Characterisation of Enzyme Evolution through Ancestral Enzyme Reconstruction

A thesis submitted in partial fulfilment

of the requirements for the degree

of

Masters in Biological Sciences

at

The University of Waikato

by

Erica Jean Prentice

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2013

Abstract

Through ancestral sequence reconstruction (ASR) techniques, ancient enzymes can be recreated and biochemically tested, giving insight into the enzymes' evolutionary history. A previous study by Hobbs et al. (2012) has shown that some ancestral 3-isopropylmalate dehydrogenase (IPMDH) enzymes of the *Bacillus* lineage are more catalytically efficient and kinetically stable than extant counterparts. Given these characteristics, this trend raises questions as to why ancestral *Bacillus* IPMDH enzymes have been superseded by catalytically slower and less kinetically stable counterparts. The homology between IPMDH and the dehydrogenases of tartrate, malate and isocitrate makes IPMDH an interesting model enzyme in terms of the evolution of substrate specificity.

Here, the reconstruction of a 2.7 billion year old enzyme has been attempted to extend the reconstruction of IPMDH back to the last common ancestor of the Firmicutes. This reconstruction tested the limits of ASR techniques in terms of time and levels of sequence divergence, especially for such a structurally complex enzyme. However, upon expression and purification, the enzyme was found to form an inactive, soluble aggregate. This suggests that current ASR techniques are too simplistic to reconstruct the complexity and divergence of IPMDH back as far as the last common ancestor of the Firmicutes. Enzyme evolution was investigated with ancestors from the *Bacillus* genus. Substrate promiscuity of ancestral enzymes was compared to a contemporary counterpart. It was concluded that the ancestral IPMDH enzymes tested do not show additional substrate promiscuity when compared to contemporary counterparts. The fitness of organisms carrying the IPMDH ancestors was assessed to establish what effects the high turnover rates and kinetic stability possessed by some ancestral IPMDH enzymes had on cells when functioning within the normal catalytic pathway for leucine. *In vivo*, the fastest and most kinetically stable ancestral IPMDH resulted in slower growth rates. This detrimental effect *in vivo* clarifies why this enzyme has been lost over evolutionary time. The X-ray crystal structure of the most recent IPMDH ancestor was also determined at 2.6 Å resolution. The structure of this ancestral IPMDH was found to be similar to other IPMDH structures, including the previously solved IPMDH from the last common ancestor of the *Bacillus*.

Acknowledgements

First, I would like to thank my supervisor Professor Vic Arcus for the guidance and support you have given, and the opportunities you have provided to me. I would also like to thank Dr Jo Hobbs for your time and expertise. From teaching me the basics in the lab, through to the writing of this thesis, your help has been invaluable.

To everyone else in the Proteins and Microbes lab, past and present: Dr Emma Andrews, Dr Judith Burrows, Dr Emma Summers, Vikas Chonira , Joel McMillan, Tifany Oulavallickal, Ali Ruthe, Abby Sharrock, and Chelsea Vickers, you all make the lab a wonderful place to work in. Your help and support (and baking) have made this process that much smoother and more enjoyable. Jo, Emma A, Tifany, Abby and Chelsea thank you for the gym trips to offset all the baking. Also, thank you to Judith for looking after us all in the lab.

Thank you also to the friends and family who have given their support and encouragement throughout this degree. Thanks to the friends within the university for the help and understanding over the past years of study, and the friends outside of university for providing the distractions from work. Also thank you to Hayden, for your continued support over the past few years.

Finally, I would also like to thank the University of Waikato for awarding me the University of Waikato Taught Postgraduate Fees Scholarship, the Science & Engineering Masters Fees Award, and the University of Waikato Masters Research Scholarship to help with funding.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

°C	degrees Celsius
Ω	ohms
3D	three-dimensional
abs	absorbance
AIC	Akaike information criterion
AMP	ampicillin
ANC	ancestor
APS	ammonium persulfate
Arg	arginine
Asp	aspartic acid
ASR	ancestral sequence reconstruction
BCVX	*Bacillus caldovelox*
BI	Bayesian inference
bp	base pair(s)
bya	billion years ago
cm	centimetre(s)
conc	concentration
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
Glu	glutamic acid
Gly	glycine
h	hour(s)
his-tag	poly histidine tag
ICD/*icd*	isocitrate dehydrogenase
Ile	isoleucine
IMAC	immobilised metal affinity chromatography
IPM	isopropylmalate
IPMDH	3-isopropylmalate dehydrogenase
IPTG	isopropyl β-D-1-thiogalactopyranoside
K	kelvin

KAN	kanamycin
kDa	kilo Daltons
KO	knock out
kV	kilovolts
L	litre
LB	Luria-Bertani
LCA	last common ancestor
Leu	leucine
LeuB	3-isopropylmalate dehydrogenase gene
Lys	lysine
M	molar
mA	milliamps
mAU	milli absorbance units
mdh	malate dehydrogenase
µF	microfarads
µg	micrograms
mg	milligrams
µl	microlitres
ml	millilitres
ML	maximum likelihood
mm	millimetre
mM	millimoles
MP	maximum parsimony
MPD	methyl pentanediol
MQ	milli Q
mya	million years ago
myr	million years
NAD^+	nicotinamide adenine dinucleotide (oxidised)
NADH	nicotinamide adenine dinucleotide (reduced)
nL	nanolitre
nm	nanometres
N-terminal	amine terminus of peptide chain
OD	optical density

OD_{600}	optical density at 600 nm
PDB	protein data bank
RMSD	root-mean-square deviation
rpm	revolutions per minute
s	second(s)
SD	standard deviation
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SE	size exclusion
SSM	secondary structure matching
TAE	tris-acetate-EDTA
TB	terrific broth
TEMED	N, N, N', N'-tetramethylethylenediamine
T_m	melting temperature
T_{opt}	optimal temperature for enzymatic activity
Tyr	tyrosine
UV	ultraviolet
Val	valine
v/v	volume per volume
w/v	weight per volume
yeaU	tartrate dehydrogenase

1 Introduction

1.1 Ancestral Sequence Reconstruction

Ancestral sequence reconstruction (ASR) has become a powerful tool in the study of evolutionary biology. The technique was first conceived by Pauling and Zuckerkandl (1963) when they proposed that it would be possible to infer ancestral states from a comparison of modern descendents, and synthesise these proteins in the laboratory for biochemical characterisation. Pauling and Zuckerkandl (1963) also performed the first inference, determining the sequence of an ancestral mammalian haemoglobin. However, technical constraints at the time prevented physical synthesis of the protein for biochemical characterisation.
It was not until almost 30 years later that the first reconstructions accompanied with synthesis and testing of the ancestral states was achieved. An ancestral ribonuclease from the bovid ruminants (Stackhouse et al. 1990) and lysozyme from the Galliformes order of birds (Malcolm et al. 1990) were both inferred and reconstructed in the laboratory. Since this time, ASR has been used to investigate a number of evolutionary questions. Improvements in DNA sequencing, computational processing and modelling of the evolutionary processes have seen the technique used in increasingly complex reconstructions, enabling access to evolutionary history that would otherwise be lost. Parallel improvements in physical reconstruction of the ancestral proteins have also furthered the technique, allowing full *de novo* reconstruction of sequences as opposed to original methods which used site directed mutagenesis of functionally important residues. This development is advantageous as the structure-function relationships of the protein do not need to be well known in advance to identify relevant residues, and the whole ancestral enzyme is able to be reconstructed (Hall 2006).

1.1.1 Ancestral Sequence Reconstruction Approaches

A number of different approaches have been utilised in ASR studies, the major methods being maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). Each of these methods differ in the criteria for optimal

phylogenetic tree selection, and as such have different assumptions and biases introduced into the final tree used for ancestral inference (Williams et al. 2006). The different assumptions introduced in each method make each technique suited to different reconstruction time frames and rates of evolution. All three of the methods utilise an alignment of contemporary sequences, and a phylogenetic tree summarising the relationships between the species; ML and BI also incorporate models of evolution into the inference stages (Zhang & Nei 1997; Cai et al. 2004). MP was used in early reconstructions due to the low computational cost, but has since been superseded by ML methods which incorporate models to encapsulate the complexity of the evolutionary processes being studied. BI has been the most recent method to be proposed, however it is still debated if this approach improves the accuracy of the reconstruction (Hall 2006; Hanson-Smith et al. 2010; Hobbs et al. 2012).

1.1.1.1 *Maximum Parsimony Methods*

Initial ancestral reconstructions employed the only technique available at the time: maximum parsimony inference methods (Fitch 1971). Examples of reconstructions utilising this method include reconstructions of ribonucleases from bovid ruminants (Stackhouse et al. 1990) and Artiodactyla (Jermann et al. 1995), lysozyme from *Galliformes* (Malcolm et al. 1990) and the reconstruction of a retrotransposon promoter sequence (Adey et al. 1994). In the MP method, evolution is considered to follow the most parsimonious route, i.e. the model with the least character state changes necessary to explain the contemporary diversity (Camin & Sokal 1965). Each amino acid site is considered separately, and ancestral states at the node points are assigned so as to result in the least number of character state changes over the tree (Fitch 1971). As this method does not take account of tree branch lengths or biases in amino acid substitution rates, final inferences are generally unreliable, especially with high levels of sequence divergence (Collins et al. 1994b). With sequences of low divergence, where different rates of evolution and branch lengths are not an issue, MP is sufficient to model the phylogeny accurately. However, MP fails to distinguish between equally parsimonious reconstructions when more than one state is predicted with equal probability under the model (Zhang & Nei 1997; Cai et al. 2004). Despite

the methods early use, these issues with the basis of this method have seen it superseded by more complex approaches which address these errors (Liberles 2007).

1.1.1.2 *Maximum Likelihood Methods*

Maximum likelihood methods for ancestral reconstructions were developed by Yang, Kumar, and Nei (1995). The major advancement in ML methodology was the incorporation of a measurement of accuracy of the inferred ancestral sequence. For each possible amino acid at every sequence position in the ancestral nodes, the likelihood of the extant sequences given the ancestral states, tree and model of evolution is calculated (Hanson-Smith et al. 2010). The inferred sequence is then comprised of the most likely amino acid at each position, and the uncertainty at each site gives an overall accuracy of the ancestral sequence. Having such a quantification of the uncertainty is essential to establish confidence in the accuracy of ancestral sequences and any other conclusions that may be drawn from the reconstruction.

Unlike MP, ML also considers branch lengths and amino acid or nucleotide substitution rates in the reconstruction methods (Yang 2006). Taking account of these parameters by incorporating models of evolution, a wide variety of which are available, has allowed ML to be applied to a range of evolutionary questions (Galtier & Gouy 1998). As such, compared to MP methods ML offers a more realistic approach to the evolutionary process, as well as giving a measure of accuracy of the final reconstruction.

One shortcoming of ML is that the method fails to account for sampling errors when estimating parameters. This issue is amplified in smaller datasets, where sampling errors become more apparent (Yang 2006).

1.1.1.3 *Hierarchical Bayesian Inference Methods*

Hierarchical BI methods were proposed as a solution to the assumption in ML that the alignment, tree, model and model parameters are correct. BI achieves this by summing likelihoods over a range of possible trees and parameters (Hanson-Smith

et al. 2010). BI methods have been developed to integrate uncertainty in the tree topology (Pagel et al. 2004), parameters of the model (Schultz & Churchill 1999), and combinations of both (Huelsenbeck & Bollback 2001). These methods have also been advanced to incorporate non-constant rates of evolution, and further incorporated with the evolutionary models developed for ML methods (Yang 2006). Integrating these uncertainties into the reconstruction may improve inferences, however this has not yet been conclusively established (Hanson-Smith et al. 2010).

1.1.2 Ancestral Reconstruction Accuracy

Accuracy in ASR is crucial to the reliability of the technique, and any other conclusions that are drawn from the study (Yang 2006). However, assessment of the accuracy of any inferred sequence is by definition difficult, as the actual ancestral sequence is not known. One exception to this has been a known bacteriophage T7 phylogeny (Hillis et al. 1992). In this study, a mutagen was used in serial propagation to speed genetic change and create a phylogeny where topology and ancestral states were known. From the descendant viral sequences, a phylogenetic tree was constructed and used to infer the ancestral node states by the MP method. This allowed restriction digest maps of the inferred and actual ancestors to be compared. Of the sites this approach covered, 97.3 % of ancestral sites were correctly inferred, 1.4 % were incorrectly inferred, and 1.3 % were ambiguous. Of the eighteen incorrectly inferred sites, three involved ancestral states that were no longer represented in the descendent sequences. Such occurrences of lost information are likely to be a common occurrence in ASR, but are not apparent if ancestral sequences are not available. In contrast, a similar study using a bacteriophage phylogeny and similar methodology gave conflicting results (Oakley & Cunningham 2000). Comparison of inferred and actual viral ancestral states revealed high inaccuracy in the inferred sequences for various approaches of MP, even when the known parent sequence was used to improve the analysis. Computer simulations indicated that the reconstruction was prone to errors due to the fast rates of evolution in the viral sequences and directional bias of the character evolution. To date, this system has not been applied to ML and

BI inference methods. The relevance of this artificially accelerated viral system to slowly evolving natural phylogenies over millions of years is also not known.

Despite the recent dominance of ML inference in ASR studies, the accuracy of this approach has been questioned. A number of studies have concluded that BI produces more reliable and realistic ancestral sequences. In a study of mitochondrial cytochrome b and cytochrome oxidase subunit 1 from primates, BI was found to infer ancestral sequences with nucleotide frequencies more similar to extant ratios than ML. Simulations determined this effect in the ML inference to be likely caused by an inherent deterministic bias in the reconstruction. The functionality of ancestral mitochondrial primate tRNAs was also assessed theoretically, based on *in silico* stabilities due to complementary base pairing in the folded structure. Sequences inferred by ML were found to be theoretically less functional based on *in silico* stability than those from BI due to the deterministic biases in the ML reconstruction (Krishnan et al. 2004). Similar conclusions have been drawn with regards to the accuracy of BI over ML in the assessment of phylogenies produced by sequence evolution simulation programs, although the two methods were noted to be similar in accuracy (Hall 2006). In a more complex computer simulation of evolution, Williams, et al (2006) also concluded in favour of BI for ancestral reconstruction. Computational protein evolution was used, with purifying selection for variants retaining stability in a specified target structure. Following ASR, *in silico* thermodynamic properties of true and inferred ancestors were compared. ML, although being more accurate overall in the reconstruction, was found to overestimate the thermal stability of the ancestors, whereas BI did not. This effect was concluded to be due to the tendency for ML to eliminate residues that are detrimental to stability due to their infrequent spread over the tree. In the BI, the selection of less probable residues from the posterior probability at some sites counteracted any overestimation.

The above studies must be interpreted with caution, as they all involve theoretical phylogenies and/or theoretical measures of physical parameters of the proteins used to approximate fitness. A test of the accuracy of ML using a true bacteriophage phylogeny has been performed, using nested PCR to increase the rate of evolution (Sanson et al. 2002). Predicted ancestors showed 99.46 % –

99.87 % sequence similarity to the actual ancestors. The reconstructions were more error prone at deeper branching ancestors, suggesting sequences become less reliable with increased age. As stated in the previous bacteriophage phylogeny studies, the relevance of artificially accelerated neutral evolution to actual phylogenies must also be considered.

ML has been shown to outperform MP significantly in situations of high sequence divergence, especially when the phylogenetic tree includes long branches (Zhang & Nei 1997). However, the accuracy of ML and BI methods is still debated. Studies have reported contradictory results about the relative accuracies of ML and BI methods. Using simulated phylogenies, Hanson-Smith et al. (2010) concluded that the capacity of BI to incorporate phylogenetic uncertainty into the reconstruction was neither necessary or beneficial to the final sequence accuracy. Assessment of the accuracy of reconstructed elongation factor Tu has demonstrated that ML methods are robust even when there is uncertainty in parameters such as the phylogeny and amino acid frequencies of the evolutionary model (Gaucher et al. 2008). Oceanic temperature trends were in agreement with the melting temperature (T_m) of the inferred enzymes, lending credibility to the accuracy of the inferred properties of the ancestral enzymes.

Only one study has compared the physical properties of enzymes reconstructed by ML and BI methods. Hobbs et al (2012) reconstructed ancestral *Bacillus* isopropylmalate dehydrogenases (IPMDH) inferred by both ML and BI inference methods. The thermodynamic and kinetic properties of the reconstructed enzymes were compared to each other, and to contemporary *Bacillus* enzymes. Although enzymes from the same ancestral node inferred with the different methods displayed similar levels of thermophily, enzymes inferred by BI had abnormally high K_M values and were kinetically unstable. These factors suggested that the BI approach was inferring biologically unrealistic ancestral states. By comparison, ancestors inferred by ML methods exhibited thermodynamic and kinetic properties similar to those exhibited by extant enzymes.

A variety of other factors can influence the accuracy of ancestral inference and need to be considered in reconstructions. The number of taxon sequences included in the phylogenetic tree and which sequences to include is important to the final result. It has been demonstrated that more taxa are not necessarily better for the reconstruction in both MP and ML inferences (Li et al. 2008), but a sufficient number of taxa must still be sampled (Salisbury & Kim 2001). In some situations, a single representative extant taxon at the end of a slow evolving deep branching lineage results in a more accurate inference compared to a full tree. Recombination has also been shown to have a significant effect on ASR. Although recombination is more common in eukaryotes and viruses, recombination events also occur in bacterial genes. If ignored, recombination events have been shown to have significant effects on the reconstruction and alter the biological predictions made from the reconstruction (Arenas & Posada 2010).

Many factors influence the accuracy of ASR. When properly considered and accounted for, ASR studies can give accurate information about ancestral molecules, their function and the environment their host inhabited. However, with any result, it must be remembered that the sequences are inferred pseudo-data, not real observed data and treated with the cautions this necessitates (Yang 2006).

1.1.3 Previous Reconstructions

ASR has been used to reconstruct a variety of ancestral proteins and investigate a range of evolutionary questions. A list of these reconstructions is given in Table 1.1.

Table 1.1: Summary of ancestral reconstructions reported in the literature.

Protein reconstructed	Reference
ATPase	(Finnigan et al. 2012)
β-lactamase	(Risso et al. 2013)
Carbohydrate binding protein	(Krishnan et al. 2004)
Dehydrogenases	(Iwabata et al. 2005)
	(Thomson et al. 2005)
	(Hobbs et al. 2012)
Elongation factor Tu	(Gaucher et al. 2003)
	(Gaucher et al. 2008)
Fluorescent protein	(Ugalde et al. 2004)
	(Field & Matz 2010)
Galectin	(Konno et al. 2007)
Glycoside hydrolase	(Malcolm et al. 1990)
	(Voordeckers et al. 2012)
Ribonuclease	(Stackhouse et al. 1990)
	(Jermann et al. 1995)
	(Zhang & Rosenberg 2002)
Serine protease	(Chandrasekharan et al. 1996)
	(Wouters et al. 2003)
Steroid hormone receptor	(Thornton et al. 2003)
	(Li et al. 2005)
	(Bridgham, Carroll, & Thornton, 2006)
	(Ortlund et al. 2007)
	(Bridgham et al. 2009)
	(Bridgham et al. 2010)
	(Carroll et al. 2011)
Thioredoxin	(Perez-Jimenez et al. 2011)
Visual pigment	(Chang et al. 2002)
	(Shi & Yokoyama 2003)
	(Chinen et al. 2005)
	(Yokoyama et al. 2008b)
	(Yokoyama et al. 2008a)

Of these studies, the majority have been on binding proteins, not enzymes. Enzyme reconstruction has the advantage that there is an internal measure of

accuracy in the activity of the reconstructed enzymes. If too many errors are introduced in the inference, the reconstructed enzyme is likely to be non-functional or exhibit aberrant properties. The enzymes that have been reconstructed to date are generally young, and thus present fewer difficulties in the reconstruction as less sequence divergence has occurred. The three exceptions to this are the studies on dehydrogenases (Hobbs et al. 2012), thioredoxins (Perez-Jimenez et al. 2011) and β-lactamases (Risso et al. 2013). These studies have given insight into the thermal environments the Precambrian ancestors lived in, and the biochemical evolution of the enzymes. Perez-Jimenez, et al. (2011) concluded that thioredoxin enzymes have retained the same catalytic mechanisms, but have adapted to environmental changes in temperature and oceanic acidity. β-lactamase has been reconstructed back to 3 billion years ago (bya), the oldest reconstruction of an enzyme (Risso et al. 2013). Ancestral proteins from this reconstruction showed thermal denaturation profiles similar to extant enzymes, and broader substrate specificity than contemporary counterparts. The reconstruction of the more structurally complex dehydrogenase (Hobbs et al. 2012) is discussed later in section 1.3.4.

1.2 Enzyme Evolution and Fitness

Enzyme evolution is a complex process, involving changes which impact upon an enzymes thermodynamic and kinetic properties, expression level and interaction with other molecules. Enzymatic activity on the native substrate can also be affected, or changes can lead to functional divergence by increasing activity toward an alternative substrate. The impact of these changes on fitness at the organism level is dictated in the context of the environment and regulatory landscape within the cell. In studying enzyme evolution, random noise and the ignorance of selection pressures acting during the adaptation all cloud a complete understanding of the processes which have occurred (Arnold et al. 2001).

The study of enzyme evolution is difficult due to the long time scales over which change occurs. Often a directed evolution approach is taken, aiming to mimic natural evolution over short time scales to identify adaptive mechanisms (Arnold et al. 2001). The spread of functionality over extant species also provides information about the history of related enzymes (Gerlt & Babbitt 2001). ASR

studies have provided a significant tool in the study of protein evolution, allowing past evolutionary histories to be reconstituted and analysed (Dean & Thornton 2007).

1.2.1 Enzyme Evolution and Substrate Promiscuity

A commonly held theory of evolution suggests that primitive metabolism was driven by a small number of enzymes with broad specificities for a range of substrates. Over time, gene duplication and substrate specialisation have resulted in the large array of highly specific enzymes present today (Jensen 1976; Depristo 2007; Hernandez-Montes et al. 2008). This diversification and specialisation allows for enhanced catalytic speed as well as finer scale control of individual metabolic pathways (Copley 2012). Enzymes today that share the same protein fold adapted to different functions or to catalyse the same reactions in different pathways provide evidence for this. For example, the α/β-hydrolase fold is present in a variety of enzymes. Conserved residues in the active sites of these enzymes catalyse reactions through nucleophilic attack of the substrate. However, substrates and bonds targeted by the fold differ markedly between different enzymes in the superfamily (Ollis et al. 1992; O'Brien & Herschlag 1999).

Substrate promiscuity, coupled with gene duplication, is suggested to be the driving factor behind the evolution of new enzymatic functions; low activity toward an alternate substrate provides a starting point for adaptive evolution to maximise a new function (O'Brien & Herschlag 1999; James & Tawfik 2001; Copley 2012). Gene duplication precedes diversification as this eliminates the functional constrains on one duplicate sequence, allowing for a new function to develop (Zhang 2003).

In light of this theory of enzyme evolution, it could be expected that an increase in substrate promiscuity would be observed in ancestral enzymes. This trend has been found in the immune defence proteases from mammals (Wouters et al. 2003). A reconstructed ancestral protease had broad specificity for the complete range of primary substrate specificities found in the descendants. From this ancestor, a range of extant immune defence proteases have arisen with differing

primary specificities. A trend from generalist to specialist activity has also been described for the reconstruction of ancestral β-lactamases (Risso et al. 2013). However, the opposite has been found for an ancestral chymase. The ancestor of the diverse range of chymases showed specific and efficient angiotensin II forming activity. Thus it would appear that in this enzymatic lineage, specificity has been lost in favour of a broader activity range (Chandrasekharan et al. 1996).

1.2.2 Enzyme Evolution in a Cellular Context

Enzyme evolution is often studied at the level of single proteins. Characterisation of enzymes *in vitro*, in terms of rate and affinity for substrate, ignores the complex interactions occurring in the cellular context. It has been shown that the evolution rate of enzymes is linked to a variety of cellular factors. Highly connected enzymes (sharing product or reactant metabolites with other enzymes) evolve more slowly than less connected enzymes, and enzymes with high metabolic flux are subject to more evolutionary constraints (Vitkup et al. 2006). Studying enzyme evolution at the single enzyme level provides an oversimplified account of these multifaceted influences on the overall fitness of an enzyme in the context of a regulated enzymatic pathway (Vitkup et al. 2006).

The issues with this oversimplified view have been illustrated by studies of the biological breakdown of glucose. Despite detailed understanding of the pathway and enzymes acting at each step, *in vitro* kinetics did not correlate to an accurate description of *in vivo* activity for individual enzymes or the pathway as a whole (Teusink et al. 2000). Such discrepancies have been attributed to regulatory mechanisms (such as enzyme-enzyme interactions, enzyme modifications, and substrate channelling) and the differences between assay media and physiological conditions. Indeed, inclusion of some of these parameters has resulted in more accurate modelling of metabolite accumulation *in vivo* from data measured *in vitro* (van Eunen et al. 2012).

1.3 Isopropylmalate Dehydrogenase

A good candidate enzyme for ASR and the study of enzyme evolution is IPMDH (EC 1.1.1.85) from the leucine biosynthetic pathway. IPMDH is appropriate for ASR as it occupies a genomic region that shows little evidence of recombination in Bacillus (Didelot et al. 2010). Due to the essential nature of the enzyme across all bacteria, this is likely true across the Firmicutes. The gene shows moderate sequence conservation overall, with interspersed highly conserved regions to aid sequence alignment. Many structures of the enzyme are also available. The biosynthetic pathway and IPMDH enzyme are also conserved over a large number of organisms that can synthesise leucine (Stieglitz & Calvo 1974).

1.3.1 Biosynthetic Pathway of Leucine

Branched chain amino acids (leucine, isoleucine and valine) are synthesised using a core set of substrates and enzymes (Leavitt & Umbarger 1961; Freundlich et al. 1962). The pathway for the synthesis of these three amino acids is illustrated in Figure 1.1.

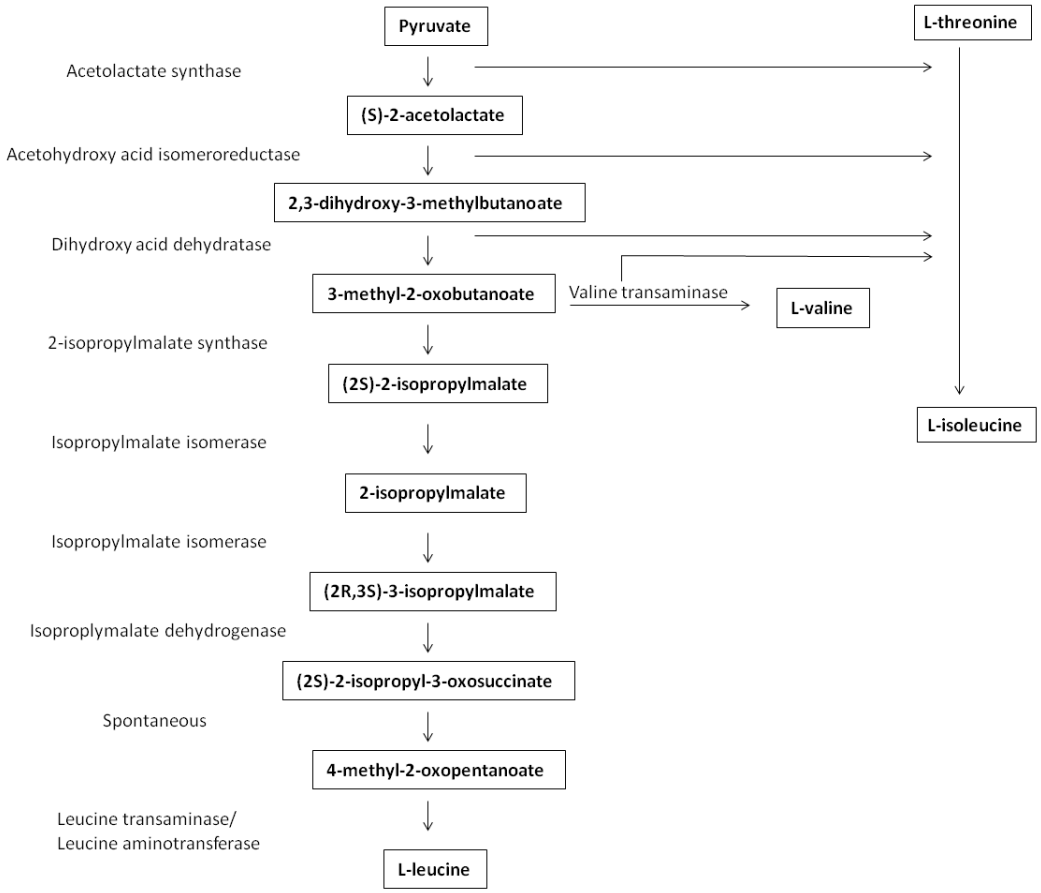


Figure 1.1: Biosynthetic pathway of the branched chain amino acids. Steps in the biosynthesis of L-valine and L-isoleucine have been omitted for clarity.

The synthesis of valine and leucine both proceed from pyruvate, which is converted to 3-methyl-2-oxobutanoate via three enzymatically controlled steps. The enzymes which act at these reactions also catalyse the early steps in the biosynthesis of isoleucine, converting 2-oxobutanoate and pyruvate to (S)-3-methyl-2-oxopentanoate. The same transaminase enzyme is also used in the biosynthesis of the two amino acids. From 3-methyl-2-oxobutanoate, valine and leucine are synthesised via separate routes. In the first step specific to the synthesis of leucine, 3-methyl-2-oxobutanoate is converted to (2S)-2-isopropylmalate by 2-isopropylmalate synthase (Calvo et al. 1962). Successive dehydration and addition reactions isomerise isopropylmalate, producing (2R,3S)-3-isopropylmalate (IPM) via an unsaturated intermediate (Gross et al. 1963). Isomerisation is catalysed by the heterodimeric protein isopropylmalate

isomerase. IPMDH catalyses the next oxidation step in the pathway, using NAD^+ in the parallel reduction reaction. The product of this reaction undergoes spontaneous decarboxylation, producing 4-methyl-2-oxopentanoate. Transamination of this structure produces the product leucine, via the transfer of an amine group from L-glutamate (Powell & Morrison 1978).

1.3.1.1 *Regulation of the Pathway*

The pathway to branched chain amino acids is regulated at a variety of points. As the pathway feeds carbon into three different amino acids, tight regulation of the pathway is necessary to ensure the correct ratio of the three molecules is made relative to the current abundance of each (Singh & Shaner 1995). Two mechanisms to specifically control the rate of leucine biosynthesis have been identified in *B. subtilis* (Ward & Zahler 1973). 2-isopropylmalate synthase activity is inhibited by the presence of L-leucine by up to 50 % at pH 7.5. The concentration of the enzymes 2-isopropylmalate synthase, isopropylmalate isomerase and IPMDH have also been shown to be decreased in *B. subtilis* when L-leucine is supplied (Ward & Zahler 1973). Initial enzymes in the process which are necessary for the production of all three amino acids in the pathway are only repressed when all three amino acids are present in excess (Freundlich et al. 1962). This system prevents an excess of any one amino acid inhibiting the biosynthesis of another, and specific control over the production of leucine is achieved at the steps specific to synthesis of this amino acid.

1.3.2 Catalysis by Isopropylmalate Dehydrogenase

IPMDH catalyses the penultimate step in the biosynthesis of leucine, converting (2R, 3S)-3-IPM to 4-methyl-2-oxopentanoate in a oxidative decarboxylation reaction (Fujita et al. 2001). The conversion proceeds in a two step process. Firstly the substrate is oxidised, in a reaction catalysed by IPMDH. Analysis of products from deuterated substrate have revealed that this reaction proceeds with retention of the stereochemistry at C3 (Kakinuma et al. 1989). Oxidation of is IPM is coupled to the reduction of NAD^+ to NADH. The (2S)-2-isopropyl-3-oxosuccinate formed in this reaction undergoes spontaneous decarboxylation to

produce 2-oxoisocaproate. The two steps of this reaction are illustrated in Figure 1.2.

