Creation and evaluation of a pyruvate decarboxylase dependent ethanol fermentation pathway in Geobacillus thermoglucosidasius

Buddrus, Lisa

Award date:
2017

Awarding institution:
University of Bath

Creation and evaluation of a pyruvate decarboxylase dependent ethanol fermentation pathway in *Geobacillus thermoglucosidasius*

Lisa Buddrus

A thesis submitted for the degree of Doctor of Philosophy

University of Bath
Department of Biology and Biochemistry

August 2016

ABSTRACT

Bioethanol, produced from organic waste as a second-generation biofuel, is an important renewable energy source. Here, recalcitrant carbohydrate sources, such as municipal and agricultural waste, and plants grown on land not suitable for food crops, are exploited.

The thermophilic, Gram-positive bacterium *Geobacillus thermoglucosidasius* is naturally very flexible in its growth substrates and produces a variety of fermentation products, including lactate, formate, acetate and ethanol. TMO Renewables Ltd. used metabolic engineering to enhance ethanol production, creating the production strain TM242 (NCIMB 11955 Δ*ldh*, Δ*pfl*, pdh^{up}). Ethanol yield has been increased to 82% of the theoretical maximum on glucose and up to 92% of the theoretical maximum on cellobiose. However, this strain still produces acetate, presumably derived from the overproduction of acetyl-CoA through the upregulated *pdh* gene encoding the pyruvate dehydrogenase complex.

An alternative to the mixed fermentation pathway found in *G. thermoglucosidasius* is to introduce a homoethanologenic pathway. Yeast and a very limited range of mesophilic bacteria use the homoethanol fermentation pathway, which employs pyruvate decarboxylase (PDC) in conjunction with alcohol dehydrogenase (ADH), to convert pyruvate to ethanol. Despite extensive screening, no PDC has yet been identified in a thermophilic organism.

Using the thermophile *G. thermoglucosidasius* as a host platform, we endeavoured to develop a thermophilic version of the homoethanol pathway for use in *Geobacillus* spp.

This Thesis reports the *in vitro* characterization and crystal structure of one of the most thermostable bacterial PDCs from the mesophile *Zymobacter palmae* (ZpPDC) and describes strategies to improve expression of active PDC at high growth temperatures. This includes codon harmonization and the successful development of a PET (producer of ethanol) operon. Furthermore, ancestral sequence reconstruction was explored as an alternative engineering approach, but did not yield a PDC more thermostable than ZpPDC.

In vitro ZpPDC is most active at 65°C with a denaturation temperature of 70°C, when sourced from a recombinant mesophilic host. Codon harmonization improved detectable PDC activity in *G. thermoglucosidasius* cultures grown up to 65°C by up to 42%. Pairing this PDC with *G. thermoglucosidasius* ADH6 produced a PET functional up to 65°C with ethanol yields of 87% of the theoretical maximum on glucose. This increase in yield at temperatures of up to 15°C higher than previously reported for any PDC expressed in a thermophilic host could make a significant difference for industrial-scale production.

Wohin ich auch gehe, ich werde niemals vergessen
wer mir half dorthin zu kommen.

ACKNOWLEDGEMENTS

Firstly, I would like to thank the sponsors of this KTN Biosciences coordinated BBSRC CASE studentship together with TMO Renewables Ltd., Guildford, UK, and my supervisors, David J. Leak, Michael J. Danson and Kirstin Eley, for making this project possible and for their invaluable guidance and support.

Furthermore, I am very grateful to all members – past and present – of the Leak, Danson, Pudney and Mason labs, as well as the teaching lab and technical support team for their endless supportive advice, discussions and assistance, including Carolyn Williamson, Steve Bowden, Emanuele Kendrick, Alex Holland, Giannina Espina Silva, Chris Vennard, Shyam Maskapalli, Leann Bacon, Micaela Chacon, Matthew Styles, Ali Hussein, Jeremy Bartosiak-Jentys, Maria Ortenzi, Alex Lathbridge, Daniel Baxter, Dragana Catici, Hannah Jones, Ed Nesbitt, Ana Panek – to name but a few, and my friends Hazel Roberts and Nathaniel Storey.

