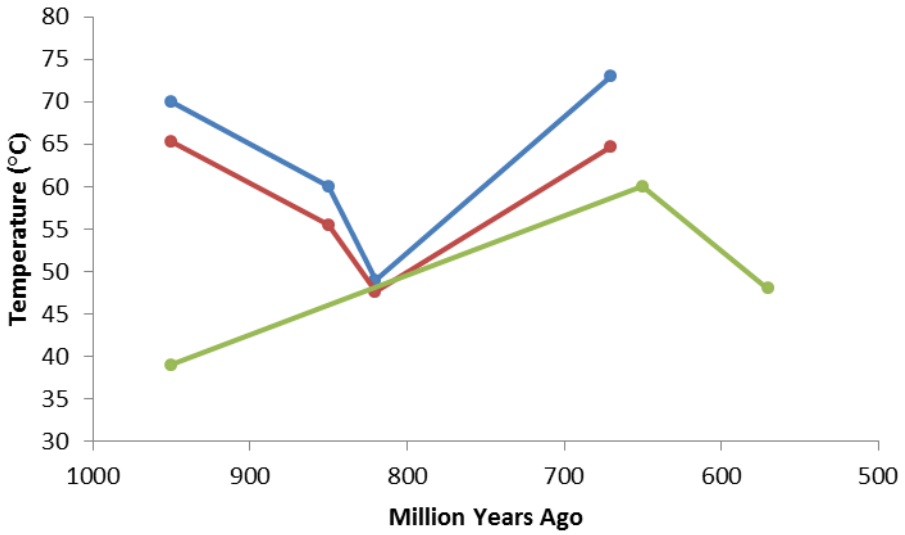

Figure 4.9 Trends in the thermal adaptation of reconstructed ancestral *Bacillus* CM-DAH7PS and IPMDH enzymes over evolutionary time.
T_m of ancestral IPMDH enzymes (red), T_{opt} of ancestral IPMDH enzymes (blue) and T_{opt} of ancestral CM-DAH7PS enzymes (green) are plotted against their estimated ages. IPMDH thermal data is from Hobbs et al. (2012).

These preliminary results suggest that thermophily is not strictly a primitive trait, but rather may be the result of much more recent adaptations, with this adaptation occurring in the *Bacillus* genus between approximately 950 and 650 Mya. However, as mentioned previously, there are limitations to the accuracy of the T_{opt} measurements therefore this is only a preliminary hypothesis. In order to support the inferences made regarding the evolution of thermophily within the *Bacillus* genus from the ancestral CM-DAH7PS T_{opt}, robust T_m values for the contemporary and ancestral CM-DAH7PS need to be determined using DSC.

The possible mesophilic origin for *Bacillus* suggested by the T_{opt} determined for the *Anc4*-CM-DAH7PS enzyme would also be further supported by thermal and

kinetic data for *Anc5*-CM-DAH7PS. *Anc5*-CM-DAH7PS is only 129 Myr older than *Anc4*-CM-DAH7PS, so if *Anc5*-CM-DAH7PS displayed a similar mesophilic profile this would support a mesophilic origin for the *Bacillus* genus and increase the confidence in the accuracy of the reconstructed *Anc4*-CM-DAH7PS and its determined T_{opt}. Further support for the accuracy of *Anc4*-CM-DAH7PS is desirable, as although the kinetic properties of *Anc4*-CM-DAH7PS are similar to those of the contemporary and other ancestral CM-DAH7PS enzymes, it does appear to be slower and less catalytically efficient than the other enzymes, although not drastically so. It could be argued from these results, and the fact that stable *Anc5*-CM-DAH7PS has yet to be purified, that small errors may have accumulated during the inference of these older ancestral enzymes, resulting in inaccurately inferred enzymes which are less stable as a result of these errors, rather than indicating an actual mesophilic origin for the *Bacillus* genus. However, the sequence conservation for CM-DAH7PS is high compared with other reconstructed enzymes so this is unlikely. This might then suggest that there are limits to how structurally complex ancestral enzymes can be before errors in ancestral inference begin to significantly affect the properties of reconstructed Precambrian enzymes. From the kinetic and thermal data collected thus far this possibility cannot be excluded. If the mesophilic nature inferred by the T_{opt} of *Anc4*-CM-DAH7PS is found to likely be spurious, then the thermal data of the ancestral CM-DAH7PS enzymes would not conflict with the data of Hobbs et al. (2012). However, *Bsel*-CM-DAH7PS appears to display similar properties with *Anc4*-CM-DAH7PS and *Anc5*-CM-DAH7PS in regards to stability and enzyme aggregation which suggests that the stability issues associated with these ancestral enzymes are unlikely to be due to inaccuracies in the ancestral inference. If the mesophilic nature of *Anc4*-CM-DAH7PS suggested by its T_{opt} is supported by a robust $T_{m,}$ and if stable *Anc5*-CM-DAH7PS is able to be characterised and found to be biologically feasible, the likelihood of the suggested mesophilic properties of the *Anc4*-CM-DAH7PS being due to errors in the ancestral inference will be low. If the mesophilic nature of *Anc4*-CM-DAH7PS is not due to errors in ancestral inference, this has a number of implications for ASR studies which have used the thermostabilities of ancestral proteins to infer the evolution of thermophily. This could indicate either that evolutionary trends observed from ASR studies are highly sensitive to the contemporary species used in the ancestral

inferences, or that any trends observed are protein dependent and reconstruction of different proteins may lead to different inferences being made regarding the origin and evolution of thermophily. Li et al. (2010) found that the posterior probabilities of ML inferred ancestral sequences did not change significantly when a few taxa were added to, or removed, from ancestral inferences. However, the effect that incorporating sequences from more or different species with different biochemical properties in ancestral inferences has on the biochemical properties of reconstructed ancestral proteins has yet to be investigated. There are a number of differences in the contemporary *Bacillus* species used to infer the ancestral CM-DAH7PSs in the present study and the ancestral IPMDHs in the study of Hobbs et al. (2012), which might explain the difference in the proposed temperature profiles of the *Anc4*-CM-DAH7PS and *Anc4*-IPMDH enzymes. If the biochemical properties of ancestral enzymes is shown to be highly dependent on the biochemical properties of the contemporary proteins used to infer the ancestral sequences, this would suggest that unless sequences are available for all of the different species which have diverged from a particular ancestral node, any inferences regarding the evolution of thermophily will be biased by species selection. If reconstructed ancestral enzymes from different enzyme families are shown to result in different inferences being made about the evolution of thermophily, then either reconstructed proteins should not be used to infer how thermophily may have evolved or it may be that, in order to have a high level of confidence in any inferences made regarding the evolution of thermophily, a number of different types of ancestral enzymes must be reconstructed and biochemically characterised to present a more robust inference.

Determination of the 3D X-ray crystal structures of the ancestral and contemporary *Bacillus* CM-DAH7PS could potentially provide further indication as to the accuracy of the reconstructed ancestral CM-DAH7PS enzymes, as errors in ancestral inference may be revealed by any major structural changes between the ancestral enzymes, the contemporary *Bacillus* enzymes and other DAH7PS structures in the PDB.

5 X-Ray Crystallography of CM-DAH7PS

5.1 Introduction

X-ray crystallography is the technique most commonly used to determine 3D structures of proteins. To perform X-ray crystallography a protein crystal that diffracts X-rays is required. Photons from an X-ray beam are diffracted by the electrons of atoms within the protein crystal and these diffraction patterns are collected at many different angles. Three parameters define each diffraction spot: amplitude, position and phase. The amplitude and position of each spot is easily determined from the X-ray diffraction pattern, however the phase cannot be determined directly from the data. One method for determining the phase of the data is molecular replacement, which uses the phase information from a related, previously solved 3D protein structure to infer the phase of the data. Molecular replacement can usually be used to provide phase information if the sequence identity between the model protein structure and the target protein is greater than 30%. An electron density map can then be reconstructed from this data and a 3D structural model of the target protein can then be built into the data.

There are currently no 3D structures of *Bacillus* CM-DAH7PS enzymes in the PDB. X-ray crystallography was performed in order to determine how the different domains are orientated in CM-DAH7PS from *Bacillus* species and if any differences exist between the ancestral and contemporary structures. Determination of the X-ray crystal structures of the ancestral enzymes would also enable the accuracy of ancestral reconstruction to be assessed in terms of their structure, by comparing them to the structures of the contemporary *Bacillus* enzymes and previously determined DAH7PS structures in the PDB from other contemporary species. Changes in enzyme thermostabilities may also be able to be structurally rationalised through structural comparisons of enzymes with different thermostabilities.

5.2 Results and Discussion

5.2.1 Crystallisation of CM-DAH7PS Enzymes

Crystal screens were set up for *Bstr*-CM-DAH7PS, *Anc2*-CM-DAH7PS, and *Anc4*-CM-DAH7PS. Several promising conditions were observed in the initial crystallisation screens with *Bstr*-CM-DAH7PS and *Anc2*-CM-DAH7PS. Crystallisation conditions were optimised by screening around these conditions using the hanging drop method (section 2.5.3). An example of CM-DAH7PS crystal growth is shown in Figure 5.1. Large *Bstr*-CM-DAH7PS crystals were observed in 0.1 M trisodium citrate pH 5.0, 27% (v/v) PEG MME 550 (Figure 5.1). These crystals diffracted to ~7 Å resolution. *Anc2*-CM-DAH7PS crystals formed in screens around the following conditions: 0.1 M Bicine pH 8.0, 40% (w/v) PEG 5000; and 0.2 M ammonium acetate, 0.1 M BIS-TRIS pH 6.5, 20% (w/v) PEG 3350.

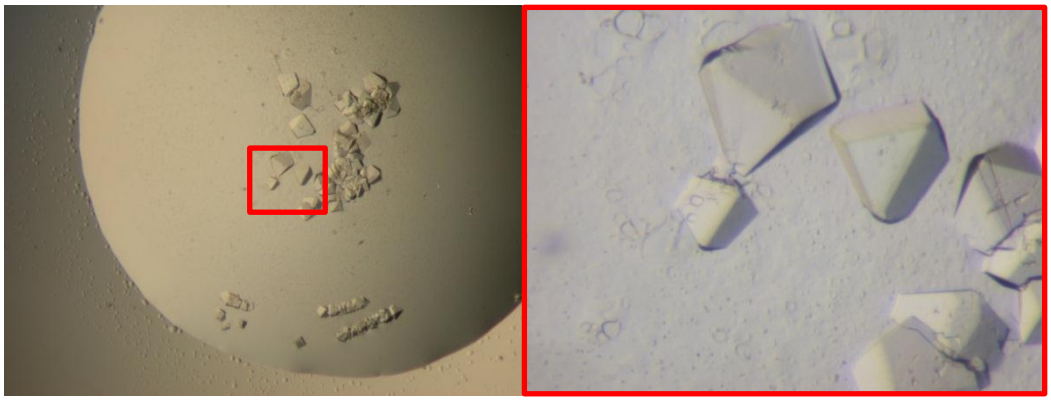

Figure 5.1 *Bstr*-CM-DAH7PS crystals.
Bstr-CM-DAH7PS crystals grown in 0.1 M trisodium citrate pH 5.0, 27% (v/v) PEG MME 550

No promising crystallisation conditions were observed in the initial *Anc4*-CM-DAH7PS crystallisation screen and heavy protein precipitation was observed in a high percentage of the conditions. The high percentage of conditions with heavy protein precipitate is likely due to too high a protein concentration, as the same phenomenon was observed when highly concentrated *Bstr*-CM-DAH7PS was screened. Further screens should be performed with lower concentrations of *Anc4*-CM-DAH7PS. *Anc4*-CM-DAH7PS co-expressed with GroES and GroEL should also be used in future trials instead of

Anc4-CM-DAH7PS expressed using IPTG-induction as the high level of precipitation could also have been caused by the instability of the *Anc4*-CM-DAH7PS enzyme that was observed shortly after this screen was performed.

5.2.2 X-Ray Diffraction

Anc2-CM-DAH7PS crystals from the precipitant condition 0.2 M ammonium acetate, 0.1 M BIS-TRIS pH 7.0, 20% (w/v) PEG 3350 were sent to the Australian Synchrotron and a data set was collected as described in section 2.5.5, with one of the crystals producing X-ray diffraction data to 2.2 Å resolution (Figure 5.2). Images were taken over 720° in 1° increments with 2 s exposures at a detector distance of 180 mm. The space group was determined with MOSFLM using images 90° apart. The unit cell was found to be monoclinic with the space group $P2_1$. No deterioration in the X-ray diffraction was observed over the entire 720 images, so all of the images were used to integrate the data set with MOSFLM.