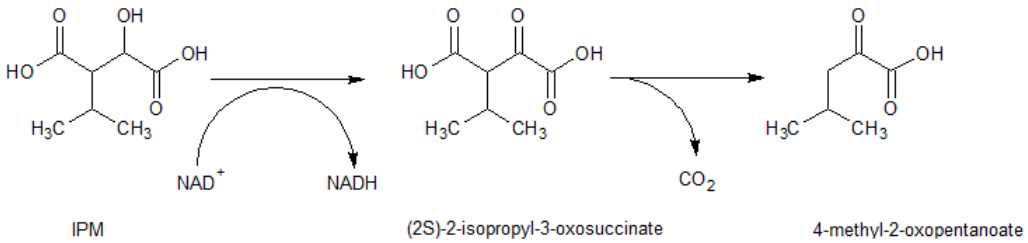


Figure 1.2: Two step reaction catalysed by IPMDH in the biosynthesis of leucine.

The activity of IPMDH is dependent on a divalent cation cofactor, typically Mn^{2+} or Mg^{2+}. Supplementation of reactions with Mn^{2+} results in optimal activity of the enzyme (Wallon et al. 1997a). This divalent metal ion aids the domain closure that drives the reaction in conjunction with IPM. Neither of the components in isolation induce a significant closure (0-5 %), and only with both components bound is a significant movement induced (Graczer et al. 2011b). Thus, the metal ion appears to be involved in structural movements rather than the actual chemistry of the catalysis. This would also be consistent with kinetic observations that NAD^+ binds before IPM, allowing all components to enter the active site prior to domain closure (Pirrung et al. 1994).

1.3.3 Crystal Structures of Isopropylmalate Dehydrogenase

Structures of IPMDH from a variety of organisms have been solved using X-ray crystallography. The initial structure from *Thermus thermophilus* revealed a dimeric structure held together by extensive hydrophobic interactions between the two subunits, as shown in Figure 1.3. Each monomer consists of a β sheet between two α/β domains, and can be divided into two domains on either side of a cleft in the structure. This structure is homologous to that of isocitrate dehydrogenase (ICD), but not to any other known structures. A hydrophobic pocket situated between the two domains includes the residues essential for substrate binding, consistent with the observed activity of the enzyme (Imada et al. 1991).

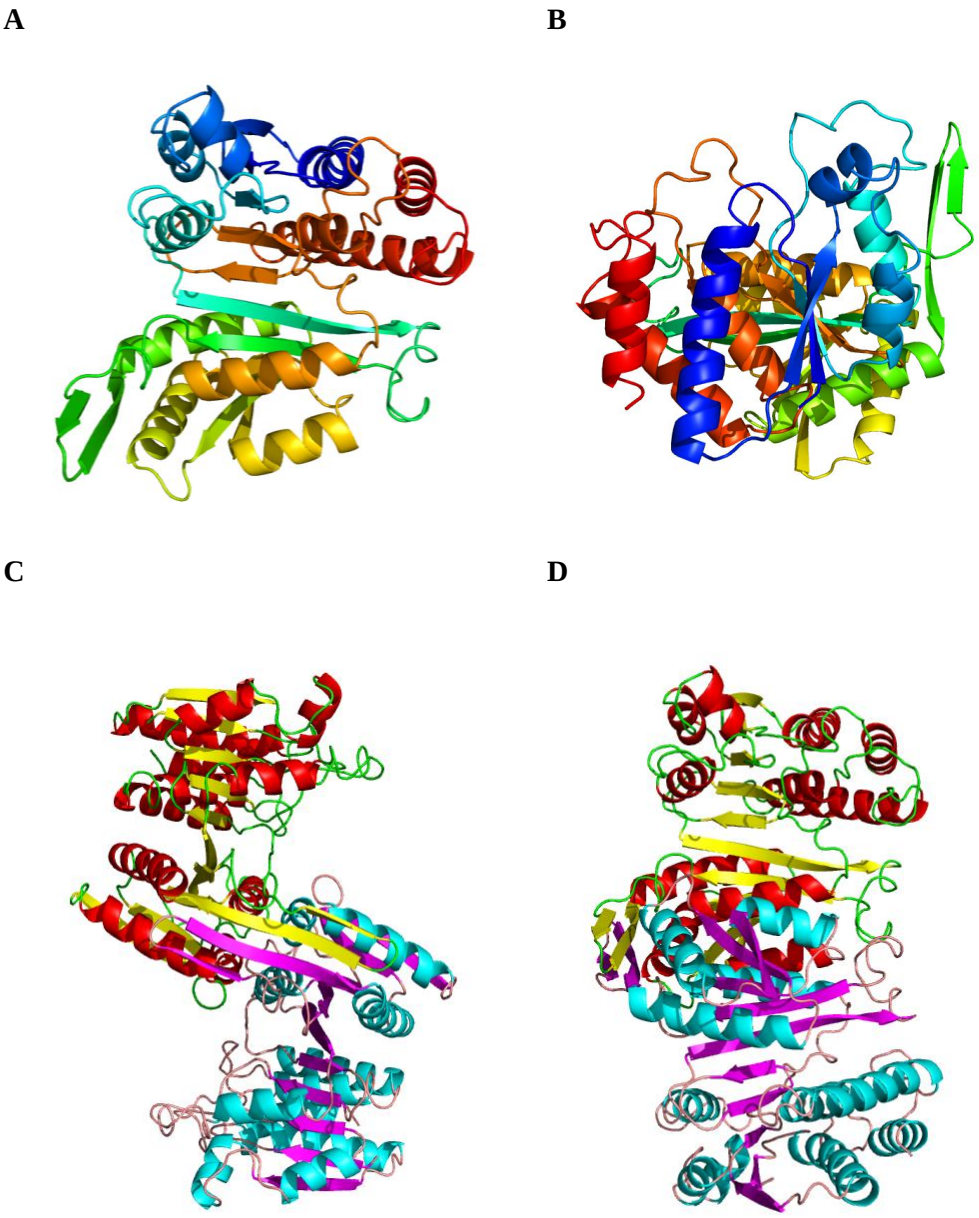


Figure 1.3: Structure of IPMDH from *T. Thermophilus* (PDB code 1IPD). (A) and (B) show the monomeric unit from the front (A) and rotated 90 ° to look down the structure (B); (C) shows the structure in the active dimeric form. Monomers are each coloured based on secondary structure. Active sites can be seen in the cleft in each monomer; (D) shows the dimer rotated 90 ° in the plane of the page.

From this structure, IPMDH structures have been solved from a variety of organisms. Structures with substrate and cofactors bound have also been solved.

A summary of the range of IPMDH structures that have been solved from bacterial species is provided in Table 1.2.

Table 1.2: Summary of solved structures for wild type IPMDH from bacteria.

Species	Bound components	PDB code	Reference
T. thermophilus		1IPD	(Imada et al. 1991)
	NAD^+	1HEX	(Hurley & Dean 1994)
	Inhibitor and NAD^+	2ZTW	(Nango et al. 2009)
	NADH and Mn^{2+}	2Y42	(Graczer et al. 2011a)
	IPM and Mn^{2+}	2Y41	(Graczer et al. 2011a)
	Mn^{2+}	2Y40	(Graczer et al. 2011a)
B. coagulans		1V53	(Tsuchiya et al. 1997)
A. ferrooxidans	IPM	1A05	(Imada et al. 1998)
S. typhimurium		1CNZ	(Wallon et al. 1997b)
E. coli		1CM7	(Wallon et al. 1997b)
M. tuberculosis		1W0D	(Singh et al. 2005)
A. thaliana		3R8W	(He et al. 2011)
T. maritima		1VLC	-
S. typhimurium	Mn^{2+}	1CNZ	-
Bacillus LCA		3U1H	(Hobbs et al. 2012)

1.3.3.1 *NAD^+ Binding Pocket*

NAD^+ binds into IPMDH near where IPM also associates with the structure, so that the two units are about 3.5 Å apart by edge to edge distances (Graczer et al. 2011a). The structure from *T. thermophilus* with NAD^+ bound (Hurley & Dean 1994) shows that NAD^+ is in an extended conformation, and the nicotinamide ring arranges in both syn and anti conformations. Interactions between IPMDH and NAD^+ form predominantly through the adenine moiety (see Figure 1.4). A hydrophobic pocket is formed from residues Ile11, Val15, Gly255, Leu254,

Ile279 and a portion of Asp326 (residue numbering for the *T. thermophilus* sequence). Hydrogen bonds are also formed between N2 and N6 and the main chain backbone at 286. The ribose sugar of the adenine moiety forms two hydrogen bonds to Asp278; Asp78 forms similar hydrogen bonds to the ribose at the nicotinamide end of the structure. This interaction is likely a major influence on cofactor specificity. Additional interactions to the nicotinamide ring (through Glu87) and the phosphate group hydrogen to main chain amide groups at residues 274 and 276 are present.

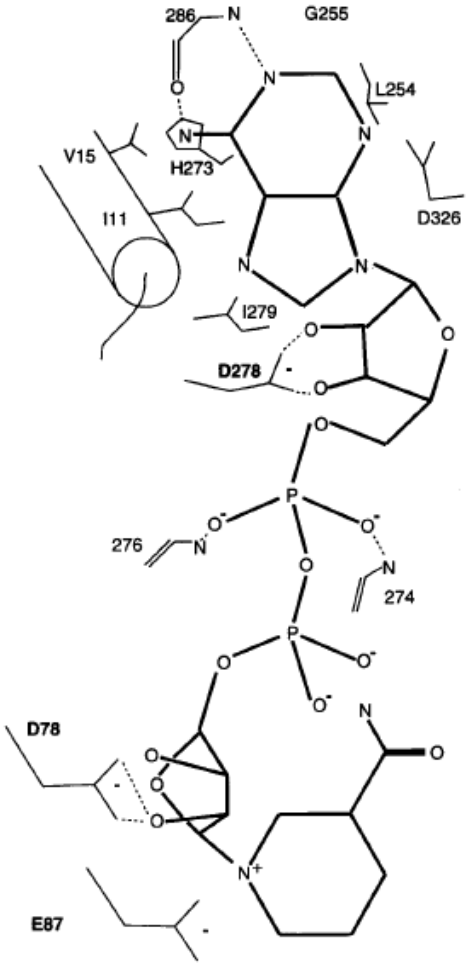


Figure 1.4: Interactions between NAD^+ and IPMDH. Depiction shows the syn conformation for the nicotinamide ring only. Hydrogen bonds are shown as dotted lines. Image taken from Hurley & Dean (1994).

Compared to structures without NAD^+ bound (Imada et al. 1991), this structure adopts a partially closed conformation. Domain closure is mainly affected by the binding of Mn^{2+} and IPM, with influences also from NAD^+ (Graczer et al. 2011b;

Graczer et al. 2011a). Binding of the adenine moiety to residues from four different regions of the enzyme appears to be the driving force behind this partial domain closure. These four binding segments are pulled inwards to form the hydrophobic binding pocket, closing the structure around the bound NAD^+ (Hurley & Dean 1994). Domain closure is only fully complete when in the dimeric form with IPM also bound, where the loop from one subunit is inserted into the cavity between the two domains of the other subunit, extending the extent of the hydrophobic interactions (Graczer et al. 2011a).

1.3.3.2 *Substrate Binding Pocket*

A structure of IPMDH from *Acidothiobacillus ferrooxidans* with the substrate IPM bound also shows a partial domain closure (Imada et al. 1998). IPM binds into a pocket formed between the two domains of the monomeric unit. The malate backbone of IPM hydrogen bonds to Arg95, Arg105, Arg133, Tyr140, Asp246 and Asp250, as well as residues Lys190 and Asp222 from the second subunit in the dimer. Glu88, Leu91 and Leu92 and Val193 from the second subunit form a hydrophobic surface into which the isopropyl moiety fits. The two carbons in the side chain of Glu88 contribute to the hydrophobic pocket, interacting in the space formed between the fork of the isopropyl group. The terminal carboxyl of the side chain is situated away from the substrate, potentially interacting with NAD^+. This is consistent with the observed decrease in binding of NAD^+ when Glu88 is mutated (Dean & Dvorak 1995). This set up of the active site also likely aids electron transfer from IPM to NAD^+ as the two constituents are in close proximity to each other. The catalytically essential divalent cation is also in close proximity to the two redox reactants (Graczer et al. 2011a).

1.3.3.2.1 Substrate Specificity

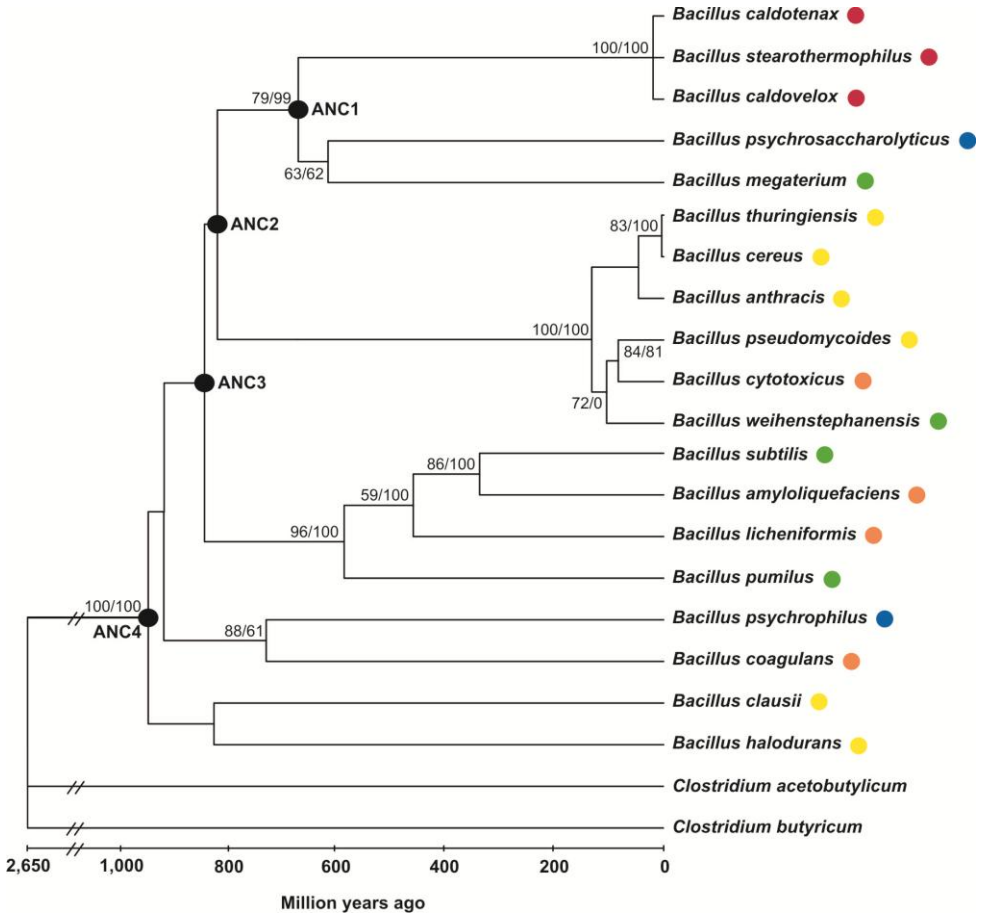
IPMDH shows activity towards a range of substrates. Activity has been measured against a variety of (2R,3S)-3-alkylmalates, with hydrogen, methyl, ethyl, iso-butyl, *tert*-butyl and iso-pentyl functional groups substituted in place of the C3 isopropyl group on the malate backbone of the natural substrate (Miyazaki et al. 1993; Fujita et al. 2001). Despite the homology of the structure of IPMDH to ICD, little to no catalysis of the oxidation of isocitrate is observed (Miyazaki et al.

1993). Given the hydrophobic pocket formed to accommodate the C3 isopropyl group in the native substrate, the lack of activity on isocitrate with a C3 carboxyl group is expected. Consistent with this idea, mutation of residues forming the hydrophobic pocket to ionisable groups accommodates the binding of isocitrate (Dean & Dvorak 1995). Glu88 has been suggested to be crucial to the activity against a variety of alkylmalates. The flexibility of this side chain allows accommodation of a range of hydrophobic C3 side groups into the active site (Imada et al. 1998).

1.3.4 Isopropylmalate Dehydrogenase in ASR Studies

IPMDH has previously been the subject of an ASR study and reconstructed back to the last common ancestor (LCA) of the *Bacillus* genus which existed approximately 950 million years ago [mya; (Hobbs et al. 2012)]. Four time points in this ancestry were reconstructed by ML methods and biochemically characterised. The chronogram and biochemical data for the enzymes characterised are given in Figure 1.5. This reconstruction suggested that the evolution of IPMDH over the last billion years in the *Bacillus* lineage has not been a linear process. The T_{opt} of the reconstructed enzymes fluctuated from thermophilic to mesophilic, and back to thermophilic over relatively short time scales. The biophysical properties of the ancestral enzymes also suggest that the evolution of IPMDH is complex. Both ANC1 and ANC4 exhibited significantly faster substrate turnover rates than extant IPMDHs, being three and seven times faster than IPMDH from the contemporary species *B. caldovelox* (BCVX). ANC4 also showed a higher kinetic barrier to unfolding than any of the other enzymes. This would suggest that in the evolution of IPMDH, factors other than the catalytic rate and stability have been maximised.

A



B

Enzyme	K_M (IPM) (mM)	K_M (NAD+) (mM)	k_{cat} (s^{-1})	T_{opt} (°C)	$\Delta G^{\ddagger}_{N\text{-}U}$ (kJ mol^{-1})	T_m (°C)
BPSYC	0.18	0.61	6.5	47	94.9	-
BSUB	0.65	8.05	48.7	53	95.9	-
BCVX	1.12	0.81	53.8	69	100.7	61
ANC1	1.32	0.52	114.8	73	100.9	64.7
ANC2	0.96	0.93	41.7	49	91.9	47.6
ANC3	2.65	0.93	102.3	60	95.6	55.5
ANC4	1.69	0.97	362.2	70	110.8	65.3

Figure 1.5: Characterisation of ancestral IPMDH from *Bacillus*. (A) ML chronogram based on IPMDH protein sequences. Ancestral enzymes are indicated at nodes. Bootstrap values for 1024 replicates are also indicated. Contemporary species are coloured based on optimal growth temperature [60-80°C (red), 45-50°C (orange), 37°C (yellow), 25-30°C (green), 20°C (blue)]; (B) physical constants for contemporary and ancestral IPMDH enzymes. BPSYC (*B. psychrosaccharolyticus*), BSUB (*B. subtilis*), BCVX (*B. caldovelox*).

1.3.5 Evolutionary History and Promiscuity

The closest known paralogues to IPMDH are the NAD-dependent and NADP-dependent ICD. These enzymes share the same fold and catalytic mechanism (Imada et al. 1998; Chen & Jeong 2000). Differences in the substrate of these enzymes occur at the 2R position of the malate backbone. This difference is evident in the crystal structures of both enzymes, which share largely similar substrate binding pockets except in the section accommodating the 2R moiety (Imada et al. 1998). It has been proposed that the ancestral enzyme of these paralogues possessed a broad specificity, with activity against a range of structurally similar molecules. Duplication, likely in the progenitor of all extant organisms, and diversification of this gene then lead to the paralogues present today (Dean & Golding 1997; Chen & Jeong 2000). Tartrate dehydrogenase also belongs to this enzyme family (Tipton & Beecher 1994).

Contemporary IPMDHs show promiscuous activity against a range of (2R, 3S)-3-alkylmalates. IPMDH shows activity against structures with a 2R hydrogen (malate), methyl, ethyl, iso-butyl, *tert*-butyl and iso-pentyl moieties on the malate backbone. By comparison, activity is not seen against the alkylmalate isocitrate (Miyazaki et al. 1993; Matsunami et al. 1998; Fujita et al. 2001). Given the substrate binding pocket and the charged 2R moiety of isocitrate, this lack of activity is expected. Tartrate, with a 2R carboxyl group, is also an alkylmalate. Activity against the (2R, 3S)-3-alkylmalates has been shown to be malleable, showing significant changes in rate against the different substrates with single amino acid changes near the active site (Fujita et al. 2001).

1.4 Research Objectives

IPMDH has previously been reconstructed back to the LCA of *Bacillus*, giving insight into the last billion years of the enzymes' evolution (Hobbs et al. 2012). The limitations of ASR techniques in terms of time and sequence divergence for reconstructions of longer time periods than this are unknown with such a large and complex enzyme as IPMDH. This project aimed to take the reconstruction of

IPMDH back further to the LCA of the Firmicutes, and test the capability of ASR techniques over this time scale.

The reconstruction by Hobbs et al. (2012) indicated interesting trends in the evolution of IPMDH with respect to the kinetic barrier to unfolding and catalytic rate. Ancestral enzymes showed faster catalytic rates and more kinetic stability than extant counterparts. This trend raises questions about why ancestral IPMDH enzymes have been superseded by slower and less stable counterparts. A second aim was to assess ancestral enzymes *in vivo* to gauge what impact these enzymes have on growth while functioning within the biosynthetic pathway for leucine in a cellular context, and assess the fitness of each enzyme. Substrate promiscuity of the contemporary and ancestral enzymes was also assessed, and used to test the hypothesis that ancestral enzymes exhibited broader activity which has been narrowed down to more efficient specific activity over evolutionary time.

The structure of ANC4, the LCA of the *Bacillus* has previously been solved, revealing structural homology to contemporary thermophilic IPMDH enzymes. The final aim of this project was to solve the structure of ANC1 to allow structural comparison to ANC4, as well as to contemporary enzymes.

2 Materials and Methods

Additional information on bacterial strains, plasmids, and the compositions of reagents, buffers and growth media are given in the appendices.

2.1 Phylogenetics and Ancestral Inference

2.1.1 Sequences and Alignment

All amino acid and nucleotide sequences were obtained from the GenBank data base. Gene accession numbers and species information are available in Appendix B1.

IPMDH sequences were collected in a BLAST (Altschul et al. 1990) search against the BCVX protein sequence. Sequences were collated in Geneious version 6.1.5 (Drummond et al. 2011) and aligned using the ClustalW2 package (Larkin et al. 2007). The protein alignment was manually checked, and the nucleotide sequence edited to match the protein alignment. Amino acid and nucleotide sequence alignments were exported in PHYLIP format.

2.1.2 Phylogenetic Analysis

2.1.2.1 *Determination of Included Species*

For each genus within the Firmicutes phylum, available sequences were collected from GenBank and aligned using ClustalW2 (Larkin et al. 2007). GARLI version 1.0 (Zwickl 2006) was used to construct a maximum likelihood tree for each genus. A species representing each deep rooted node of these trees was represented in the final tree used for ancestral inference.

2.1.2.2 *Determination of the Models of Evolution*

ProtTest version 2.4 (Abascal et al. 2005) was used to establish the appropriate model of protein evolution for the IPMDH protein alignment. For the nucleotide alignment, jModelTest version 0.1.1 (Posada 2008) was used to determine the appropriate model of nucleotide evolution.

2.1.2.3 *Phylogenetic Tree Construction and Validation*

The final protein sequence alignment of the Firmicutes and two *Mycobacterium* out-group species, along with the appropriate model of protein evolution were implemented in GARLI version 1.0 (Zwickl 2006) to produce eight maximum likelihood trees. These trees were viewed and rooted in Geneious version 6.1.5 (Drummond et al. 2011). The best tree from this set was chosen based on the log likelihood values for each tree and overall consistency of clades between the eight replicates.

Bootstrapping was performed by running GARLI version 1.0 (Zwickl 2006) to construct 1024 pseudoreplicate trees. This set of trees was opened in Geneious version 6.1.5 (Drummond et al. 2011) and a consensus tree generated.

2.1.3 Ancestral Sequence Inference

The best tree as determined in section 2.1.2.3 was exported from Geneious version 6.1.5 (Drummond et al. 2011) in Nexus format, manually edited into Newick format and saved as a .tre file. Amino acid and nucleotide alignments were exported in PHYLIP format and saved as .aa and .nuc files respectively.

Three ancestral inference methods (nucleotide, amino acid and codon) were performed in PAML version 4.3 (Yang 2007). For the nucleotide inference the best tree and nucleotide alignment were used in the program BASEML, along with the best model of nucleotide evolution as determined by jModelTest [(Posada 2008); Section 2.1.2.2]. CODEML was used for the codon and amino acid inferences, with the nucleotide and protein alignments respectively, along

with the tree and best model of protein evolution as determined by ProtTest [(Abascal et al. 2005); section 2.1.2.2].

2.1.3.1 *Consensus Ancestral Sequence Determination*

The tree description contained within the output rst files from the reconstructions were saved as .tre files and visualised in TreeView version 1.6.6 (Page 1996). Node labelling from these trees was used to extract the corresponding node sequences from the three inferences rst output files. The inferred sequences from the three methods were transferred to Geneious version 6.1.5 (Drummond et al. 2011), and the nucleotide inference translated into a protein sequence. The three ancestral amino acid sequences for the node of interest were aligned using ClustalW2 (Larkin et al. 2007). A consensus sequence was compiled based on the following criteria: if all three methods chose the same amino acid, this amino acid was used; if the codon method and one other method chose the same amino acid, this amino acid was used. Any ambiguous sites which did not fit these criteria were manually resolved upon examination of the protein alignment based on the following criteria: if one of the inferred amino acids was spread evenly over contemporary species in the tree, this amino acid was generally chosen; based on the JTT model of amino acid classification (Taylor & Jones 1993), amino acids with conserved physiochemical characteristics over contemporary species were chosen; if the nucleotide and amino acid inferences predicted the same amino acid, this amino acid was generally chosen after considering the first two criteria. At ambiguous sites, sequence bias from contemporary species branching close to the node was also avoided.

2.2 Cloning

2.2.1 Gene Synthesis

The ancestral gene (LCA) was optimised for expression in *E. coli* and synthesised by GENEART (Regensburg, Germany), flanked by the restriction sites *Xho*I and *Pst*I in pMA-T plasmid.

2.2.2 Plasmid Extraction

pPROEX HTb plasmid (Invitrogen, USA) was isolated from DH5α cells grown overnight at 37 °C in 5 ml cultures of LB with 100 µg/ml ampicillin (AMP). Plasmid was isolated from cells using a QIAprep Spin Miniprep Kit (Qiagen, Netherlands) according to the manufacturer's protocols.

2.2.3 Plasmid Preparation

The GENEART plasmid containing the LCA gene was resuspended in 25 µL MQ H_2O. A 15 µl aliquot of suspended plasmid was digested with 1 µl each of *Xho*I and *Pst*I and 2 µl of Buffer K (Invitrogen, USA), made up to 20 µl with MQ H_2O. Digestions were incubated at 37 °C for 3 hours.

Extracted pPROEX (section 2.2.2) was also digested, with adjusted volumes of 25 µl plasmid, 1.5 µl *Xho*I and *Pst*I and 3.3 µl Buffer K (Invitrogen, USA), made up to 33 µl with H_2O.

2.2.3.1 *Agarose Gel Electrophoresis*

Agarose gels (1 % w/v) were made up in TAE buffer (Appendix C1) and contained 2-5 µL SYBR safe DNA gel stain (Invitrogen, USA). DNA samples were mixed with 10 x loading dye (Invitrogen, USA) prior to loading. Gels were run at 100V for 40 minutes to achieve DNA separation. Gels were visualised on a blue light box and DNA size determined against a 1 kb plus DNA ladder (Invitrogen, USA).

2.2.3.2 *Purification of Digested pPROEX and LCA*

The excised LCA gene was separated from the plasmid fragment by agarose gel electrophoresis (section 2.2.3.1). The band corresponding to the gene was carefully excised from the gel with a scalpel and the DNA was extracted using the QIAquick Gel Extraction Kit (Qiagen, Netherlands). Digested pPROEX was

purified from solution using the QIAquick PCR purification kit (Qiagen, Netherlands).

2.2.3.3 *DNA Quantification*

DNA concentrations were measured via absorbance at 260 nm using a Nanodrop ND-2000 spectrophotometer (Nanodrop Technologies, USA).

2.2.3.4 *Ligation*

The ligation of the LCA gene into pPROEX was set up using a 3:1 molar ratio of insert to vector based on the measured concentrations of the respective solutions (section 2.2.3.3). The reaction mixture also consisted of T4 ligase and buffer (Invitrogen, USA), as per the manufacturer's instructions. The reaction was incubated at room temperature for 6 hours.

A negative control ligation was also set up with the same composition, minus the gene insert. This control was treated the same throughout the ligation and transformation steps of the protocol.

2.2.4 Transformation

2.2.4.1 *Electrocompetent Cell Preparation*

Starter cultures were set up in 10 ml of LB in a 50 ml Falcon tube, inoculated with a single colony of cells plated from glycerol stocks and incubated at 37 °C overnight with shaking. The 10 ml overnight starter culture was 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 log phase was reached [an OD_{600} between 0.5 and 0.7 as measured by a Heλios spectrophotometer (Thermo Scientific, USA)]. Once an appropriate optical density (OD) was reached, the flask was chilled on ice for 30 minutes then the culture transferred to sterile centrifuge bottles and cells isolated by centrifugation at 4,600 rpm for 15 minutes at 4 °C. The supernatant was removed, and the cell pellet resuspended in a total of 1 L cold, sterilised 10% glycerol, and

then centrifuged as before. This centrifugation and resuspension process was repeated in sequentially decreasing volumes of ice-cold sterile
10% glycerol (0.5 L, 20 ml, and 2.5 ml). Aliquots of 50 µL from the final 2.5 ml resuspension were transferred to individual tubes and immediately frozen for storage at -80 °C.

2.2.4.2 *Electroporation*

Plasmids were transformed into electrocompetent DH5α cells using the Bio-Rad Gene Pulser (Bio-Rad Laboratories, USA). Chilled electro-cuvettes (Bio-Rad Laboratories, USA) were filled with 10 µl of plasmid, 50 µl of thawed electrocompetent DH5α cells and 30 µL of 10 % glycerol, and electroporated with 2.5 kV at 25 µF capacitance and 200 Ω resistance. Immediately after electroporation, 1ml of LB was added to the cells. Cells were left at 37 °C for 1 hour to recover.

2.2.4.3 *Transformant Selection*

After recovery, 100 µL of transformation cell suspension was plated on LB agar plates containing 100 µg/ml AMP. The remainder of the cells were pelleted by centrifugation (13,000 rpm, 3 minutes) and streaked onto LB agar plates containing 100 µg/ml AMP. Plates were incubated overnight at 37 °C, and the number of colonies compared between the control and ligation plate.

2.2.4.4 *Gene Insert Screening*

Four colonies were selected and grown in 5 ml LB containing 100 µg/ml AMP overnight with shaking at 37 °C. Colonies were checked for the presence of insert by purifying the plasmid from 3 ml of the overnight culture as described in section 2.2.3.2. Plasmids were then digested as described in 2.2.3, and the resultant fragments run on an agarose gel as described in 2.2.3.1.

2.2.4.5 *Glycerol Stocks*

A culture with a positive screen for the LCA insert was then used to make up a glycerol stock. Glycerol stocks were made of 900 µL of overnight culture with 100 µL of sterile 80 % glycerol. Glycerol stocks were frozen and stored at -80 °C.

2.3 *In vitro* Enzyme Characterisation

2.3.1 Protein Expression

Cell starter cultures were grown from glycerol stocks in 10 ml LB cultures with 100 µg/ml AMP overnight at 37 °C. Starter cultures were used to inoculate 1 L of TB with 100 µg/ml AMP in a 2 L baffled flask, and grown at 37 °C while shaking at 200 rpm. OD was measured at various intervals at 600 nm using a ThermoSpectronic Heλios spectrophotometer (Thermo Scientific, USA) till log phase was reached (OD from 0.5 to 0.7). Once in log phase, protein expression was induced with the addition of IPTG to a final concentration of 1 mM, and expressed overnight at either 37 °C or 18 °C. Cells were isolated by centrifugation (4600 rpm, 15 minutes, 4 °C) and stored at -80 °C.

2.3.2 Protein Purification

2.3.2.1 *Cell Lysis*

Cells were resuspended in 30 ml of lysis buffer (50 mM sodium phosphate buffer, pH 8.0, 300 mM NaCl, and 50 mM imidazole) and sonicated on ice for six 15 second intervals separated by 30 second cooling periods. Cell debris was removed by centrifugation at 15,000 rpm at 4 °C for 20 minutes.