A special thanks also has to go to Lesley Chapman, Martin White, Colin Cooper, and Ewan Basterfield.

My deepest gratitude goes to Charlie Hamley-Bennett, Beata Lisowska, Chris Hills, Chris Ibenegbu, Alice Marriott, Nick Morant, Andrew Preston, Luke Williams, and Susan Crennell, without whom much of the project would not exist, as well as Vic Arcus and his group at the University of Waikato, NZ, in particular Emma Andrews, Konny Shim, and Judith Burrows, whom I met on my research visit funded by the Microbiology Society and the Biochemical Society.

Special recognition must go to my friends and mentors Charlie Hamley-Bennett, Beata Lisowska, Chris Hills, Skevoulla Christou, Jennifer Zaslona, Mike Bushell and Tim Dawson for their inspiration and endless encouragement.

Of course a very special thanks has to go to my family for their support and encouragement throughout my studies and in life.

Finally, I would like to thank my 4-pawed friend - Dexter - for keeping me sane and driving me mad in almost equal measures.

CONTENTS

LIST OF ABBREVIATIONS

ADH, *adh*	alcohol dehydrogenase, alcohol dehydrogenase gene
AIC	Akaike information criterion
AK	acetate kinase
ALS, *als*	acetolactate synthase, acetolactate synthase gene
APS	ammonium persulphate
ASR	ancestral sequence reconstruction
AU	absorbance unit
BIC	Bayesian information criterion
Bis-Tris	2-[Bisamino]-2-1,3-propanediol
BMC	bacterial microcompartment
bp	base pair
BSA	bovine serum albumin
C_T	threshold cycle
CUF	codon usage frequency
DMSO	dimethyl sulfoxide
dNTP	deoxyribonucleotide triphosphate (A: adenine, C: cytosine, G: guanine, T: thymine)
DTT	dithiothreitol
E. coli	*Escherichia coli*
EDO	1,2-ethanediol
EDTA	ethylenediaminetetraacetic acid
EMP	Embden-Meyerhof-Parnas pathway
F	structure factor
F_c	$F_{calculated}$
Fd	ferredoxin
F_o	$F_{observed}$
Gk	*Geobacillus kaustophilus*
Gst	*Geobacillus stearothermophilus*
Gt	*Geobacillus thermoglucosidasius*
GTR	general time-reversible model of evolution (REV is the same)
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPLC	high-performance liquid chromatography
I	intensity
IPD	indole-pyruvate decarboxylase
Kb	kilobases
K_M	Michaelis constant (concentration of substrate when reaction rate is ½ of V_{max})
LB	lysogeny broth
LDH, *ldh*	lactate dehydrogenase, lactate dehydrogenase gene
MCS	multiple cloning site
Mb(p)	mega bases (pairs)
MES	2-(N-morpholino)ethanesulfonic acid
ML	maximum likelihood
MOPS	3-(N-morpholino)propanesulfonic acid