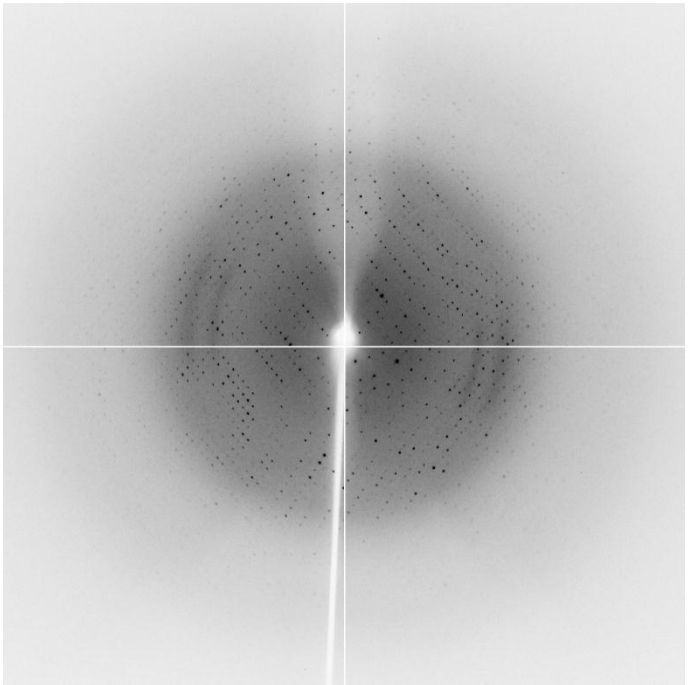

Figure 5.2 X-ray diffraction pattern of *Anc2*-CM-DAH7PS. The image boundary represents 1.9 Å.

5.2.3 Data Processing

Matthews coefficient analysis predicted four molecules in the asymmetric unit. The integrated reflections were scaled and merged in SCALA within CCP4 with

an upper resolution limit of 2.2 Å. Data collection statistics are provided in Table 5.1.

Table 5.1 Data collection statistics for *Anc2*-CM-DAH7PS.

Data Collection Statistics	
Space group	$P2_1$
Wavelength (Å)	0.9536
Cell dimensions	
a b c	81.97/92.26/106.10
α β γ	90.0/110.82/90.0
Molecules in asymmetric unit	4
Mosaicity	0.91
Resolution range (Å)	34.10-2.20 (2.32-2.20)
No. of observed reflections	1107604 (159328)
No. of unique reflections	75064 (10916)
R_{merge}	0.133 (0.517)
Mean I/σI	19.4 (5.8)
Completeness	100 (100)
Multiplicity	14.8 (14.6)

Values in parentheses are for the highest resolution shell

5.2.4 Molecular Replacement

The structure of *Anc2*-CM-DAH7PS was determined by molecular replacement using the two CM-DAH7PS structures from *L. monocytogenes* as models (PDB codes 3NVT and 3TFC). *Lm*-CM-DAH7PS shares 73.8% sequence identity with *Anc2*-CM-DAH7PS. The CM (1-87) and DAH7PS (103-366) domains were truncated and modelled separately, as the two domains are connected by a linker region which was suspected to be highly flexible. The DAH7PS domains were modelled first in PHASER within PHENIX, using the DAH7PS domain of 3NVT, followed by molecular replacement of the CM domains. Four DAH7PS domains were successfully modelled but only one CM domain was placed successfully. Molecular replacement was then attempted by fixing the four modelled *Anc2*-DAH7PS domains first and then searching for the CM domains, however this was not possible within PHENIX so molecular replacement of the CM domains was then attempted using PHASER within CCP4. Again, only one CM domain was modelled poorly. It was thought that the three helices within the CM domains may be arranged in a different manner in *Anc2*-CM-DAH7PS compared to *Lm*-CM-DAH7PS, so modelling of the three CM helices was attempted separately, however this was also unsuccessful. Molecular replacement of the CM

domains was then attempted using the CM domain of 3TFC as a model and on this occasion a number of CM domains were successfully modelled. There are no obvious structural differences in the orientation or organisation of the CM domains between the 3NVT and 3TFC structures, so the reason why molecular replacement was successful with the CM domain of 3TFC but not 3NVT is unclear. Two CM domains were modelled correctly in the initial molecular replacement, while a third was modelled using the symmetry mate of one of the CM domains. This resulted in a model with a CM dimer on one side of the core DAH7PS tetramer and a single CM domain on the opposite side. In order to model the remaining CM domain, the CM dimer from one side of the catalytic DAH7PS core was superimposed onto the CM monomer on the opposite side by secondary structure matching (SSM). This resulted in an initial model with a tetrameric catalytic DAH7PS core with two symmetrical six-helix CM dimers positioned on opposite sides of the DAH7PS core without any domain-domain linker regions modelled.

5.2.5 Model Building and Refinement

The DAH7PS domains were built using the PHENIX AutoBuild wizard and the CM domains were modelled separately using PHASER within CCP4. The linker regions between the CM and DAH7PS domains were built manually within COOT into $2|F_O|-|F_C|$ and $|F_O|-|F_C|$ maps contoured to 1 σ and 3 σ, respectively. N- and C-terminal residues, and a number of residues from the CM domain loops, were also manually built into the model. The model was refined using Refmac 5.0.

The final model has an R-factor of 17.8% and an R_{free} of 21.6% (Table 5.2). There were a large number of residues from the CM domains that were not able to be resolved. These residues were primarily part of the α1'-α2' loop, the α2' helix of the CM domains, or the domain-domain linker regions. There was also a lack of side chain density for many of the residues in the CM domains. In total, 1276 out of 1464 residues were able to be successfully modelled. *Anc2*-CM-DAH7PS is composed of four chains (monomers) which each have a different number of residues that were able to be resolved. *Anc2*-CM-DAH7PS monomers are composed of 366 amino residues, however not all of these residues were able to

be modelled from the X-ray diffraction data collected. The chain A, B, C and D models are composed of 306, 329, 294 and 347 amino residues, respectively. The 3D structure of the most complete CM-DAH7PS monomer (chain D) is shown in Figure 5.3. There were no specific metals added during expression and purification of the *Anc2*-CM-DAH7PS enzyme which was crystallised. However, there were areas of electron density observed at the metal binding sites of the DAH7PS domains. A number of different metal ions were modelled into the electron density at the metal binding sites of the four DAH7PS domains. It was found that Mn^{2+} ions fit the density the best out of all the metals modelled at all of the active sites. The average B-factor of the overall structure is 24.2 $Å^2$. Ramachandran analysis of the φ and ψ backbone torsion angles revealed that 97.05% residues are in preferred regions, 2.87% are in allowed regions and 0.08% are in disallowed regions (Table 5.2).

Table 5.2 Refinement and model statistics.

Refinement and Model Statistics	
R-factor	17.8
R_{free}	21.6
Total No. of atoms	10154
No. of protein atoms	9700
Other molecules/ions	20
No. of waters	367
RMSD	
Bond lengths (Å)	0.025
Bond angles (°)	1.938
Average B-factors ($Å^2$)	
Overall	24.2
Chain A	23.2
Chain B	24.9
Chain C	22.7
Chain D	25.6
Waters	22.0
Mn	41.5
Ramachandran analysis	
Percentage in favoured regions	97.05
Percentage in allowed regions	2.87
Percentage in disallowed regions	0.08

5.2.6 Structure of *Anc2*-CM-DAH7PS

The X-ray crystal structure of *Anc2*-CM-DAH7PS was solved at 2.2 Å resolution. The enzyme has a homotetramic quaternary structure with Mn^{2+} bound at the four separate DAH7PS active sites (Figure 5.4). The four DAH7PS domains are assembled in a similar manner to *Tm*-DAH7PS and in a near identical fashion to *Lm*-CM-DAH7PS, but in a different manner to *Pf*-DAH7PS and *Ap*-DAH7PS which form homotetramers when crystallised, but are homodimeric in solution (Schofield et al., 2005, Zhou et al., 2012). Like the *Lm*-CM-DAH7PS structures (PDB codes 3NVT and 3TFC) the CM domains in *Anc2*-CM-DAH7PS from diagonally opposite DAH7PS domains form dimers positioned on opposite sides of the DAH7PS core tetramer.

The core $(\beta/\alpha)_8$-barrel is very similar to the previously solved structures of type Iβ DAH7PS enzymes from *L. monocytogenes* (PDB codes 3TFC and 3NVT), *P. furiosus* (PDB code 1ZCO), *A. pernix* (PDB code 1VS1) and *T. maritima* (1RZM), which share 73.8%, 52.8%, 45.9% and 43.9% sequence identity with *Anc2*-CM-DAH7PS, respectively. The active site is located at the C-terminal end of the $(\beta/\alpha)_8$-barrel as demonstrated by the positioning of the Mn^{2+} ions within the structure. As can be seen in Figure 5.4, the six helix CM dimers are positioned asymmetrically relative to the DAH7PS tetrameric core. The surface helices of the DAH7PS domains and the CM domains in particular have high B-factors (Figure 5.5). These high B-factors are expected as surface atoms are generally expected to have a higher B-factors than atoms packed within the core of the protein, and high B-factors have been observed for the CM domains of the CM-DAH7PS from *L. monocytogenes* (Light et al., 2012). The CM domains of chains A and C have especially high B-factors with large portions of these domains being unresolved (Figure 5.5). The high B-factors may be the result of disorder in the crystals with respect to the CM domains.

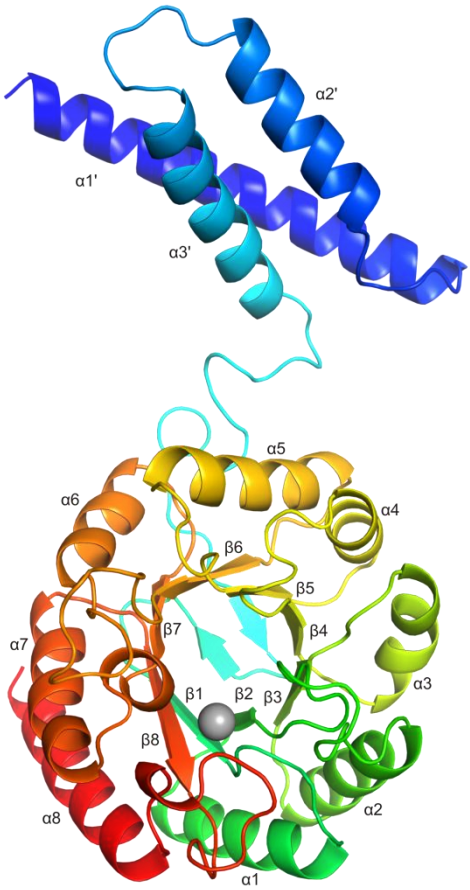

Figure 5.3 Cartoon representation of *Anc2*-CM-DAH7PS monomer.
Rainbow-coloured view of *Anc2*-CM-DAH7PS monomer (chain D) looking in from the C-terminal end of the $(\beta/\alpha)_8$-barrel. Rainbow colouration starts with the N-terminal α1' CM helix coloured blue and ends with the α8 DAH7PS helix coloured red. The grey sphere is the bound Mn^{2+} ion.

In chain A, four N-terminal residues, the C-terminal half of the CM α1' helix, the α1'-α2' loop and the α2' helix (residues 26-61), and most of the domain-domain linker region (residues 82-91) are absent. In chain B, 15 N-terminal residues, the end of α1' and the majority of the α1'-α2' loop (residues 34-43), and three residues from the α2'-α3' loop (residues 63-65) are absent. In chain C, nine N-terminal residues, the C-terminal half of α1', the α1'-α2' loop and α2' helix (residues 24-63), and the domain-domain linker region (residues 79-92) are absent. In chain D, only the first ten N-terminal residues are absent from the CM domain (Figure 5.3). In all of the chains, the 11 C-terminal residues (residues 358-366) are absent, of which eight are additional residues which were added to allow

cloning and purification of the enzyme. There were 67 residues in the final model for which there was insufficient density to support the orientation of the side chain rotamers so the side chains of these residues were modelled with zero occupancy.

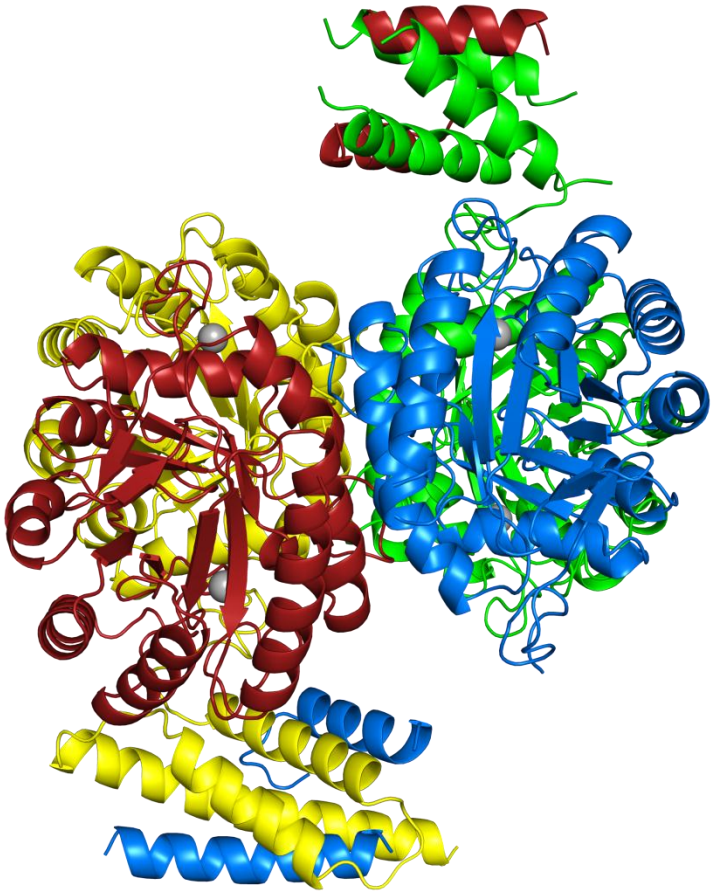

Figure 5.4 Cartoon representation of *Anc2*-CM-DAH7PS homotetramer. Chains A, B, C and D are coloured blue, green, red and yellow, respectively and the grey spheres represent Mn^{2+} ions.