2.3.2.2 *Immobilised Metal Affinity Chromatography Purification*

HisTrap (GE Healthcare, UK) nickel columns were prepared before use by passing 2 column volumes of 100 mM EDTA pH 7.5 to remove Ni^{2+} ions, and

recharged with 1 column volume of 100 mM $NiCl_2$. Washing with 2 column volumes of H_2O followed the Ni^{2+} removal and recharging steps. The column was equilibrated with lysis buffer prior to use.

Supernatant from the sonication process was passed in succession through 1.2 μm, 0.45 μm, and 0.2 μm filters before loading onto a prepared 5 ml HisTrap FF nickel affinity column (GE Healthcare, UK) for immobilised metal affinity chromatography (IMAC) purification. Filtered supernatant was loaded onto a prepared nickel column equilibrated with two column volumes of lysis buffer.

Bound protein was eluted from the nickel column on an ÄKTA Basic, Prime, or Purifier system (GE Healthcare, Sweden). Protein bound with low affinity was removed via washing at 1 ml per minute with a solution of 4 % elution buffer (50 mM sodium phosphate buffer, pH 8.0, 300 mM NaCl, and 1 M imidazole) until the 280 nm absorbance dropped to baseline and levelled off. A gradient from 0-100% elution buffer at a rate of 1 ml per minute over 50 ml was used to elute the target protein from the column and collected as 2 ml fractions.

2.3.2.3 *Size Exclusion Purification*

2.3.2.3.1 Standard Protocol

Pooled nickel fractions with desired protein present as identified by SDS-PAGE (section 2.3.2.6) were pooled and concentrated (section 2.3.2.4) down to 20 mg/ml (measured as described in section 2.3.2.5). Concentrated protein was passed through a 0.2 μm filter and loaded onto a Superdex 200 10/300 GL column (GE Healthcare Life Science, UK) equilibrated with size exclusion (SE) buffer (20 mM potassium phosphate buffer, pH 7.6). Protein was eluted by flowing elution buffer at 0.5 ml/minute through the column, with fractionated collection in 0.5 ml aliquots. Protein elution was followed by absorbance measurement at 280 nm, and fractions identified as containing protein from these measures further assessed by SDS-PAGE (Appendix C3). Protein containing fractions were pooled and concentrated to the necessary concentration for further use.

2.3.2.3.2 Reductant Based Protocol

When disulfide bridges were potentially an issue in protein folding, SE buffer was supplemented with 1 mM of the reductant β-mercaptoethanol. β-mercaptoethanol to a final concentration of 1 mM was also added to the protein prior to loading onto the column.

2.3.2.4 *Protein Concentration*

Proteins were concentrated by centrifugation in 0.5 ml, 2 ml, or 20 ml Vivaspin concentrators with a 10 kDa cut off (Sartorius AG, Germany) at 4000 rpm at 4 °C.

2.3.2.5 *Protein Concentration Measurement*

Protein concentration was measured using a Nanodrop 2000 UV-vis spectrophotometer (Thermo scientific, USA) and the extinction coefficient calculated by ProtParam (http://web.expasy.org/protparam/)

2.3.2.6 *SDS-PAGE Gels*

Protein samples were mixed with 4 x loading buffer and incubated at 95 °C for five minutes. Aliquots of 15 µL were loaded into wells of 12 % SDS gels (see Appendix C3), along with 10 µL of Precision Plus Protein Unstained Ladder (Bio-Rad Laboratories, USA). Gels were run at 15 mA through the stacker layer, and at 30 mA through the rest of the gel until the dye front reached the end of the gel.

For staining, gels were transferred to microwavable containers, covered with coomassie stain (Appendix C1) and microwave heated for 30 seconds. Gels were left at room temperature with shaking for 10 to 20 minutes to complete dyeing. Coomassie stain was decanted off, the gel covered in destain solution (10 % v/v acetic acid) and microwave heated for 30 seconds and shaken at room temperature for at least half an hour. Destaining steps were repeated till protein bands were well defined.

2.3.3 Enzyme Assays

Assays were conducted in a ThermoSpectronic Heλios spectrophotometer (Thermo Scientific, USA), with a single cell peltier-effect cuvette holder to control reaction temperatures. Absorbance values at 340nm were taken at 0.25 second intervals over a reaction time of one minute to follow the reaction through the formation of the cofactor reduction product, NADH (Hobbs et al. 2012).

Substrates were made up as stock solutions in assay buffer (20 mM potassium phosphate buffer, pH 7.6, 300 mM KCl and 0.2 mM $MnCl_2$). The cofactor, NAD^+, was made up as an 80 mM stock solution in MQ H_20. The substrate was diluted to the desired concentration in assay buffer, and 350 µL of was heated in the cuvette to the desired temperature before the addition of 50 µL of NAD^+ (10 mM final concentration) and reheated to the reaction temperature. Reactions were started by the addition of 2 µl of enzyme in SE buffer, rapidly mixed, and the course of the reaction followed for the following minute.

2.3.3.1 *Michaelis-Menten Kinetic Analysis*

Michaelis-Menten kinetics were measured at the T_{opt} of each enzyme. The reaction rate was measured at a variety of substrate concentrations, with an excess concentration of the cofactor NAD^+ (10 mM). Details of the reaction protocol are given in section 2.3.3. The initial rate for each reaction was manually assessed from the resultant plots of time versus absorbance at 340 nm.

Graphpad Prism version 5.01 (GraphPad Software, USA) was used to fit Michaelis-Menten plots to the data and determine K_M and V_{max} values. To calculate k_{cat}, abs/s at V_{max} was converted to conc/s using the Beer-Lambert Law and molar absorptivity of NADH (6,220 L M^{-1} cm^{-1}). The increase in molar concentration of NADH was divided by the molar concentration of enzyme added to the reaction to determine the k_{cat} of the reaction.

2.3.3.2 *Specific Activity Analysis*

The specific activity of enzymes was measured as described in section 2.3.3 at the T_{opt} for each enzyme. Substrate concentration was kept constant over replicates and different enzymes. Reaction rates in abs/s were converted to mM product/s/mM protein based on the absorbance change and concentration of enzyme used.

2.4 *In vivo* Enzyme Characterisation

2.4.1 Cell Strains

E. coli K-12 BW25113 single gene knockout strains of the genes *leuB*, *icd*, *mdh* and *yeaU* (IPMDH, ICD, malate dehydrogenase and tartrate dehydrogenase respectively), as well as the parent strain, were obtained from the Keio collection (Baba et al. 2006).

Cells were made electrocompetent and transformed as described in section 2.2.4. Plasmids used in this transformation (pPROEX containing the *leuB* gene) were purified as described in section 2.2.3.2 from DH5α cells previously screened for the correct plasmid insert. The parent strain and each knockout strain were also transformed with an empty pPROEX plasmid to act as controls. A full list of transformants used in this study is available in Appendix A2.

2.4.2 Growth on Solid Media

Cells were grown at 37 °C with shaking in 1 ml of LB media with 100 µg/ml AMP and 30 µg/ml kanamycin (KAN) in 10 ml Falcon tubes. Parent strains were grown under the same conditions, minus the KAN antibiotic. To remove residual LB, cells were pelleted by centrifugation (10 minutes, 4000 rpm, 4 °C) and resuspended in 1 ml M9 minimal media supplemented with 100 µg/ml AMP. Cells were pelleted and resuspended a second time before plating on M9 agar plates containing 100 µg/ml AMP and 1 mM IPTG. Plates were grown at 37 °C for 24 to 36 hours depending on the rate of growth.

2.4.2.1 *Determination of Carbon Source Utilisation*

Cells were washed and resuspended in M9 media, and streaked out on M9 agar plates containing 100 µg/ml AMP, 1 mM IPTG and various carbon sources (glucose 10 g/L; malate 2 g/L and 5 g/L; sodium acetate 5 g/L; succinate 5 g/L).

Media appropriate for testing the enzyme activity *in vivo* was determined by growth of the parent strain with empty pPROEX vector, and no growth of the knockout strain with empty pPROEX vector.

2.4.2.2 *Semi-quantification of Growth Rate*

Cells were prepared as detailed in section 2.4.2. Cells were serially diluted to concentrations of 1/10, 1/100, 1/1,000 and 1/10,000 in M9 media containing 100 µg/ml AMP, 1 mM IPTG and the relevant carbon source for each strain. Each dilution plus the undiluted culture were spotted out in 20 µL drops on 100 µg/ml M9 agar plates (Juhas et al. 2012).

2.4.3 Growth Rate Determination

2.4.3.1 *Growth Measurement*

To determine growth rate, cells were grown in liquid M9 media. Inoculation cultures were made up from glycerol stocks in 1 ml of LB supplemented with 100 µg/ml AMP and 30 µg/ml KAN. Cultures were grown overnight at 37 °C with shaking. The LB culture was used to inoculate (1 in 100) 1 ml of M9 media, supplemented with 100 µg/ml AMP and 30 µg/ml KAN, and grown at 37 °C for 24 hours; this was repeated into a second M9 starter culture and grown for an additional 48 hours.

The final starter culture was used for a 1 in 100 inoculation of 100 µl of M9 media containing 100 µg/ml AMP, 1 mM IPTG and 30 µg/ml KAN. Cells were grown in 96 well plates. Plates were incubated at 37 °C with shaking in a Multiskan GO Microplate Spectrophotometer (Thermo Scientific, USA). OD was measured in

20 minutes intervals, at which times shaking was stopped. Measurements were continued for up to 200 measurements.

2.4.3.2 *Analysis of Growth Rates*

Changes in OD over the course of the experiment were plotted as time versus log OD graphs. The slope of the log increase in OD over the section of maximum growth rate of these plots was taken as the specific growth rate (Breidt et al. 1994) and converted to the doubling time of the culture.

2.5 Protein Crystallography

2.5.1 Protein Preparation

Protein was expressed as in section 2.3.1, and isolated and purified as in section 2.3.2. The buffer used for the SE was changed to 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid. Purified protein was concentrated (section 2.3.2.4) to between 50 and 60 mg/ml.

2.5.2 Initial Crystallisation Condition Determination

General crystallisation conditions were determined by the sitting drop method. From the four 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 (96 per crystallisation screen) were pipetted into the large wells of four 96-2 low profile Intelli-Plate protein crystallisation plates (Hampton Research, USA). A Mosquito crystallisation robot (TTP LabTech Ltd., USA) was used to lay each screen, dispensing 100 nL of protein and 100 nL of mother liquor into the sitting drop wells of the Intelli-Plates. Plates were sealed with ClearSeal film (Hampton Research, USA) and left at 18 °C on shock proof shelving. Conditions were checked regularly for crystal growth.

2.5.2.1 *Additive Crystallisation Condition Screening*

Sitting drop crystallisation screens were also performed as in section 2.5.2 with the addition of cofactors of the reaction to stabilise the protein. NAD^+ to a final concentration in the drop of 1 mM and $MnCl_2$ to a final concentration of 0.2 mM were added to the protein prior to the screens being laid down.

2.5.3 Fine Screening to Optimise Crystallisation Conditions

2.5.3.1 *Standard Fine Screen*

The hanging drop method was used to screen potential crystallisation conditions from the four crystal screens. Variants of the mother liquor pH and component compositions were set up in 24-well VDX plates (Hampton Research, USA). Each well top was lined with grease for air tight sealing, and filled with 500 µL of mother liquor. Aliquots of 1 µL of protein solution were pipetted onto a siliconised glass cover slip, and mixed with1 µL of the mother liquor by pipetting. The cover slip with protein-mother liquor drop was then inverted over the pre-greased well and gently pressed to seal. Plates were stored at 18 °C on shock proof shelving.

2.5.3.2 *Seeding Screens*

2.5.3.2.1 Batch Seeding

Batch seeding was performed when crystals were of low quality. A sacrificial drop from a previous fine screen was diluted 10-fold in mother liquor and vortexed to produce a stock seeder solution. This solution was further diluted to 1 in 100, 1 in 1,000 and 1 in 10,000. The three lower dilutions were used as the mother liquor in the protein droplet over standard mother liquor in hanging drop fine screens, as described in section 2.5.3.1.

2.5.3.2.2 Streak Seeding

Streak seeding was performed when crystals were of low quality. Fine screens were set up as described in section 2.5.3.1 and left at 18 °C for four hours. A cat's whisker was used to streak through a crystal containing droplet from a previous fine screen, and streaked through the seeding droplet to transfer crystals. Plates were stored at 18 °C on shock proof shelving. Whiskers were cleaned with ethanol between use.

2.5.4 Crystal Preparation for Data Collection

Protein crystals were transferred via a cryo-loop (Hampton research, USA) to a cryo- protectant solution of mother liquor containing 20 % v/v glycerol. Crystals were left in this solution for approximately 30 seconds. Alternately, a series of cryo-protectants of increasing glycerol concentration (5%, 10%, 15%, 20%) were used to soak the crystal. Soaked crystals were frozen in liquid or gaseous nitrogen depending on the data collection procedure used.

2.5.5 Crystal Diffraction Testing

Crystals for X-ray diffraction testing were prepared as described in section 2.5.4, mounted in a SuperNova X-ray diffractometer (Agilent, USA) and frozen by a stream of gaseous nitrogen at 100 K. Diffracting crystals were transferred to liquid nitrogen for storage and full data collection.

2.5.6 Data Collection

X-ray diffraction data were collected at the Australian Synchrotron, Melbourne, Australia using the MX1 and MX2 beam lines. To measure reflections, an ADSC Quantum 210r detector (Area Detector Systems Corp., USA) was used. The MOSFLM (Leslie & Powell 2007) strategy function was used in conjunction with the collection process to assist optimal data collection.

2.5.7 Data Processing

2.5.7.1 *Indexing and Integration*

Diffraction images were visualised, scaled and integrated in MOSFLM (Leslie & Powell 2007). Unit cell parameters were determined through the auto indexing function, and integrated in the appropriate space group for the data. When issues were encountered in integration due to low intensity spots at high resolution, a resolution cut off was applied prior to indexing.

2.5.7.2 *Combining of Two Data Sets*

Indexed data sets were combined using Sortmtz (P. J. Daly, Daresbury) within the CCP4 program suite (Winn et al. 2011). One data set was renumbered, and added onto the first image set so that a continuous series of images was produced.

2.5.7.3 *Scaling*

Data was scaled in SCALA (Evans 2006) within the CCP4 program suite (Winn et al. 2011). SCALA outputs were examined, and the data rescaled at decreasing resolution to optimise R_{merge} to below 80 % in the outer shell of the data.

2.5.7.4 *Matthews Coefficient*

The number of monomers in the asymmetric unit was calculated from indexed output files using MATTHEWS_COEF (Matthews 1968) in CCP4 (Winn et al. 2011). The number of monomers in the asymmetric unit was selected based on calculated solvent occupancy of the unit cell.

2.5.7.5 *Molecular Replacement*

Molecular replacement was performed using the structure of the LCA IPMDH of *Bacillus* (PDB code 3U1H). Molecular replacement was performed using PHASER (McCoy et al. 2007) within PHENIX (Adams et al. 2010).

2.5.7.6 *Model Refinement*

Automated refinement of the model was executed in Autobuild (Terwilliger et al. 2008) within PHENIX (Adams et al. 2010). Manual building and refinement of the model was performed in COOT (Emsley & Cowtan 2004) with electron density maps contoured to 1 σ. Model refinement was performed using phenix.refine (Afonine et al. 2012).

2.5.8 Structural Analysis

All structures image in this work were generated in PyMOL version 2.5.4 (Delano 2002). Ramachandran analysis was performed in Procheck in the CCP4 program suite (Winn et al. 2011). Average B factors were calculated in Baverage in the CCP4 program suite. Structural homologues were determined in PDBeFold (Krissinel & Henrick 2004). PDBePISA (Krissinel & Henrick 2007) was used to calculate the buried surface area of the dimer interface.

3 Phylogenetics, Ancestral Inference and Activity Assessment

3.1 Introduction

IPMDH has been the focus of previous ancestral reconstruction studies (Miyazaki et al. 2001; Hobbs et al. 2012). IPMDH is an ideal candidate for ASR as it is easily over expressed and biochemical assay protocols are well defined, aiding characterisation of the enzyme (Wallon et al. 1997a). Biochemical characterisation of the reconstructed enzymes also acts as a good indicator of a successful inference, as inference errors in ancestral sequences are likely to result in inactive or biologically unrealistic enzymes.

Previous ASR studies of IPMDH have successfully reconstructed the enzyme back 950 million years (myr) to the LCA of *Bacillus* (Hobbs et al. 2012). Taking ancestral reconstructions back further than this is questionable due to the levels of sequence divergence. The only enzyme that has been reconstructed back past this time period is β-lactamase (Risso et al. 2013). In this case, enzymes from 2–3 bya showed biologically realistic thermal denaturation and catalytic rates. However the limitations of ASR in terms of time and sequence divergence are not known for such a large and complex enzyme as IPMDH. Here, the ability for ASR techniques to reconstruct IPMDH back to the LCA of the Firmicutes, about 2.7 bya (Battistuzzi et al. 2004), was assessed.

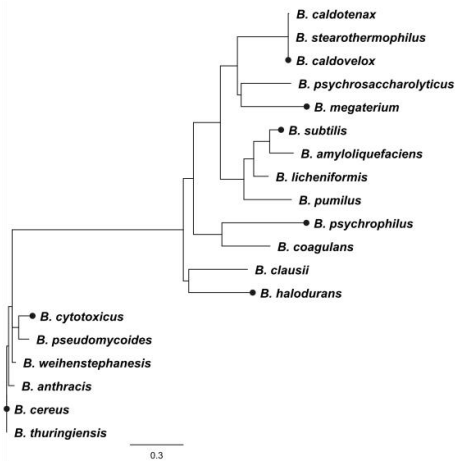
3.2 Results and Discussion

3.2.1 Selection of Representative Species

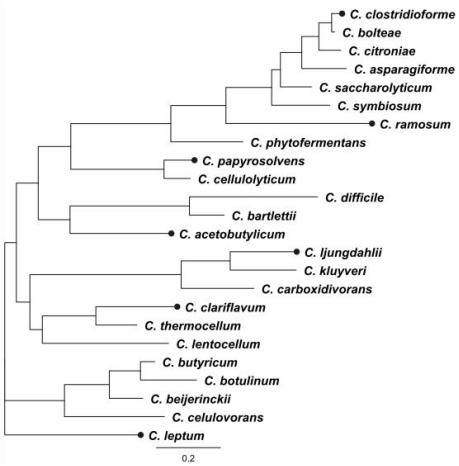
In order to perform an ancestral inference, alignments of extant nucleotide and amino acid sequences are necessary. IPMDH sequences for use in these alignments were obtained from the GenBank database following a BLAST (Altschul et al. 1990) search against the BCVX IPMDH protein sequence.

Accession numbers for all the sequences used in this study are given in Appendix B1.

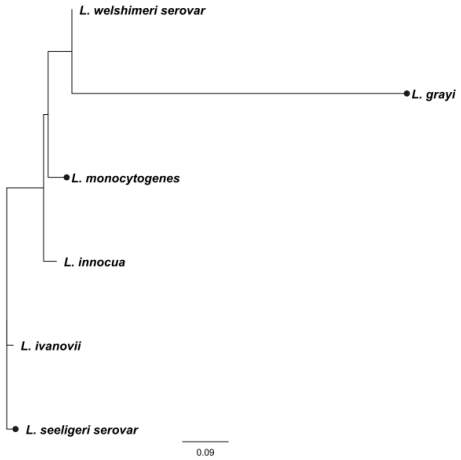
A common issue with ASR is determining how many species to include in the analysis, and which species should be incorporated to give a good representation of overall diversity. It has been shown that the inclusion of more taxa in a tree is not necessarily better for ancestral inference, assuming the tree topology is correct (Li et al. 2008). However, species which accurately represent the range of extant diversity must be selected. To choose sequences which represented the size and diversity of each major genus within the Firmicutes as included in the phylogeny (Battistuzzi et al. 2004), phylogenetic trees were constructed of each genus (*Bacillus, Clostridium, Lactococcus, Listeria, Staphylococcus, Streptococcus and Thermoanaerobacter*). *Mycoplasma* and *Ureaplasma* were unable to be included due to the reduction of the genome size which has removed all the amino acid biosynthesis genes in these species (Battistuzzi et al. 2004). Trees were constructed in GARLI (Zwickl 2006) from ClustalW (Larkin et al. 2007) protein alignments of all species within each genus available from the GenBank database. Sequences representing main branches of each tree were selected for inclusion in the full Firmicutes tree as described previously (Iwabata et al. 2005). Trees can be found in Figure 3.1. For the lactococci, all sequences available were identical over 99 % of residues. No tree was constructed in this case. The *Bacillus* tree illustrated in Figure 3.1 shows similar species groupings to the previously described tree for *Bacillus* IPMDHs (Hobbs et al. 2012).



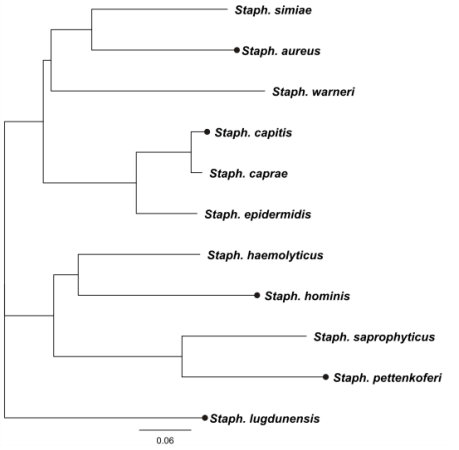
Bacillus tree



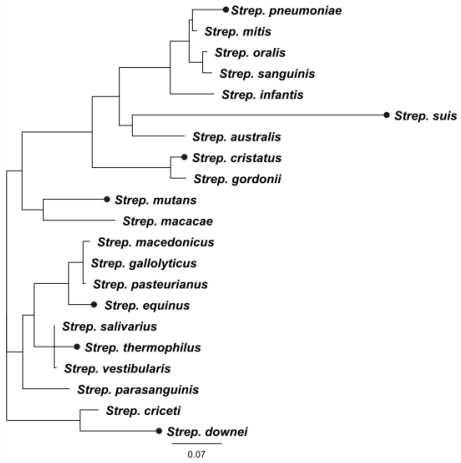
Clostridium tree



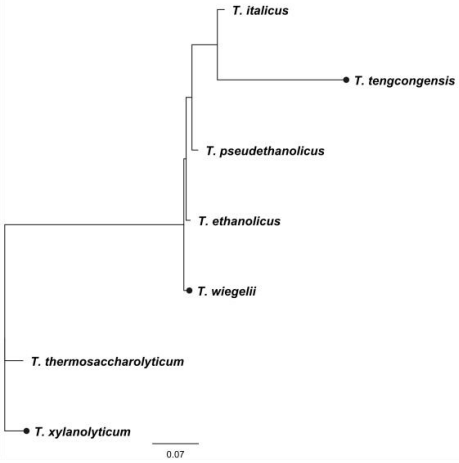
Listeria tree



Staphylococcus tree



Streptococcus tree



Thermoanaerobacter tree

Figure 3.1: Phylogenetic trees used for the selection of representative species from each major genus of the Firmicutes.

The scale for substitution rate per site per unit time is indicated at the bottom of each tree. Species selected for inclusion in the final phylogenetic tree are indicated with circles.

Mycobacterium was selected as the outgroup based on the known phylogenetic relationship of this group to the Firmicutes (Battistuzzi et al. 2004).

Representative species of each genus selected for inclusion in the full Firmicutes tree and as the outgroup are listed in Table 3.1.

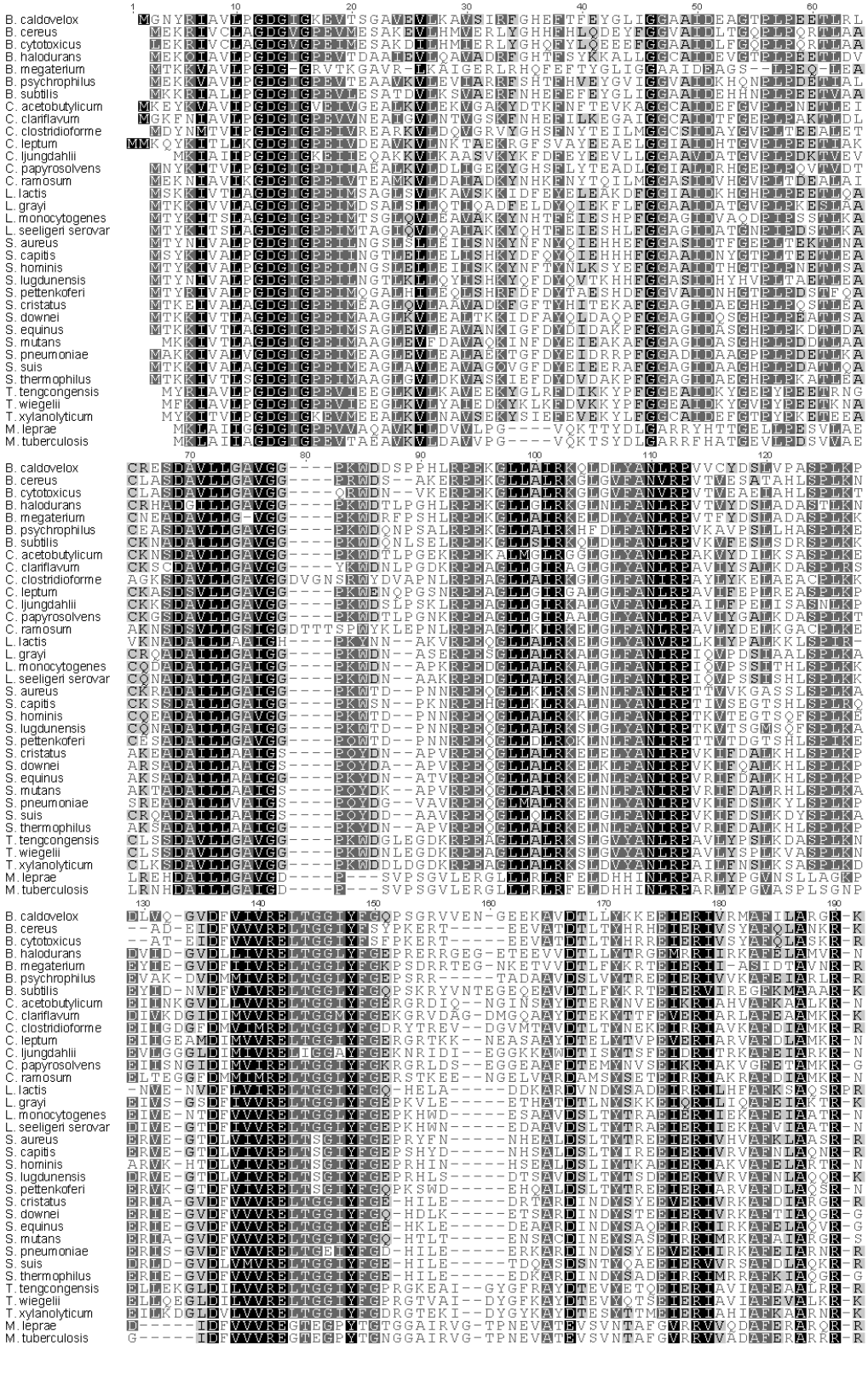
Table 3.1: List of species selected for inclusion into the full Firmicutes tree.

Genus	**Species**
Bacillus	*B. caldovelox*
	B. cereus
	B. cytotoxicus
	B. halodurans
	B. megaterium
	B. psychrophilus
	B. subtilis
Clostridia	*C. acetobutylicum*
	C. clariflavum
	C. clostridioforme
	C. leptum
	C. ljungdahlii
	C. papyrosolvens
	C. ramosum
Lactococcus	*L. lactis*
Listeria	*L. grayi*
	L. monocytogenes
	L. seeligeri serovar
Staphylococcus	*Staph. aureus*
	Staph. capitis
	Staph. hominis
	Staph. lugdunensis
	Staph. pettenkoferi
Streptococcus	*Strep. cristatus*
	Strep. downei
	Strep. equinus
	Strep. mutans
	Strep. pneumoniae
	Strep. suis
	Strep. thermophilus
Thermoanaerobacter	*T. tengcongensis*
	T. wiegelii
	T. xylanolyticum
Mycobacteria (outgroup)	*M. tuberculosis*
	M. leprae

3.2.2 Sequence Alignment

In order to construct a phylogenetic tree of the representative Firmicutes sequences, alignments of the nucleotide and amino acid sequences must be made

to match equivalent codons/amino acids from each sequence. Alignment was performed in ClustalW (Larkin et al. 2007) and manually checked and adjusted. Overall alignment was aided by highly conserved regions dispersed throughout the gene (see Figure 3.2). A number of these conserved regions correspond to the substrate binding domain in the crystal structure (Imada et al. 1998). For example, Arg102, Arg112, Arg141, Tyr148, Lys200, Asp233, Asp257 and Asp261 (numbering corresponding to the sequence alignment in Figure 3.2) bind to the malate backbone of IPM, and are universally conserved over all known IPMDHs. Universally conserved residues Glu95, Leu98, Leu99 and Val203 make up the binding pocket for the isopropyl group of the substrate. Residues in the NAD^+ binding pocket are also conserved over most species in the alignment, with the more variable sites showing conservation of hydrophobic amino acids. These residues include the universally conserved Asp346, and highly conserved Ile15, Val19, Leu270, Gly271 and Ile298 (Hurley & Dean 1994). Gaps between these conserved regions were edited manually to resolve any obvious errors in the alignment. Overall the protein alignment seen in Figure 3.2 has 52.6 % pairwise identity and 13.6 % identical sites. The percentage of identical sites in the alignment is low, likely due to the range of species included in the alignment. However, the pairwise identity of the alignment is considerably higher than the percentage of identical sites. Areas with lower conservation over the alignment fulfil structural functions rather than having direct involvement in the chemistry of catalysis.


B. caldovelox
B. cereus
B. cytotoxicus
B. halodurans
B. megaterium
B. psychrophilus
B. subtilis
C. acetobutylicum
C. clariflavum
C. clostridioforme
C. leptum
C. ljungdahlii
C. papyrosolvens
C. ramosum
L. lactis
L. grayi
L. monocytogenes
L. seeligeri serovar
S. aureus
S. capitis
S. hominis
S. lugdunensis
S. pettenkoferi
S. cristatus
S. downei
S. equinus
S. mutans
S. pneumoniae
S. suis
S. thermophilus
T. tengcongensis
T. wiegelii
T. xylanolyticum
M. leprae
M. tuberculosis

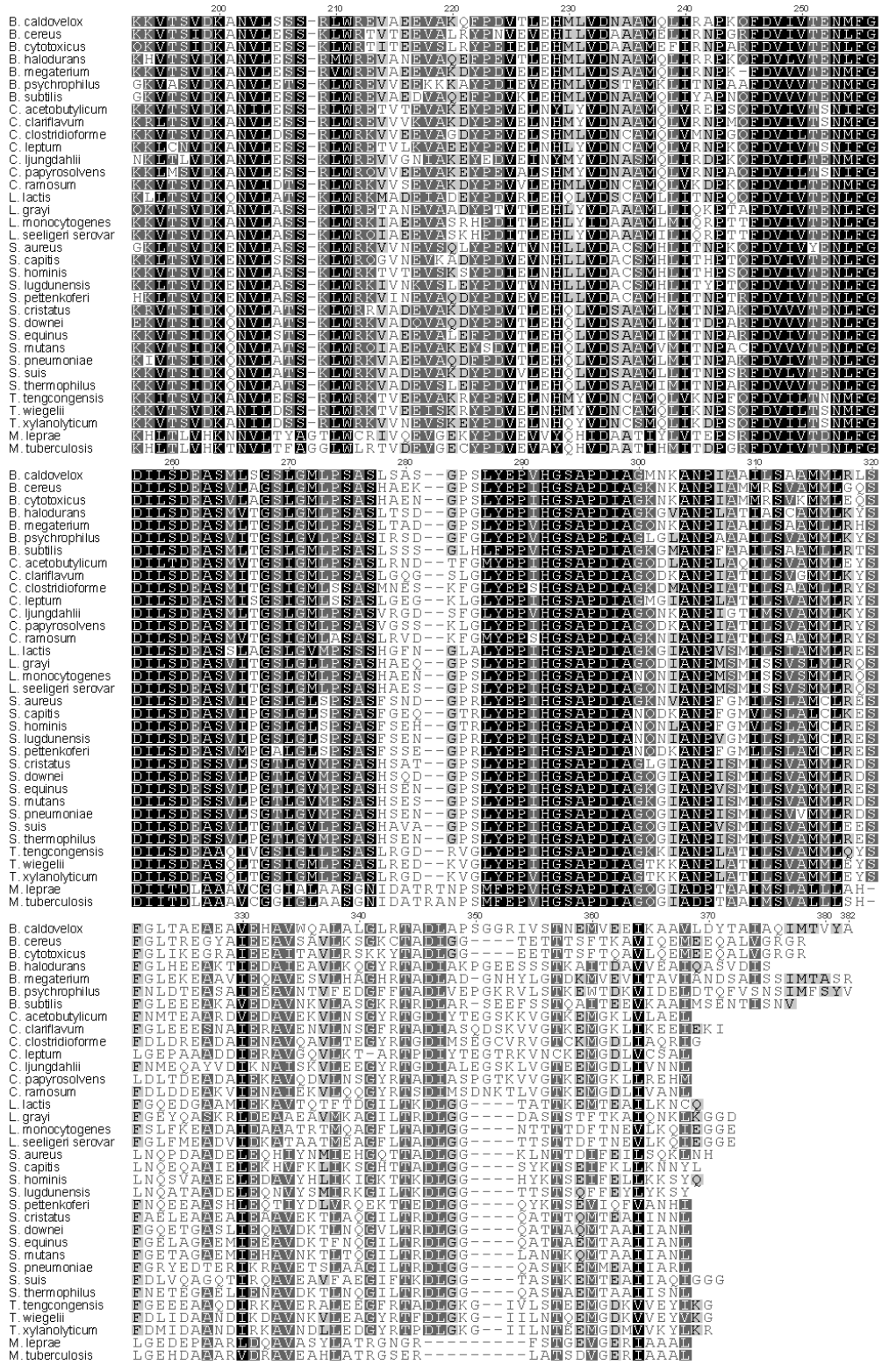


Figure 3.2: Firmicutes IMPDH protein sequence alignment.