Mr	relative molecular mass
NAD^+, NADH	nicotinamide adenine dinucleotide oxidised form, reduced form
OD	optical density
PCR	polymerase chain reaction
PDC, *pdc*	pyruvate decarboxylase, pyruvate decarboxylase gene
PDH, *pdh*	pyruvate dehydrogenase, pyruvate dehydrogenase gene
PEG	polyethylene glycol
PET	producer of ethanol operon
PFL, *pfl*	pyruvate formate lyase, pyruvate formate lyase gene
POR	pyruvate ferredoxin oxidoreductase
PTA	phosphotransacetylase
RBS	ribosome binding site
rRNA	ribosomal RNA
RT-qPCR	reverse transcription quantitative polymerase chain reaction
SDS-PAGE	sodium dodecyl sulphate-polyacrylamide gel electrophoresis
spp.	species (plural)
TAE	Tris-acetic acid-EDTA buffer
T_d	denaturing temperature
TEMED	tetramethylethylenediamine
T_m	melting temperature
TPP	thiamine pyrophosphate
tRNA	transfer RNA
T_{room}	room temperature (~21°C)
U	µmol/min
V_{max}	maximum velocity of reaction
WAG	Whelan and Goldman model of evolution
wt	wild type
Zp	*Zymobacter palmae* (*Z. palmae*)
ZpPDC 2.0	enzyme encoded by the codon harmonized Zp*pdc*
Ɛ	absorption coefficient
θ	Bragg angle
λ	wavelength

1. General Introduction

1.1 Bioethanol: Waste to Energy

In an effort to reduce dependency on finite fossil fuel resources and to reduce carbon dioxide emissions attributed to their use, the search for commercially-viable and environmentally-sustainable renewable fuels has become a major industrial goal (Gronenberg *et al.* 2013). Biofuels, such as diesel, butanol, hydrogen, and ethanol are important contenders, as they are compatible with the currently existing infrastructure (Liao *et al.* 2016). They are made from renewable, plant-based carbohydrate sources and have the potential to be carbon neutral, as burning them only releases the carbon that was captured by photosynthesis during plant growth (Antoni 2007, Liao *et al.* 2016).

Bioethanol is one of the major biofuel products, due to its ease of production and access to market. In 2015 over 52 billion litres were produced in the USA alone, with global production reaching 90 billion litres (RFA 2016 Ethanol industry outlook, RFA 2016 Industry statistics). Traditionally, bioethanol is produced through fermentation of starch-rich feedstocks, such as corn or wheat, and sucrose-rich sugar cane. When starch-rich or sucrose-rich feedstocks are broken down, C6 sugars are released that can be readily converted to ethanol by yeast, such as *Saccharomyces cerevisiae* (Sanchez & Cardona 2008). Although this is now a mature technology, industrial-scale production of first-generation bioethanol from food crops has sparked a "food vs. fuel" debate. Around 40% of the corn currently grown in the USA is used in fuel production (Hood 2016) and it has been argued that the land would be better used to grow crops to feed the world's growing population (Taylor *et al.* 2009).

As an alternative to using food crops as fermentation feedstock, the production process of second-generation biofuels turns lignocellulosic feedstocks into energy. This approach exploits cheap and readily-available organic material, such as municipal and agricultural waste, and plants grown on land not suitable for food crop production. Typical feedstocks include *Miscanthus siensis*, wheat straw (*Triticum*), *Zea mays* stalks, sugar cane bagasse (*Saccharum officinarum*), sugar beet pulp (*Beta vulgaris*), brewery and paper mill waste, forestry residues and municipal solid waste. According to some estimates, this non-food biomass makes up 50% of the biomass on Earth (10 to 50 billion tons according to Claassen *et al.* 1999). This type of production can enhance the value of waste products, while avoiding the use of land suitable for the production of food crops, and reducing landfill and green-house gas emission, making it environmentally favourable (Liao *et al.* 2016).

Production of biofuels from lignocellulosic feedstocks requires extensive feedstock pre-treatment to release carbohydrates (Lynd *et al.* 2008, Mosier *et al.* 2005). This is due to the complex nature of the feedstock. Lignocellulosic feedstock is typically made up of 38-50% cellulose (a polymer of β 1-4 linked glucose molecules), encapsulated by 17-32% hemicellulose (mixed polymers of xylose, mannose, galactose, and arabinose) and 15-30% lignin (Ritter 2008) (Figure 1.1). The exact composition depends on the feedstock and may also include pectin.