5.2.6.1 Oligomeric Assembly

The oligomeric assembly of the DAH7PS domains occurs in a similar fashion to the DAH7PS from *T. maritima* and the CM-DAH7PS from *L. monocytogenes* in particular (Light et al., 2012, Shumilin et al., 2004). *Anc2*-CM-DAH7PS forms a dimer-of-dimers quaternary structure. Monomers associate to form homodimers through extensive interactions between the α4 and α5 helices, and the β2-α2, β3-α3, β4-α4, β5-α5 and β6-α6 loops at the C-terminal end of the $(\beta/\alpha)_8$-barrels of the DAH7PS domains over a total surface area of 1581.1 $Å^2$. Dimers associate to form a homotetrameric structure through extensive interactions between the α5,

α6 and α7 helices and the β5-α5, β6-α6 and β7-α7 loops of the DAH7PS domains burying a total surface area of 1279.6 Å^2. Substantial inter-subunit interactions also occur between the CM domains from diagonally opposite CM-DAH7PS monomers. All of the CM helices contribute to these inter-subunit interactions burying a total surface area of 1883.8 Å^2. The substantial interactions between CM domains likely stabilises the tetrameric state as has been demonstrated recently with the FL domain of *Tm*-DAH7PS (Cross et al., 2011).

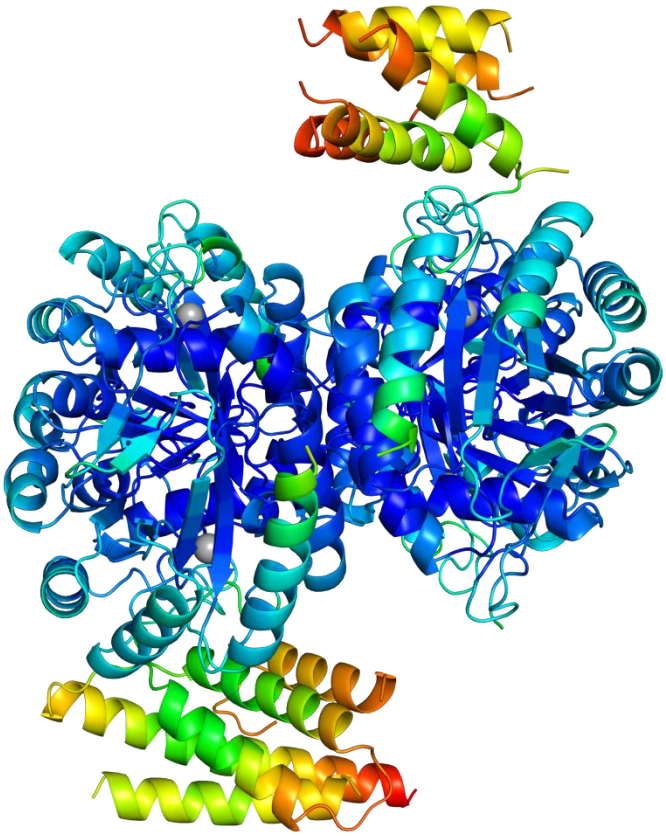

Figure 5.5 Cartoon representation of *Anc2*-CM-DAH7PS coloured by B-factor. B-factors are demonstrated by rainbow colouration, with red representing residues with high B-factors and dark blue representing low B-factors. Grey spheres represent Mn^{2+} ions.

5.2.6.2 Metal Binding Site

The Mn^{2+} ions are coordinated by the four metal binding ligands Cys126, His296, Glu322 and Asp333, which are conserved in all identified DAH7PS enzymes. The bond distances between the Mn^{2+} ions and the side chain oxygen, nitrogen and sulphur atoms are very similar to typical Mn-O, Mn-S and Mn-N bond distances (Rupp, 2010). These four ligands appear to occupy the two axial and two of the equatorial positions of a trigonal bipyramidal metal coordination site which leaves

the other equatorial site free to bind either E4P or water in a similar fashion to other DAH7PS enzymes (Webby et al., 2005a). E4P was not included in the crystallisation condition or soaked into the crystal. At two of the metal binding sites in the structure, there are water molecules in relatively close proximity to the metals which could potentially occupy the remaining equatorial site some of the time. In addition, two of the metal binding sites have areas of density near the modelled Mn^{2+} ions which were not large enough for atoms to be modelled into them. This also suggests that a solvent molecule, such as water, may be coordinated with the remaining equatorial binding site of the metal ions.

The binding of Mn^{2+} ions at the active site of *Anc2*-CM-DAH7PS is interesting and unexpected based on the results from high-resolution inductively coupled plasma mass spectrometry (ICPMS) of *Bsub*-CM-DAH7PS performed by Wu et al. (2005). Their results showed that as-isolated recombinant *Bsub*-CM-DAH7PS expressed in *E. coli* had mostly Zn^{2+} and Fe^{2+} bound at its active sites, with no Mn^{2+} detected. Wu et al. (2005) also removed the metal ions with the chelating agent DPA, then analysed the ability of *Bsub*-CM-DAH7PS to bind a number of different metal ions and also measured the amount of activity recovered upon binding of the different metal ions. Out of all the metal ions tested, *Bsub*-CM-DAH7PS was found to have the lowest affinity for Mn^{2+} and *Bsub*-CM-DAH7PS with Mn^{2+} bound displayed minimal enzymatic activity. Zn^{2+} was found to bind with the highest affinity of any of the metals and restored activity to almost the same level as the as-isolated enzyme, while binding of Cd^{2+} resulted in the highest level of activity of any of the metal ions tested. This may indicate that the metal affinity of *Anc2*-CM-DAH7PS is different to that of *Bsub*-CM-DAH7PS, however ICPMS analysis of *Anc2*-CM-DAH7PS, in addition to determination of the optimal metal for activity would be required to determine whether the metal affinity of *Anc2*-CM-DAH7PS is indeed different from that of *Bsub*-CM-DAH7PS.

5.2.6.3 Structural Alignment

Structural alignments of the most complete *Anc2*-CM-DAH7PS monomer, chain D, with structures in the PDB revealed that the closest structural homologue is the CM-DAH7PS from *L. monocytogenes* with PEP bound at the active sites (PDB code 3TFC), with a Z-score of 22.2, an RMSD of 0.80 Å and high Q- and P-scores over almost the entire monomer (340 residues). Q-score represents the quality of the alignment, where a value of 1 represents an identical structure. P-score is the negative logarithm of the probability of obtaining the same or a better quality match by chance. Higher P-scores indicate more statistically significant matches. Z-score is also a measure of the statistical significance of a match, with values greater than 3 considered to be significant. Significant structural similarity was also observed with all of the type Iβ DAH7PS enzyme structures in the PDB (Table 5.3).

Table 5.3 PDBeFold structural alignment.

Chain D	PDB code	Q-score	P-score	Z-score	RMSD	N_{align}	Species	Reference
1	3TFC	0.91	54.5	22.2	0.80	340	*L. monocytogenes*	Light et al. (2012)
2	3NVT	0.88	52.6	21.8	0.93	340	*L. monocytogenes*	Light et al. (2012)
3	1ZCO	0.67	40.2	19.1	0.99	260	*P. furiosus*	Schofield et al. (2005)
4	1VS1	0.65	35.6	17.9	1.07	260	*A. pernix*	Zhou et al. (2012)
5	3PG8	0.61	23.7	15.3	1.25	256	*T. maritima*	Cross et al. (2011)

Summary of the structural alignment statistics of the almost complete chain D *Anc2*-CM-DAH7PS monomer with the closest structural homologues in the PDB. N_{align} represents the number of aligned residues.

The high degree of structural similarity between *Anc2*-CM-DAH7PS and two of the closest structural homologues is demonstrated in Figures 5.6 and 5.7, which show SSM overlays of the *Anc2*-CM-DAH7PS monomer with *Lm*-CM-DAH7PS (PDB code 3TFC) and *Pf*-DAH7PS monomer structures (PDB code 1ZCO), respectively. There are very few structural differences between the *Anc2*-CM-DAH7PS and *Lm*-CM-DAH7PS monomers.

The CM domain from *Anc2*-CM-DAH7PS only differs slightly from the CM domain of *Lm*-CM-DAH7PS. A 1-2 Å displacement of α1' from *Anc2*-CM-DAH7PS is observed, and some significant displacement can be observed at the C-terminal end of α2' and the α2'-α3' loop. The *Lm*-CM-DAH7PS structure is missing part of the α1'-α2' loop so could not be compared. The positioning of the α3' helix and somewhat surprisingly, the domain linker region are essentially identical in the two structures. Even greater structural homology is observed between the DAH7PS domains. The only significant positional differences between the Cα atoms of the two DAH7PS monomeric structures occur at Lys110 (first loop before β1), Glu143, Gln144 and Gly145 (C-terminal end of α1 and start of α1-β2 loop), and the C-terminal end of α8. All of these differences occur at positions within the structure which do not play an active role in catalysis or oligomeric assembly, so are unlikely to be significant and may be the result of crystal packing. The similarity between the CM and DAH7PS domain orientations of the *Lm*-CM-DAH7PS and *Anc2*-CM-DAH7PS structures is significant as the two structures were crystallised in unrelated C2 and $P2_1$ space groups, respectively, and suggests that these orientations are biologically relevant and are not the result of crystal packing forces.

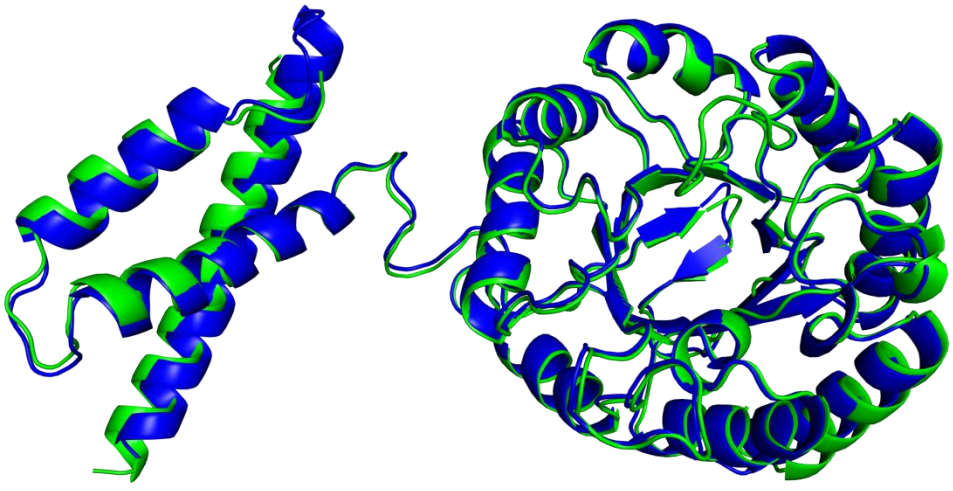

Figure 5.6 Overlay of *Anc2*-CM-DAH7PS and *Lm*-CM-DAH7PS monomers. Secondary structure matching (SSM) overlay of a *Anc2*-CM-DAH7PS monomer (blue) and a *Lm*-CM-DAH7PS monomer (green; PDB code 3TFC).

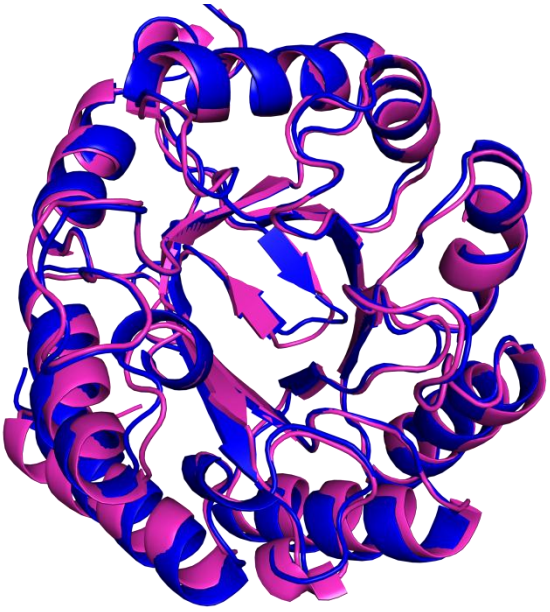

Figure 5.7 Overlay of the DAH7PS domain of *Anc2*-CM-DAH7PS and *Pf*-DAH7PS monomeric structures.
SSM overlay of the DAH7PS domain of the *Anc2*-CM-DAH7PS chain D monomer (blue) and a *Pf*-DAH7PS monomer (magenta; PDB code 1ZCO).

The DAH7PS domains of *Anc2*-CM-DAH7PS and the more distantly related *Pf*-DAH7PS are also structurally highly similar, despite sharing only 52.8% sequence identity. More tertiary structural differences exist between the peptide backbones of these two enzymes than between *Anc2*-CM-DAH7PS and *Lm*-CM-DAH7PS, but as was the case with the DAH7PS from *L. monocytogenes*, none of them occur in positions within the structure which are likely to affect catalysis. Slight differences are observed in some βn-αn loops that contain residues which form part of the active site (β3-α3 and β4-α4). The portions of the loops which contain the active site residues, however, are unaffected and there are no significant differences in the positioning of these residues.