Alignment was generated with ClustalW (Larkin et al. 2007) in Geneious (Drummond et al. 2011). Shading at each position indicates the percentage similarity based on residue chemistry of the position over the sequence alignment. Black = 100 %; dark grey = 80-100 %; light gray = 60-80 %; white = < 60 % similarity.

3.2.3 Phylogenetic Analysis

For the IPMDH amino acid alignment, the most appropriate model of evolution to describe the rate at which one amino acid replaces another was determined in ProtTest (Abascal et al. 2005). The LG model of evolution (Le & Gascuel 2008) was found to model the data the most accurately under both Akaike information criterion (AIC) and Bayesian information criterion. The model with the lowest AIC score represents the best model out of those available based on the relative likelihood. The LG model incorporates varying rates of evolution across sites in the matrix, and is considered to be the most sophisticated without being overly computationally demanding (Le & Gascuel 2008).

The LG model and alignment shown in Figure 3.2 were used in GARLI (Zwickl 2006) as described in section 2.1.2.3 to produce eight ML trees. The phylogenetic trees were rooted manually by selecting the two *Mycobacterium* species as the outgroup. Log likelihood scores (-ln) and a manual assessment of the correct association of species into genus groups was used to assess the best tree. A tree with a -ln of -13565 was selected, as shown in Figure 3.3. Despite other trees having slightly better -ln scores, this tree was selected due to the consistent association of species into the correct genera. Two inconsistencies in genera groupings are evident: *L. lactis* is grouping with the streptococci, and *C. ljungdahlii* is separate from the rest of the clostridia, separated by the *Thermoanaerobacter*. These discrepancies occurred consistently in all the tree replicates performed. *L. lactis* is closely related to the streptococci, and shares many similar features (Schleifer et al. 1985). Clostridia and *Thermoanaerobacter* are also closely related (Collins et al. 1994a; Battistuzzi et al. 2004). Thus, the grouping of these sequences together is not surprising.

Confidence in the phylogeny was assessed by bootstrapping (Felsenstein 1985; Efron et al. 1996). 1024 pseudoreplicate trees were generated in GARLI (Zwickl 2006) and combined into a consensus tree where bootstrap values are assigned to branches based on the proportion of pseudoreplicate trees including the node. This number allows the reliability of each node within the tree to be estimated. Bootstrap values are included on branches in the phylogenetic tree in Figure 3.3.

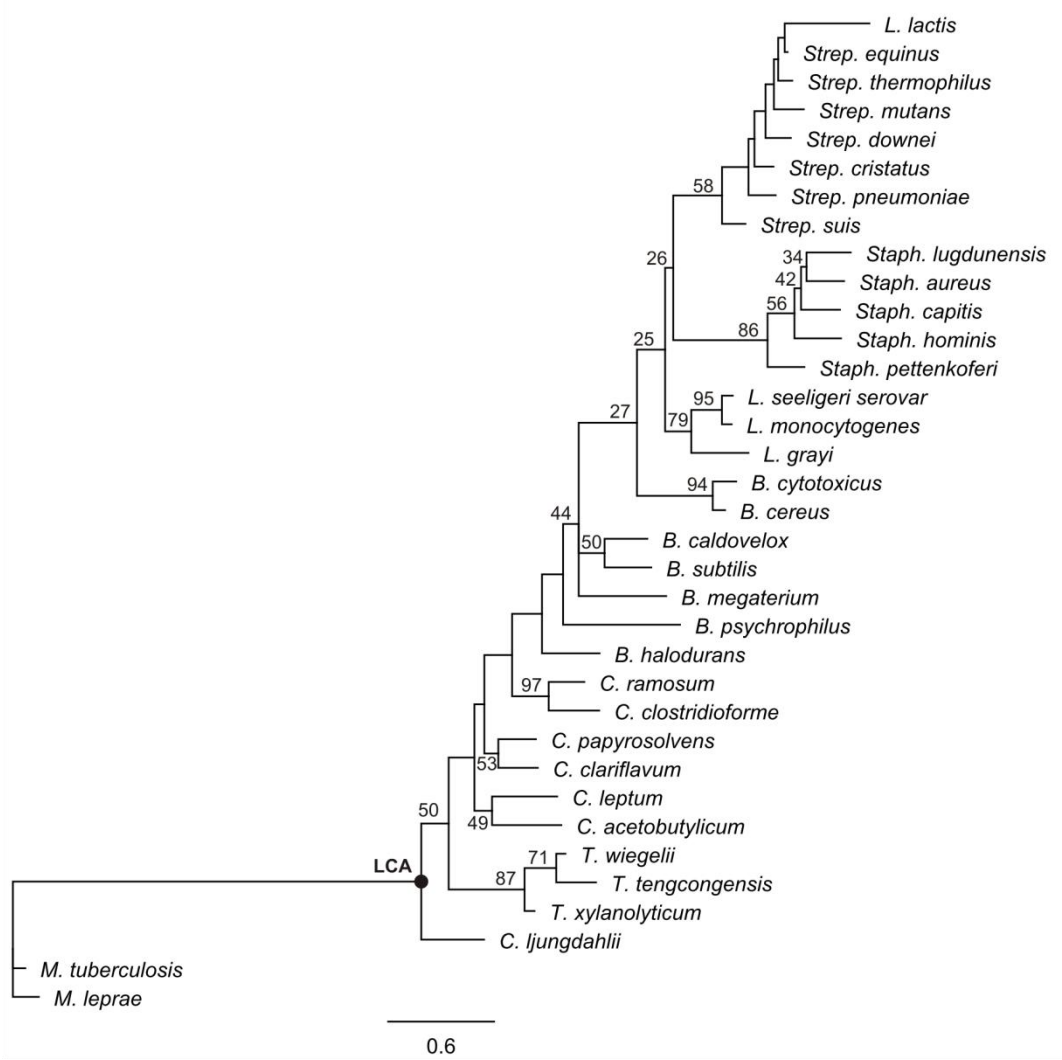


Figure 3.3: ML phylogeny of the Firmicutes based on IPMDH amino acid sequences.

The position of the reconstructed LCA of the Firmicutes is indicated as a black circle. LCA is about 2.7 billion years old (Battistuzzi et al. 2004). Numbers on branches represent bootstrap percentages. The scale for substitution rate per site per unit time is indicated at the bottom of the tree.

3.2.4 Ancestral Inference

ML was chosen to perform this analysis as previous inference of ancestral IPMDH enzymes by BI methods produced biochemically unrealistic enzymes. By comparison, ML methods produced ancestral IPMDHs with realistic functional parameters (Hobbs et al. 2012).

Inference on the basis of nucleotides and codons requires a nucleotide alignment that matches the amino acid alignment and an appropriate model of nucleotide evolution. The most appropriate model of nucleotide evolution for the alignment was determined by jModelTest version 0.1.1 (Posada 2008) to be the generalised time reversible model (Tavaré 1986). This model assumes a symmetric substitution matrix for complementary base pairs (bp), different rates of substitution for each nucleotide pair, and allows nucleotides to occur at different frequencies.

Nucleotide, amino acid and codon ancestral inferences were performed in PAML version 4.3 (Yang 2007). Average posterior probabilities for each of these inferences for the LCA of the Firmicutes were 0.842 (amino acid), 0.815 (nucleotide) and 0.693 (codon). Probabilities are calculated from the likelihood of each amino acid or nucleotide occupying a particular site given the data and evolutionary model. These probabilities are slightly lower than ideal for an ancestral reconstruction (greater than 0.9), but not unexpected given the timescale of the inference and were deemed sufficient to continue with the process. Inferred nucleotide and codon sequences for the LCA of the Firmicutes were translated to amino acid sequences, and all three inferences were aligned using ClustalW (Larkin et al. 2007). A consensus sequence was complied and any ambiguous sites resolved under the criteria listed in section 2.1.3.1. There were few ambiguous sites in the ancestral sequence, with only 22 ambiguous residues out of the 382 amino acid alignment not agreeing under all three inference methods, or the codon plus one other inference method. Of these, 13 were resolved by selecting the amino acid for which the nucleotide and amino acid reconstructions both agreed. Ambiguous sites were generally in regions of the alignment that showed high levels of sequence divergence, and often potential residues at an uncertain site were chemically similar. The C terminal end of the sequence was also truncated by 10 residues compared to the total alignment as this end region is only present in the *Bacillus* species of the alignment, and likely represents an insertion present only in this genus. In the *B. coagulans* structure (PDB code 1V53), these 10 C terminal residues are not resolved, and likely represent a

flexible region of the enzyme not essential of catalysis. The final sequence for the LCA of the Firmicutes is given below.

IPMDH sequence of the LCA of the Firmicutes

MKMKIAVIPGDGIGPEIIEEAIKVLNAVAEKYGLKFEYKEVLLGGCAIDETGVPL
PEETVEVCKKSDAVLLGAVGGPKWDNLPSNKRPEAGLLGIRKGLGVYANLRPAIL
YPALKSASPLKPEILEGIDIMVVRELTGGIYFGERGRIDIGGKKAYDTEIYTTFE
IERIARKAFEAARKRNKKLTSVDKANVLESSRLWREVVEEVAKEYPDVELNYMYV
DNASMQLIRDPKQFDVIVTSNMFGDILTDEASMLTGSIGMLPSASLRGDGPSLYE
PVHGSAPDIAGQNKANPIATIMSVAMMLKYSFDMEEAADDIKNAVEKVLEEGYRT
GDIAIEGTKIVGTEEMGDLIVEDLEKI

Comparison of the inferred ancestral sequence (see Table 3.2) to contemporary proteins gives pairwise identities between 47% and 80 % (identity to *S. pettenkoferi* and *C. ljungdahlii* respectively). These numbers are not so closely related to any one species to suggest that an extant sequence may be overly biasing the ancestral node: this is particularly significant for deep branching species such as *C. ljungdahlii*. The extant and ancestral sequences are also all reasonably related, indicating a degree of structural and functional conservation, thus increasing the likelihood of a successful reconstruction. Pairwise identities to *M. leprae* and *M. tuberculosis* are low (40 % and 43 % respectively), consistent with these groups relatedness to the rest of the tree. Sequences from the clostridia and *Thermoanaerobacter* show the highest sequence similarity to the LCA, consistent with the deep branching position of these species in the tree.

Following inference, the gene encoding the LCA sequence was synthesised by GENEART (Regensburg, Germany). The gene sequence is provided in Appendix B2.

Table 3.2: Comparison of the pairwise sequence identities between extant and ancestral protein sequences.

	ANC 4	ANC 1	BCVX	BCER	BCYT	BHAL	BMEG	BPSY	BSUB	CACE	CCLA	CCLO	CLEP	CLJU	CPAP	CRAM	LGRA	LMON	LSEE	LLAC	MLEP	MTUB	SAUR	SCAP	SHOM	SLUG	SPET	SCRI	SDOW	SEQU	SMUT	SPNE	SSUI	STHE	TTEN	TWIE	TXYL
LCA	68	65	61	54	51	61	58	56	56	71	71	62	66	80	70	62	51	53	54	49	40	43	49	50	50	49	47	53	50	52	49	51	52	51	69	70	71
ANC 4		82	75	59	55	75	70	67	75	63	65	61	60	62	63	60	56	58	59	53	35	37	52	53	53	51	52	56	54	59	56	56	57	55	62	60	61
ANC 1			83	58	56	69	79	63	72	59	61	59	55	59	60	57	55	57	57	51	37	39	51	52	52	50	52	56	54	57	55	55	55	54	58	58	60
BCVX				57	54	65	69	59	68	56	59	56	55	56	57	55	53	54	55	51	36	37	51	52	52	51	52	55	53	55	55	54	53	53	57	57	59
BCER					87	56	53	55	54	49	53	52	49	49	53	47	59	59	60	53	35	37	51	52	52	51	54	57	56	58	56	54	57	56	52	53	50
BCYT						53	52	53	52	48	50	50	48	48	51	46	58	60	59	51	33	34	52	51	52	50	53	55	54	56	54	53	55	54	50	52	49
BHAL							63	58	65	58	60	55	56	56	58	56	55	53	55	49	35	37	50	48	50	48	50	52	51	55	51	52	54	52	57	55	54
BMEG								58	65	54	55	56	48	53	55	54	52	52	54	47	36	37	47	48	47	47	48	53	51	51	50	51	52	51	53	53	53
BPSY									60	53	54	53	52	55	54	53	51	53	55	49	35	37	50	50	52	48	51	53	51	52	51	50	54	51	55	55	55
BSUB										53	55	53	51	53	54	52	52	53	54	51	33	35	50	48	50	48	51	52	52	53	54	52	52	53	53	52	53
CACE											66	59	64	61	67	60	50	48	49	47	36	38	47	49	51	48	46	48	47	47	46	46	49	49	62	64	65
CCLA												59	64	61	72	57	50	50	51	47	36	37	46	48	47	47	46	49	47	48	46	47	48	48	64	63	61
CCLO													58	57	64	72	48	49	50	44	33	35	48	46	47	48	46	47	46	47	47	45	47	47	57	57	57
CLEP														58	67	56	49	48	49	47	35	37	48	48	49	48	46	48	46	49	48	47	47	47	59	59	59
CLJU															61	59	49	48	50	45	35	38	47	48	47	47	46	47	46	47	47	46	49	48	58	61	63
CPAP																60	49	51	51	49	34	36	47	49	48	47	46	49	48	50	49	47	47	49	63	63	61
CRAM																	48	48	49	46	32	35	47	45	46	46	45	49	49	49	49	46	48	49	56	59	56
LGRA																		69	71	56	35	38	56	58	57	59	57	60	59	61	60	61	62	61	50	51	50
LMON																			89	56	32	34	57	56	60	60	57	61	60	63	59	59	60	60	52	52	50
LSEE																				56	33	36	58	57	59	61	57	60	59	62	60	59	62	60	52	52	52
LLAC																					34	36	53	54	54	55	54	67	69	72	68	64	66	70	51	50	51
MLEP																						82	34	33	34	35	34	35	35	35	34	34	34	35	36	33	34
MTUB																							34	35	35	35	39	37	38	38	37	39	37	39	36	37
SAUR																								75	75	77	71	55	56	56	54	52	56	55	50	48	50
SCAP																									75	75	72	54	54	55	52	52	54	55	49	49	50
SHOM																										77	71	54	54	55	53	53	56	55	50	51	52
SLUG																											68	55	56	56	54	53	58	58	49	50	50
SPET																												56	58	57	54	54	58	56	46	46	48
SCRI																													81	81	81	81	76	82	50	49	50
SDOW																														82	81	76	72	84	49	49	48
SEQU																															84	78	73	90	51	50	49
SMUT																																75	72	83	48	48	48
SPNE																																	74	77	50	49	49
SSUI																																		73	51	51	51
STHE																																			49	50	50
TTEN																																				83	75
TWIE																																					81

Abbreviations: Ancestor (ANC), *B. caldovelox* (BCVX), *B. cereus* (BCER), *B. cytotoxicus* (BCYT), *B. halodurans* (BHAL), *B. megaterium* (BMEG), *B. psychrophilus* (BPSY) *B. subtilis* (BSUB), *C. acetobutylicum* (CACE), *C. clariflavum* (CCLA), *C. clostridioforme* (CCLO), *C. leptum* (CLEP), *C. ljungdahlii* (CLJU), *C. papyrosolvens* (CPAP), *C. ramosum* (CRAM), *L. grayi* (LGRA), *L. monocytogenes* (LMON), *L. Seeligeri serovar* (LSEE), *L. lactis* (LLAC), *M. leprae* (MLEP), *M. tuberculosis* (MTUB), *Staph. aureus* (SAUR), *Staph. capitis* (SCAP), *Staph. hominis* (SHOM), *Staph. lugdunensis* (SLUG), *Staph. pettenkoferi* (SPET), *Strep. cristatus* (SCRI), *Strep. downei* (SDOW), *Strep. equinus* (SEQU), *Strep. mutans* (SMUT), *Strep. pneumoniae* (SPNE), *Strep. suis* (SSUI), *Strep. thermophilus* (STHE), *T. tengcongensis* (TTEN), *T. wiegelii* (TWIE), *T. xylanolyticum* (TXYL).

3.2.5 Cloning, Expression and Purification of LCA

3.2.5.1 *Cloning of IPMDH LCA*

The consensus sequence of LCA was synthesised by GENEART (Regensburg, Germany) with codons optimised for expression in *E. coli*. The gene was excised from the pMA-T vector, cleaned up and ligated into pPROEX as described in section 2.2.3. Plasmids were transformed into electrocompetent *E. coli* DH5α cells as described in section 2.2.4.2. The presence of the LCA gene in transformant colonies was checked from overnight cultures through plasmid extraction and restriction digest (section 2.2.4.4). An agarose gel of the plasmid restriction digest products is shown in Figure 3.4. For the colonies corresponding to lanes 3 and 4, a band for the excised LCA gene is present at the expected size (1150 bp) indicating successful ligation of the LCA gene. A glycerol stock (section 2.2.4.5) was prepared from the overnight of a successful transformant.

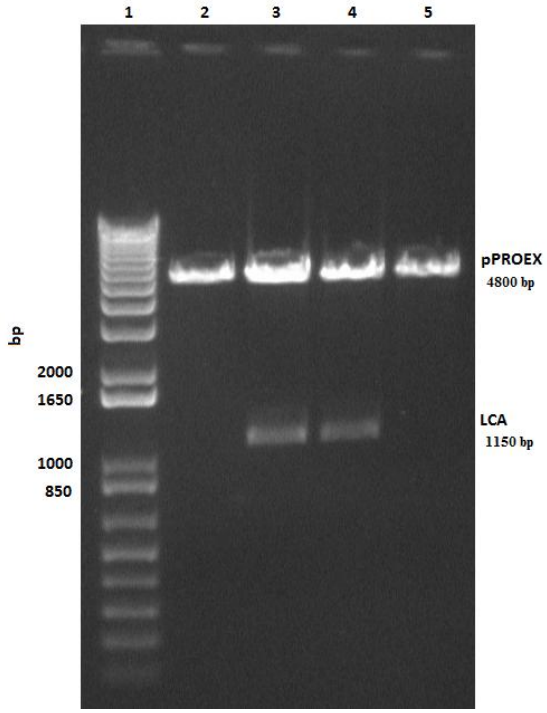


Figure 3.4: Agarose gel of the plasmid restriction digest screen for correct ligation of the LCA gene into pPROEX.

Lane 1, 1 kb plus DNA ladder; lanes 2 – 5, pPROEX plasmids purified from transformant colonies, digested with the restriction enzymes *XhoI* and *PstI*. Plasmid from colonies in lanes 3 and 4 show an excised fragment at the correct size (1150 bp) of the LCA gene.

3.2.5.2 *Expression and Purification of LCA*

Following successful cloning, the LCA enzyme was expressed and purified, as per the protocol in Hobbs et al. (2012) (see sections 2.3.1 and 2.3.2). Expression at 37 °C resulted in low protein yields following nickel affinity chromatography. Therefore, protein expression was trialled at 18 °C to attempt to increase the protein yield. Results from this expression are shown in Figure 3.5.

A SDS-PAGE gel of the insoluble fraction (Figure 3.5) showed that a predominant portion of the protein was insoluble. However, significant amounts of soluble LCA were also detected in the supernatant.

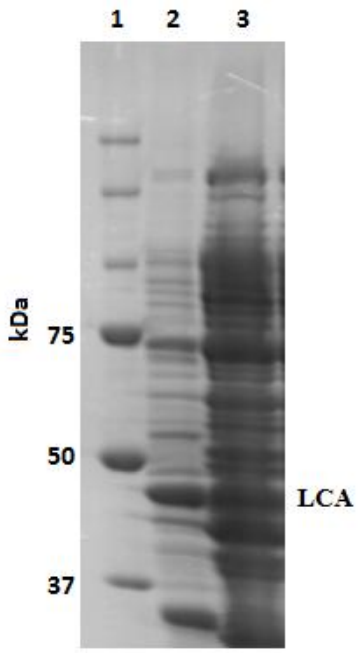
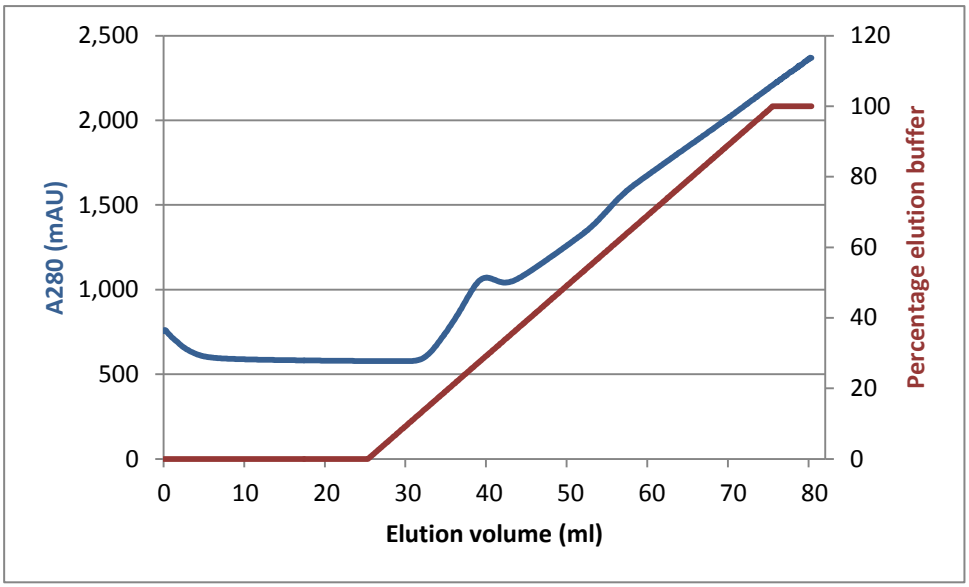


Figure 3.5: Expression trial of LCA at 18 °C.
Lane 1, protein ladder; lane 2, insoluble pellet; lane 3, supernatant

Soluble LCA protein from the supernatant was purified by nickel affinity chromatography. The trace from this purification, given in Figure 3.6, shows low protein yield compared to similar purifications of ANC1, ANC4 and BCVX (see Figure 4.1 for representative ANC4 nickel affinity chromatography results). Protein elution occurred at around 30 % elution buffer. A SDS-PAGE gel of fractions corresponding to the absorption peak confirmed that LCA protein was present in the fractions, and was reasonably pure, with only faint bands of contaminating proteins (Figure 3.6).

B



A

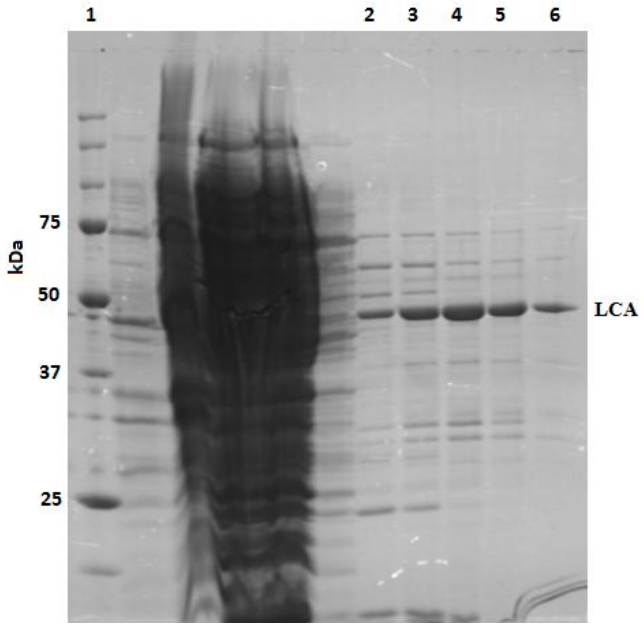
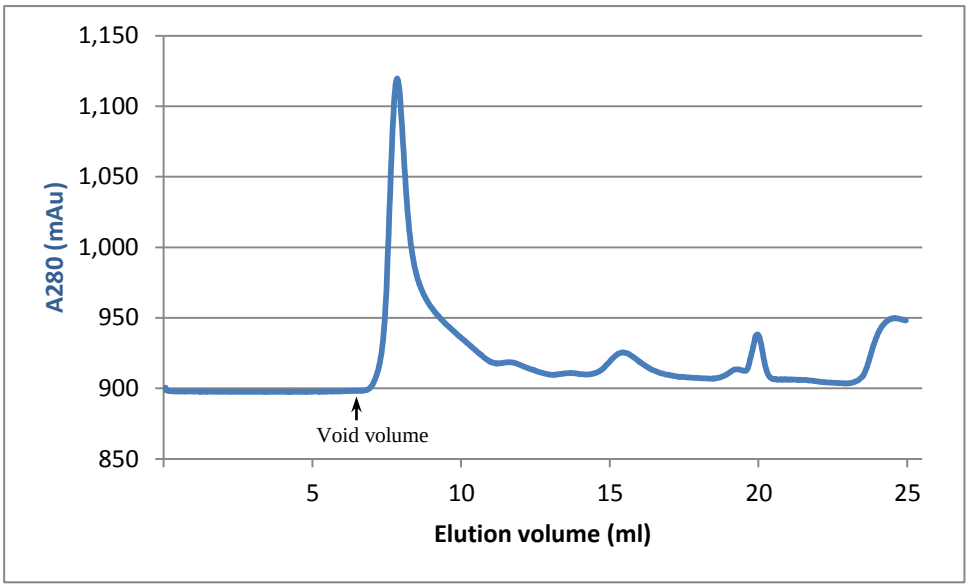


Figure 3.6: Nickel affinity chromatography of LCA expressed at 18 °C. (A) Trace from nickel affinity purification, measured at 280 nm; (B) SDS-PAGE gel of nickel trace fractions. Lane 1, ladder; lanes 2 – 6, fractions from 33 to 40 ml of elution volume.

Size exclusion chromatography of pooled nickel affinity chromatography fractions gave results inconsistent with the expected dimeric size (84 kDa) of the protein. As shown in the trace in Figure 3.7 A, protein is eluted from the column just after the void volume, characteristic of a large molecular weight. This was considerably different to the elution of other IPMDH enzymes (see Figure 4.2)

where protein is eluting at around 14 ml of elution volume. LCA was expected to form a dimer of similar molecular weight to other IPMDH enzymes, and thus have a similar SE elution volume. The small peak eluting from the column at 16 ml, consistent with the dimeric size of the protein, was shown not to be LCA based on the protein size when further analysed by SDS-PAGE (Figure 3.7 B, lanes 6 - 10). The elution characteristics, coupled with the SDS-PAGE gel (Figure 3.7 B, lanes 2 - 5) showing the correct size for the monomeric protein unit, and the clear appearance of the protein containing fractions indicated that the protein was forming a soluble aggregate. LCA proteins also showed no activity in the standard IPM oxidation assay (section 2.3.3).

A



B

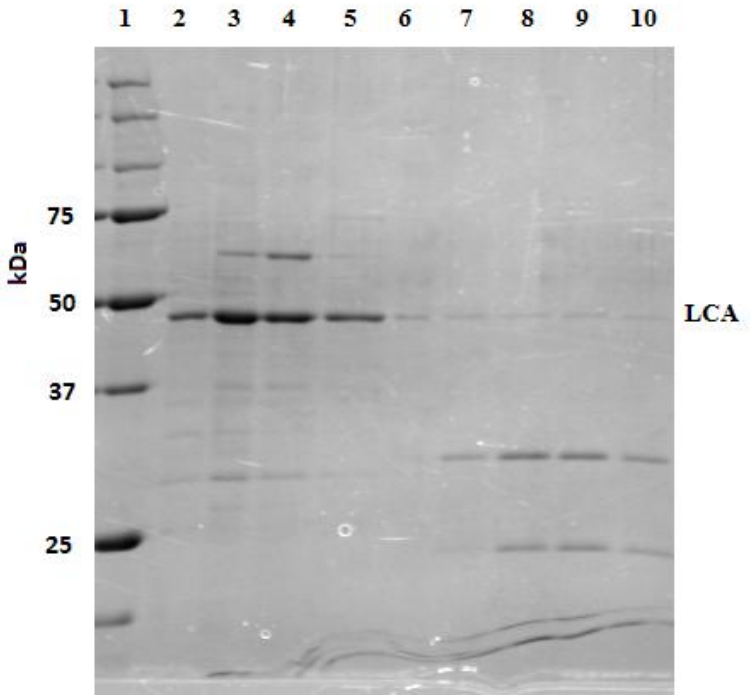


Figure 3.7: SE purification of LCA protein.

(A) SE column trace of LCA purified from culture grown at 18 °C, measured at 280 nm; (B) SDS-PAGE gel of SE fractions. Lane 1, ladder; lanes 2 – 5, SE fractions from 7 to 10 ml of elution volume; lanes 6 – 10, fractions from 15 ml to 16.5 ml of elution buffer.

Soluble aggregates can be caused by a number of factors, and are generally difficult to rectify (Cromwell et al. 2006). One cause of soluble aggregates that

was identified as a possible cause of LCA aggregation was disulfide bridges between molecules. Analysis of the sequence of LCA compared to known active dimeric enzymes (ANC1, ANC4 and BCVX) showed that an additional cysteine is present at position 46 of the LCA sequence. A cysteine is also present at position 63 in the LCA sequence. The possibility this was causing the soluble aggregate was tested by the addition of 1 µM of the reducing agent β-mercaptoethanol to the standard SE buffer and protein during purification. This yielded no change to the point the protein eluted from the column. Therefore it was concluded that disulfide bonds were not the cause of the soluble aggregate. The formation of a soluble aggregate for LCA was concluded to be most likely due to a misfolded protein due to errors in the ancestral inference.

From this result it was concluded that the inference methods used were not sufficient given the time scale and sequence divergence of the IPMDH protein over the Firmicutes. This has been further substantiated with subsequent work on the reconstruction of this ancestor (Groussin et al, unpublished). Successful inference of IPMDH back to the LCA of the Firmicutes has been achieved by way of a larger phylogenetic tree of 70 sequences reconciled against a species tree based on 16S sequences with or without a more sophisticated model of evolution incorporating heterogeneity into the evolution rate of residue sites. Incorporation of site heterogeneity more accurately accounts for the structural and functional constraints on certain sites of the sequence, and the reconciled tree accounts for events such as duplication, horizontal gene transfer and gene loss. Use of the reconciled tree with either the more sophisticated model or the LG model resulted in ancestral proteins with biological properties similar to contemporary IPMDHs. The use of the more sophisticated model with an IPMDH sequence only tree also resulted in an active enzyme, however a high k_{cat} and low kinetic barrier to unfolding despite high T_{opt} indicate errors in the inference. Thus, for reconstructions back over large time periods in complex enzymes with significant levels of sequence divergence, it seems that more rigorous inference methods are required, especially the reconciliation of the sequence tree against a species tree.