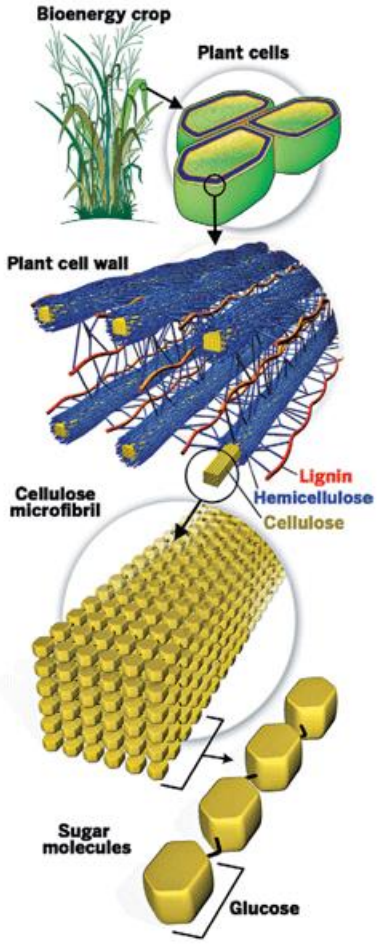


Figure 1.1 Lignocellulosic feedstocks are complex materials. The plant cell wall is made up of cellulose strands enmeshed by hemicellulose and further strengthened by lignin. From Ritter (2008).

The pre-treatment process involves breaking down the lignocellulosic material physically (through milling or steam explosion) or chemically (acid or base hydrolysis) in preparation for enzyme hydrolysis (Mosier *et al.* 2005). Enzyme hydrolysis uses industrially-produced enzyme mixes typically containing at least cellobiohydrolases, endoglucanases and β-glucosidases, and

releases short-chain polysaccharides and monosaccharides, making them accessible for utilization by fermentative microorganisms (Liao *et al.* 2016). The hydrolysed biomass is then fed into fermentation by yeasts or bacteria, from which ethanol can be distilled (Figure 1.2).

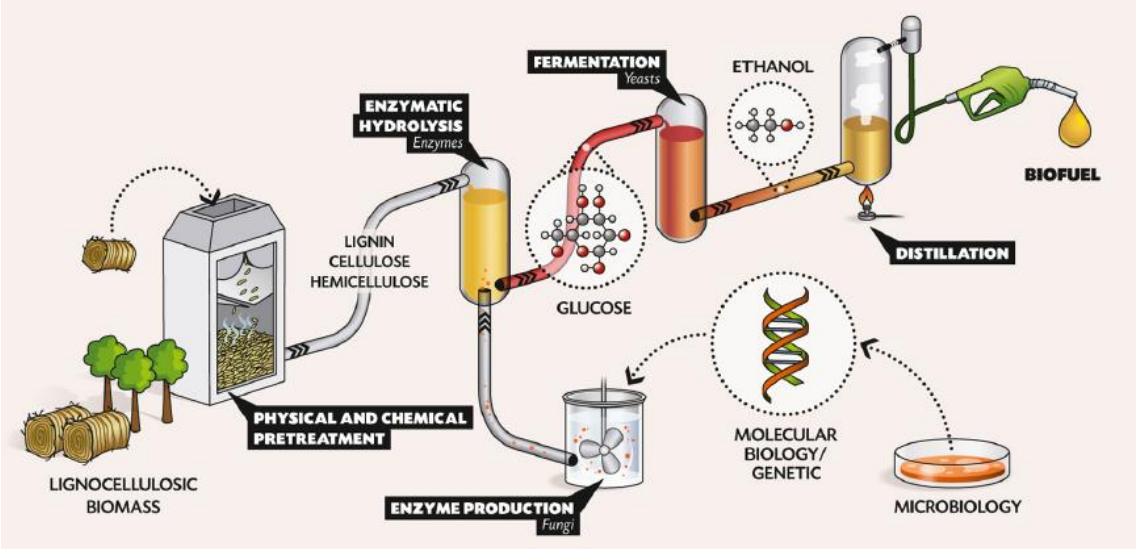


Figure 1.2 Schematic representation of biocatalysed bioethanol production from lignocellulosic biomass. The plant cell wall is made up of cellulose strands enmeshed by hemicellulose and further strengthened by lignin. Pre-treatment and enzyme hydrolysis are required to break down this complex material and release sugars for fermentation by microorganisms, such as yeasts or bacteria. From IFP Energies Nouvelles (2016).