Thus far, *Anc2*-CM-DAH7PS is the only ancestral or contemporary *Bacillus* CM-DAH7PS X-ray crystal structure that has been solved. X-ray crystal structures of type Iβ DAH7PS enzymes have been solved from species from a diverse range of phyla: *L. monocytogenes* is from the Firmicutes phylum like the *Bacillus* genus, *T. maritima* is from the deeply-branching bacterial phylum Thermatogae, and *P. furiosus* and *A. pernix* are from the archaeal phyla

Euryarchaeota and Crenarchaeota, respectively. The high level of structural similarity between *Anc2*-CM-DAH7PS and all of these structures suggests that *Anc2*-CM-DAH7PS has been accurately reconstructed, and suggests that there are very rigid structural constraints acting on the DAH7PS enzyme so any error in ancestral sequences which produced a small change in the 3D structure, particularly in the DAH7PS domain would result in an inactive enzyme. The lack of obvious differences between the *Anc2*-CM-DAH7PS structure and the other type Iβ DAH7PS structures, which might indicate an error in ancestral inference, and the very high structural similarity observed with the *Lm*-CM-DAH7PS structure, suggest that the inference of *Anc2*-CM-DAH7PS is accurate.

The closest structural homologue in the PDB is *Lm*-CM-DAH7PS (PDB code 3TFC), which is also the most closely related homologue identified in terms of sequence identity and its position in the prokaryotic phylogenetic tree relative to the *Bacillus* genus (Battistuzzi et al., 2004). The structures from *L. monocytogenes* are the only DAH7PS structures in the PDB with CM domains fused at their N-termini. The domains in the *Anc2*-CM-DAH7PS and *Lm*-CM-DAH7PS structures are orientated in a nearly identical fashion (Figure 5.8), which suggests that the observed positioning of the CM dimers in the homotetrameric enzymes may indicate that this orientation is biologically relevant and unlikely to be the result of subtle crystal packing forces. The *Lm*-CM-DAH7PS enzymes crystallised in a C2 space group which is very different to the $P2_1$ space group of the *Anc2*-CM-DAH7PS crystal, therefore, the observed domain orientations are very unlikely to be the result of crystal packing forces and likely represent a biologically relevant orientation despite the low number of interactions between the DAH7PS and CM domains.

As no other *Bacillus* CM-DAH7PS structures were able to be solved in the present study, and because no contemporary *Bacillus* CM-DAH7PS structures have been solved to date, the high thermostability of *Anc2*-CM-DAH7PS was not able to be structurally rationalised by comparison with other structures from *Bacillus* species. Also, as the *Anc2*-CM-DAH7PS and *Lm*-CM-DAH7PS are from species in different genera it would be extremely difficult to distinguish between changes which have led to differences in enzyme thermostability, and random changes or changes which have occurred for other functional reasons over the

long period of time since the *Bacillus* and *Listeria* genera diverged. Although it is likely that *Lm*-CM-DAH7PS is less thermostable than *Anc2*-CM-DAH7PS, as *L. monocytogenes* is a mesophile, it is not known with certainty that the thermostabilities of these enzymes are different.

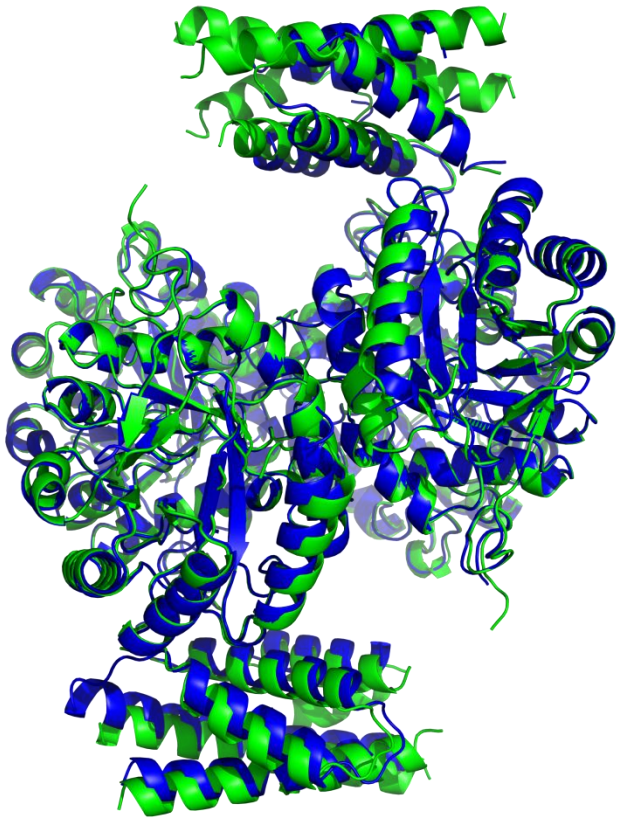

Figure 5.8 Overlay of *Anc2*-CM-DAH7PS and *Lm*-CM-DAH7PS tetramers.
SSM overlay of *Anc2*-CM-DAH7PS tetramer (blue) and *Lm*-CM-DAH7PS tetramer (green; PDB code 3TFC).

6 Discussion

Four Precambrian CM-DAH7PS enzymes from the *Bacillus* genus estimated to be between 570 and 1,079 Myr old have been reconstructed. These are the largest and most structurally complex Precambrian enzymes reconstructed thus far and test the limits of ancestral reconstruction. The three most recent ancestral enzymes have been biochemically characterised along with the CM-DAH7PS from *B. subtilis* and *B. stearothermophilus*. The oldest ancestor has been shown to display DAH7PS enzymatic activity, however the enzyme has yet to be purified in a state which is stable enough to allow biochemical characterisation of the enzyme. The similar properties observed for the CM-DAH7PS from *B. selenitireducens* and this ancestor, which share 96.4% sequence identity, suggest that this instability is not due to an error in inference of this enzyme but is rather the result of the purification not yet being optimised for these enzymes.

The major clades in the ML phylogenetic tree constructed in this study based on the CM-DAH7PS sequences of 20 *Bacillus* species are in similar positions to the corresponding clades in the *Bacillus* ML phylogenetic trees constructed by Hobbs et al. (2012) based on the IPMDH sequences from 19 *Bacillus* species and by Alcaraz et al. (2012) based on the core genomes from 20 *Bacillus* species. Some of the internal nodes in the ML phylogenetic tree based on the *Bacillus* CM-DAH7PS sequences had low bootstrap support, however as similar nodes exist in the ML phylogenetic trees constructed by Hobbs et al. (2012) and Alcaraz et al. (2010) the positioning of the ancestral nodes was deemed reliable. The four inferred ancestral CM-DAH7PS sequences have average posterior probabilities >0.87. The ancestral CM-DAH7PS sequences share between 71 and 96% sequence identities with contemporary *Bacillus* CM-DAH7PS sequences, which correspond to between 104 and 15 amino acid differences, respectively, between ancestral and contemporary enzymes. This is a significant number of amino acid differences as even a small number of amino acid changes has the potential to drastically alter the kinetic and thermal properties of the enzymes.

The three biochemically characterised ancestral CM-DAH7PS enzymes share similar kinetic properties with the CM-DAH7PS enzymes from *B. subtilis* and

B. stearothermophilus. This, together with the high level of similarity between the *Anc2*-CM-DAH7PS X-ray crystal structure and other DAH7PS structures in the PDB indicates that the ancestral inference is likely to have been accurate. The high level of structural similarities observed between *Anc2*-CM-DAH7PS and the other DAH7PS structures indicates that there are very rigid structural constraints acting on the DAH7PS enzymes so any error in ancestral inference which produced a small change in the 3D structure, particularly in the DAH7PS domain would result in an inactive enzyme. The fact that the three biochemically characterised ancestral CM-DAH7PS enzymes display similar kinetic properties to the CM-DAH7PS from *B. subtilis* and *B. stearothermophilus*, despite the large number of amino acid differences, strongly suggests that these ancestral enzymes have been accurately inferred. However, due to the four-fold lower k_{cat} value of *Anc4*-CM-DAH7PS and issues regarding the stability and aggregation which affected this enzyme and *Anc5*-CM-DAH7PS more than the two more recent ancestral enzymes, the possibility that the low T_{opt} of this enzyme is the result of inaccurate inference and that the potential mesophilic origin of the *Bacillus* genus proposed may be spurious cannot be excluded. If this T_{opt} value is incorrect, then the thermal data of the ancestral CM-DAH7PS data do not conflict with the results of Gaucher et al. (2008), Perez-Jimenez et al. (2011), and Hobbs et al. (2012). Based on all the data currently available, however, the likelihood that the T_{opt} value for *Anc4*-CM-DAH7PS is spurious is considered to be low. Therefore, a preliminary hypothesis regarding the evolution of thermophily based on the T_{opt} of the three biochemically characterised ancestral CM-DAH7PS is possible.

The T_{opt} values suggest a mesophilic origin for the *Bacillus* genus and that thermophily may be a recent adaptation rather than a primitive trait, with the T_{opt} data suggesting that thermophily evolved within the *Bacillus* genus between ~950 and 650 Mya. These results conflict with the those of Gaucher et al. (2008) and Perez-Jimenez et al. (2011) which suggested a thermophilic origin for thermophily and a general decrease in thermophily with time from this primitive thermophilic ancestor. The ancestral CM-DAH7PS T_{opt} values also conflict with the origin and evolution of thermophily inferred from the thermostabilities of reconstructed ancestral *Bacillus* IPMDH enzymes by Hobbs et al. (2012). The data of Hobbs et al. (2012) suggests a thermophilic origin for the *Bacillus* genus

and the evolution of thermophily at least twice within the *Bacillus* genus. Similar ancestral nodes were reconstructed in this study and the study of Hobbs et al. (2012) in terms of the divergence of different clades, however the ancestral enzymes reconstructed from these similar nodes have different thermal properties and so could result in different inferences being made about the evolution of thermophily. These conflicting results could indicate either that evolutionary trends observed from ASR studies are highly sensitive to the contemporary species used in the ancestral inferences, or that any trends observed are protein dependent and reconstruction of different proteins may lead to different inferences being made regarding the origin and evolution of thermophily. If the biochemical properties of ancestral enzymes is shown to be highly dependent on the biochemical properties of the contemporary proteins used to infer the ancestral sequences, this would suggest that unless sequences are available for all of the different species which have diverged from a particular node, any inferences regarding the evolution of thermophily will be biased by species selection. If reconstructed enzymes from different enzyme families are shown to result in different inferences being made about the evolution of thermophily, then either reconstructed proteins should not be used to infer how thermophily may have evolved or it may be that, in order to have a high level of confidence in any inferences made regarding the evolution of thermophily, a number of different types of ancestral enzymes must be reconstructed and biochemically characterised to present a more robust inference.

6.1 Future Research

Due to the limitations associated with determination of the T_{opt} of the ancestral enzymes, robust T_m measurements for these enzymes determined using DSC are required to be confident in the preliminary inferences made regarding the evolution of thermophily. Further work should be done to optimise the expression and purification conditions used for *Anc5*-CM-DAH7PS and *Bsel*-CM-DAH7PS to obtain stable forms of these enzymes so that these enzymes can be biochemically characterised.

In order to structurally rationalise the changes in enzyme thermostabilities, X-ray crystal structures need to be determined for the other ancestral enzymes and

ideally, the contemporary *Bacillus* enzymes characterised in this study. Structural data for these enzymes could then also be used to further support the accuracy of the ancestral inferences.

All of the work in this study has focused on the DAH7PS activity of the CM-DAH7PS enzymes. Due to the proposed evolution of the fused CM domain from a highly catalytically active CM domain to a prephenate-binding domain, whose sole purpose is to allow regulation of the shikimate pathway by inhibition of DAH7PS activity when the downstream product prephenate is bound, it will be interesting to investigate how the activity of CM and the prephenate-binding properties of the protein has evolved in the *Bacillus* genus. If the hypothesis regarding the evolution of the CM domain is correct it would be expected that as the age of the ancestral enzymes increased, the CM domains would become more catalytically active and a less efficient binder of prephenate. Observable CM catalytic activity from the ancestral CM-DAH7PS enzymes would also increase the level of confidence in the biochemical properties observed for the enzymes.

Appendices

Appendix A: Reagents, Buffers, Growth Media, Bacterial Strains and Plasmids

A1: Buffers and Reagents

Coomassie stain	0.05% (w/v) coomassie blue R-250, 25% (v/v) isopropanol, 10% (v/v) acetic acid
10 x DNA loading dye	0.4% (w/v) bromophenol blue, 0.4% (w/v) xylene, 50% (v/v) glycerol
Destain solution	10% (v/v) acetic acid
GITC	295.4 g guanidine thiocyanate, 2.5 g N-lauroyl sarcosine, 3.9 g tri-sodium citrate, 3.6 mL β-mercaptoethanol, 280 mL diethlypyrocarbonate water.
Resolving buffer	1.5 M Tris-HCl, pH 8.8

SDS-PAGE Gel Recipe

Table A.1 SDS-PAGE gel recipes.

	12% gel	Stacker
MQ water	10.05	8.5
Resolving buffer	7.5	-
Stacking buffer	-	1.6
30% acrylamide	12	2.125
10% SDS	0.3	0.125
10% APS	0.15	0.063
Temed	0.015	0.0063

Volumes are in mL. APS (ammonium persulphate),
Temed (N, N, N', N'-tetramethylethylenediamine).