In summary, ASR techniques have been applied to the reconstruction of IPMDH from the LCA of the Firmicutes. This reconstruction tested the limits of ASR techniques in terms of the time scale and level of divergence across contemporary sequences. However, once reconstructed, the inferred LCA enzyme was found to be inactive, and to form a soluble aggregate. For reconstructions of this level of sequence divergence, it seems that more sophisticated ASR methods are required to successfully infer the ancestral sequence.

4 Substrate Promiscuity and *in vivo* Fitness of IPMDH Ancestors

4.1 Introduction

Enzymes are believed to have started off as generalists, catalysing a wide range of reactions. Over time these generalists have evolved into specialists acting proficiently in the catalysis of one substrate (Jensen 1976; Depristo 2007; Hernandez-Montes et al. 2008). Given this, ancestral enzymes should show greater activity on a wider range of substrates than their contemporary counterparts.

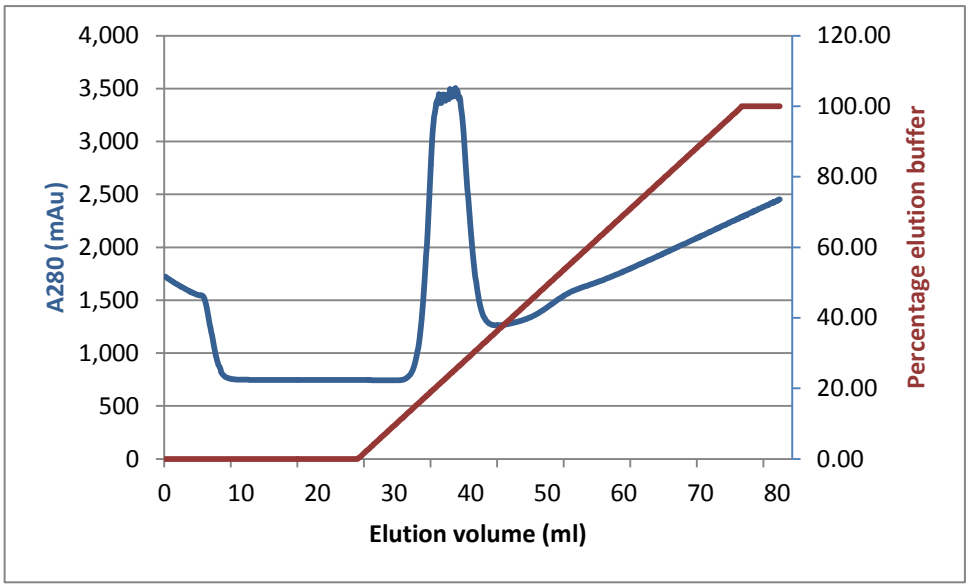
The previous reconstruction of ancestral IPMDH from *Bacillus* revealed an interesting trend regarding the evolution of enzymatic rate and kinetic stability over the history of the enzyme (Hobbs et al. 2012). ANC1 (approximately 670 mya – refer to chronogram in Figure 1.5) and ANC4 (approximately 950 mya) showed significantly faster catalytic rates than contemporary enzymes measured (k_{cat} of 141.8 s^{-1} and 362.2 s^{-1} respectively, compared to 53.8 s^{-1} for IPMDH BCVX, the fastest measured contemporary IPMDH). In addition to this, ANC4 is more kinetically stable, with a free energy of unfolding ($\Delta G^{\ddagger}_{N\text{-}U}$) of 110.8 kJ mol^{-1}, compared to 100.7 kJ mol^{-1} and 100.9 kJ mol^{-1} for BCVX and ANC1 respectively. This trend begs the question, if these ancestral enzymes were so fast and stable, why have they been lost over evolutionary time in favour of catalytically slower and less stable equivalents? This question was addressed by assessing the effect the ancestral IPMDH enzymes have in a cellular context while functioning within the normal biosynthetic pathway of leucine to see if the fast catalytic rate of ANC1, or fast catalytic rate and kinetic stability of ANC4 have any fitness costs *in vivo*.

4.2 Results and Discussion

4.2.1 Expression and Purification of ANC1, ANC4 and BCVX

ANC1, ANC4 and BCVX IPMDH proteins were expressed via IPTG induction in TB media at 37 °C as described in section 2.3.1 [as per methods from Hobbs, et al. (2012)]. Cells were pelleted from overnight culture, sonicated to achieve lysis, and centrifuged to separate cell debris. The protein was separated from the supernatant and majority of other proteins by nickel affinity chromatography via N-terminal hexa-histidine tag. Protein was removed from the nickel affinity column via gradient elution against increasing concentrations of imidazole (section 2.3.2.2). The SE trace and SDS-PAGE gel of this process is shown in Figure 4.1. SDS-PAGE lanes corresponding to the insoluble pellet and supernatant show that most of the ANC4 protein is soluble, with only a small proportion in the insoluble pellet. In the nickel affinity column flow through, ANC4 IPMDH protein has been removed from solution and separated from the other proteins evident in the solution. Protein elution from the SE column can be seen in the UV trace at 280 nm around 30 % buffer B (about 300 mM imidazole). The SDS-PAGE gel shows the SE fractions collected are comprised predominantly of the target ANC4 protein, with a few minor protein impurities present.

A



B

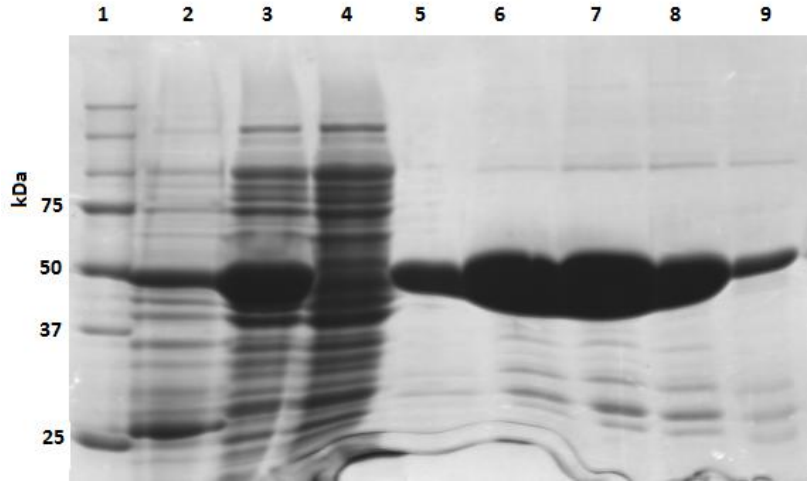


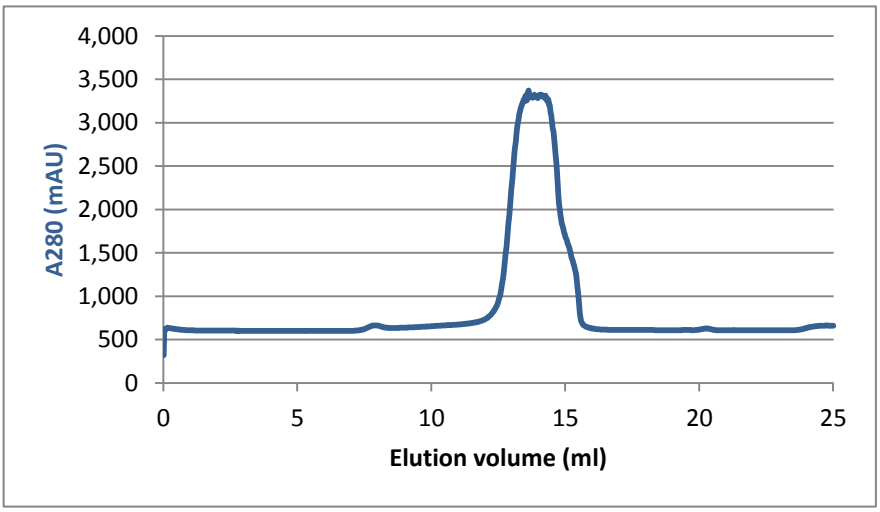
Figure 4.1: Nickel affinity purification of ANC4 protein.

(A) Elution trace of ANC4 during nickel affinity chromatography against an increasing concentration of imidazole. Protein elution was followed at 280 nm; (B) SDS-PAGE gel from nickel affinity purification of ANC4. Lane 1, protein ladder; lane 2, insoluble pellet; lane 3, supernatant; lane 4, nickel affinity column flow through; lanes 5 – 9, fractions from 34 to 44 ml of elution volume from the nickel column elution trace.

Pooled protein containing fractions from the nickel affinity column were further purified by SE chromatography as described in section 2.3.2.3. A trace of the SE process for ANC4 and the corresponding SDS-PAGE gel are given in Figure 4.2. Elution of the protein occurs at around 15 ml, consistent with a protein dimer.

The SDS-PAGE gel of protein containing fractions shows that the protein is largely pure at this stage, with only minor impurities present in the fractions.

A



B

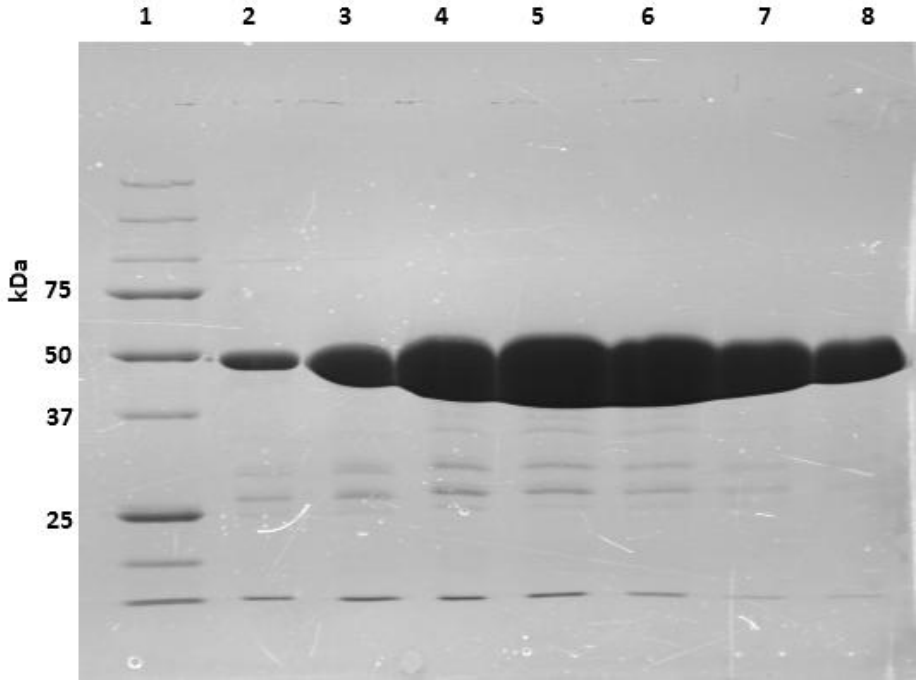


Figure 4.2: SE column purification of ANC4 protein.

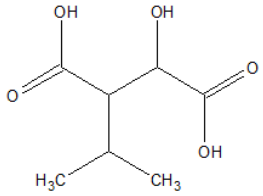
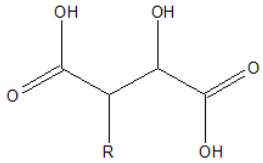
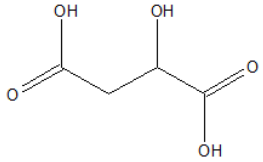
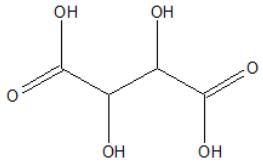
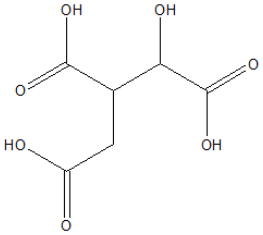
(A) SE trace at 280 nm of ANC4 elution; (B) SDS-PAGE gel of selected fractions from SE purification of ANC4. Lane 1, protein ladder; lanes 2 – 8, fractions from 12.5 to 15.5 ml of elution volume from the SE elution trace.

ANC1 and BCVX proteins were successfully purified by the same methods described here. Nickel affinity chromatography and SE column traces, and SDS-PAGE gels were similar to the representative examples given for ANC4.

4.2.2 Characterisation of Activity with Alternative Substrates

Contemporary IPMDHs catalyse oxidative decarboxylation of a range of alkylmalates (Miyazaki et al. 1993; Matsunami et al. 1998). The general structure of alkylmalates, as well as the native and alternative substrates of IPMDH are given in Table 4.1. Alternative substrates for IPMDH were assessed to test the hypothesis than substrate promiscuity will increase with older ancestral enzymes.

Table 4.1: Chemical structures of the native substrate of IPMDH, and the alternative substrates assayed in this study.

IPM	
Alkylmalate	 Where R is an alkyl chain
Malate	
Tartrate	
Isocitrate	

4.2.2.1 *Substrate specificity of Alternative Substrates*

Kinetic analysis of ANC1, ANC4 and BCVX enzymes with alternative substrates was performed as described in section 2.3.3.1. BCVX was chosen as the extant species for comparison due to its similar T_{opt} to the two ancestral enzymes being investigated. Kinetic analyses of alternative substrates for the IMPDH enzyme were performed at the previously determined T_{opt} of each enzyme (Hobbs et al. 2012).

Initially, K_M determinations for each enzyme with the alternative substrates was attempted. These were performed with an excess of the cofactor NAD^+, and increasing concentrations of the substrate till a plateau in rate occurred. These data is shown in Figure 4.3. Data was plotted in GraphPad Prism Version 5 (GraphPad Software, USA) using the Michaelis Menten function, however this function gave a poor fit to the data in most cases.

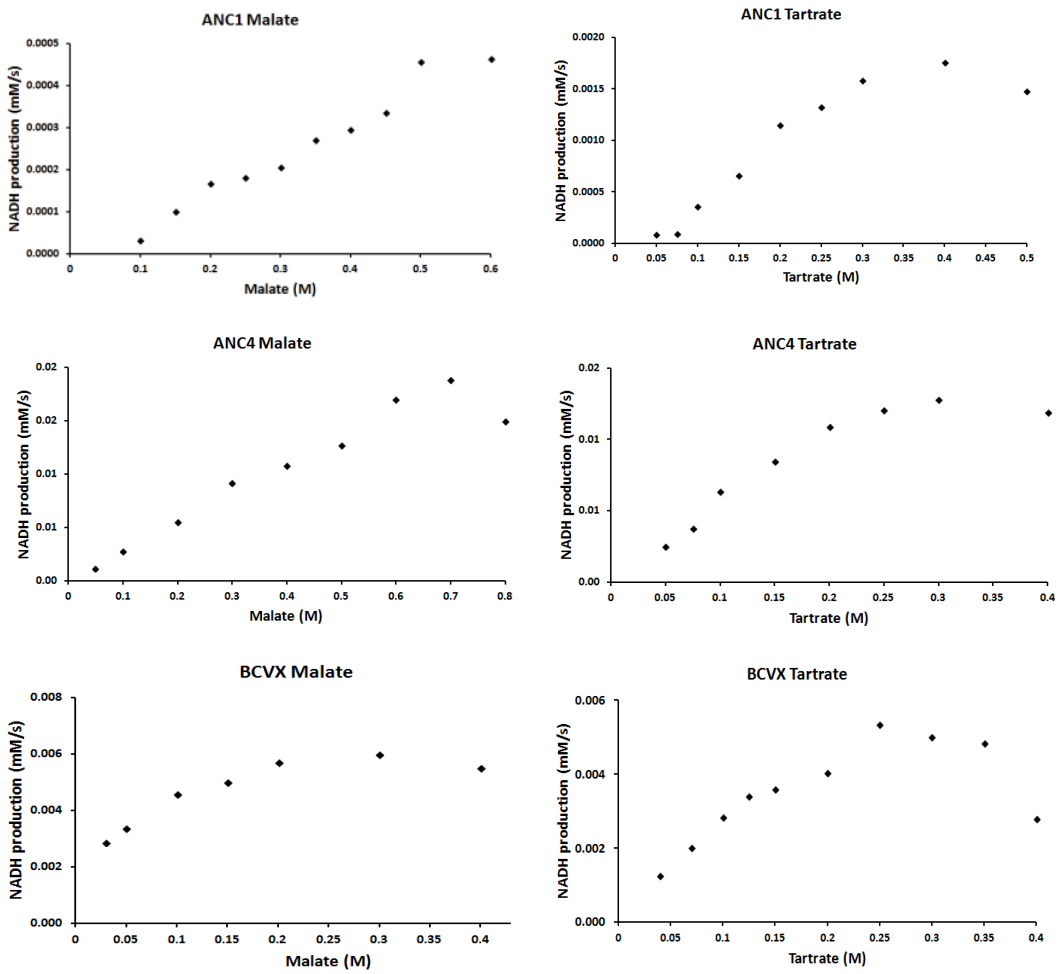


Figure 4.3: Activity assays performed for the alternative substrates for the enzymes ANC1, ANC4 and BCVX at the T_{opt} of each respective enzyme. Data was concluded not to be suitable for Michaelis Menten kinetic analysis.

Plots which fit Michaelis Menten curves well were produced by BCXV with both substrates and ANC4 with tartrate. However issues are apparent in the remainder of the plots which resulted in ill fitting Michaelis Menten curves. In these cases, issues are apparent with either a failure of the data to plateau off (often due to substrate solubility issues, especially with tartrate) or the data not passing through zero. ANC1 with tartrate, and to a lesser extent ANC4 with tartrate, show a slight lag in activity at low substrate concentrations. Substrate inhibition at high concentrations of malate is also evident for BCVX. These effects resulted in inaccurate K_M and V_{max} values, which were not suitable for comparison to evaluate the efficiency of each enzyme with the alternative substrates.

Due to the issues encountered with the Michaelis Menten kinetic analysis, a measurement of specific activity was used instead.

4.2.2.2 *Specific Activity Determination for the Alternative Substrates*

Specific activities were measured by the process in section 2.3.3.1 at the T_{opt} of each enzyme. Substrate was kept at a constant concentration between each enzyme. Substrate concentration was chosen based on the activity seen in the assay curves in Figure 4.3 so that a reasonable level of activity would be measured for each enzyme. Use of specific activity to compare enzymes also allowed isocitrate to be included in the analysis, as low solubility rendered Michaelis Menten analysis of this substrate unfeasible.

Table 4.2: Specific activities of ANC1, ANC4 and BCVX on the alternative substrates at the T_{opt} of the respective enzymes.

	ANC1 (mM product s^{-1} mM $protein^{-1}$)	**ANC4 (mM product s^{-1} mM $protein^{-1}$)**	**BCVX (mM product s^{-1} mM $protein^{-1}$)**
Malate (0.2 M)	1.6 ± 0.2	10.3 ± 0.1	3.2 ± 0.2
Tartrate (0.15 M)	0.7 ± 0.1	14 ± 1	2.9 ± 0.3
Isocitrate (0.2 M)	0.90 ± 0.06	0.084 ± 0.004	1.03 ± 0.05

Errors are reported as the SD of the replicates. Substrate concentrations used are given under the respective substrates.

ANC1 and BCVX showed similar low levels of catalysis against all three alternative substrates tested. ANC4 was expected to be faster with alternative substrates than ANC1 and BCVX as it is already known to be significantly faster on the natural substrate, IPM (k_{cat} of 362.2 s^{-1} compared to 141.8 s^{-1} and 53.8 s^{-1} and ANC1 and BCVX respectively). This was observed for malate and tartrate, however not with isocitrate, where a ten-fold decrease in activity compared to ANC1 and BCVX was found. Based on the age of the citrate cycle from which isocitrate is derived (Cunchillos & Lecointre 2003), isocitrate would have been an

encountered substrate 950 mya, at the time ANC4 was active. This does not support the idea that substrate specificity increases with age, as the most distant enzyme in the tree shows significant levels of activity with fewer substrates than the more recent descendents. This is similar to the conclusion drawn by Wouters, et al. (2003) where it was found that enzyme specificity over time is not a linear process. These results are not in agreement with the trend from generalist to specialist that has been found in ancestral β-lactamases (Risso et al. 2013). Without knowing the exact environment and compounds the ancestral enzymes were in contact with and the availability and necessity of these compounds, it is difficult to rationalise the activity of ancestral enzymes on various substrates. It is also possible that the reconstruction has not been taken back to old enough ancestors to measure increases in substrate promiscuity.

4.2.3 In vivo Characterisation

To investigate the possible reasons why ANC1 and ANC4 have been lost over evolutionary time and superseded by catalytically slower and kinetically less stable equivalents, the fitness of these enzymes *in vivo* was investigated. Growth rates were used to determine if the fast catalysis of ANC1 and ANC4 and the stability of ANC4 correlated to *in vivo* fitness, or if these enzymes have growth rate costs that are not reflected by the *in vitro* parameters.

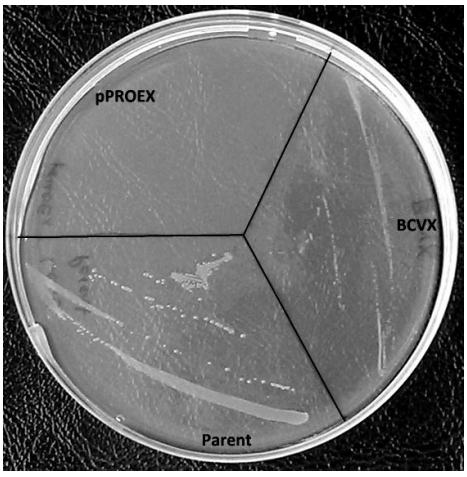
4.2.3.1 *Strain Construction*

To investigate the effect the enzymes ANC1, ANC4 and BCVX have on the growth rates of cells, the pPROEX expression constructs were transformed into Keio collection *E. coli* cells which have had the natural IPMDH gene knocked out (referred to as the *leuB* KO). Although the use of *E. coli* and an over expression vector have limitations, this system offered a preliminary set up to investigate the effect of the ancestral and extant enzymes on growth rate. The constructs were also transformed into *E. coli* strains with the tartrate dehydrogenase (*yeaU*), malate dehydrogenase (*mdh*) and isocitrate dehydrogenase (*icd*) genes knocked out to investigate how well each IPMDH enzyme complemented the normal

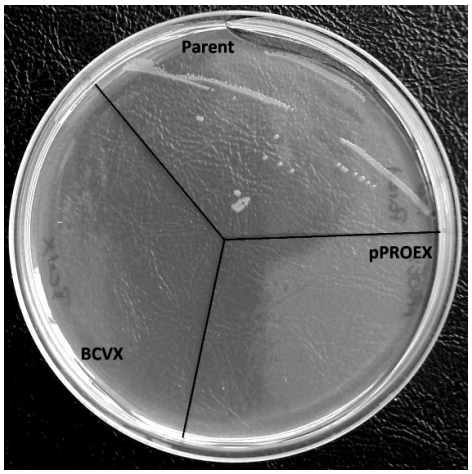
oxidation enzyme for the alternative substrates. A table of the cell lines created is available in Appendix A2.

4.2.3.2 *Plate Growth Trials*

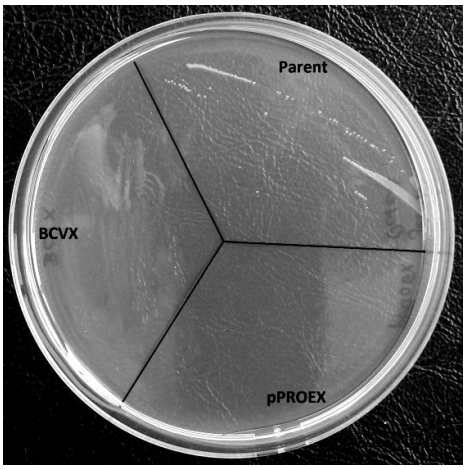
To determine which carbon source to grow each KO on to distinguish between successfully and unsuccessfully complemented strains, cells were streaked from glycerol stocks onto M9 agar (Appendix C2) supplemented with various carbon sources, 100 µg/ml AMP and 1 mM IPTG as described in section 2.4.2.1. An appropriate carbon source was determined by growth of the parent strain with no gene knocked out, and no growth of the KO complemented with empty pPROEX vector. Results of this are illustrated in Figure 4.4.



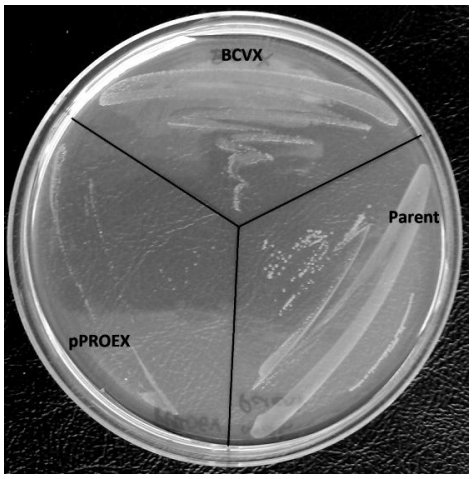
A) *leuB* KO, glucose carbon source



B) *icd* KO, glucose carbon source



C) *yeaU* KO, malate carbon source



D) *mdh* KO, malate carbon source

Figure 4.4: KO controls grown on M9 agar supplemented with various carbon sources as indicated, 100 µg/ml AMP and 1 mM IPTG.

Glucose (10 g/L) was found to be an appropriate carbon source for the *leuB* (Figure 4.4 A) and *icd* (Figure 4.4 B) KO strains by the growth of the parent strain, and lack of growth of the KO strain complemented with pPROEX only, consistent with previously reported results (Kabir & Shimizu 2004; Chen et al. 2011). D-malate (2 g/L) was successful for growing the *yeaU* KO strain [Figure 4.4 C; (Reed et al. 2006)]. Despite trialling a number carbon sources (malate, succinate and acetate) reported as appropriate for growth of the *mdh* KO (van der Rest et al. 2000), no source was found which eliminated growth of the KO strain complemented with pPROEX (Figure 4.4 D).

It is possible that the negative control for the *mdh* KO in which the malate dehydrogenase enzyme is removed and not complemented with any replacement gene is showing growth due to gene compensation. It is well documented that many enzymes can catalyse multiple reactions with different substrates. *In vivo* this can result in another enzyme performing the role of missing gene, thus compensating for the function and mitigating the KO phenotype (Khersonsky & Tawfik 2010). As no carbon source was found which gave growth dependent on the plasmid introduction of a gene to complement the KO, no further analysis of the *mdh* KO was able to be performed.

Relative growth rates of each complemented KO were determined on M9 medium by plating out a series of diluted overnight culture as described in section 2.4.3.1. Overnights were grown in LB to achieve equal growth over all complemented KOs. Prior to plating, cells were washed twice in M9 medium to remove any residual LB and components of this medium which may facilitate growth in a manner not reliant on the activity of the complemented gene. Growth results are shown in Figure 4.5. For the *leuB* KO, all three introduced genes rescued the KO. Growth rates for ANC1 and BCVX are comparable, showing growth right through to the lowest dilution. ANC4, however, showed significantly poorer growth compared to the other two strains. Complemented *icd* and *yeaU* KOs all showed only slight levels of growth. Less growth of these *yeaU* and *icd* KOs compared to the *leuB* KO was expected given the low levels of catalysis measured for IPMDH against the alternative substrates *in vitro*. Both ancestral IPMDHs showed slightly

better rescue of growth than the contemporary BCVX in the *icd* KO. In the *yeaU* KO, the ancestral and contemporary IPMDH genes showed approximately equal levels of growth.

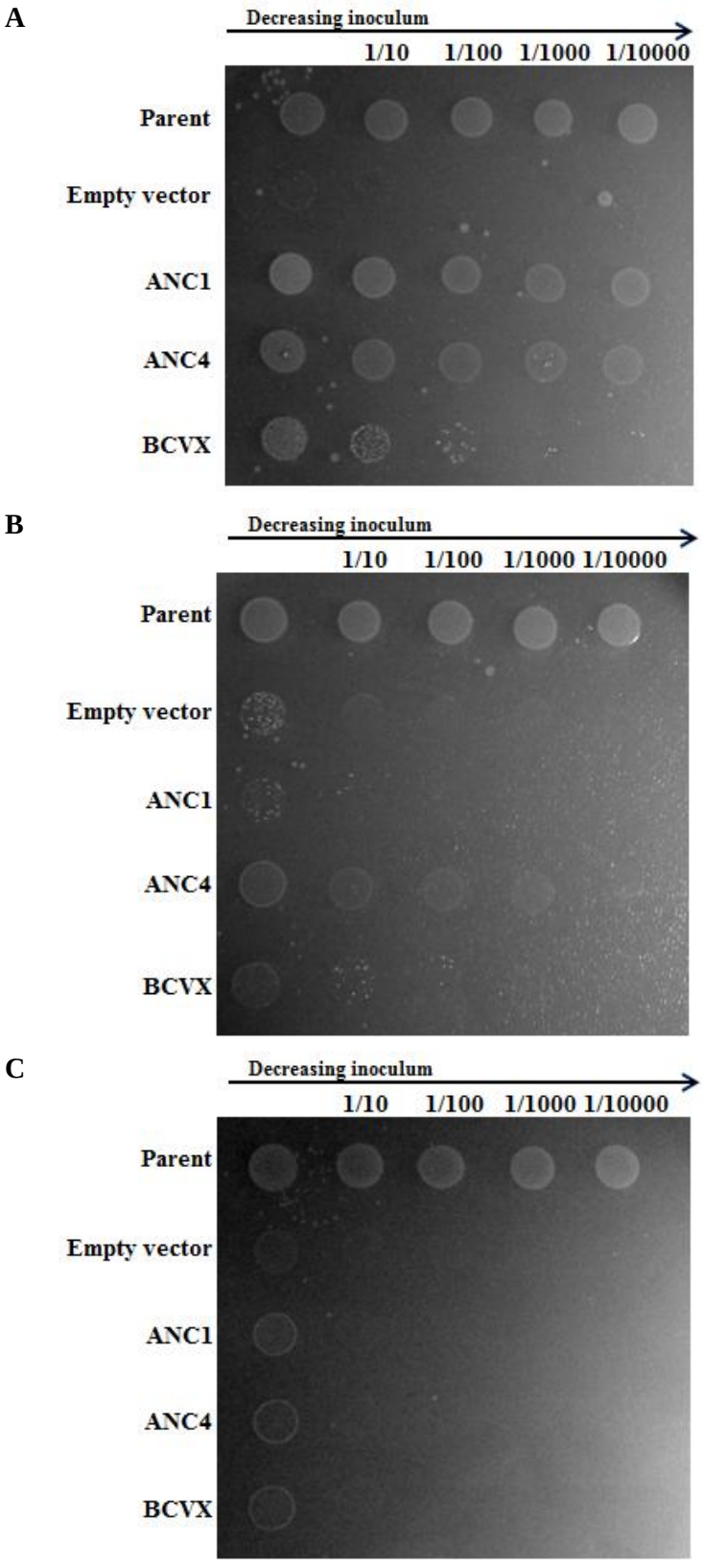


Figure 4.5: Relative growth rates on M9 agar of the Keio collection KOs complemented with extant and ancestral IPMDH genes.
(A) *leuB* KO with glucose carbon supply; (B) *icd* KO with glucose carbon supply; (C) *yeaU* KO with malate carbon supply.

4.2.3.3 *Growth Rate Determination*

As plate growth is only a semi-quantitative measure, growth rates in liquid M9 medium were determined. Starter cultures in M9 medium were inoculated at a 1 in 100 dilution from LB overnights and grown for 48 hours at 37 °C to achieve sufficient cell density. These were then used to inoculate cultures for growth rate determination. Growth rate determination was performed with 100 µl cultures in a shaking plate incubator with OD measurements at 600 nm performed every 20 minutes until stationary phase was entered as described in section 2.4.3.1. This method was found to give consistent growth curves for the *leuB* KOs, as shown in Figure 4.6.