Traditional fermentation organisms, such as S*. cerevisiae* or *Zymomonas mobilis*, are only able to utilize simple C6 sugars (glucose and mannose). They cannot ferment C5 monomers (xylose and arabinose) or short-chain polymers (cellobiose and xylobiose) which are also released from lignocellulosic biomass; therefore, fermentation is only partially completed, adding to the cost of second-generation bioethanol production (Mosier *et al.* 2005).

Converting the lignin to biofuels is incredibly challenging due to its structural diversity. However, lignin can be converted to heat and electricity or gasified for further use (Liao *et al.* 2016).

For a lignocellulosic bioethanol production process to be economically viable, the costs of extensive feedstock pre-treatment need to be reduced. This requires the production process to be extremely efficient. An ideal production microorganism would be robust against stresses associated with industrial-scale production, including high ethanol and moderate acid and pre-

treatment derived toxin levels. The ability to withstand high temperatures while maintaining balanced growth with high production yields and being resistant to contamination in minimally sterilized conditions would also be advantageous (Liao *et al.* 2016). Genetic tractability is also of considerable importance, as no single organism is likely to fit every criterion. Recent advances in molecular biology readily allow metabolic engineering to explore every opportunity to increase productivity and yields, often including overexpression of desired pathways and knock-out of competing ones (Liao *et al.* 2016).

In an effort to reduce enzyme hydrolysis costs, consolidated bioprocessing strategies have been developed. These aim at simultaneous, one-reactor cellulose hydrolysis and fuel production without the addition of supplementary enzymes. In one strategy, ethanol producers have been metabolically engineered to produce the required enzymes to digest cellulose. *Z. mobilis*, for example, was engineered by Vasan *et al.* (2011) to express an endoglucanase. In an alternative approach, naturally cellulolytic organisms, such as *Caldicellulosiruptor bescii*, have been engineered to produce ethanol (Chung *et al.* 2014).

In an attempt to improve ethanol yields, traditional fermentation strains have been metabolically engineered to utilize a wider variety of carbohydrates. Several research groups have metabolically engineered *Z. mobilis* (Agrawal *et al.* 2012, Yanase *et al.* 2012) or S*. cerevisiae* (Bettiga *et al.* 2009, Brat *et al.* 2009) for xylose utilization, for example. However, catabolite repression effects observed will need to be overcome in order to achieve simultaneous utilization of hexose and pentose sugars (Liao *et al.* 2016).

A possible alternative to these engineering approaches is the use of a novel ethanologenic species with wider substrate ranges. Thermophilic species of *Clostridium, Thermoanaerobacter* and *Geobacillus* are possible candidates for biofuel production from a variety of feedstocks (Barnard *et al.* 2010, Sanchez & Cardona 2008, Taylor *et al.* 2009).

1.2 Thermophilic Microorganisms for Bioethanol Production

Thermophiles are a sub-category of extremophilic microorganisms, thriving at environmental temperatures between 40 and 70°C. They are potentially useful as microbial cellular factories, because they have a number of advantages over mesophilic microorganisms in industrial-scale bioethanol production processes.

Thermophiles are commonly able to ferment a broad range of substrates not limited to hexose and pentose monomers, but often including short-chain polysaccharides or structurally complex materials such as cellulose (Sommer *et al.* 2004, Zaldivar *et al.* 2001). Their robust metabolic system can also be less sensitive to fluctuations in temperature and pH (Barnard *et al.* 2010).