4 x SDS loading dye	200 mM Tris-HCl pH 6.8, 8% (w/v) SDS, 40% (v/v) glycerol, 0.4% (w/v) bromophenol blue, 400 mM β-mercaptoethanol
Stacking buffer	1.0 M Tris-HCl, pH 6.8

10 x TAE buffer	400 mM Tris-acetate, 20 mM EDTA
1 x TAE buffer	10 x TAE + 900 mL H_2O
TE	10 mM Tris-HCl pH 8.8, 1 mM EDTA pH 8.0
TG-SDS running buffer	25 mM Tris-HCl pH 8.3, 250 mM glycine, 0.1% (w/v) SDS

A2: Growth Media

LB	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, pH 8.0
LB-agar	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 15 g/L agar**,** pH 8.0
NB	0.8% (w/v) nutrient broth
NB-agar	0.8% (w/v) nutrient broth, 1.5% (w/v) agar
PA-0.5G	50 mM Na_2HPO_4, 50 mM KH_2PO_4, 25 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 0.5% (w/v) glucose, 0.1 x metals mix*, 200 µg/mL each of 17 amino acids (no C, Y or M). Individual components were autoclaved or filtered to sterilise before adding to sterile water.
TB	1.2% (w/v) bactotryptone, 2.4% (w/v) yeast extract, 1 mL glycerol, 900 mL water. Autoclave. Combine with 0.17 M
ZYP-5052	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 50 mM Na_2HPO_4, 50 mM KH_2PO_4, 25 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 0.5% (w/v) glycerol, 0.05% glucose, 0.2% (w/v) α-lactose, 1 x metals mix

* 1000 x metals mix made up from sterile stocks of each component to give the following concentrations: 50 µM $FeCl_3$ in 0.12 M HCl (filter sterile), 20 µM $CaCl_2$, 10 µM $MnCl_2$, 10 µM $ZnSO_4$, 2 µM $CoCl_2$, 2 µM $NiCl_2$, 2 µM $NaMoO_4$, 2 µM Na_2SeO_3, 2 µM H_3BO_3.

A3: Bacterial Strains Used and Transformants Generated in this Study

Table A.1 Strains of *E. coli* used in this study.

Strain	Description
DH5α	*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*
BL21 (DE3)	F-*omptT hsdS*$_B$ (r_B-m_B-) *gal dcm* (DE3)

Table A.2 Plasmids used in this study.

Plasmids	Description
pET28b	*E. coli* expression vector with T7 promoter and encoding N- and C-terminal His-tags: Kan^r
pMA-T	GENEART vector with T7 promoter: Amp^r
pGroESL	*E. coli* expression vector with T7 promoter. Contains *E. coli* GroES and GroEL genes: Cam^r

Table A.3 Transformants generated in this study.

Strain	Plasmid	pGroESL
Escherichia coli DH5α	pET28b	-
	pET28b-*Anc2*-CM-DAH7PS	-
	pET28b-*Anc3*-CM-DAH7PS	-
	pET28b-*Anc4*-CM-DAH7PS	-
	pET28b-*Anc5*-CM-DAH7PS	-
	pET28b-*Bsub*-CM-DAH7PS	-
	pET28b-*Bstr*-CM-DAH7PS	-
	pET28b-*Bsel*-CM-DAH7PS	-
Escherichia coli BL21	pET28b-*Anc2*-CM-DAH7PS	-
	pET28b-*Anc3*-CM-DAH7PS	-
	pET28b-*Anc4*-CM-DAH7PS	-
	pET28b-*Anc5*-CM-DAH7PS	-
	pET28b-*Bsel*-CM-DAH7PS	-
	pET28b-*Bsub*-CM-DAH7PS	-
	pET28b-*Bstr*-CM-DAH7PS	-
	pET28b-*Anc2*-CM-DAH7PS	+
	pET28b-*Anc3*-CM-DAH7PS	+
	pET28b-*Anc4*-CM-DAH7PS	+
	pET28b-*Anc5*-CM-DAH7PS	+
	pET28b-*Bsel*-CM-DAH7PS	+

+ indicates that transformants contain pGroESL in addition to pET28b.

Table A. 4 Bacterial strains and accession numbers of *cm-dah7ps* used in ancestral inference.

Organism	**Strain/Serovar**	**Accession number/Source**
Bacillus amyloliquefaciens	FZB42	NC_009725
Bacillus anthracis	Ames	NC_003997
Bacillus atrophaeus	1942	NC_014639
Bacillus cellulosilyticus	DSM 2522	CP002394
Bacillus cereus	ATCC 10987	NC_003909
Bacillus clausii	KSM-K16	NC_006582
Bacillus coagulans	36D1	CP003056
Bacillus cytotoxicus	NVH 391-98	NC_009674
Bacillus halodurans	C-125	NC_002570
Bacillus licheniformis	ATCC 14580	NC_006270
Bacillus megaterium	QM B1551	NC_014019
Bacillus mycoides	Rock3-17	NZ_ACMW01000093
Bacillus pseudofirmus	OF4	NC_013791
Bacillus pseudomycoides	DSM 12442	NZ_ACMX01000055
Bacillus pumilus	ATCC 7061	NZ_ABRX01000009
Bacillus selenitireducens	MLS10	NC_014219
Bacillus stearothermophilus	DSM 22	http://www.genome.ou.edu/bstearo.html (contig 501)
Bacillus subtilis	Marburg 168	This study
Bacillus thuringiensis	konkukian 97-27	NC_005957
Bacillus weihenstephanensis	KBAB4	NC_010184
Clostridium acetobutylicum	ATCC 824	NC_003030
Clostridium butyricum	5521	NZ_ABDT01000092

Appendix B: Gene and Protein Information

B1: Nucleotide and Amino Acid Sequence Details

All nucleotide sequences, except for *Bsub-cm-dah7ps*, were codon optimised for expression in *E. coli*. These sequences also include the C-terminal His-tag at the end of their sequence.

Bsel-cm-dah7ps (1119 bp)

```
ATGGGAAATGAACAGCTGGAAGAACTGCGTGGTCAGCTGGATGAAGTTAATGAAA
AACTGGTGGAAATGATGAATGAACGTGCCCGTCTGGCACAAGAAATTGGTCGTGT
TAAAAGCAGCCAGGGTATGAATCGTTTTGATCCGGTTCGTGAACGTAAAATGCTG
GATATGATCCAAGAAAAAAACGAGGGTCCGTTTGAAACCGCAACCCTGCAGCACC
TGTTCAAACAAATCTTTAAAGCAAGCCTGGAACTGCAAGAAGATGATCACCGTAA
AGCACTGCTGGTTAGCCGTAAAAAACATCCGGAAGATACCATTGTTGATGTGAAT
GGCACCAAACTGGGTAATGGTGAACAGCATCTGATTGCAGGTCCGTGTAGCGTTG
AAAGCTATGAACAGGTTGAAGCAGTTGCCAAAGAACTGAAAGCACGTGGTCTGAC
CATGATGCGTGGTGGTGCATATAAACCGCGTACCAGCCCGTATGATTTTCAGGGT
CTGGGTCAAGAAGGTCTGGAAATTCTGAAAGATATCAGCGATAAATATGGCCTGA
GCGTGATTAGCGAAATTGTTACACCGGGTGATATTCAGAATGCCGTTGATTATCT
GGATGTGATTCAGATTGGTGCACGCAATATGCAGAATTTTGAACTGCTGAAAGAA
GCCGGTCGTACCAATAAACCGATTCTGCTGAAACGCGGTCTGAGCGCAACCATTG
AAGAATTTATCAATGCAGCCGAATACATTCATAGCCAGGGCAATGGTCAGATTAT
TCTGTGTGAACGTGGTATTCGCACCTATGAAAAAGCAACCCGTAATACCCTGGAT
ATTAGCGCAGTTCCGATTCTGAAACAAGAAACCCATCTGCCGGTTTTTGTGGATG
TTACCCATAGCACCGGTCGTCGTGATCTGCTGCTGCCGACCGCAAAAGCAGCATT
TGCAGTTGGTGCAGATGGTGTTATGACCGAAGTTCATCCGGATCCTGCAGTTGCA
CTGAGCGATAGCGCACAGCAGATGGATATTCCGCAGTTTGGCGAATTTCTGAAAA
ATCTGGAAGAAAGCGGTCTGTTTAAAGTGAATAAAGCAGCAAGCAAAAGCCTCGA
GCACCACCACCACCACCAC
```

Bsel-CM-DAH7PS

```
MGNEQLEELRGQLDEVNEKLVEMMNERARLAQEIGRVKSSQGMNRFDPVRERKML
DMIQEKNEGPFETATLQHLFKQIFKASLELQEDDHRKALLVSRKKHPEDTIVDVN
GTKLGNGEQHLIAGPCSVESYEQVEAVAKELKARGLTMMRGGAYKPRTSPYDFQG
LGQEGLEILKDISDKYGLSVISEIVTPGDIQNAVDYLDVIQIGARNMQNFELLKE
AGRTNKPILLKRGLSATIEEFINAAEYIHSQGNGQIILCERGIRTYEKATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAAFAVGADGVMTEVHPDPAVA
LSDSAQQMDIPQFGEFLKNLEESGLFKVNKAASKSLEHHHHHH
```

```
373 amino acids
Predicted molecular weight: 41,680 Da
Theoretical pI: 5.35
```

***Bstr-cm-dah7ps* (1104 bp)**

```
ATGGGTAATGAACGTCTGGATGAACTGCGTGCACGTGTTGATGAAATTAATCTGC
AGCTGCTGAAACTGATTAATGAACGTGGTCGTCTGGTTCAAGAAATTGGCAAAAT
TAAAGAAGCACAGGGCACCCATCGTTATGATCCGGTTCGTGAACGTAAAATGCTG
GATCTGATTAGCGAACATAATGATGGTCCGTTTGAAACCAGCACCCTGCAGCATA
TTTTTAAAGAAATTTTTAAAGCAGCCCTGGAACTGCAAGAAGATGATCATCGTAA
AGCACTGCTGGTTAGCCGTAAAAAACATCCGGAAAATACCATTGTGGAAGTTAAA
GGCGAACGTATTGGTGATGGCAATCAGTATTTTGTTATGGGTCCGTGTGCCGTTG
AAAGCTATGAACAGGTTGCAGCAGTTGCAGAAGCAGTTAAAAAACAGGGTATTAA
ACTGCTGCGTGGTGGTGCATATAAACCGCGTACCAGCCCGTATGATTTTCAGGGT
CTGGGTGTTGAAGGTCTGAAAATTCTGAAACGTATTGCCGATGAATTTGATCTGG
CCGTTATTAGCGAAATTGTTACACCGGCAGATATTGAAATTGCCCTGGATTATAT
TGATGTGATTCAGATTGGCGCACGCAATATGCAGAATTTTGAACTGCTGAAAGCA
GCAGGTCAGGTTAATAAACCGATTCTGCTGAAACGTGGTCTGGCAGCAACCATTG
AAGAATTTATTAATGCAGCCGAATATATTATGAGCCAGGGTAATGGCCAGATTAT
TCTGTGTGAACGTGGTATTCGTACCTATGAACGTGCAACCCGTAATACCCTGGAT
ATTAGCGCAGTTCCGATTCTGAAAAAAGAAACCCATCTGCCGGTTTTTGTTGATG
TTACCCATAGCACCGGTCGTCGTGATCTGCTGATTCCGTGTGCAAAAGCAGCACT
GGCAATTGGTGCAGATGGTGTTATGGCAGAAGTTCATCCGGATCCGGCAGTTGCA
CTGAGCGATAGCGCACAGCAGATGGATATTGCACAGTTTAATGAATTTATGGAAG
AAGTTCGCGCATTTCAGCGTCAGTTTGTTCGTGCACTCGAGCACCACCACCACCA
CCAC
```

***Bstr*-CM-DAH7PS**

```
MGNERLDELRARVDEINLQLLKLINERGRLVQEIGKIKEAQGTHRYDPVRERKML
DLISEHNDGPFETSTLQHIFKEIFKAALELQEDDHRKALLVSRKKHPENTIVEVK
GERIGDGNQYFVMGPCAVESYEQVAAVAEAVKKQGIKLLRGGAYKPRTSPYDFQG
LGVEGLKILKRIADEFDLAVISEIVTPADIEIALDYIDVIQIGARNMQNFELLKA
AGQVNKPILLKRGLAATIEEFINAAEYIMSQGNGQIILCERGIRTYERATRNTLD
ISAVPILKKETHLPVFVDVTHSTGRRDLLIPCAKAALAIGADGVMAEVHPDPAVA
LSDSAQQMDIAQFNEFMEEVRAFQRQFVRALEHHHHHH
```

368 amino acids
Predicted molecular weight: 41,374 Da
Theoretical pI: 5.99

***Bsub-cm-dah7ps* (1095 bp)**

```
ATGGGCAACACAGAGTTAGAGCTTTTAAGGCAGAAAGCAGACGAATTAAACCTAC
AAATTTTAAAATTAATCAACGAACGCGGCAATGTTGTAAAAGAGATCGGTAAAGC
GAAGGAAGCACAGGGTGTCAACCGATTTGACCCTGTCAGAGAACGCACAATGTTA
AACAATATCATTGAAAACAATGACGGGCCGTTCGAAAATTCAACCATCCAGCACA
TTTTTAAAGAGATATTCAAAGCCGGTTTAGAGCTTCAGGAAGAAGATCACAGCAA
AGCGCTGCTTGTCTCCCGCAAGAAAAAACCTGAAGATACAATTGTTGATATCAAA
GGCGAAAAAATCGGAGACGGCCAGCAAAGATTCATTGTCGGCCCATGTGCGGTAG
AGAGCTATGAGCAGGTAGCTGAAGTCGCTGCAGCTGCGAAAAAACAAGGGATTAA
AATTTTGCGCGGTGGAGCCTTTAAGCCTCGTACGAGCCCATACGATTTCCAAGGG
CTTGGTGTTGAAGGCCTTCAAATTTTAAAACGTGTAGCGGATGAATTTGATCTGG
CGGTTATCAGTGAAATCGTAACTCCGGCTCATATCGAAGAAGCGCTGGACTACAT
TGATGTCATTCAAATCGGAGCGCGCAACATGCAAAACTTCGAATTGCTGAAAGCG
```