4.2.3.3.1 Growth Rates of icd and yeaU Knock Outs

Complemented strains of the *icd* and *yeaU* KOs showed insignificant levels of growth over up to 70 hour growth periods. This indicates that IPMDH is too inefficient at catalysing the oxidation of isocitrate and tartrate, respectively, to complement the knocked out gene effectively and allow growth. This is especially significant considering that in this system IPMDH is being expressed at intracellular concentrations higher than normal levels. If expressed at natural intracellular concentrations, all three variants of IPMDH tested would be even less able to complement the *icd* and *yeaU* KOs.

4.2.3.3.2 Growth Rates of leuB Knock Outs

This method was found to produce consistent growth curves for the *leuB* KOs. Representative growth curves are given in Figure 4.6. Of interest is the large over growth of ANC4 before dropping back to the final cell density. All three variants displayed approximately the same OD at final density, however ANC1 and BCVX showed no or minor growth above this level before stabilising. Although the significance of this is unknown, the overshoot suggests unregulated cell growth, and may be related to the greater ability of ANC4 to turnover IPM. However, it appears the cells over use resources in the media, resulting in a drastic drop in OD back to similar levels as ANC1 and BCVX complemented strains. The greater lag observed for BCVX compared to ANC1 and ANC4 despite the similar levels of

exponential phase growth is also of interest. The significance of this is also unknown, and would require further investigation.

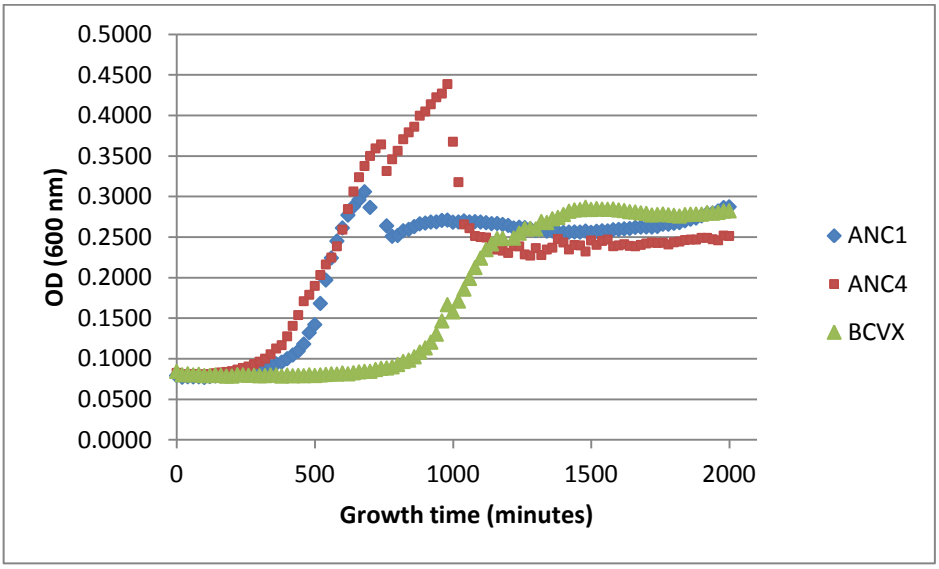


Figure 4.6: Representative growth curves for Keio collection *leuB* KOs complemented with extant and ancestral IPMDH genes.

From the growth curves, data from the log phase of growth was used to determine the specific growth rate (μ) by taking the slope of a ln OD versus time graph (Breidt et al. 1994). Specific growth rates were then converted to the doubling time of the culture using the following equation:

Specific growth rate = ln2/μ

Doubling times are reported in Table 4.3. A complete set of growth curves for complemented *leuB KOs* is available in Appendix D. Raw data is available on the attached disk.

Table 4.3: Doubling times of *leuB* KOs complemented with ancestral and contemporary IPMDH genes in M9 medium with 10 g/L glucose as a carbon source.

Complemented strain	Doubling time (h)
	Average
ANC1	2.6 ± 0.7
ANC4	2.6 ± 0.6
BCVX	3.1 ± 0.4

Doubling times are the mean of 3 biological replicates, each with 3 internal replicates. Errors are reported as the pooled SD of the replicates.

The determined doubling times showed high levels of variability. Internal replicates of growth rate from the same inoculum generally gave consistent doubling times, however larger variability was present between biological replicates. Overall the results do not show statistically significant differences in doubling times between the different ancestral and extant enzymes, and contradict the observed plate growth trends.

Subsequently, significant differences between *in vivo* growth rates for these IPMDH variants has been successfully shown (Hobbs et al, unpublished). Ancestral and extant IPMDH expression was under the control of the natural *leu* operon promoter, cloned into the pUC19 plasmid. Growth rates correlated with the trends seen in the plate growth trials in section 4.2.3.2, where ANC1 and BCVX showed greater levels of growth than the KO complemented with the ANC4 gene. The inability to observe this trend in growth rate with the pPROEX system is likely due to the artificially high levels of enzyme resulting from the overexpression vector. Despite the equal addition of IPTG to induce enzyme expression, this system is not guaranteed to give equal expression over all variants or replicates. The abnormal cellular compositions created by IPMDH overexpression also has the potential to disrupt normal cellular functions and obscure growth rate differences present between the variants.

The decreased fitness of ANC4 *in vivo* seen in plate growth and by Hobbs et al (unpublished) explains the loss of this enzyme over evolutionary time despite the faster rate and greater kinetic stability measured *in vitro*. This lowered *in vivo*

fitness could be due to a variety of factors. The fast turnover of substrate by ANC4 potentially consumes too much of the cofactor NAD^+, disrupting other reactions dependent on NAD^+. Excessive build up of reaction product could also have adverse effects on cells. The kinetic stability of ANC4 may also cause issues in the regulation of the pathway if slower turnover rates of the enzyme affect the fine scale control needed in the biosynthetic pathway. Further investigation is necessary to determine the reasons for slower growth of ANC4 than ANC1 or the contemporary BCVX.

5 The X-ray Crystal Structure of ANC1

X-ray crystallography is a technique used to determine the 3D structure of proteins. This technique relies on the crystallisation of a protein into a well ordered lattice. X-ray beam photons can then be diffracted by the electrons of atoms within the protein crystal, resulting in a diffraction pattern. This diffraction pattern contains information in the amplitude and position of each spot. However, phase information, which is also required for transforming the data into a 3D protein structure, cannot be determined directly from the diffraction pattern. Phase information can be recovered directly from the crystal by various anomalous diffraction methods, or through molecular replacement techniques. Molecular replacement relies on a similar protein structure from which phase information can be inferred. With phase information, the diffraction pattern can be transformed into an electron density map, into which the protein sequence can be built.

A number of structures for IPMDH have previously been solved, including that of ANC4. The structure of ANC1 would give a more complete structural view of IPMDH evolution in *Bacillus*, and the differences present between ANC1 and ANC4.

5.1 Results and Discussion

5.1.1 Crystallisation of IPMDH ANC1

Crystallisation screens of 384 conditions (Hampton Research, USA) were set up for ANC1 with and without additives (NAD^+ and Mn^{2+}). Crystallisation screens were performed by the sitting drop method as described in section 2.5.3.1. For the crystallisation screen with additives, small crystals were observed in a number of conditions. Three conditions were carried through to fine screens, where the pH and concentration of precipitant were varied. Crystals resulting from these fine

screens were too small for data collection, therefore crystal seeding (section 2.5.3.2) was performed from a fine screen around a condition with 0.1 M sodium citrate tribasic, pH 5.6, 20 % v/v propan-2-ol and 20 % w/v PEG 4,000. Batch seeding and streak seeding were attempted from various conditions from the initial fine screen. Batch seeding from a 1 in 10,000 dilution of the original fine screen crystal drop resulted in larger crystals (Figure 5.1). Crystals were tested for diffraction on a SuperNova X-ray diffractometer (Agilent, USA; section 2.5.5), however no diffraction resulted from these crystals.

A **B**

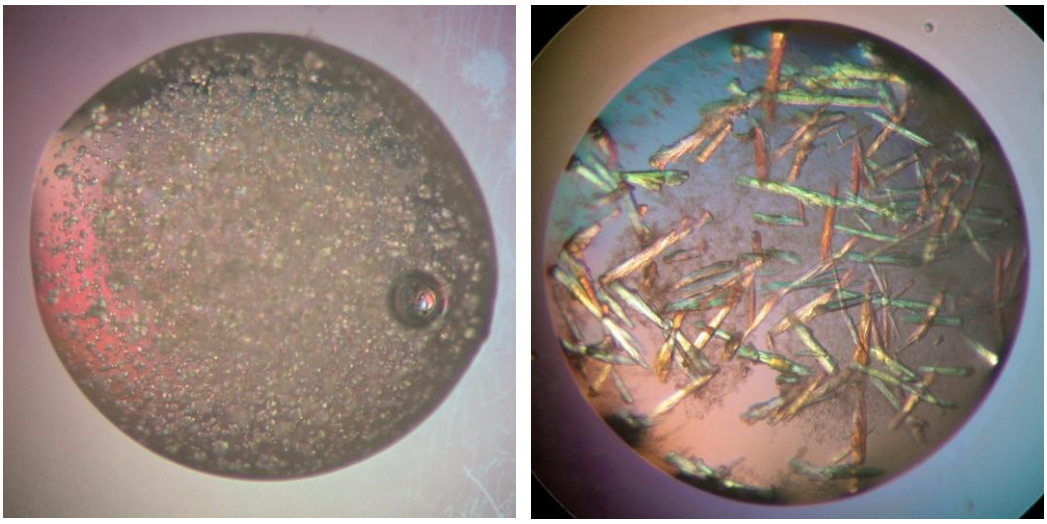


Figure 5.1: IPMDH ANC1 crystals with additives.
Crystals were grown in 0.1 M sodium citrate tribasic, pH 5.2, 20 % v/v propan-2-ol and 20 % w/v PEG 4,000, with additives of NAD^+ (1 mM) and $MnCl_2$ (0.2 mM). (A) initial crystals formed in a hanging drop fine screen; (B) crystals formed from batch seeding (1 in 10,000) from crystals in (A).

In the crystallisation screen without additives, crystals formed in a solution containing 0.2 M NaCl, 0.1 M sodium acetate, pH 4.6 and 30 % v/v 2-methyl-2,4-pentanediol (MPD). The pH and concentration of the precipitant (MPD) were altered in the fine screening process. Good crystals were formed across many of the fine screen conditions, an example of which is given in Figure 5.2. Testing of these crystals on a SuperNova X-ray diffractometer (Agilent, USA; section 2.5.5) showed good diffraction out to 2.5 Å resolution.

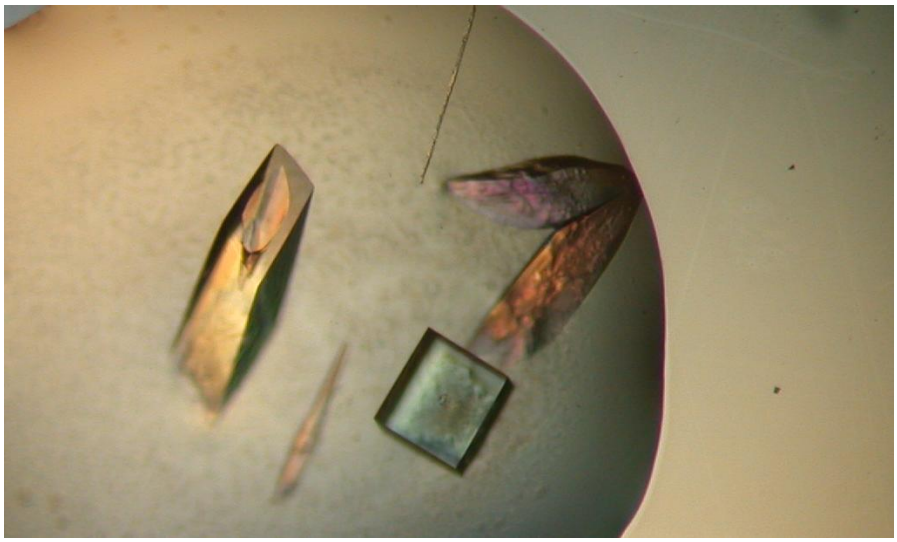


Figure 5.2: IPMDH ANC1 crystals.
Crystals were grown in 0.2 M NaCl, 0.1 M sodium acetate trihydrate, pH 4.8 and 30 % v/v MPD. The final crystal structure was solved from the cubic crystal situated to the bottom right of the image.

5.1.2 X-ray Data Collection

Crystals giving good diffraction were sent to the Australian Synchrotron for data collection. Data were collected for the cubic crystal shown in Figure 5.2 to 2.3 Å resolution. Diffraction data are shown in Figure 5.3. Data were also collected with increased attenuation giving a low resolution data set due to the presence of overloaded spots in the high resolution data set. Diffraction data were collected over 360° in 1° increments for both data sets, with detector distances of 180 mm and 220 mm for the high resolution and low resolution data sets, respectively. No deterioration of the crystal diffraction was observed over the collection of the two data sets.

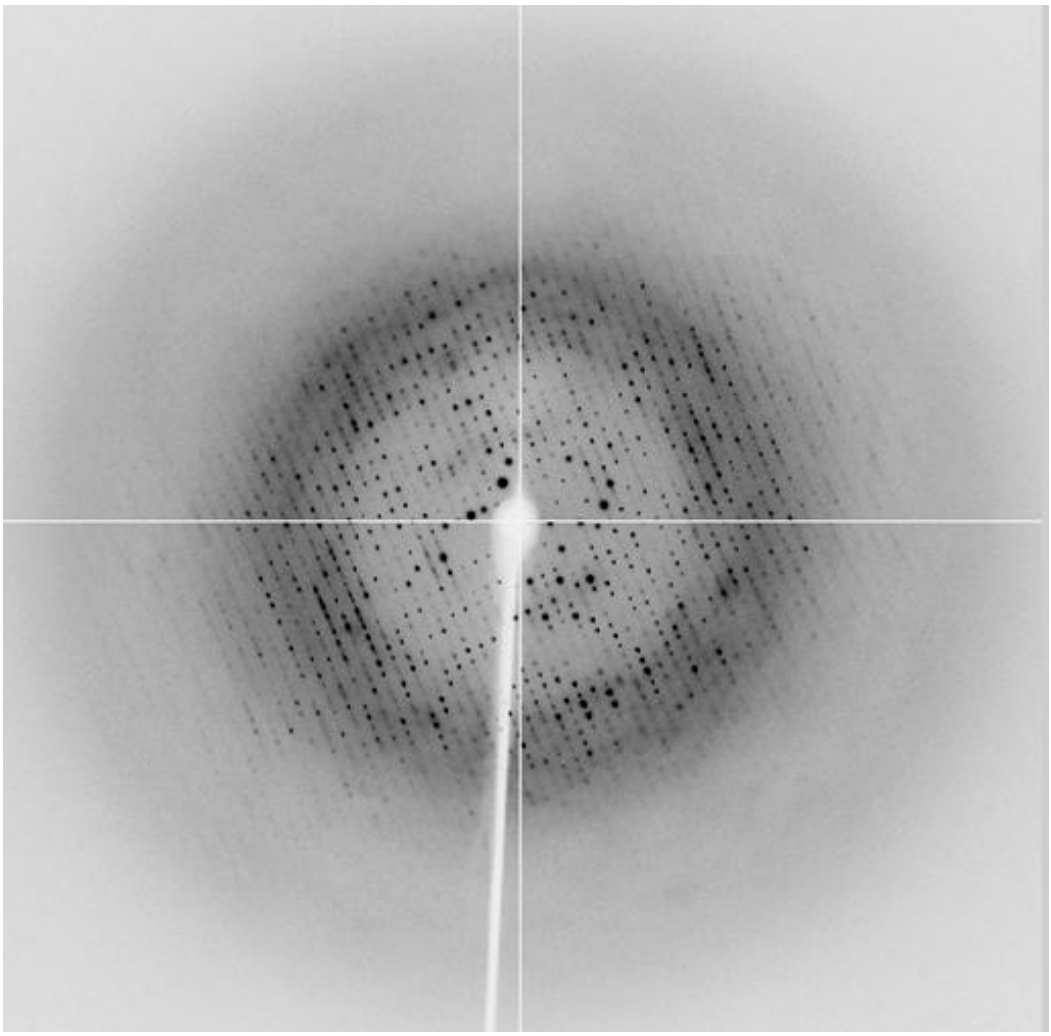


Figure 5.3: X-ray diffraction pattern of IPMDH ANC1.
The image boundary represents 1.85 Å resolution.

5.1.3 Data Processing

The space group for the data was determined in MOSFLM (Leslie & Powell 2007) to be $P4_22_12$, with a tetragonal unit cell. The high and low resolution data sets were integrated in MOSFLM separately. The two integrated data sets were combined in Sortmtz (P. J. Daly, Daresbury) within the CCP4 program suite (Winn et al. 2011). The integrated and combined reflections were scaled and merged in SCALA (Evans 2006) with a resolution cut off at 2.6 Å. Matthews coefficient analysis (Matthews 1968) predicted two molecules in the asymmetric unit. Data collection statistics are provided in Table 5.1.

Table 5.1: Data collection statistics for IPMDH ANC1.

Data statistics	**Overall**	**Outer shell**
Space group	$P4_22_12$	
Wavelength (Å)	0.9686	
Cell dimensions		
a b c (Å)	111.92/119.92/62.26	
α β γ	90/90/90	
Mosaicity	0.957 (1.013)	
Monomers in the asymmetric unit	2	
Resolution range (Å)	41.65-2.6	2.74-2.6
Number of observed reflections	624430	86097
Number of unique reflections	12703	1808
R_{merge}	0.112	0.697
Mean I/σI	28.9	7.3
Completeness	100	100
Multiplicity	49.2	47.6

Values in parenthesis are for the low resolution dataset. Statistics with only one value given are the same value for the two sets or for the combined data.

5.1.3.1 *Molecular Replacement, Model Building and Refinement of ANC1*

Molecular replacement was performed in Phenix (Adams et al. 2010) using the previously solved ANC4 structure (PDB code 3U1H). The two proteins are closely related, with 82.4 % sequence similarity. The model was further improved using AutoBuild within the Phenix software. The model was built manually in COOT (Emsley & Cowtan 2004) into $2|F_O|-|F_C|$ and $|F_O|-|F_C|$ maps contoured to 1 σ and 3 σ, respectively. Refinement of the model was performed in Phenix.

The final model has an R-factor of 23.21 % and an R_{free} of 29.01 % (see Table 5.2 for a full list of refinement statistics). There were a number of residues in the structure that were not able to be resolved. Areas which were unable to be resolved were the N and C termini, as well as three loop regions situated at the periphery of the structure. A total of 320 of the 369 amino acids were able to be

modelled successfully. A number of side chains also lacked density throughout the structure. Electron density was also found to accommodate acetate and glycerol molecules. The average B factor for the protein is 49.3 $Å^2$. Ramachandran analysis by Procheck in the CCP4 program suite (Winn et al. 2011) revealed that 91.0 % of residues are in favoured regions, 8.7 % are in allowed regions and 0.3 % are in disallowed regions.

Table 5.2: Refinement and model statistics for ANC1.

Refinement and model statistic	
R-factor	23.21
R_{free}	29.01
Total number of atoms	2518
Total number of protein atoms	2441
Other ions/molecules	2
Number of waters	56
RMSD	
Bond lengths (Å)	0.008
Bond angles (°)	1.157
Average B factors ($Å^2$)	
Protein monomer	49.3
Waters	48.0
Ramachandran analysis	
Percentage in favoured regions	91.0
Percentage in allowed regions	8.7
Percentage in disallowed regions	0.3

5.1.4 Structure of IPMDH ANC1

The X-ray crystal structure of ANC1 was solved at a resolution of 2.6 Å. The structure of ANC1 is shown in Figure 5.4. ANC1 is similar to previously solved IPMDH structures, comprising of two domains of helices with a twisted central plane of ten β strands. This conformation results in two clefts in the structure. The active site is situated in one of these cavities, at the front of the image in

Figure 5.4 A. Missing regions can be seen on the periphery of the structure in the loops connecting α helix and β sheet structures.

B-factors of the structure are higher in domain 1, as illustrated in Figure 5.5. Peripheral loops show the highest B-factors. By comparison domain 2, where the dimer interaction occurs, is more static. B-factors throughout the centre of the protein are considerably lower than the rest of the structure.

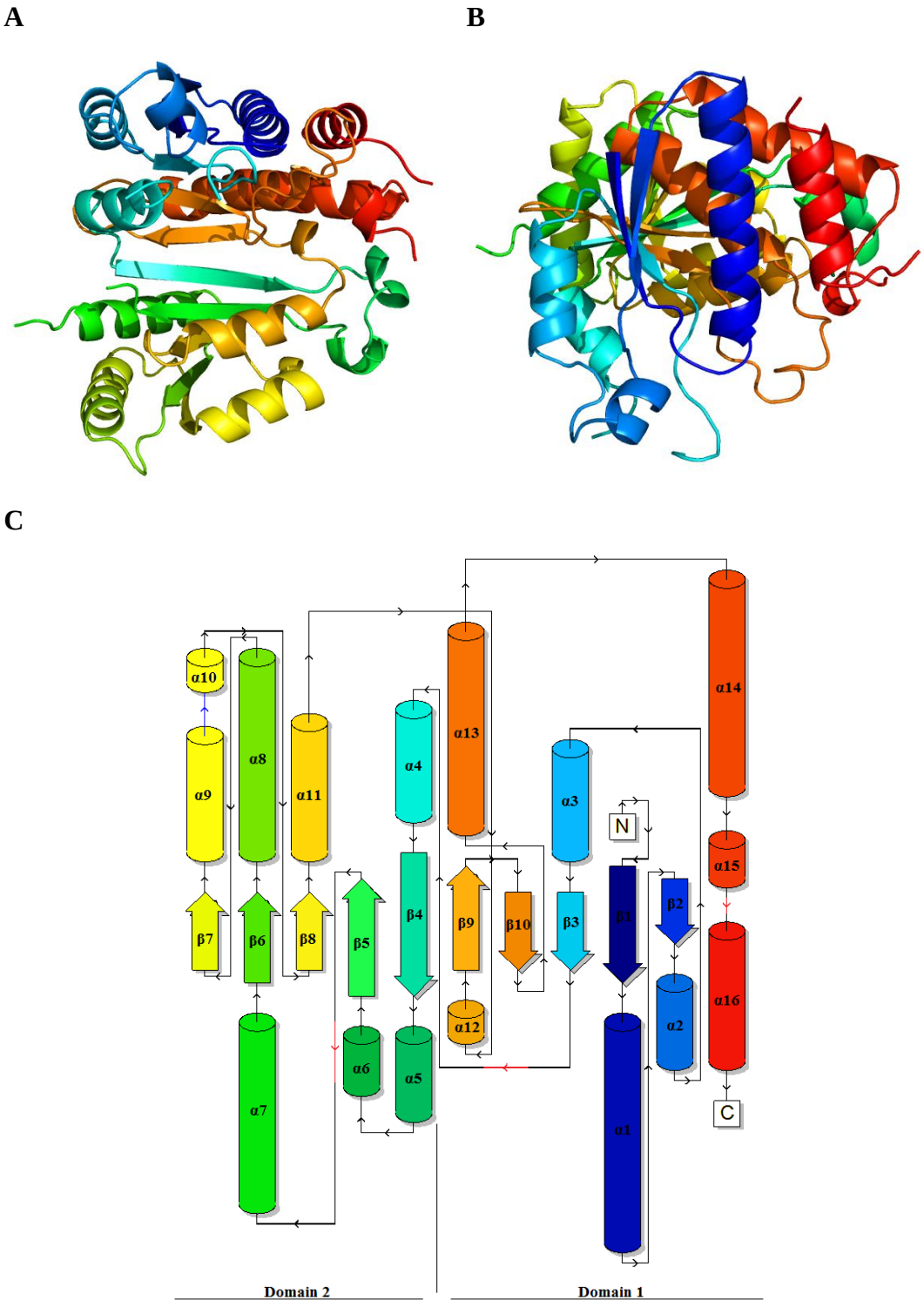


Figure 5.4: Cartoon representation of the monomeric unit of ANC1 IPMDH. (A) side view of ANC1; (B) top view of ANC1. Spectrum colouring starts at the N terminus in dark blue, and ends at the C terminus in red; (C) schematic topology of ANC1 IPMDH structure. α helix and β sheet numbering is indicated. Components are coloured to correspond to the colouration used in (A) and (B). Gaps in the structure are indicated with red lines.

Figure 5.5: Cartoon representation of ANC1 coloured according to B-factors. B-factors are indicated based on spectrum colouration, where red represents high B-factors, and dark blue represents low B-factors. The second monomer in the dimeric unit has also been included in purple to illustrate the dimer interface.

Active site residues in ANC1 were determined based on a sequence alignment with IPMDH from *T. thermophilus* and *A. ferrooxidans*, for which structures have been determined bound with NAD^+ [(Hurley & Dean 1994); PDB code 1HEX] and IMP [(Imada et al. 1998); PDB code 1A05] respectively. Active site residues are highlighted and numbered in Figure 5.6. The active site is present in the cleft between the two domains of the structure, between α11 and α4 and backed by β4 and β9. NAD^+ and IPM interacting residues, coloured red and blue respectively,

show the two areas in close proximity where the cofactor and substrate bind. Glu88, which is central to the recognition of the isopropyl moiety of the substrate as well as interacting with NAD^+ (Imada et al. 1998), is indicated in purple. Glu80 and Arg148, interacting with NAD^+ and IPM respectively, are positioned in loops for which electron density was not present in ANC1, and are thus not present in the structure. Lys190 and Asp222 in domain 2 contribute to substrate binding between subunits, interacting with the malate backbone of IPM in the active site of the second subunit of the dimer.

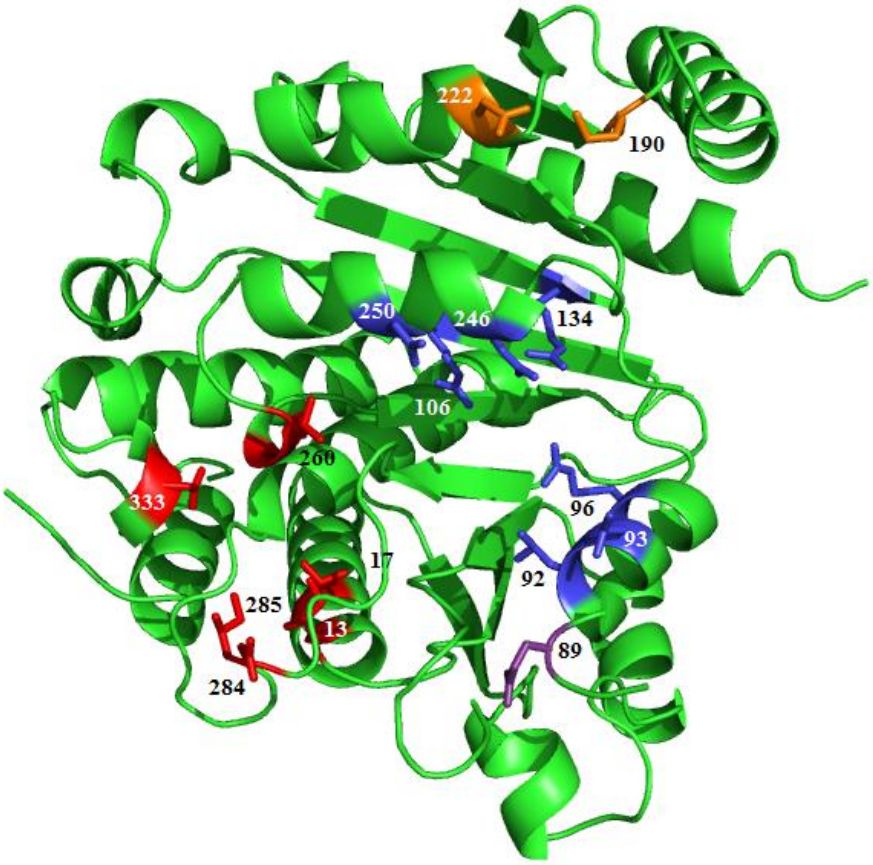


Figure 5.6: Carton representation of ANC1 showing the position of active site residues.

Active site residues were identified based on the *T. thermophilus* structure crystallised with NAD^+ [(Hurley & Dean 1994); PDB code 1HEX] and the *A. ferrooxidans* structure containing IPM [(Imada et al. 1998); PBD code 1A05]. Residues are coloured to indicate the interactions they partake in: red, NAD^+ binding residues; blue, IPM binding residues; orange, residues interacting with IPM between subunits of the dimer. Glu88 (purple) interacts with both NAD^+ and IPM. Residue numbers are as indicated.

Compared to the *A. ferrooxidans* structure with IPM bound [(Imada et al. 1998); PDB code 1A05], there is a difference in the angle of the two domains about the active site. When aligned by domain 1 (RMSD = 0.57 Å), domain 2 of ANC1 is more open by 4.9°, as measured from α11. This open conformation is consistent with the crystallisation state of ANC1 without any substrate, cofactor or metal bound in the active site (Graczer et al. 2011a).

The ANC1 dimer can be described based on the crystallographic symmetry of the two units comprising the dimer (Figure 5.7). The two units interact through α9 and α11, as well as a section of the loop connecting α5 and α6 in domain two. The two domains are in contact over an area of 971 $Å^2$ as calculated by PDBePISA (Krissinel & Henrick 2007).

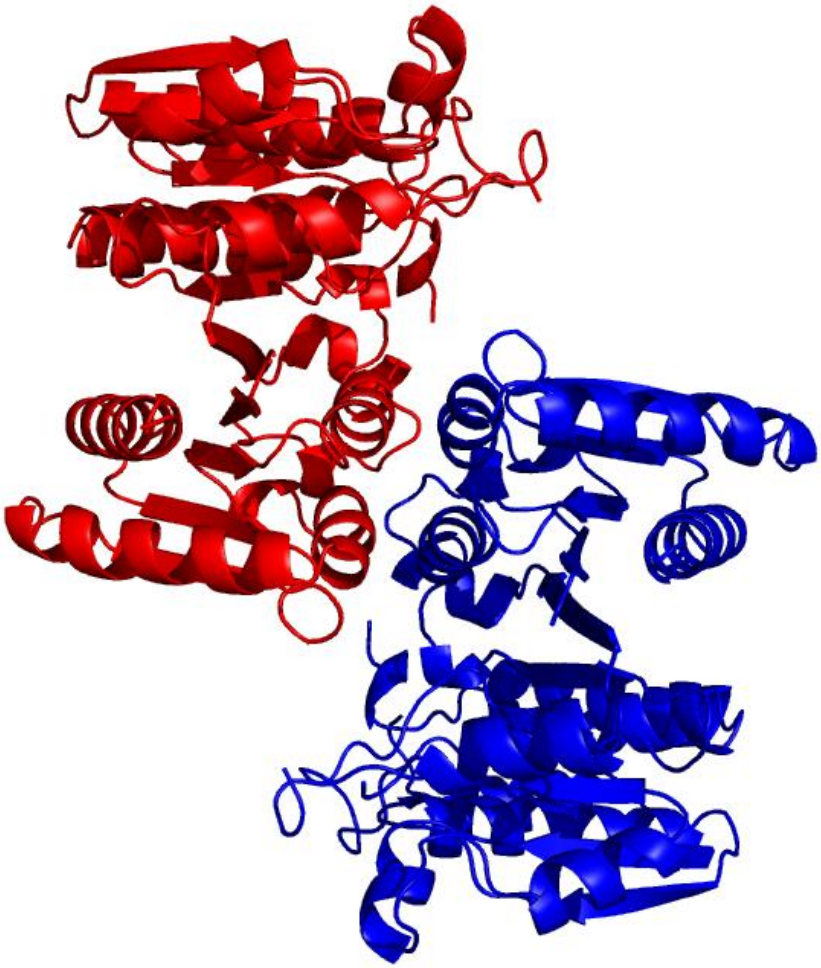


Figure 5.7: Cartoon representation of the dimeric structure of ANC1.