Furthermore, running high temperature fermentations limits the risk of contamination. The most common microbial contaminants of mesophilic fermentations are feedstock-derived species of lactobacilli. These contaminants reduce ethanol yields and drive up costs, as the addition of antibiotics to the fermentation process is necessary to limit contamination (Taylor *et al.* 2009). The optimum growth temperature of thermophilic production strains is much higher than the maximum permissible growth temperature of these kinds of contaminants, therefore production losses and costs are reduced. Furthermore, as gas solubility is significantly lower at 60°C than at 37°C, an anaerobic environment is readily maintained, favouring fermentative metabolism and limiting contamination by obligate aerobic microorganisms (Taylor *et al.* 2009).

High temperature processes allow further reductions in running costs, as cooling between the enzyme hydrolysis step (typically run at 50-55°C) and the fermentation is not required. Furthermore, high temperatures accelerate chemical reaction rates, promote solubility and efficient mixing of the substrate, and facilitate the removal of ethanol (Barnard *et al.* 2010.) Aqueous ethanol readily evaporates at temperatures above 50°C, so applying a mild vacuum might allow continuous “stripping” reducing the build-up of ethanol to toxic levels (Taylor *et al.* 2009).

Despite these advantages to the production process, thermophilic ethanologens are often limited in their production yields due to their mixed acid fermentation metabolism. This is an inefficient use of carbohydrate sources, and in addition, unwanted organic acids may affect growth negatively (Taylor *et al.* 2009).

Furthermore, thermophilic ethanologens often lack high ethanol tolerance, generally less than 2% (v/v) exogenous ethanol (Georgieva *et al.* 2007, Zaldivar *et al.* 2001). *Geobacillus thermoglucosidasius* shows tolerance up to 4% (v/v) exogenous ethanol (TMO Renewables Ltd., personal communication). Some *Geobacillus* strains have been reported to have a tolerance to exogenous ethanol concentrations of up to 10% (v/v) (Fong *et al.* 2006); although, their experimental set-up does not exclude that this may be the ethanol tolerance of

spores rather than viable cells. However, these figures are still considerably less than the 16% (v/v) reported for *Z. mobilis* (Swings & De Ley 1977) or the 20% (v/v) reported for *S. cerevisiae* (Hosaka *et al.* 1998). The figures reported are, however, concentrations of exogenous ethanol, not endogenously produced ethanol, which appears more toxic to the cell (Georgieva *et al.* 2007). Maximum concentrations of endogenously produced ethanol are generally lower than tolerated exogenous ethanol concentrations, leading to the "titre gap" (Olson *et al.* 2015). High ethanol concentrations inhibit glycolysis and alter membrane organisation and permeability. The membrane integrity impairment is further enhanced by the increased growth temperature, as high temperatures destabilise membrane organisation and increase membrane fluidity (Georgieva *et al.* 2007, Zaldivar *et al.* 2001). The ethanol concentration tolerated by the fermentation organism dictates the allowable substrate concentration, and *vice versa*, the tolerated substrate concentration limits ethanol yields. A commercially viable production process requires an ethanol concentration of at least 4% (v/v) in the fermentation broth (Georgieva *et al.* 2007, Olson *et al.* 2015).

The limitations of low ethanol tolerance and inefficient carbohydrate utilization through mixed acid fermentation require genetic engineering in order to generate an efficient thermophilic production strain.