GCCGGCGCCGTGAAAAAGCCAGTGCTTCTGAAGCGCGGTCTTGCTGCAACGATCT
CTGAATTCATCAATGCTGCTGAATACATCATGTCACAAGGAAATGACCAAATTAT
CCTTTGTGAGCGCGGAATCAGAACATATGAAACAGCAACGAGAAACACGCTGGAT
ATTTCAGCTGTGCCGATTTTGAAACAAGAAACGCATTTGCCAGTCTTTGTTGATG
TTACGCATTCAACAGGCCGCCGTGACCTCTTGCTTCCGACAGCTAAAGCCGCTTT
AGCGATCGGTGCTGATGGCGTAATGGCTGAGGTTCACCCTGATCCGTCAGTCGCA
CTTTCTGACTCTGCTCAGCAAATGGCGATTCCTGAATTCGAAAAATGGCTGAATG
AACTGAAGCCAATGGTGAAAGTCAACCTCGAGCACCACCACCACCACCAC

Bsub-CM-DAH7PS

MGNTELELLRQKADELNLQILKLINERGNVVKEIGKAKEAQGVNRFDPVRERTML
NNIIENNDGPFENSTIQHIFKEIFKAGLELQEEDHSKALLVSRKKKPEDTIVDIK
GEKIGDGQQRFIVGPCAVESYEQVAEVAAAAKKQGIKILRGGAFKPRTSPYDFQG
LGVEGLQILKRVADEFDLAVISEIVTPAHIEEALDYIDVIQIGARNMQNFELLKA
AGAVKKPVLLKRGLAATISEFINAAEYIMSQGNDQIILCERGIRTYETATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAALAIGADGVMAEVHPDPSVA
LSDSAQQMAIPEFEKWLNELKPMVKVNLEHHHHHH

365 amino acids
Predicted molecular weight: 40,503 Da
Theoretical pI: 5.80

Anc2-cm-dah7ps (1098 bp)

ATGGGTAATAAAGAACTGGAACAGCTGCGTGAACAGGTGGATGAAATCAACCTGC
AGATTCTGGAGCTGCTGAATGAACGTGGTCGTATTGTTCAGGAAATTGGCAAAGT
TAAAGAAGCACAGGGCGTTAATCGTTTTGATCCGGTTCGTGAACGTAAAATGCTG
GATCTGATTGCCGAAAATAACGATGGTCCGTTTGAAACCAGCACCCTGCAGCATA
TCTTCAAAGAAATCTTTAAAGCCAGCCTGGAACTGCAGGAAGATGATCATCGTAA
AGCACTGCTGGTTAGCCGCAAAAAGAAACCGGAAAATACCATTGTGGATATCAAA
GGCGAAAAAATTGGTGATGGCAACCAGCAGTTTATTATGGGTCCGTGTGCAGTTG
AAAGCTATGAACAGGTTCGTGAAGTTGCAGAAGCAATGAAAGAACAGGGTCTGAA
ACTGATGCGTGGTGGTGCATTTAAACCGCGTACCAGCCCGTATGATTTTCAGGGT
CTGGGTGTTGAAGGTCTGCAGATCCTGCGTCAGGTTGCAGATGAATTTGATCTGG
CAGTGATTAGCGAAATTGTTACCCCGAACGATATTGAAATGGCCCTGGATTATGT
GGATGTGATTCAGATTGGTGCACGCAATATGCAGAATTTTGAACTGCTGAAAGCA
GCAGGTAGCGTTAATAAACCGGTTCTGCTGAAACGTGGTCTGGCAGCAACCATTG
AAGAATTTATCAATGCAGCCGAATATATCATGAGCCAGGGTAATGGCCAGATTAT
TCTGTGTGAACGTGGTATTCGTACCTATGAACGTGCAACCCGTAATACCCTGGAT
ATTAGCGCAGTTCCGATCCTGAAAAAAGAAACCCATCTGCCGGTTGTTGTTGATG
TTACCCATAGCACCGGTCGTCGTGATCTGCTGCTGCCGACCGCAAAAGCAGCACT
GGCAATTGGTGCAGATGCAGTTATGGCAGAAGTTCATCCGGACCCTGCAGTTGCA
CTGAGCGATAGCGCACAGCAGATGGATATTCCGGAATTTAACAAATTTATGGAAG
AACTGAAAGCCTTTGGCAACAAACTGAGCCTCGAGCACCACCACCACCACCAC

***Anc2*-CM-DAH7PS**

MGNKELEQLREQVDEINLQILELLNERGRIVQEIGKVKEAQGVNRFDPVRERKML
DLIAENNDGPFETSTLQHIFKEIFKASLELQEDDHRKALLVSRKKKPENTIVDIK
GEKIGDGNQQFIMGPCAVESYEQVREVAEAMKEQGLKLMRGGAFKPRTSPYDFQG
LGVEGLQILRQVADEFDLAVISEIVTPNDIEMALDYVDVIQIGARNMQNFELLKA
AGSVNKPVLLKRGLAATIEEFINAAEYIMSQGNGQIILCERGIRTYERATRNTLD
ISAVPILKKETHLPVVVDVTHSTGRRDLLLPTAKAALAIGADAVMAEVHPDPAVA
LSDSAQQMDIPEFNKFMEELKAFGNKLSLEHHHHHH

366 amino acids
Predicted molecular weight: 40,100 Da
Theoretical pI: 5.35

***Anc3-cm-dah7ps* (1098 bp)**

ATGGGTAATAAAGAACTGGAACAGCTGCGTGAACAGGTGGATGAAATTAACCTGC
AGATTCTGGAACTGATTAATGAACGTGGTCGTATTGTGCAGGAAATCGGTAAAGT
TAAAGAAGCACAGGGCGTTAATCGTTTTGATCCGGTTCGTGAACGTAAAATGCTG
GATCTGATTGCCGAAAATAACGATGGTCCGTTTGAAACCAGCACCCTGCAGCATA
TCTTTAAAGAAATCTTCAAAGCCAGCCTGGAACTGCAGGAAGATGATCATCGTAA
AGCACTGCTGGTTAGCCGCAAAAAGAAACCGGAAAATACCATTGTGGATATCAAA
GGCGAAAAAATTGGTGATGGCAACCAGCAGTTTATTATGGGTCCGTGTGCAGTTG
AAAGCTATGAACAGGTTGCAGAAGTTGCCGAAGCAGTTAAAGAACAGGGTCTGAA
ACTGCTGCGTGGTGGTGCATTTAAACCGCGTACCAGCCCGTATGATTTTCAGGGT
CTGGGTGTTGAAGGTCTGCAGATCCTGAAACGTGTTGCAGATGAATTTGATCTGG
CCGTTATTAGCGAAATTGTTACACCGGCAGATATTGAAAAAGCCCTGGATTATGT
GGATGTGATTCAGATTGGTGCACGCAATATGCAGAATTTTGAACTGCTGAAAGCA
GCAGGTAGCGTTAATAAACCGGTTCTGCTGAAACGTGGTCTGGCAGCAACCATTG
AAGAATTTATCAATGCAGCCGAATATATCATGAGCCAGGGTAATGGCCAGATTAT
TCTGTGTGAACGTGGTATTCGTACCTATGAACGTGCAACCCGTAATACCCTGGAT
ATTAGCGCAGTTCCGATTCTGAAACAGGAAACCCATCTGCCGGTTTTTGTTGATG
TTACCCATAGCACCGGTCGTCGTGATCTGCTGCTGCCGACCGCAAAAGCAGCACT
GGCAATTGGTGCAGATGGTGTTATGGCAGAAGTTCATCCGGATCCTGCAGTTGCA
CTGAGCGATAGCGCACAGCAGATGGATATTCCGCAGTTTAACAAATTTATGGAAG
AACTGAAAGCCTTTGGCAACAAAAAAGCACTCGAGCACCACCACCACCACCAC

***Anc3*-CM-DAH7PS**

MGNKELEQLREQVDEINLQILELINERGRIVQEIGKVKEAQGVNRFDPVRERKML
DLIAENNDGPFETSTLQHIFKEIFKASLELQEDDHRKALLVSRKKKPENTIVDIK
GEKIGDGNQQFIMGPCAVESYEQVAEVAEAVKEQGLKLLRGGAFKPRTSPYDFQG
LGVEGLQILKRVADEFDLAVISEIVTPADIEKALDYVDVIQIGARNMQNFELLKA
AGSVNKPVLLKRGLAATIEEFINAAEYIMSQGNGQIILCERGIRTYERATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAALAIGADGVMAEVHPDPAVA
LSDSAQQMDIPQFNKFMEELKAFGNKKALEHHHHHH

366 amino acids
Predicted molecular weight: 40,851 Da
Theoretical pI: 5.49

***Anc4-cm-dah7ps* (1122 bp)**

ATGGGTAATGAACAGCTGGAAGAACTGCGTGATCAGCTGGATGAAGTTAATCTGA
AACTGCTGGAACTGATTAATGAACGTGCACGTCTGGTTCAGGAAATTGGTAAAGT
TAAAAGCGCACAGGGTGTGAATCGTTTTGATCCGGTTCGTGAACGTAAAATGCTG
GATCTGATTGCCGAAAATAACAAAGGTCCGTTTGAAACCAGCACCCTGCAGCATA
TCTTTAAACAAATCTTTAAAGCCAGCCTGGAACTGCAGGAAGATGATCATCGTAA
AGCACTGCTGGTTAGCCGTAAAAAACATCCGGAAAATACCATCGTTGATGTGAAA
GGCGAAAAAGTTGGTGATGGTAAACAGCGTCTGATTATGGGTCCGTGTGCAGTTG
AAAGCTATGAACAGGTTGCAGCAGTTGCAAAAGCAGTTAAAGAACGTGGCCTGAA
ACTGCTGCGTGGTGGTGCATTTAAACCGCGTACCAGCCCGTATGATTTTCAGGGT
CTGGGTCTGGAAGGTCTGAAAATTCTGAAACGTGTTGCCGATGAATTTGATCTGG
CAGTTATTAGCGAAATTGTTACACCGGCAGATATTGAAGAAGCCCTGGATTACGT
TGATGTGATTCAGATTGGTGCCCGTAATATGCAGAATTTTGAACTGCTGAAAGCA
GCAGGTAGCGTTAATAAACCGGTTCTGCTGAAACGTGGTCTGAGCGCAACCATTG
AAGAGTTTATCAATGCAGCCGAATATATTGTGAGCCAGGGTAATGGCCAGATTAT
GCTGTGTGAACGTGGTATTCGCACCTATGAAAAAGCAACCCGTAATACCCTGGAT
ATTAGCGCAGTTCCGATTCTGAAACAGGAAACCCATCTGCCGGTTTTTGTTGATG
TTACCCATAGCACCGGTCGTCGTGATCTGCTGCTGCCGACCGCAAAAGCAGCACT
GGCAATTGGTGCAGATGGTGTTATGGCAGAAGTTCATCCGGATCCTGCAGTTGCA
CTGAGCGATAGCGCACAGCAGATGGATATTCCGCAGTTTAATGAATTTGTGGATG
ATCTGATTGCGAGCGGTCTGTATAAAGCAGCAACCAAAACCGCTCAGCAGAAACT
CGAGCACCACCACCACCACCAC

***Anc4*-CM-DAH7PS**

MGNEQLEELRDQLDEVNLKLLELINERARLVQEIGKVKSAQGVNRFDPVRERKML
DLIAENNKGPFETSTLQHIFKQIFKASLELQEDDHRKALLVSRKKHPENTIVDVK
GEKVGDGKQRLIMGPCAVESYEQVAAVAKAVKERGLKLLRGGAFKPRTSPYDFQG
LGLEGLKILKRVADEFDLAVISEIVTPADIEEALDYVDVIQIGARNMQNFELLKA
AGSVNKPVLLKRGLSATIEEFINAAEYIVSQGNGQIMLCERGIRTYEKATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAALAIGADGVMAEVHPDPAVA
LSDSAQQMDIPQFNEFVDDLIASGLYKAATKTAQQKLEHHHHHH

374 amino acids
Predicted molecular weight: 41,526 Da
Theoretical pI: 5.99

***Anc5-cm-dah7ps* (1122 bp)**

ATGGGTAATGAACAGCTGGAAGAACTGCGTGATCAGCTGGATGAAGTTAATCTGA
AACTGGTGGAAATGATGAATGAACGTGCACGTCTGGCACAGGAAATTGGTCGCGT
TAAAAGCAGCCAGGGTATGAATCGTTTTGATCCGGTTCGTGAACGTAAAATGCTG
GATATGATCGCCGAAAAAAACGAAGGTCCGTTTGAAACCGCAACCCTGCAGCACC
TGTTTAAACAAATCTTTAAAGCAAGCCTGGAACTGCAGGAAGATGATCATCGTAA
AGCACTGCTGGTTAGCCGTAAAAAACATCCGGAAGATACCATTGTTGATGTGAAT
GGCACCAAAATTGGTGATGGTGAACAGCATCTGATTGCAGGTCCGTGTAGCGTTG
AAAGCTATGAACAGGTTGAAGCAGTTGCCAAAGAACTGAAAGAACGTGGTCTGAA
ACTGCTGCGTGGTGGTGCATTTAAACCGCGTACCAGCCCGTATGATTTTCAGGGT
CTGGGTCAGGAAGGTCTGGAAATTCTGAAAGATGTGGCCGATAAATATGGTCTGA
GCGTTATTAGCGAAATTGTTACACCGGGTGATATTGAAAACGCCGTTGATTATGT
GGATGTGATTCAGATTGGTGCACGCAATATGCAGAATTTTGAACTGCTGAAAGAA