5.1.4.1 *Comparison to Other Structures*

Structural alignments of the ANC1 monomer with structures in the PDB revealed that ANC4 is the closest structural homologue of ANC1 (PDB code 3U1H). Structural alignment of the two enzymes had a Z-score of 20.5, an RMSD of 0.85 Å, and high P- and Q- scores (46.6 and 0.81 respectively) over 317 aligned residues (N_{align}). The Z-score is a measure of the statistical significance of the structural alignment, with values greater than 3 considered to be significant. The P-score is a measure of the negative logarithm of the probability of obtaining an identical or better structural alignment by chance, thus a higher P-score indicates a more significant structural match. Q-scores give the quality of the alignment, where 1 indicates an identical structure. Structural similarity was observed over a range of IPMDH structures in the PDB database. The most similar bacterial structures are summarised in Table 5.3. Given the relatively low sequence identity to *T. thermophilus* and *A. ferrooxidans* (56.6 % and 55.7 % respectively) there is a high level of structural correlation between ANC1 and these structures. ANC1 is more closely related to the IPMDH from the deep branching *A. ferrooxidans* than the contemporary *B. coagulans*. This similarity to deep branching IPMDH structures supports the ancestral nature of the ANC1 enzyme. With the exception of *C. jejuni*, the most similar structures are from thermophilic species, suggesting a commonality between the IPMDHs adapted to function at higher temperatures.

ANC1 and ANC4 are both structurally similar to *T. thermophilus*, *A. ferrooxidans*, and *B. coagulans*. RMSD values for the structural alignment of ANC4 with these structures are 1.04 Å, 1.22 Å and 1.44 Å respectively. ANC1 is more closely related to these contemporary enzymes than ANC4, as expected given the ages of the two ancestral enzymes.

Table 5.3: PDBeFold structural alignment of PDB structures closely related to ANC1.

Structure	PBD code	Q-score	P-score	Z-score	RMSD	N_{align}	Reference
ANC4	3U1H	0.81	46.6	20.5	0.85	317	(Hobbs et al. 2012)
T. thermophilus	2Y41	0.79	39.9	19.1	1.10	314	(Graczer et al. 2011a)
C. jejuni	3UDO	0.77	39.8	18.9	1.09	315	-
A. ferrooxidans	1A05	0.75	36.6	18.8	1.18	315	(Imada et al. 1998)
B. coagulans	1V53	0.73	38.0	18.6	1.25	313	-

Secondary structure matching (SSM) overlays for the two most similar structures to ANC1, ANC4 (PDB code 3U1H) and *T. thermophilus* (PDB code 2Y41), are illustrated in Figure 5.8 and Figure 5.9 respectively. Both structures have very few structural differences compared to ANC1. The major difference evident in both is the missing β loop from ANC1 that would be situated between β5 and α7 if the region had been able to be modelled. A smaller section of this loop is also missing in ANC4. In the *T. thermophilus* dimer, these loops align in an antiparallel manner between the two monomeric units of this dimer, creating another intermolecular contact. In the overlay of ANC1 and ANC4, domain 2 is very structurally similar. Differences between the structures include β1 and β2 which are angled in ANC1 so as to come into closer proximity to α1 than in ANC4. A 1-2 Å displacement in α16 is also evident between the two structures. The angle that β9 and β10 orientate in the central plane of β sheets differs, positioning the loop that links these two components differently between the two structures. Overall ANC1 an ANC4 are structurally very similar. No differences between the two structures were identified to rationalise the biochemical differences between the two enzymes. Nor was any commonality found between the two structures to explain the shared thermophilic nature of the two enzymes. This lack of thermophilic 'signature' between ANC1 and ANC4 has already been recognised from the amino acid compositions of the two enzymes (Hobbs et al. 2012). This suggests that thermophily has arisen twice, independently, over the evolution of IPMDH in *Bacillus*.

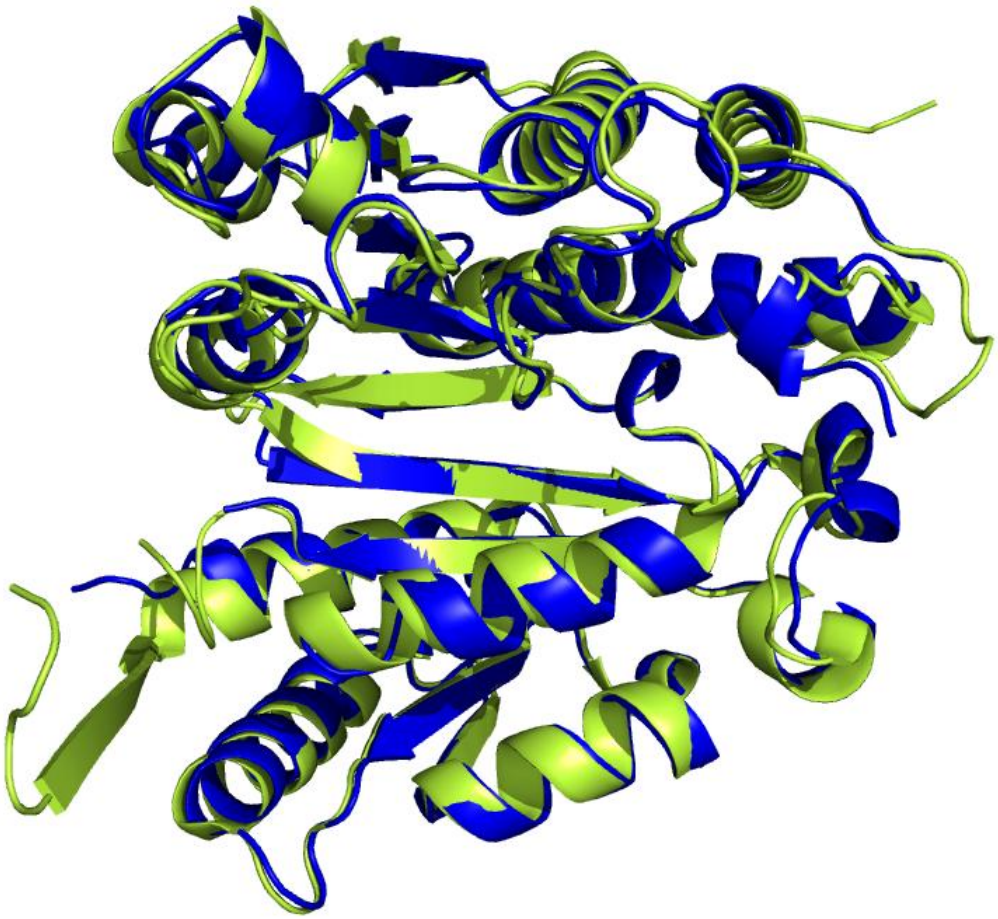


Figure 5.8: SSM overlay of ANC1 and ANC4 IPMDH structures.
The ANC1 monomer is coloured with blue, while the ANC4 monomer (PDB code 3U1H) is coloured green.

Domain 2 of ANC1 and *T. thermophilus* are also very structurally similar. In domain 1, α1 and α16 are each displaced by between 2-3 Å. The major difference between the two structures is in β9 and β10, which are considerably shorter in the *T. thermophilus* structure compared to ANC1 (β9 - 6.3 Å and 13.2 Å; β10 - 3.8 Å and 10.3 Å respectively). The differences between ANC1 and both structures, with the exception of the shortening of β9 and β10, are on the periphery of the structure. These regions are not likely to have significant effects on enzymatic catalysis, and differences may be the result of crystal packing.

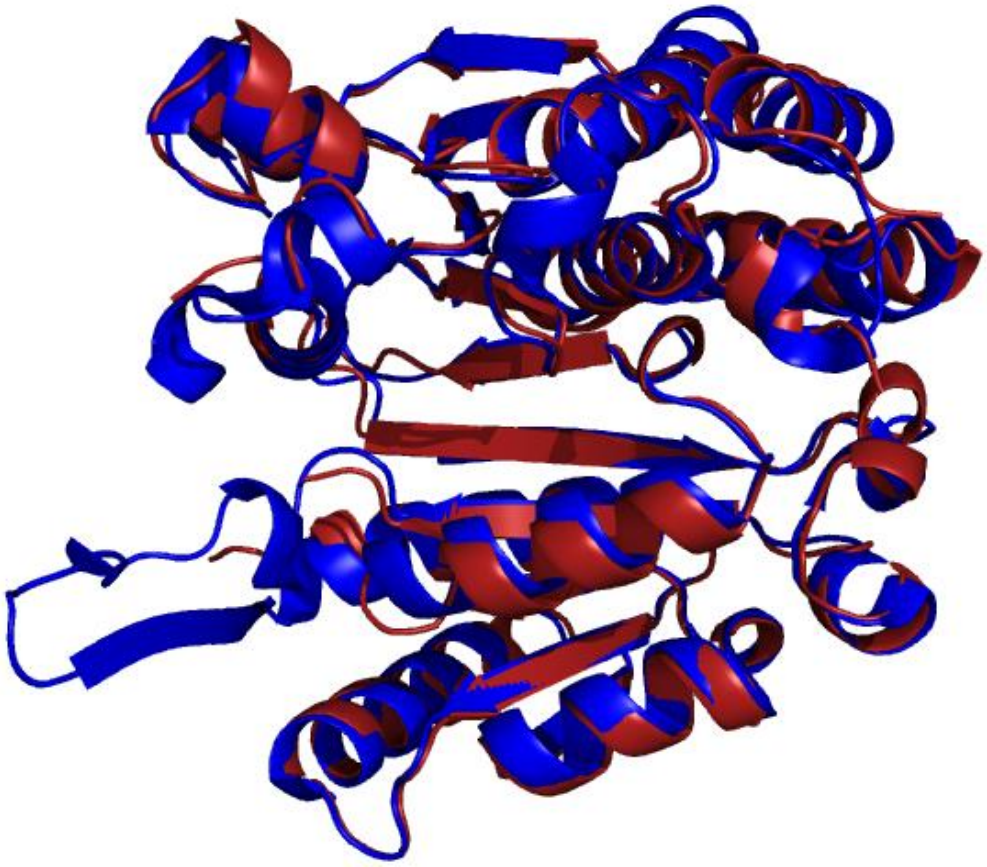


Figure 5.9: SSM overlay of ANC1 and *T. thermophilus* IPMDH structures. The ANC1 monomer is coloured with blue, while the *T. thermophilus* monomer (PDB code 2T41) is coloured purple.

In conclusion, the structure of ANC1 has been solved to 2.6 Å resolution. ANC1 was found to be structurally similar to previously solved contemporary IPMDHs, and ANC4, the LCA of *Bacillus*. ANC1 was more structurally similar to contemporary structures than ANC4, consistent with the relative ages of the two enzymes. No structural rational was found in a comparison of the two ancestral IPMDHs for the differences in biochemistry, or similarity in thermal activity of the two enzymes.

Discussion

ASR techniques involving a site homogeneous model of evolution and protein sequence only phylogenetic tree have been applied to the inference of IPMDH from the LCA of the Firmicutes. IPMDH has previously been successfully reconstructed back 950 myr to the LCA of *Bacillus* (Hobbs et al. 2012). Taking this reconstruction of IPMDH back further to 2.7 billion years ago to the LCA of the Firmicutes tested the limits of ASR techniques in terms of time and sequence divergence. In the ASR process, difficulties were encountered in the selection of species for inclusion in the ancestral inference, and in the consistency of phylogenetic trees produced. These issues were overcome by selecting species representative of major groups within each of the main genera of the Firmicutes and selecting a phylogenetic tree with consistent grouping of species into the correct genera. For the ancestral sequence inferences of the LCA of the Firmicutes, likelihood scores were slightly lower than optimal for the process (0.842 for the amino acid reconstruction, 0.815 for the nucleotide reconstruction and 0.693 for the codon reconstruction). However, given the time scale and sequence variation of the reconstruction, these likelihood scores were deemed adequate. The final LCA sequence shared between 47 % and 80 % sequence identity to contemporary and previously reconstructed IPMDH enzymes. This indicates that no one contemporary enzyme is disproportionately influencing the LCA sequence, and that there is a reasonable level of sequence conservation over the IPMDH enzymes in the study. The LCA gene was synthesised for biochemical characterisation. During purification of the LCA enzyme, the protein was found to be forming soluble aggregates with no measurable catalytic activity. From this, it was concluded that the ASR techniques employed were not sufficient to reconstruct over the large time scale given the sequence divergence and structural complexity of IPMDH.

Subsequently, the reconstruction of IPMDH from the LCA of the Firmicutes has been achieved (Groussin et al, unpublished). This was achieved by the use of a larger phylogenetic tree reconciled against a species tree based on 16S sequences,

with or without a more sophisticated model of evolution incorporating site heterogeneity into the model of evolution. It seems that to reconstruct over large time scales and levels of sequence divergence, more complex reconstruction methods are necessary, especially the use of a reconciled tree that combines information from both a species tree and a gene tree.

As the reconstruction of IPMDH could not be taken back to the LCA of the Firmicutes, ancestral enzymes from the *Bacillus* phylogeny were tested for activity on alternative substrates for IPMDH to assess how substrate promiscuity may have evolved in this enzyme over the last billion years. IPMDH appears to be evolutionarily related to ICD, malate dehydrogenase and tartrate dehydrogenase, and exhibits slow turnover of the substrates for these related dehydrogenases. Activity of ancestral IPMDHs was tested against these alternative substrates to characterise the evolution of substrate promiscuity in IPMDH. Due to issues encountered in measuring Michaelis-Menten kinetics with the alternative substrates, specific activity was measured instead. ANC1 and the contemporary BCVX IPMDH had similar rates of substrate turn over for the alternative substrates malate and tartrate. For ANC4, turnover rates of malate and tartrate were significantly faster than for ANC1 and BCVX as expected given the faster turnover of natural substrate by ANC4. However, ANC4 was tenfold slower than ANC1 or BCVX at turning over isocitrate. Given the age of the citrate cycle (Cunchillos & Lecointre 2003), isocitrate would likely have been an encountered at the time ANC4 was active. However, without more knowledge of the activity demands of ANC4 at the time it was active it is difficult to rationalise the slow rate of the enzyme with isocitrate compared to its younger counterparts. Overall, these results do not suggest that there is any increase in substrate promiscuity in ancestral IPMDH enzymes. Taken in conjunction with the results of Risso et al. (2013) and Wouters et al. (2003), who have shown increased substrate promiscuity in ancestral β-lactamases and that enzyme specificity over time is not linear in ancestral granzymes respectively, it seems that the evolution of substrate promiscuity varies between different enzymes and is not necessarily a linear process from generalist to specialist. It is also possible that this reconstruction back to the LCA of the *Bacillus* has not been taken far enough to see the original generalist enzyme that has subsequently specialised in IPM

oxidative decarboxylation. Promiscuity testing of more ancient ancestors, such as the LCA of the Firmicutes, may reveal this original generalist activity.

Previous characterisation of the ancestral IPMDHs from *Bacillus* (Hobbs et al. 2012) has shown that ANC1 and ANC4 are catalytically faster, and ANC4 is more kinetically stable, than their extant counterparts. To assess why these faster and more stable enzymes have been lost over evolutionary time, the effect the ancestral enzymes have on growth while functioning within the biosynthetic pathway for leucine in a cellular context was measured. IPMDH enzymes were expressed from pPROEX with IPTG induction in the Keio collection *E. coli leuB* KO strain (Baba et al. 2006). This system has an obvious draw back in the use of an overexpression vector, resulting in abnormally high intracellular levels of IPMDH. Ideally genes would be cloned into the genomic DNA to be expressed under control of the natural regulation system of the *leu* operon. This overexpression system was trialled as an initial test for growth rate differences between the ancestral and contemporary IPMDH enzymes. When grown on M9 agar, obvious differences in the extent of growth were observed. ANC1 and BCVX showed similar levels of growth, while ANC4 had considerably poorer growth. An attempt to fully quantify the growth rates in liquid M9 medium was made, however no significant differences in growth rates were able to be determined. However, the growth trend observed in plate growth has subsequently been corroborated using a more realistic expression system where IPMDH expression under the control of the *leu* operon promoter was measured (Hobbs et al, unpublished). This trend was likely obscured in the growth rate trials with IPMDH overexpression due to the artificially high levels of intracellular enzyme. Abnormal cellular compositions created by IPMDH overexpression potentially disrupt normal cellular functions and reduce the growth rate differences present between the IPMDH variants. The system is also not guaranteed to give even levels of IPMDH expression across the strains, despite equal addition of IPTG. The decreased growth rate of ANC4 seen in plate growth could be due to a number of effects. The greater rate of substrate turnover by ANC4 may deplete cellular concentrations of the cofactor NAD^+, affecting other metabolic pathways dependent on NAD^+. The rate of ANC4 may also result in the build up of product to levels which have adverse effects on the cell. It may

also be the kinetic stability of ANC4 which affects the *in vivo* fitness, disrupting enzyme turnover and thus the overall regulation of the pathway for leucine biosynthesis. The OD overshoot before dropping down to a stable OD that was observed in ANC4 growth but not ANC1 or BCVX also suggests a regulation issue in the cells. Although the exact cause of the decreased fitness for ANC4 is not known, the decreased *in vivo* fitness compared to contemporary counterparts clarifies why this enzyme has been lost over evolutionary time.

In vivo growth rates were also attempted in *E. coli* strains with the natural dehydrogenases of the alternative substrates tartrate and isocitrate knocked out to determine if IPMDH enzymes could complement these losses. Small amounts of growth were observed on M9 agar, however significant growth was not measurable for quantitation in liquid medium. Given the low rates of substrate turnover *in vitro* for the alternative substrates, this inability to complement the KO *in vivo* is not unexpected. This is especially significant given that the IPMDH enzymes in this system are overexpressed, and the higher than natural levels were still unable to achieve the substrate turnover necessary to complement the KO and support growth.

In the context of the *in vivo* and *in vitro* differences between ANC1 and ANC4, the X-ray crystal structure of ANC1 was determined to try structurally rationalise these differences. The structure of ANC1 shows close structural homology to previously solved IPMDH structures. Closest structural homology is to ANC4. Between the two structures, no differences were found to rationalise the activity differences observed between ANC1 and ANC4. ANC1 is structurally closer to contemporary IPMDH structures than ANC4, as expected given the respective ages of the two enzymes. ANC1 also shows the closest structural homology to IPMDHs from the thermophilic organisms *T. thermophilus*, *A. ferrooxidans*, and *B. coagulans*, suggesting a commonality between IPMDHs of thermophilic origin.

5.2 Future research

Following the successful reconstruction of IPMDH back to the LCA of the Firmicutes, this enzyme can now be characterised as ANC1 and ANC4 have been in this study in terms of substrate promiscuity and *in vivo* fitness. To further complete our understanding of IPMDH evolution, ANC2 and ANC3 from Hobbs et al. (2012) should also be characterised. X-ray crystal structures for these enzymes, as well as contemporary *Bacillus* IPMDHs, would also give a more complete structural understanding of the evolution of IPMDH.

Given the issues with the *in vivo* expression system described here, better systems are needed to more accurately measure the fitness effects of ancestral IPMDHs. Ideally this would involve allelic replacement of the IPMDH genes into a thermophilic *Bacillus* strain. This would better replicate the cellular conditions the enzymes were active within, the temperature that ANC1, ANC4, and BCVX are most active at and the natural regulation of the enzyme in the *leu* operon. *In vivo* evolution experiments could also complement this fitness work. This would involve following the changes in the *leuB* gene as well as the whole genome over successive generations as adaptive evolution takes place. This would allow any changes arising in ANC4 that increase fitness, to be characterised as the strain evolves.

References

Abascal F, Zardoya R, Posada D 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21(9): 2104-2105.

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW and others 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D 66(2): 213-221.

Adey NB, Tollefsbol TO, Sparks AB, Edgell MH, Hutchison CA, 3rd 1994. Molecular Resurrection of an Extinct Ancestral Promoter for Mouse L1. PNAS 91(4): 1569-1573.

Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD 2012. Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallographica Section D 68(4): 352-367.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215(3): 403-410.

Arenas M, Posada D 2010. The Effect of Recombination on the Reconstruction of Ancestral Sequences. Genetics 184(4): 1133-1139.

Arnold FH, Wintrode PL, Miyazaki K, Gershenson A 2001. How enzymes adapt: lessons from directed evolution. Trends in Biochemical Sciences 26(2): 100-106.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio Collection. Molecular Systems Biology 2(2006): 21-32.

Battistuzzi FU, Feijao A, Hedges SB 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol 4: 44-57.

Breidt F, Romick TL, Fleming HP 1994. A rapid method for the determination of bacterial growth kinetics. Journal of Rapid Methods & Automation in Microbiology 3(1): 59-68.

Bridgham JT, Ortlund EA, Thornton JW 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461: 515-519.

Bridgham JT, Eick GN, Larroux C, Deshpande K, Harms MJ, Gauthier MEA, Ortlund EA, Degnan BM, Thornton JW 2010. Protein Evolution by Molecular Tinkering: Diversification of the Nuclear Receptor Superfamily from a Ligand-Dependent Ancestor. PLoS Biology 8(10): e1000497.

Cai W, Pei J, Grishin NV 2004. Reconstruction of ancestral protein sequences and its applications. BMC Evolutionary Biology 4(33).

Calvo JM, Kalyanpur MG, Stevens CM 1962. 2-Isopropylmalate and 3-Isopropylmalate as Intermediates in Leucine Biosynthesis. Biochemistry 1(6): 1157-1161.

Camin JH, Sokal RR 1965. A Method for Deducing Branching Sequences in Phylogeny. Evolution 19(3): 311-326.

Carroll SM, Ortlund EA, Thornton JW 2011. Mechanisms for the Evolution of a Derived Function in the Ancestral Glucocorticoid Receptor. PLoS Genetics 7(6): e1002117.

Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, Husain A 1996. Angiotensin II-Forming Activity in a Reconstructed Ancestral Chymase. Science 271(5248): 502-505.

Chang BSW, Jönsson K, Kazmi MA, Donoghue MJ, Sakmar TP 2002. Recreating a functional ancestral archosaur visual pigment. Molecular Biology and Evolution 19(9): 1483-1489.

Chen R, Jeong S-S 2000. Functional prediction: Identification of protein orthologs and paralogs. Protein Science 9(12): 2344-2353.

Chen Y, Apolinario E, Brachova L, Kelman Z, Li Z, Nikolau BJ, Showman L, Sowers K, Orban J 2011. A nuclear magnetic resonance based approach to accurate functional annotation of putative enzymes in the methanogen *Methanosarcina acetivorans*. BMC Genomics 12(Suppl 1).

Chinen A, Matsumoto Y, Kawamura S 2005. Reconstitution of Ancestral Green Visual Pigments of Zebrafish and Molecular Mechanism of Their Spectral Differentiation. Molecular Biology and Evolution 22(4): 1001-1010.

Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia P, Cai J, Hippe H, Farrow JAE 1994a. The Phylogeny of the Genus *Clostridium*: Proposal of Five New Genera and Eleven New Species Combinations. International Journal of Systematic Bacteriology 44(4): 812-826.

Collins TM, Wimberger PH, Naylor GJP 1994b. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. Systematic Biology 43(4): 482-496.

Copley SD 2012. Toward a Systems Biology Perspective on Enzyme Evolution. Journal of Biological Chemistry 287(1): 3-10.

Cromwell ME, Hilario E, Jacobson F 2006. Protein aggregation and bioprocessing. AAPS 8(3): E572-579.

Cunchillos C, Lecointre G 2003. Evolution of Amino Acid Metabolism Inferred through Cladistic Analysis. Journal of Biological Chemistry 278(48): 47960-47970.

Dean AM, Dvorak L 1995. The role of glutamate 87 in the kinetic mechanism of *Thermus thermophilus* isopropylmalate dehydrogenase. Protein Science 4(10): 2156-2167.

Dean AM, Golding GB 1997. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proceedings of the National Academy of Sciences 94(7): 3104-3109.

Dean AM, Thornton JW 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nature Reviews: Genetics 8(9): 675-688.

Delano WL 2002. The PyMOL Molecular Graphics System. DeLano scientific

Depristo MA 2007. The subtle benefits of being promiscuous: adaptive evolution potentiated by enzyme promiscuity. HFSP Journal 1(2): 94-98.

Didelot X, Lawson D, Darling A, Falush D 2010. Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences. Genetics 186(4): 1435-1449.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S and others 2011. Geneious. Version 5.4.

Efron B, Halloran E, Holmes S 1996. Bootstrap confidence levels for phylogenetic trees. PNAS 93(23): 13429-13434.

Emsley P, Cowtan K 2004. Coot: model-building tools for molecular graphics. Acta Crystallographica Section D 60: 2126-2132.

Evans P 2006. Scaling and assessment of data quality. Acta Crystallographica Section D 62(1): 72-82.

Felsenstein J 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39(4): 783-791.

Field SF, Matz MV 2010. Retracing Evolution of Red Fluorescence in GFP-Like Proteins from Faviina Corals. Molecular Biology and Evolution 27(2): 225-233.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW 2012. Evolution of increased complexity in a molecular machine. Nature 481(7381): 360-364.

Fitch WM 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Zoology 20(4): 406-416.

Freundlich M, Burns RO, Umbarger HE 1962. Control if isoleucine, valine, and leucine biosynthesis, I. multi-valent repression. PNAS 48: 1804-1808.

Fujita M, Tamegai H, Eguchi T, Kakinuma K 2001. Novel substrate specificity of designer 3-isopropylmalate dehydrogenase derived from *Thermus thermophilus* HB8. Bioscience, Biotechnology, and Biochemistry 65(12): 2695-2700.

Galtier N, Gouy M 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Molecular Biology and Evolution 15(7): 871-879.

Gaucher EA, Govindarajan S, Ganesh OK 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451(7179): 704-707.

Gaucher EA, Thomson JM, Burgan MF, Benner SA 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425(6955): 285-288.

Gerlt JA, Babbitt PC 2001. Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. Annual Review of Biochemistry 70: 209-246.

Graczer E, Merli A, Singh RK, Karuppasamy M, Zavodszky P, Weiss MS, Vas M 2011a. Atomic level description of the domain closure in a dimeric enzyme: *Thermus thermophilus* 3-isopropylmalate dehydrogenase. Molecular BioSystems 7(5): 1646-1659.

Graczer E, Konarev PV, Szimler T, Bacso A, Bodonyi A, Svergun DI, Zavodszky P, Vas M 2011b. Essential role of the metal-ion in the IPM-assisted domain closure of 3-isopropylmalate dehydrogenase. FEBS Letters 585(20): 3297-3302.

Gross SR, Burns RO, Umbarger HE 1963. The Biosynthesis of Leucine. II. The Enzymatic Isomerisation of β-Carboxy-β-Hydroxyisocaproate and α-Hydroxy-β-Carboxyisocaproate. Biochemistry 2(4): 1046-1052.

Hall BG 2006. Simple and accurate estimation of ancestral protein sequences. PNAS 103(14): 5431-5436.

Hanson-Smith V, Kolaczkowski B, Thornton JW 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. Molecular Biology and Evolution 27(9): 1988-1999.

He Y, Galant A, Pang Q, Strul JM, Balogun SF, Jez JM, Chen S 2011. Structural and Functional Evolution of Isopropylmalate Dehydrogenases in the Leucine and Glucosinolate Pathways of *Arabidopsis thaliana*. Journal of Biological Chemistry 286(33): 28794-28801.

Hernandez-Montes G, Diaz-Mejia JJ, Perez-Rueda E, Segovia L 2008. The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. Genome Biology 9(6): 2008-2009.

Hillis DM, Bull JJ, White ME, Badgett MR, Molineux IJ 1992. Experimental Phylogenetics: Generation of a Known Phylogeny. Science 255(5044): 589-592.

Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, Arcus VL 2012. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of *Bacillus*. Molecular Biology and Evolution 29(2): 825-835.

Huelsenbeck JP, Bollback JP 2001. Empirical and Hierarchical Bayesian Estimation of Ancestral States. Systematic Biology 50(3): 351-366.

Hurley JH, Dean AM 1994. Structure of 3 isopropylmalate dehydrogenase in complex with NAD^+: ligand induced loop closing and mechanism for cofactor specificity. Structure 2(11): 1007-1016.

Imada K, Sato M, Tanaka N, Katsube Y, Matsuura Y, Oshima T 1991. Three-dimensional structure of a highly thermostable enzyme, 3-isopropylmalate dehydrogenase of *Thermus thermophilus* at 2.2 A resolution. Journal of Molecular Biology 222(3): 725-738.

Imada K, Inagaki K, Matsunami H, Kawaguchi H, Tanaka H, Tanaka N, Namba K 1998. Structure of 3-isopropylmalate dehydrogenase in complex with 3-isopropylmalate at 2.0 å resolution: the role of Glu88 in the unique substrate-recognition mechanism. Structure 6(8): 971-982.

Iwabata H, Watanabe K, Ohkuri T, Yokobori S-i, Yamagishi A 2005. Thermostability of ancestral mutants of *Caldococcus noboribetus* isocitrate dehydrogenase. FEMS Microbiology Letters 243(2): 393-398.

James LC, Tawfik DS 2001. Catalytic and binding poly-reactivities shared by two unrelated proteins: The potential role of promiscuity in enzyme evolution. Protein Science 10(12): 2600-2607.

Jensen RA 1976. Enzyme recruitment in evolution of new function. Annual Review of Microbiology 30: 409-425.

Jermann TM, Opltz JG, Stackhouse J, Benner SA 1995. Reconstructing the Evolutionary History of the Artiodactyl ribonuclease superfamily. Nature 374: 57-59.

Juhas M, Stark M, von Mering C, Lumjiaktase P, Crook DW, Valvano MA, Eberl L 2012. High Confidence Prediction of Essential Genes in *Burkholderia Cenocepacia*. PLoS ONE 7(6): e40064.

Kabir MM, Shimizu K 2004. Metabolic regulation analysis of *icd*-gene knockout *Escherichia coli* based on 2D electrophoresis with MALDI-TOF mass spectrometry and enzyme activity measurements. Applied Microbiology and Biotechnology 65(1): 84-96.

Kakinuma K, Ozawa K, Fujimoto Y, Akutsu N, Oshima T 1989. Stereochemistry of the decarboxylation reaction catalysed by 3-isopropylmalate dehydrogenase from the thermophilic bacterium *Thermus thermophilus*. Journal of the Chemical Society, Chemical Communications(17): 1190-1192.

Khersonsky O, Tawfik DS 2010. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annual Review of Biochemistry 79(1): 471-505.

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. Molecular Biology and Evolution 24(11): 2504-2514.

Krishnan NM, Seligmann H, Stewart C-B, de Koning APJ, Pollock DD 2004. Ancestral Sequence Reconstruction in Primate Mitochondrial DNA: Compositional Bias and Effect on Functional Inference. Molecular Biology and Evolution 21(10): 1871-1883.

Krissinel E, Henrick K 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D 60: 2256-2268.

Krissinel E, Henrick K 2007. Inference of macromolecular assemblies from crystalline state. Journal of Molecular Biology 372(3): 774-797.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R and others 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21): 2947-2948.

Le SQ, Gascuel O 2008. An Improved General Amino Acid Replacement Matrix. Molecular Biology and Evolution 25(7): 1307-1320.

Leavitt RI, Umbarger HE 1961. Isoleucine and Valine Metabolism in *Escherichia coli* : X. THE ENZYMATIC FORMATION OF ACETOHYDROXYBUTYRATE. Journal of Biological Chemistry 236(9): 2486-2491.

Leslie AW, Powell H 2007. Processing diffraction data with mosflm. Evolving Methods for Macromolecular Crystallography, Springer Netherlands. Pp. 41-51.

Li G, Steel M, Zhang L 2008. More Taxa Are Not Necessarily Better for the Reconstruction of Ancestral Character States. Systematic Biology 57(4): 647-653.

Li Y, Suino K, Daugherty J, Xu HE 2005. Structural and Biochemical Mechanisms for the Specificity of Hormone Binding and Coactivator Assembly by Mineralocorticoid Receptor. Molecular Cell 19(3): 367-380.

Liberles DA 2007. Ancestral Sequence Reconstruction. 1 ed. USA, Oxford University Press.

Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC 1990. Ancestral Lysozymes Reconstructed, Neutrality Tested, and Thermostability Linked to Hydrocarbon Packing. Nature 345(6270): 86-89.