GCAGGTCGTACCAATAAACCGATTCTGCTGAAACGTGGTCTGTCAGCAACCATTG
AAGAATTTATCAACGCAGCCGAATATATTCATAGCCAGGGTAATGGCCAGATTAT
TCTGTGTGAACGTGGTATTCGCACCTATGAAAAAGCAACCCGTAATACCCTGGAT
ATTAGCGCAGTTCCGATTCTGAAACAGGAAACCCATCTGCCGGTTTTTGTTGATG
TTACCCATAGCACCGGTCGTCGTGATCTGCTGCTGCCGACCGCAAAAGCAGCATT
TGCAGTTGGTGCAGATGGTGTTATGACCGAAGTTCATCCGGATCCTGCAGTTGCA
CTGAGCGATAGCGCACAGCAGATGGATATTCCGCAGTTTGATGAATTTCTGAAAA
ATCTGGAAGAAAGCGGTCTGTTTAAAGTGAAAAAAGCAGCGAGCAAAAGCAAACT
CGAGCACCACCACCACCACCAC

Anc5-CM-DAH7PS

MGNEQLEELRDQLDEVNLKLVEMMNERARLAQEIGRVKSSQGMNRFDPVRERKML
DMIAEKNEGPFETATLQHLFKQIFKASLELQEDDHRKALLVSRKKHPEDTIVDVN
GTKIGDGEQHLIAGPCSVESYEQVEAVAKELKERGLKLLRGGAFKPRTSPYDFQG
LGQEGLEILKDVADKYGLSVISEIVTPGDIENAVDYVDVIQIGARNMQNFELLKE
AGRTNKPILLKRGLSATIEEFINAAEYIHSQGNGQIILCERGIRTYEKATRNTLD
ISAVPILKQETHLPVFVDVTHSTGRRDLLLPTAKAAFAVGADGVMTEVHPDPAVA
LSDSAQQMDIPQFDEFLKNLEESGLFKVKKAASKSKLEHHHHHH

374 amino acids
Predicted molecular weight: 41,857 Da
Theoretical pI: 5.67

References

ABASCAL, F., ZARDOYA, R. & POSADA, D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics,* 21**,** 2104-5.

ALCARAZ, L. D., MORENO-HAGELSIEB, G., EGUIARTE, L. E., SOUZA, V., HERRERA-ESTRELLA, L. & OLMEDO, G. 2010. Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. *BMC Genomics,* 11**,** 332.

ARENAS, M. & POSADA, D. 2010. The effect of recombination on the reconstruction of ancestral sequences. *Genetics,* 184**,** 1133-9.

BAILEY, S. 1994. The CCP4 Suite - Programs for protein crystallography. *Acta Crystallographica Section D-Biological Crystallography,* 50**,** 760-763.

BATTISTUZZI, F. U., FEIJAO, A. & HEDGES, S. B. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol,* 4**,** 44.

BECERRA, A., DELAYE, L., LAZCANO, A. & ORGEL, L. E. 2007. Protein disulfide oxidoreductases and the evolution of thermophily: was the last common ancestor a heat-loving microbe? *J Mol Evol,* 65**,** 296-303.

BLUM, J. S., BINDI, A. B., BUZZELLI, J., STOLZ, J. F. & OREMLAND, R. S. 1998. *Bacillus arsenicoselenatis,* sp nov, and *Bacillus selenitireducens,* sp nov: two haloalkaliphiles from Mono Lake, California that respire oxyanions of selenium and arsenic. *Archives of Microbiology,* 171**,** 19-30.

BRIDGHAM, J. T., CARROLL, S. M. & THORNTON, J. W. 2006. Evolution of hormone-receptor complexity by molecular exploitation. *Science,* 312**,** 97-101.

BRIDGHAM, J. T., EICK, G. N., LARROUX, C., DESHPANDE, K., HARMS, M. J., GAUTHIER, M. E., ORTLUND, E. A., DEGNAN, B. M. & THORNTON, J. W. 2010. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol,* 8.

BRIDGHAM, J. T., ORTLUND, E. A. & THORNTON, J. W. 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature,* 461**,** 515-9.

BROCHIER, C. & PHILIPPE, H. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature,* 417**,** 244.

CAI, W., PEI, J. & GRISHIN, N. V. 2004. Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol,* 4**,** 33.

CAMPBELL, S. A., RICHARDS, T. A., MUI, E. J., SAMUEL, B. U., COGGINS, J. R., MCLEOD, R. & ROBERTS, C. W. 2004. A complete shikimate pathway in Toxoplasma gondii: an ancient eukaryotic innovation. *International Journal for Parasitology,* 34**,** 5-13.

CARROLL, S. M., ORTLUND, E. A. & THORNTON, J. W. 2011. Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genet,* 7**,** e1002117.

CHANDRASEKHARAN, U. M., SANKER, S., GLYNIAS, M. J., KARNIK, S. S. & HUSAIN, A. 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science,* 271**,** 502-5.

CHANG, B. S., JONSSON, K., KAZMI, M. A., DONOGHUE, M. J. & SAKMAR, T. P. 2002. Recreating a functional ancestral archosaur visual pigment. *Mol Biol Evol,* 19**,** 1483-9.

CHINEN, A., MATSUMOTO, Y. & KAWAMURA, S. 2005. Reconstitution of ancestral green visual pigments of zebrafish and molecular mechanism of their spectral differentiation. *Mol Biol Evol,* 22**,** 1001-10.

CROSS, P. J., DOBSON, R. C., PATCHETT, M. L. & PARKER, E. J. 2011. Tyrosine latching of a regulatory gate affords allosteric control of aromatic amino acid biosynthesis. *J Biol Chem,* 286**,** 10216-24.

DEAN, A. M. & DVORAK, L. 1995. The role of glutamate 87 in the kinetic mechanism of *Thermus thermophilus* isopropylmalate dehydrogenase. *Protein Science,* 4**,** 2156-2167.

DELANO, W. L. 2002. The PyMOL Molecular Graphics System. *DeLano Scientific.*

DELEO, A. B., DAYAN, J. & SPRINSON, D. B. 1973. Purification and kinetics of tyrosine-sensitive 3-deoxy-D-*arabino*-heptulosonic acid 7-phosphate synthetase from *Salmonella*. *Journal of Biological Chemistry,* 248**,** 2344-2353.

DELEO, A. B. & SPRINSON, D. B. 1968. Mechanism of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP) synthetase. *Biochemical and Biophysical Research Communications,* 32**,** 873-&.

DENIZCI, A. A., KAZAN, D., ABELN, E. C. & ERARSLAN, A. 2004. Newly isolated *Bacillus clausii* GMBAE 42: an alkaline protease producer capable to grow under higly alkaline conditions. *J Appl Microbiol,* 96**,** 320-7.

DEWICK, P. M. 1995. The biosynthesis of shikimate metabolites. *Natural Product Reports,* 12**,** 579-607.

DIDELOT, X., LAWSON, D., DARLING, A. & FALUSH, D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics,* 186**,** 1435-49.

DRUMMOND, A., ASHTON, B., BUXTON, S., CHEUNG, M., COOPER, A., DURAN, C., FIELD, M., HELED, J., KEARSE, M., MARKOWITZ, S., MOIR, R., STONES-HAVAS, S., STURROCK, S., THIERER, T. & WILSON, A. 2011. Geneious v5.4, Available from http://www.genious.com.

DUKE, C. C., MACLEOD, J. K. & WILLIAMS, J. F. 1981. Nuclear magnetic-resonance studies of D-erythrose 4-phosphate in aqueous-solution - structures of the major contributing monomeric and dimeric forms. *Carbohydrate Research,* 95**,** 1-26.

EASTER, A. D. 2010. *Decoupling Enzyme Catalysis from Thermal Denaturation.* MSc, University of Waikato.

EMSLEY, P. & COWTAN, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography,* 60**,** 2126-2132.

ERICSSON, U. B., HALLBERG, B. M., DETITTA, G. T., DEKKER, N. & NORDLUND, P. 2006. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem,* 357**,** 289-98.

EVANS, D. A., BEUKES, N. J. & KIRSCHVINK, J. L. 1997. Low-latitude glaciation in the Palaeoproterozoic era. *Nature,* 386**,** 262-266.

FIELD, S. F. & MATZ, M. V. 2010. Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. *Mol Biol Evol,* 27**,** 225-33.

FRITZE, D. & PUKALL, R. 2001. Reclassification of bioindicator strains *Bacillus subtilis* DSM 675 and *Bacillus subtilis* DSM 2277 as Bacillus *atrophaeus*. *Int J Syst Evol Microbiol,* 51**,** 35-7.

GALTIER, N., TOURASSE, N. & GOUY, M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science,* 283**,** 220-221.

GAUCHER, E. A., GOVINDARAJAN, S. & GANESH, O. K. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature,* 451**,** 704-7.

GAUCHER, E. A., THOMSON, J. M., BURGAN, M. F. & BENNER, S. A. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature,* 425**,** 285-8.

GORDON, R. E. 1972. The Genus *Bacillus*. *In:* LASKIN, A. I. & LECHEVALIER, H. (eds.) *CRC Handbook of Microbiology vol 1: Organismic Biology.* Cleveland: CRC Press.

GOSSET, G., BONNER, C. A. & JENSEN, R. A. 2001. Microbial origin of plant-type 2-keto-3-deoxy-D-arabino-heptulosonate 7-phosphate synthases, exemplified by the chorismate- and tryptophan-regulated enzyme from *Xanthomonas campestris*. *Journal of Bacteriology,* 183**,** 4061-4070.

GROMIHA, M. M., OOBATAKE, M. & SARAI, A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem,* 82**,** 51-67.

HALL, B. G. 2006. Simple and accurate estimation of ancestral protein sequences. *Proc Natl Acad Sci U S A,* 103**,** 5431-6.

HANSON-SMITH, V., KOLACZKOWSKI, B. & THORNTON, J. W. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol,* 27**,** 1988-99.

HARTMANN, M., SCHNEIDER, T. R., PFEIL, A., HEINRICH, G., LIPSCOMB, W. N. & BRAUS, G. H. 2003. Evolution of feedback-inhibited β/α barrel isoenzymes by gene duplication and a single mutation. *Proc Natl Acad Sci U S A,* 100**,** 862-7.

HILLIS, D. M., BULL, J. J., WHITE, M. E., BADGETT, M. R. & MOLINEUX, I. J. 1992. Experimental phylogenetics - generation of a known phylogeny. *Science,* 255**,** 589-592.

HOBBS, J. K., SHEPHERD, C., SAUL, D. J., DEMETRAS, N. J., HAANING, S., MONK, C. R., DANIEL, R. M. & ARCUS, V. L. 2012. On the origin and evolution of thermophily: reconstruction of functional Precambrian enzymes from ancestors of *Bacillus*. *Mol Biol Evol,* 29**,** 825-35.

HOFFMANN, P. J., DOY, C. H. & CATCHESIDE, D. E. 1972. The separation of three allosterically inhibitable 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthases from extracts of *Neurospora crassa* and the purification of the tyrosine inhibitable isoenzyme. *Biochim Biophys Acta,* 268**,** 550-61.

HOLMGREN, A. 1995. Thioredoxin structure and mechanism: conformational changes on oxidation of the active site sulfhydryls to a disulfide. *Structure,* 3**,** 239-243.

HORWICH, A. L., FENTON, W. A., CHAPMAN, E. & FARR, G. W. 2007. Two families of chaperonin: physiology and mechanism. *Annu Rev Cell Dev Biol,* 23**,** 115-45.

HUANG, L., MONTOYA, A. L. & NESTER, E. W. 1974. Characterization of the functional activities of the subunits of 3-deoxy-D-arabinoheptulosonate 7-

phosphate synthetase-chorismate mutase from *Bacillus subtilis* 168. *J Biol Chem,* 249**,** 4473-0.

HUELSENBECK, J. P. & BOLLBACK, J. P. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic Biology,* 50**,** 351-366.

JAFFRES, J. B. D., SHIELDS, G. A. & WALLMANN, K. 2007. The oxygen isotope evolution of seawater: A critical review of a long-standing controversy and an improved geological water cycle model for the past 3.4 billion years. *Earth-Science Reviews,* 83**,** 83-122.

JENSEN, R. A. & NESTER, E. W. 1966. Regulatory enzymes of aromatic amino acid biosynthesis in *Bacillus subtilis*. I. Purification and properties of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthetase. *J Biol Chem,* 241**,** 3365-72.

JENSEN, R. A., XIE, G., CALHOUN, D. H. & BONNER, C. A. 2002. The correct phylogenetic relationship of KdsA (3-deoxy-D-*manno*-octulosonate 8-phosphate synthase) with one of two independently evolved classes of AroA (3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase). *Journal of Molecular Evolution,* 54**,** 416-423.

JERMANN, T. M., OPITZ, J. G., STACKHOUSE, J. & BENNER, S. A. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature,* 374**,** 57-9.

JIAO, W., HUTTON, R. D., CROSS, P. J., JAMESON, G. B. & PARKER, E. J. 2012. Dynamic cross-talk among remote binding sites: the molecular basis for unusual synergistic allostery. *J Mol Biol,* 415**,** 716-26.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci,* 8**,** 275-82.

JU, J., XU, S., WEN, J., LI, G., OHNISHI, K., XUE, Y. & MA, Y. 2009. Characterization of endogenous pyridoxal 5'-phosphate-dependent alanine racemase from *Bacillus pseudofirmus* OF4. *J Biosci Bioeng,* 107**,** 225-9.

KNAGGS, A. R. 2001. The biosynthesis of shikimate metabolites. *Natural Product Reports,* 18**,** 334-355.

KNAUTH, L. P. 2005. Temperature and salinity history of the Precambrian ocean: implications for the course of microbial evolution. *Palaeogeography Palaeoclimatology Palaeoecology,* 219**,** 53-69.

KONNO, A., OGAWA, T., SHIRAI, T. & MURAMOTO, K. 2007. Reconstruction of a probable ancestral form of conger eel galectins revealed their rapid adaptive evolution process for specific carbohydrate recognition. *Mol Biol Evol,* 24**,** 2504-14.

KRISHNAN, N. M., SELIGMANN, H., STEWART, C. B., DE KONING, A. P. & POLLOCK, D. D. 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference. *Mol Biol Evol,* 21**,** 1871-83.

KRISSINEL, E. & HENRICK, K. 2004. Secondary-structure matching (PDBeFold), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D-Biological Crystallography,* 60**,** 2256-2268.

KRISSINEL, E. & HENRICK, K. 2007. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.,* 372**,** 774-797.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. &

HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-8.

LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S. & THORNTON, J. M. 1993. PROCHECK - a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography,* 26**,** 283-291.

LE, S. Q. & GASCUEL, O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution,* 25**,** 1307-1320.

LESLIE, A. G. W. 1992. Recent changes in the MOSFLM package for processing film and image data. *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography,* 26.

LI, G., MA, J. & ZHANG, L. 2010. Greedy selection of species for ancestral state reconstruction on phylogenies: elimination is better than insertion. *PLoS One,* 5**,** e8985.

LI, Y., SUINO, K., DAUGHERTY, J. & XU, H. E. 2005. Structural and biochemical mechanisms for the specificity of hormone binding and coactivator assembly by mineralocorticoid receptor. *Mol Cell,* 19**,** 367-80.

LIGHT, S. H., HALAVATY, A. S., MINASOV, G., SHUVALOVA, L. & ANDERSON, W. F. 2012. Structural analysis of a 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase with an N-terminal chorismate mutase-like regulatory domain. *Protein Sci,* 21**,** 887-95.

MA, N., WEI, L., FAN, Y. & Q., H. 2012. Heterologous expression and characterization of soluble recombinant 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase from *Actinosynnema pretiosum* ssp. *auranticum* ATCC31565 through co-expression with chaperones in *Escherichia coli*. *Protein Expression and Purification,* 82**,** 263-269.

MALCOLM, B. A., WILSON, K. P., MATTHEWS, B. W., KIRSCH, J. F. & WILSON, A. C. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature,* 345**,** 86-9.

MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, L. C. & READ, R. J. 2007. Phaser crystallographic software. *Journal of Applied Crystallography,* 40**,** 658-674.

MOHAN, C. 2003. *Buffers: A Guide for the Preparation and Use of Buffers in Biological Systems,* Darmstadt, EMD Biosciences Inc.

MURSHUDOV, G. N., VAGIN, A. A. & DODSON, E. J. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D-Biological Crystallography,* 53**,** 240-255.

NAKAMURA, L. K. 1998. *Bacillus pseudomycoides* sp. nov. *Int J Syst Bacteriol,* 48 Pt 3**,** 1031-5.

NAKAMURA, L. K. & JACKSON, M. A. 1995. Clarification of the taxonomy of *Bacillus mycoides*. *International Journal of Systematic Bacteriology,* 45**,** 46-49.

NIMMO, G. A. & COGGINS, J. R. 1981. Some kinetic properties of the tryptophan-sensitive 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Neurospora crassa*. *Biochem J,* 199**,** 657-65.

NOGI, Y., TAKAMI, H. & HORIKOSHI, K. 2005. Characterization of alkaliphilic *Bacillus* strains used in industry: proposal of five novel species. *Int J Syst Evol Microbiol,* 55**,** 2309-15.

OAKLEY, T. H. & CUNNINGHAM, C. W. 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution,* 54**,** 397-405.

OGINO, T., GARNER, C., MARKLEY, J. L. & HERRMANN, K. M. 1982. Biosynthesis of aromatic compounds: ^{13}C NMR spectroscopy of whole *Escherichia coli* cells. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences,* 79**,** 5828-5832.

ONDERKA, D. K. & FLOSS, H. G. 1969. Stereospecificity of the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthetase reaction. *Biochem Biophys Res Commun,* 35**,** 801-4.

ORTLUND, E. A., BRIDGHAM, J. T., REDINBO, M. R. & THORNTON, J. W. 2007. Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science,* 317**,** 1544-8.

PACE, N. R. 1991. Origin of life - facing up to the physical setting. *Cell,* 65**,** 531-533.

PAGE, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci,* 12**,** 357-8.

PAULING, L. & ZUCKERKANDL, E. 1963. Chemical paleogenetics molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.,* 17**,** S9-S16.

PEREZ-JIMENEZ, R., INGLES-PRIETO, A., ZHAO, Z. M., SANCHEZ-ROMERO, I., ALEGRE-CEBOLLADA, J., KOSURI, P., GARCIA-MANYES, S., KAPPOCK, T. J., TANOKURA, M., HOLMGREN, A., SANCHEZ-RUIZ, J. M., GAUCHER, E. A. & FERNANDEZ, J. M. 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol,* 18**,** 592-6.

POSADA, D. 2003. Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Curr Protoc Bioinformatics,* Chapter 6**,** Unit 6 5.

POSADA, D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol,* 25**,** 1253-6.

RAMBAUT, A. 2009. *http://tree.bio.ed.ac.uk/software/figtree/* [Online].

REICHAU, S., JIAO, W. T., WALKER, S. R., HUTTON, R. D., BAKER, E. N. & PARKER, E. J. 2011. Potent inhibitors of a shikimate pathway enzyme from *Mycobacterium tuberculosis*: combining mechanism- and modeling-based design. *Journal of Biological Chemistry,* 286.

ROBERT, F. & CHAUSSIDON, M. 2006. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature,* 443**,** 969-972.

ROBERTS, F., ROBERTS, C. W., JOHNSON, J. J., KYLE, D. E., KRELL, T., COGGINS, J. R., COOMBS, G. H., MILHOUS, W. K., TZIPORI, S., FERGUSON, D. J. P., CHAKRABARTI, D. & MCLEOD, R. 1998. Evidence for the shikimate pathway in apicomplexan parasites. *Nature,* 393**,** 801-805.

RUNNEGAR, B. 2000. Loophole for snowball earth. *Nature,* 405**,** 403-404.

RUPP, B. 2010. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology,* New York, Garland Science.

SANDERSON, M. J. 2003. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics,* 19**,** 301-2.

SCHNAPPAUF, G., HARTMANN, M., KUNZLER, M. & BRAUS, G. H. 1998. The two 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase isoenzymes from *Saccharomyces cerevisiae* show different kinetic modes of inhibition. *Archives of Microbiology,* 169**,** 517-524.

SCHOFIELD, L. R., ANDERSON, B. F., PATCHETT, M. L., NORRIS, G. E., JAMESON, G. B. & PARKER, E. J. 2005. Substrate ambiguity and crystal structure of *Pyrococcus furiosus* 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase: An ancestral 3-deoxyald-2-ulosonate-phosphate synthase? *Biochemistry,* 44**,** 11950-62.

SCHOFIELD, L. R., PATCHETT, M. L. & PARKER, E. J. 2004. Expression, purification, and characterization of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Pyrococcus furiosus*. *Protein Expression and Purification,* 34**,** 17-27.

SCHONER, R. & HERRMANN, K. M. 1976. 3-Deoxy-D-*arabino*-heptulosonate 7-phosphate synthase. Purification, properties, and kinetics of the tyrosine-sensitive isoenzyme from *Escherichia coli*. *J Biol Chem,* 251**,** 5440-7.

SHI, Y. & YOKOYAMA, S. 2003. Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proc Natl Acad Sci U S A,* 100**,** 8308-13.

SHUMILIN, I. A., BAUERLE, R., WU, J., WOODARD, R. W. & KRETSINGER, R. H. 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. *J Mol Biol,* 341**,** 455-66.

SHUMILIN, I. A., KRETSINGER, R. H. & BAUERLE, R. H. 1999. Crystal structure of phenylalanine-regulated 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase from *Escherichia coli*. *Structure with Folding & Design,* 7**,** 865-875.

STACKHOUSE, J., PRESNELL, S. R., MCGEEHAN, G. M., NAMBIAR, K. P. & BENNER, S. A. 1990. The ribonuclease from an extinct bovid ruminant. *FEBS Lett,* 262**,** 104-6.

STETTER, K. O. 2006. Hyperthermophiles in the history of life. *Philos Trans R Soc Lond B Biol Sci,* 361**,** 1837-42.

TAYLOR, W. R. & JONES, D. T. 1993. Deriving an amino acid distance matrix. *J Theor Biol,* 164**,** 65-83.

TERWILLIGER, T. C., GROSSE-KUNSTLEVE, R. W., AFONINE, P. V., MORIARTY, N. W., ZWART, P. H., HUNG, L. W., READ, R. J. & ADAMS, P. D. 2008. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D-Biological Crystallography,* 64**,** 61-69.

THOMSON, J. M., GAUCHER, E. A., BURGAN, M. F., DE KEE, D. W., LI, T., ARIS, J. P. & BENNER, S. A. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet,* 37**,** 630-5.

THORNTON, J. W., NEED, E. & CREWS, D. 2003. Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. *Science,* 301**,** 1714-7.

TRIBE, D. E., CAMAKARIS, H. & PITTARD, J. 1976. Constitutive and repressible enzymes of common pathway of aromatic biosynthesis in *Escherichia coli* K-12: Regulation of enzyme synthesis at different growth rates. *Journal of Bacteriology,* 127**,** 1085-1097.

UGALDE, J. A., CHANG, B. S. & MATZ, M. V. 2004. Evolution of coral pigments recreated. *Science,* 305**,** 1433.

VAGENENDE, V., YAP, M. G. & TROUT, B. L. 2009. Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. *Biochemistry,* 48**,** 11084-96.

WEBBY, C. J., BAKER, H. M., LOTT, J. S., BAKER, E. N. & PARKER, E. J. 2005a. The structure of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Mycobacterium tuberculosis* reveals a common catalytic scaffold and ancestry for type I and type II enzymes. *Journal of Molecular Biology,* 354**,** 927-939.

WEBBY, C. J., JIAO, W., HUTTON, R. D., BLACKMORE, N. J., BAKER, H. M., BAKER, E. N., JAMESON, G. B. & PARKER, E. J. 2010. Synergistic allostery, a sophisticated regulatory network for the control of aromatic amino acid biosynthesis in *Mycobacterium tuberculosis*. *J Biol Chem,* 285**,** 30567-76.

WEBBY, C. J., PATCHETT, M. L. & PARKER, E. J. 2005b. Characterization of a recombinant type II 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase from *Helicobacter pylori*. *Biochem J,* 390**,** 223-30.

WILLIAMS, P. D., POLLOCK, D. D., BLACKBURNE, B. P. & GOLDSTEIN, R. A. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol,* 2**,** e69.

WOESE, C. R. 1987. Bacterial evolution. *Microbiol Rev,* 51**,** 221-71.

WU, J., HOWE, D. L. & WOODARD, R. W. 2003. *Thermotoga maritima* 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP) synthase: The ancestral eubacterial DAHP synthase? *J Biol Chem,* 278**,** 27525-31.

WU, J., SHEFLYAN, G. Y. & WOODARD, R. W. 2005. *Bacillus subtilis* 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase revisited: resolution of two long-standing enigmas. *Biochem J,* 390**,** 583-90.

WU, J. & WOODARD, R. W. 2006. New insights into the evolutionary links relating to the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase subfamilies. *J Biol Chem,* 281**,** 4042-8.

YANG, Z. 2006. *Computational Molecular Evolution,* New York, Oxford University Press Inc.

YANG, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

YANG, Z., KUMAR, S. & NEI, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics,* 141**,** 1641-50.

YOKOYAMA, S., YANG, H. & STARMER, W. T. 2008. Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. *Genetics,* 179**,** 2037-43.

ZHANG, J. & ROSENBERG, H. F. 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc Natl Acad Sci U S A,* 99**,** 5486-91.

ZHOU, L., WU, J., JANAKIRAMAN, V., SHUMILIN, I. A., BAUERLE, R., KRETSINGER, R. H. & WOODARD, R. W. 2012. Structure and characterization of the 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase from *Aeropyrum pernix*. *Bioorg Chem,* 40**,** 79-86.

ZWICKL, D. 2006. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* PhD dissertation, The University of Texas at Austin.