Matsunami H, Kawaguchi H, Inagaki K, Eguchi T, Kakinuma K, Tanaka H 1998. Overproduction and substrate specificity of 3-isopropylmalate dehydrogenase from *Thiobacillus ferrooxidans*. Bioscience, Biotechnology, and Biochemistry 62(2): 372-373.

Matthews BW 1968. Solvent content of protein crystals. Journal of Molecular Biology 33(2): 491-497.

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ 2007. Phaser crystallographic software. Journal of Applied Crystallography 40(4): 658-674.

Miyazaki J, Nakaya S, Suzuki T, Tamakoshi M, Oshima T, Yamagishi A 2001. Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: experimental evidence supporting the thermophilic common ancestor hypothesis. Journal of Biochemistry 129(5): 777-782.

Miyazaki K, Kakinuma K, Terasawa H, Oshima T 1993. Kinetic analysis on the substrate specificity of 3-isopropylmalate dehydrogenase. FEBS Letters 332(1–2): 35-36.

Nango E, Yamamoto T, Kumasaka T, Eguchi T 2009. Crystal structure of 3-isopropylmalate dehydrogenase in complex with NAD^+ and a designed inhibitor. Bioorganic & Medicinal Chemistry 17(22): 7789-7794.

O'Brien PJ, Herschlag D 1999. Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry & Biology 6(4): R91-R105.

Oakley TH, Cunningham CW 2000. Independent Contrasts Succeed Where Ancestor Reconstruction Fails in a Known Bacteriophage Phylogeny. Evolution 54(2): 397-405.

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J and others 1992. The alpha/beta hydrolase fold. Protein Engineering 5(3): 197-211.

Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW 2007. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. Science 317(5844): 1544-1548.

Page RD 1996. TreeView: an application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12(4): 357-358.

Pagel M, Meade A, Barker D 2004. Bayesian Estimation of Ancestral Character States on Phylogenies. Systematic Biology 53(5): 673-684.

Pauling L, Zuckerkandl E 1963. Chemical paleogenetics molecular "restoration studies" of extinct forms of life. Acta Chemica Scandinavica 17: S9-S16.

Perez-Jimenez P, Inglés-Prieto A, Zhao Z, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A and others 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nature Structural & Molecular Biology 18: 592-596.

Pirrung MC, Han H, Nunn DS 1994. Kinetic Mechanism and Reaction Pathway of *Thermus thermophilus* Isopropylmalate Dehydrogenase. The Journal of Organic Chemistry 59(9): 2423-2429.

Posada D 2008. jModelTest: phylogenetic model averaging, Molecular Biology and Evolution

Powell JT, Morrison JF 1978. Role of the *Escherichia coli* aromatic amino acid aminotransferase in leucine biosynthesis. Journal of Bacteriology 136(1): 1-4.

Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO 2006. Systems approach to refining genome annotation. PNAS 103(46): 17480-17484.

Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM 2013. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β-Lactamases. Journal of the American Chemical Society 135(8): 2899-2902.

Salisbury BA, Kim J 2001. Ancestral state estimation and taxon sampling density. Systematic Biology 50(4): 557-564.

Sanson GFO, Kawashita SY, Brunstein A, Briones MRS 2002. Experimental Phylogeny of Neutrally Evolving DNA Sequences Generated by a Bifurcate Series of Nested Polymerase Chain Reactions. Molecular Biology and Evolution 19(2): 170-178.

Schleifer KH, Kraus J, Dvorak C, Kilpper-Bälz R, Collins MD, Fischer W 1985. Transfer of *Streptococcus lactis* and Related Streptococci to the Genus *Lactococcus* gen. nov. Systematic and Applied Microbiology 6(2): 183-195.

Schultz TR, Churchill GA 1999. The Role of Subjectivity in Reconstructing Ancestral Character States: A Bayesian Approach to Unknown Rates, States, and Transformation Asymmetries. Systematic Biology 48(3): 651-664.

Shi Y, Yokoyama S 2003. Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. PNAS 100(14): 8308-8313.

Singh BK, Shaner DL 1995. Biosynthesis of Branched Chain Amino Acids: From Test Tube to Field. Plant Cell 7(7): 935-944.

Singh RK, Kefala G, Janowski R, Mueller-Dieckmann C, von Kries J-P, Weiss MS 2005. The High-resolution Structure of LeuB (Rv2995c) from *Mycobacterium tuberculosis*. Journal of Molecular Biology 346(1): 1-11.

Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA 1990. The ribonuclease from an extinct bovid ruminant. FEBS Letters 262(1): 104-106.

Stieglitz BI, Calvo JM 1974. Distribution of the Isopropylmalate Pathway to Leucine Among Diverse Bacteria. Journal of Bacteriology 118(3): 935-941.

Tavaré S 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. American Mathematical Society: Lectures on Mathematics in the Life Sciences, Amer Mathematical Society. Pp. 57-86.

Taylor WR, Jones DT 1993. Deriving an amino acid distance matrix. Journal of Theoretical Biology 164(1): 65-83.

Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung L-W, Read RJ, Adams PD 2008. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallographica Section D 64(1): 61-69.

Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV and others 2000. Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry. European Journal of Biochemistry 267(17): 5313-5329.

Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nature Genetics 37(6): 630-635.

Thornton JW, Need E, Crews D 2003. Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling. Science 301(5640): 1714-1717.

Tipton PA, Beecher BS 1994. Tartrate dehydrogenase, a new member of the family of metal-dependent decarboxylating R-hydroxyacid dehydrogenases. Archives of Biochemistry and Biophysics 313(1): 15-21.

Tsuchiya D, Sekiguchi T, Takenaka A 1997. Crystal Structure of 3-Isopropylmalate Dehydrogenase from the Moderate Facultative Thermophile, *Bacillus coagulans*: Two Strategies for Thermostabilization of Protein Structures. Journal of Biochemistry 122(6): 1092-1104.

Ugalde JA, Chang BSW, Matz MV 2004. Evolution of coral pigments recreated. Science 305(5689): 1433-1433.

van der Rest ME, Frank C, Molenaar D 2000. Functions of the membrane-associated and cytoplasmic malate dehydrogenases in the citric acid cycle of *Escherichia coli*. Journal of Bacteriology 182(24): 6892-6899.

van Eunen K, Kiewiet JAL, Westerhoff HV, Bakker BM 2012. *Testing Biochemistry* Revisited: How *In Vivo* Metabolism Can Be Understood from *In Vitro* Enzyme Kinetics. PLoS Computational Biology 8(4): e1002483.

Vitkup D, Kharchenko P, Wagner A 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biology 7(5): R39.

Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, Verstrepen KJ 2012. Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. PLoS Biology 10(12): 1-17.

Wallon G, Yamamoto K, Kirino H, Yamagishi A 1997a. Purification, catalytic properties and thermostability of 3-isopropylmalate dehydrogenase from *Escherichia coli*. Biochimica et Biophysica Acta 1337: 105-112.

Wallon G, Kryger G, Lovett ST, Oshima T, Ringe D, Petsko GA 1997b. Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-isopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus*. Journal of Molecular Biology 266(5): 1016-1031.

Ward JB, Jr., Zahler SA 1973. Regulation of leucine biosynthesis in *Bacillus subtilis*. Journal of Bacteriology 116(2): 727-735.

Williams PD, Pollock DD, Blackburne BP, Goldstein RA 2006. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. PLoS Computational Biology 2(6): e69.

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A and others 2011. Overview of the CCP4 suite and current developments. Acta Crystallographica 67(4): 235-242.

Wouters MA, Liu K, Riek P, Husain A 2003. A Despecialization Step Underlying Evolution of a Family of Serine Proteases. Molecular cell 12(2): 343-354.

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

Yang Z 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24(8): 1586-1591.

Yang Z, Kumar S, Nei M 1995. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. Genetics 141: 1641-1650.

Yokoyama S, Yang H, Starmer WT 2008a. Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. Genetics 179(4): 2037-2043.

Yokoyama S, Tada T, Zhang H, Britt L 2008b. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. PNAS 105(36): 13480-13485.

Zhang J 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18(6): 292-298.

Zhang J, Nei M 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. Journal of Molecular Evolution 44(1): S139-S146.

Zhang J, Rosenberg HF 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. PNAS 99(8): 5486-5491.

Zwickl DJ 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Unpublished PhD dissertation thesis, The University of Texas at Austin.

Appendices

Appendix A: Bacterial Strains and Plasmids

A1: Plasmids and cell strains

Table A.1: List of plasmids used in study.

Plasmid	**Description**
pPROEX HTb	*E. coli* expression vector with trc promotor, encoding a N-terminal His-tag: Amp^r

Table A.2: List of *E. coli* cell strains used in study.

Cell strain	**Description**
DH5α	F– Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rK–, mK+) *pho*A *sup*E44 λ– *thi*-1*gyr*A96 *rel*A1
Keio collection – parent strain: *E. coli* BW25141	(*rrnB3* DE*lacZ4787* DE*phoBR580hsdR514* DE(*araBAD*)*567DE(rhaBAD)568galU95 DEendA9::FRT DEuidA3::pir(wt) recA1 rph-1*
E. coli BW25141 – *leuB* KO	(*rrnB3* DE*lacZ4787* DE*phoBR580hsdR514* DE(*araBAD*)*567DE(rhaBAD)568galU95 DEendA9::FRT DEuidA3::pir(wt) recA1 rph-1*Δ*leuB*
E. coli BW25141 – *icd* KO	(*rrnB3* DE*lacZ4787* DE*phoBR580hsdR514* DE(*araBAD*)*567DE(rhaBAD)568galU95 DEendA9::FRT DEuidA3::pir(wt) recA1 rph-1* Δ*icd*
E. coli BW25141 – *yeaU* KO	(*rrnB3* DE*lacZ4787* DE*phoBR580hsdR514* DE(*araBAD*)*567DE(rhaBAD)568galU95 DEendA9::FRT DEuidA3::pir(wt) recA1 rph-1* Δ*yeaU*
E. coli BW25141 – *mdh* KO	(*rrnB3* DE*lacZ4787* DE*phoBR580hsdR514* DE(*araBAD*)*567DE(rhaBAD)568galU95 DEendA9::FRT DEuidA3::pir(wt) recA1 rph-1* Δ*mdh*

A2: Genetically modified organisms used in study

Table A.3: List of transformant cell lines used in this study.

Cell strain	**Plasmid**
E. coli DH5α	pPROEX-*BCVX*
	pPROEX-*ANC1*
	pPROEX-*ANC4*
E. coli BW25141	pPROEX
E. coli BW25141 – *leuB* KO	pPROEX
	pPROEX-*BCVX*
	pPROEX-*ANC1*
	pPROEX-*ANC4*
E. coli BW25141 – *icd* KO	pPROEX
	pPROEX-*BCVX*
	pPROEX-*ANC1*
	pPROEX-*ANC4*
E. coli BW25141 – *yeaU* KO	pPROEX
	pPROEX-*BCVX*
	pPROEX-*ANC1*
	pPROEX-*ANC4*
E. coli BW25141 – *mdh* KO	pPROEX
	pPROEX-*BCVX*
	pPROEX-*ANC1*
	pPROEX-*ANC4*

Appendix B: Accession Numbers and Gene Sequences

B1: Accession numbers of sequences used in study

Table A.4: Bacterial strains and IPMDH accession numbers.

Organism	Strain/serovar	Accession number/source
B. amyloliquefaciens	FZB42	NC_009725
B. anthracis	Ames	AE016879
B. caldotenax	Unknown	X04762
B. caldovelox	DSM 411	HQ625361
B. cereus	ATCC 14579	AE017002
B. clausii	KSM-K16	NC_006582
B. coagulans	ATCC 7051	M33099
B. cytotoxicus	NVH 391-98	NC_009674
B. halodurans	C-125	AP0001517
B. licheniformis	ATCC 14580	NC_006322
B. megaterium	DSM 319	X65184
B. pseudomycoides	DSM 12442	NZ_ACMX01000029
B. psychrophilus	DSM 3	HQ625362
B. psychrosaccharolyticus	DSM 6	HQ625363
B. pumilus	SAFR-032	NC_009848
B. stearothermophilus	DSM 22	http://www.genome.ou.edu/bstearo.html (contig 446)

B. subtilis	168	ZP_03592614
B. thuringiensis	Kurstaki T03a001	ZP_04113977
B. weihenstephanesis	KBAB4	CP000903
C. acetobutylicum	ATCC 824	NC_003030
C. asparagiforme	DSM 15981	NZ_ACCJ01000143
C. bartlettii	DSM 16795	NZ_ABEZ02000017
C. beijerinckii	NCIMB 8052	NC_009617
C. bolteae	ATCC BAA-613	NZ_ABCC02000023
C. botulinum	E3 Alaska E43	NC_010723
C. butyricum	5521	NZ_ABDT01000094
C. carboxidivorans	P7	NZ_ACVI01000012
C. cellulolyticum	H10	NC_011898
C. celulovorans	743B	NC_014393
C. citroniae	WAL-17108	NZ_ADLJ01000014
C. clariflavum	DSM 19732	NC_016627
C. clostridioforme	2_1_49FAA	NZ_ADLL01000027
C. difficile	002-P50-2011	AGAA01000028
C. kluyveri	DSM 555	NC_009706
C. lentocellum	DSM 5427	NC_015275
C. leptum	DSM 753	NZ_ABCB02000020
C. ljungdahlii	DSM 13528	NC_014328
C. papyrosolvens	DSM 2782	NZ_ACXX02000006
C. phytofermentans	ISDg	NC_010001
C. ramosum	DSM 1402	NZ_ABFX02000013
C. saccharolyticum	WM1	NC_014376

C. symbiosum	WAL-14163	NZ_ADLQ01000063
C. thermocellum	ATCC 27405	NC_009012
L. lactis	KF147	NC_013656
L. grayi	DSM 20601	NZ_ACCR02000003
L. innocua	Clip11262	AL596171
L. ivanovii	PAM 55	NC_016011
L. monocytogenes	EGD-e	AL591981
L. seeligeri serovar	1/2b SLCC3954	NC_013891
L. welshimeri serovar	6b SLCC5334	NC_008555
Staph. aureus	H116	DQ413537
Staph. capitis	SK14	NZ_ACFR01000019
Staph. caprae	C87	NZ_GL545274
Staph. epidermidis	M23864:W1	NZ_ACJB01000008
Staph. haemolyticus	JCSC1435	NC_007168
Staph. hominis	C80	NZ_GL545258
Staph. lugdunensis	VCU139	AHLK01000030
Staph. pettenkoferi	VCU012	AGUA01000088
Staph. saprophyticus	ATCC 15305	NC_007350
Staph. simiae	CCM 7213	NZ_AEUN01000493
Staph. warneri	L37603	NZ_ACPZ01000065
Strep. australis	ATCC 700641	NZ_AEQR01000019
Strep. criceti	HS-6	NZ_AEUV02000002
Strep. cristatus	ATCC 51100	NZ_AEVC01000007
Strep. downei	F0415	NZ_AEKN01000006
Strep. equinus	ATCC 9812	NZ_AEVB01000006

Strep. gallolyticus	UCN34	NC_013798
Strep. gordonii	CH1	NC_009785
Strep. infantis	X	AFUQ01000003
Strep. macacae	NCTC 11558	NZ_AEUW02000001
Strep. macedonicus	ACA-DC 198	NC_016749
Strep. mitis	SK597	NZ_AEDV01000044
Strep. mutans	UA159	NC_004350
Strep. oralis	ATCC 35037	NZ_ADMV01000020
Strep. parasanguinis	SK236	AFUC01000006
Strep. pasteurianus	ATCC 43144	NC_015600
Strep. pneumoniae	D39	NC_008533
Strep. salivarius	CCHSS3	NC_015760
Strep. sanguinis	ATCC 49296	NZ_AEPO01000014
Strep. suis	R61	AEYY01000041
Strep. thermophilus	LMD-9	NC_008532
Strep. vestibularis	F0396	NZ_AEKO01000005
T. ethanolicus	JW 200	NZ_AEYS01000010
T. italicus	Ab9	NC_013921
T. pseudethanolicus	ATCC 33223	NC_010321
T. tengcongensis	MB4	NC_003869
T. thermosaccharolyticum	DSM 571	NC_014410
T. wiegelii	Rt8.B1	NC_015958
T. xylanolyticum	LX-11	NC_015555

B2: LCA sequence information

The LCA protein sequence was codon optimised for expression in *E. coli*. The sequence also included an N-terminal hexa-histidine tag for purification purposes.

LCA IPMDH gene sequence optimised for expression in *E. coli*

GGATCCATGAAAATGAAAATTGCCGTTATTCCGGGTGATGGTATTGGTCCGGAAA
TTATTGAAGAAGCCATCAAAGTTCTGAATGCCGTGGCAGAAAAATATGGCCTGAA
ATTCGAATATAAAGAGGTTCTGCTGGGAGGTTGTGCAATTGATGAAACCGGTGTT
CCGCTGCCGGAAGAAACCGTTGAAGTTTGTAAAAAAAGTGATGCAGTGCTGCTGG
GTGCAGTTGGTGGTCCGAAATGGGATAATCTGCCGAGCAATAAACGTCCGGAAGC
AGGTCTGCTGGGTATTCGTAAAGGTCTGGGTGTTTATGCAAATCTGCGTCCGGCA
ATTCTGTATCCGGCACTGAAAAGCGCAAGTCCGCTGAAACCGGAAATTCTGGAAG
GTATTGATATTATGGTTGTGCGTGAACTGACCGGTGGTATCTATTTTGGTGAACG
TGGTCGCATTGATATCGGTGGTAAAAAAGCATATGATACCGAGATCTATACCACC
TTTGAAATTGAACGTATTGCCCGTAAAGCATTTGAAGCAGCACGTAAACGTAACA
AAAAACTGACCAGCGTTGATAAAGCCAATGTTCTGGAAAGCAGCCGTCTGTGGCG
TGAAGTTGTTGAAGAAGTTGCAAAAGAATATCCGGATGTGGAACTGAACTATATG
TATGTTGATAATGCAAGCATGCAGCTGATTCGTGATCCGAAACAGTTTGATGTTA
TTGTGACCAGCAATATGTTTGGCGATATTCTGACCGATGAAGCGAGCATGCTGAC
CGGTTCAATTGGTATGCTGCCGAGCGCAAGCCTGCGTGGTGATGGTCCGGGTCTG
TATGAACCGGTGCATGGTAGCGCACCGGATATTGCAGGCCAGAATAAAGCAAATC
CGATTGCAACCATTATGAGCGTTGCAATGATGCTGAAATATAGCTTCGATATGGA
AGAGGCAGCCGATGATATCAAAAATGCCGTTGAAAAAGTTCTGGAAGAGGGTTAT
CGTACCGGTGATATTGCAATTGAAGGCACCAAAATTGTTGGCACCGAAGAAATGG
GTGATCTGATTGTGGAAGATCTGGAAAAAATCTAATAACTGCAG

LCA IPMDH protein sequence

MKMKIAVIPGDGIGPEIIEEAIKVLNAVAEKYGLKFEYKEVLLGGCAIDETGVPL
PEETVEVCKKSDAVLLGAVGGPKWDNLPSNKRPEAGLLGIRKGLGVYANLRPAIL
YPALKSASPLKPEILEGIDIMVVRELTGGIYFGERGRIDIGGKKAYDTEIYTTFE
IERIARKAFEAARKRNKKLTSVDKANVLESSRLWREVVEEVAKEYPDVELNYMYV
DNASMQLIRDPKQFDVIVTSNMFGDILTDEASMLTGSIGMLPSASLRGDGPSLYE
PVHGSAPDIAGQNKANPIATIMSVAMMLKYSFDMEEAADDIKNAVEKVLEEGYRT
GDIAIEGTKIVGTEEMGDLIVEDLEKI

Appendix C: Reagents, Buffers, Gels and Growth Media

C1: Reagents and buffers

10 x DNA loading dye	0.4 % w/v bromophenol blue, 0.4 % w/v xylene, 50 % v/v glycerol
Assay buffer	20 mM potassium phosphate buffer, pH 7.6, 0.3 M KCl, 0.2 mM $MnCl_2$
Coomassie stain	0.05 % w/v coomassie blue R-250, 25 % v/v isopropanol, 10% v/v acetic acid
Elution buffer	50 mM sodium phosphate buffer, pH 8.0, 300 mM NaCl, 1 M imidazole
Lysis buffer	50 mM sodium phosphate buffer, pH 8.0, 300 mM NaCl, 50 mM imidazole
SE buffer	20 mM potassium phosphate buffer, pH 7.6
TAE buffer	40 mM Tris-acetate, 2 mM EDTA.

C2: Growth media

LB	10 g/L bacto tryptone, 5 g/L yeast extract, 10 g/L NaCl, pH 7.0
LB agar	10 g/L bacto tryptone, 5 g/L yeast extract, 10 g/L NaCl, 15 g/L agar**,** pH 7.0

M9 minimal media	5 x M9 salts:	64 g/L $Na_2HPO_4.7H20$
		15 g/L KH_2PO_4
		2.5 g/L NaCl
		5.0 g/L NH_4Cl
	M9 media:	200 ml/L 5x M9 salts
		2 ml/L 1 M $MgSO_4$
		20 ml/L carbon source (0.2 µm filter sterilised)
		0.1 ml/L $CaCl_2$
		780 ml/L H_2O
M9 agar	200 ml/L 5x M9 salts	
	2 ml/L 1 M $MgSO_4$	
	20 ml/L carbon source (0.2 µm filter sterilised)	
	0.1 ml/L $CaCl_2$	
	780 ml/L H_2O with 15 g agar	
TB	Solution A	12 g tryptone
		24 g yeast extract
		4 ml glycerol
		Made up to 900 ml with H_2O
	Solution B	2.31 g KH_2PO_4
		12.54 g K_2HPO_4
		Made up to 100 ml with H_20

C3: Gels

Agarose gel 1 % w/v agarose, in TAE buffer. 2-5 µL SYBR safe DNA gel stain (Invitrogen, USA) per gel.

SDS-PAGE gel

Table A.5: SDS-PAGE gel composition

	Stacker (ml)	12 % gel (ml)
MQ H_2O	8.5	10.05
Stacking buffer	1.6	-
Resolving buffer	-	7.5
30 % acrylamide	2.125	12
10 % SDS	0.125	0.3
10 % APS *	0.063	0.15
TEMED **	0.0063	0.015

* APS (ammonium persulphate)

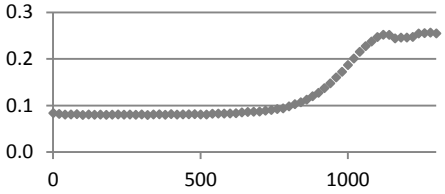
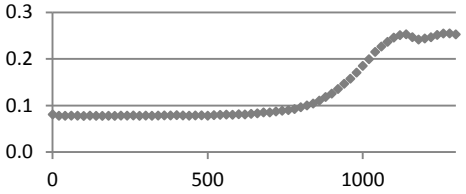
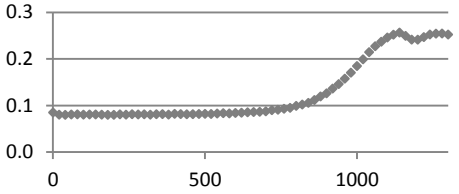
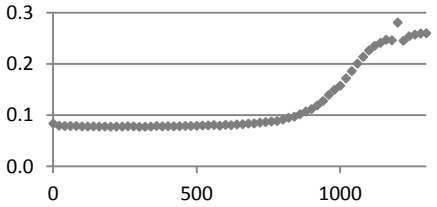
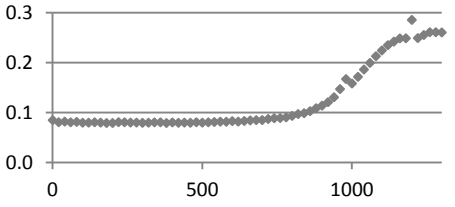
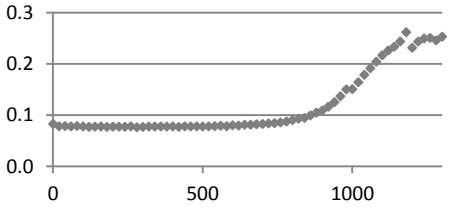
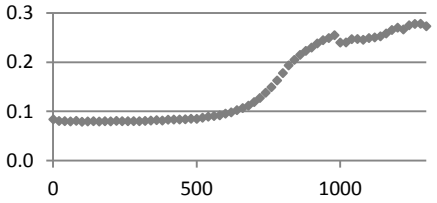
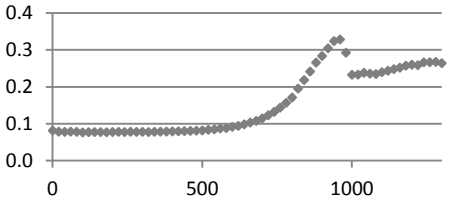
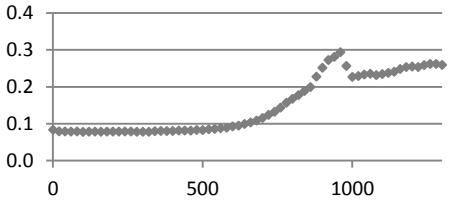
** TEMED (N, N, N', N'-tetramethylethylenediamine)

Stacking buffer 1.0 M Tris-HCl, pH 6.8

Resolving buffer 1.5 M Tris-HCl, pH 8.8

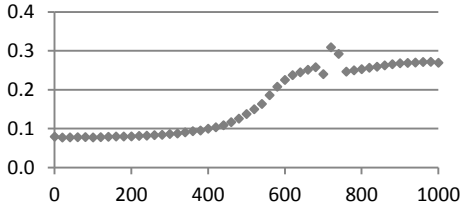
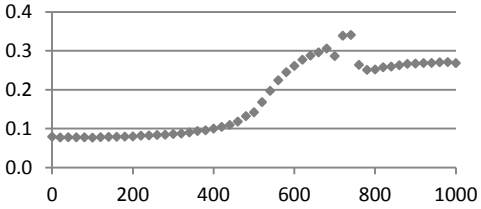
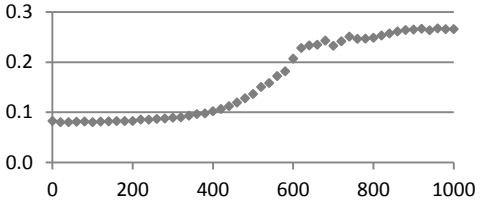
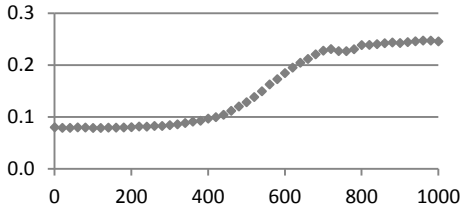
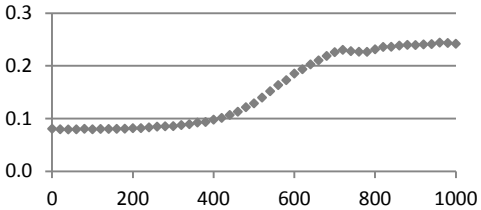
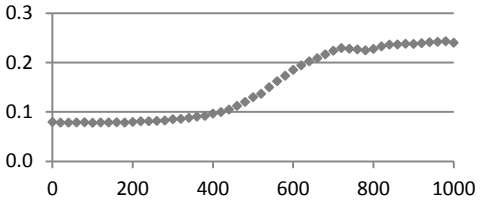
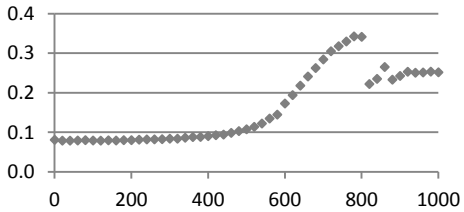
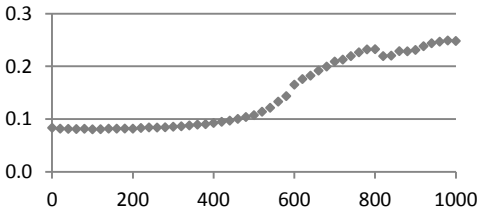
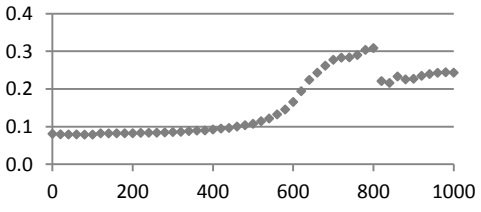
Appendix D: Growth curves for complemented *leuB* KOs

Table A.6: Growth curves of *E. coli leuB* KOs complemented with BCVX IPMDH

Internal replicate	Biological replicate one	Biological replicate two	Biological replicate three
1			
2			
3			

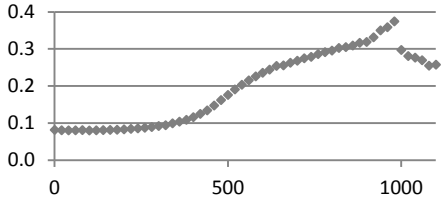
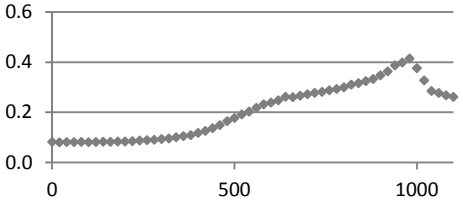
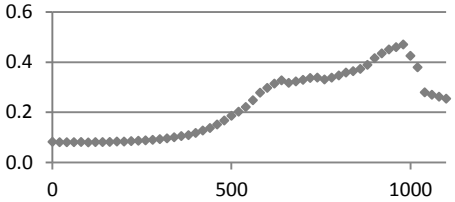
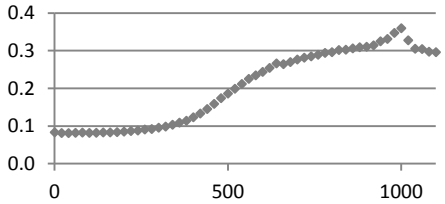
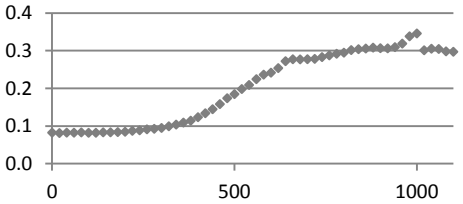
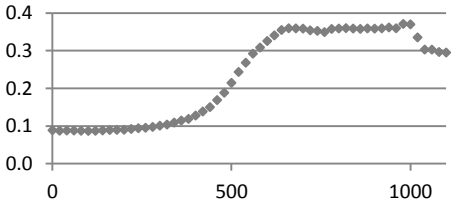
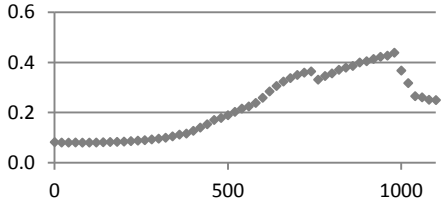
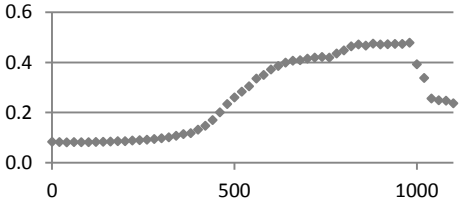
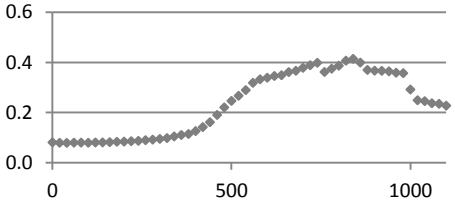
Axes are growth time (minutes) and OD (600 nm) for the horizontal and vertical axes respectively.

Table A.7: Growth curves of *E. coli leuB* KOs complemented with ANC1 IPMDH

Internal replicate	Biological replicate one	Biological replicate two	Biological replicate three
1			
2			
3			

Axes are growth time (minutes) and OD (600 nm) for the horizontal and vertical axes respectively.

Table A.8: Growth curves of *E. coli leuB* KOs complemented with ANC4 IPMDH

Internal replicate	Biological replicate one	Biological replicate two	Biological replicate three
1			
2			
3			

Axes are growth time (minutes) and OD (600 nm) for the horizontal and vertical axes respectively.