Fermentative ethanol production can occur through one of four possible pathways: through pyruvate dehydrogenase (PDH), pyruvate formate lyase (PFL), pyruvate decarboxylase (PDC), and pyruvate ferredoxin oxidoreductase (POR), together with their associated enzymes (generally including at least an alcohol dehydrogenase (ADH)) to convert either acetyl-CoA or acetaldehyde to ethanol (Olson *et al.* 2015). While the PDC pathway converts pyruvate directly to acetaldehyde and hence, via a single ADH-catalysed step to ethanol, the PDH, PFL and POR pathways produce acetaldehyde from acetyl-CoA, requiring an acetaldehyde dehydrogenase (AcDH). The AcDH and ADH functions are often mediated by a single, bifunctional enzyme, ADHE. To date there are examples of engineering efficient thermophilic ethanol producers for PDC, PDH and POR, with yields >90% of the theoretical maximum, but not for PFL. However, an artificial PFL-catalysed route has been patented by Biocaldol Ltd. (Bioconversion Technologies Ltd. 2007, US20090226992 A1), in which they envisaged using a formate dehydrogenase to cover the reducing equivalents lost as formate, which are essential for redox balanced ethanol production.

Recent developments in genome sequencing and metabolic modelling allow access to information required for targeted metabolic engineering. Strain development has been used to

generate some efficient thermophilic ethanologens (Olson *et al.* 2015). Obligate anaerobes, such as the archaeon *Pyrococcus furiosus* and the bacteria *Thermoanaerobacter mathranii, Thermoanaerobacterium saccharolyticum* and *Clostridium thermocellum* often use the POR mode. The POR pathway generates reduced ferredoxin from which electrons are transferred to NAD^+ or $NADP^+$ to allow high ethanol production yields (Olson *et al.* 2015). *T. saccharolyticum* has been engineered for improved ethanol yields up to 100% of the theoretical maximum on xylose and cellobiose at 55°C through deletion of lactate and acetate producing pathways (Shaw *et al.* 2008 and 2009). *T. mathranii* engineering included the deletion of lactate producing pathways and regulation of ADHE by a xylose-inducible promoter to achieve 95% of the theoretical maximum yield on xylose at 70°C (Yao & Mikkelsen 2010). *C. thermocellum* does not consume pentose sugars, which *T. saccharolyticum* and *T. mathranii* do, but it does have the advantage over those species that it is capable of utilizing crystalline cellulose (Olson *et al.* 2015). This makes it a leading candidate for consolidated bioprocessing (Liao *et al.* 2016). Deletion of a number of hydrogenases increased ethanol yield to 61% of the theoretical maximum at 55°C on cellobiose (Olson *et al.* 2015).

Some thermophiles lack an ADHE, an AcDH or an enzyme with a similar acetaldehyde-producing function and can therefore not produce ethanol. Expression of a functional ADHE can turn them into efficient ethanologenic strains. Due to its ability to consume complex polysaccharides, similarly to *C. thermocellum*, *C. bescii* has attracted considerable interest, despite the fact that it is not a natural ethanol producer. By expressing the *C. thermocellum adh*E this organism was engineered to produce ethanol yields up to 33% of the theoretical maximum from cellobiose at 65°C (Chung *et al.* 2014). Similarly, the hyperthermophilic, anaerobic archaeon *P. furiosus* was modified to express the *adhA* from *Thermoanaerobacter* spp. X514, resulting in a yield of 35% of the theoretical maximum from cellobiose at 72°C (Basen *et al.* 2014).

The methylotrophic, thermotolerant yeast *Ogataea polymorpha* uses a PDC-ADH pathway, which has not been found in thermophilic bacteria as of yet (Professor D. A. Cowan, University of Pretoria, South Africa; unpublished results). Over-expressing γ-glutamylcysteine synthase to improve tolerance to endogenously produced ethanol in this organism has increased ethanol production at 48°C to 96% of the theoretical maximum (Grabek-Lejko *et al.* 2011).

The facultatively anaerobic bacterium *G. thermoglucosidasius* possesses genes for both the PFL and the PDH mode. It is not entirely certain which mode is the major natural ethanol production pathway, but engineering with a focus on the PDH mode has produced an efficient

ethanol producing strain (Cripps *et al.* 2009, Olson *et al.* 2015). The second-generation bioethanol production company TMO Renewables Ltd., Guildford, UK (relaunched as ReBio Technologies Ltd. in 2014) selected *G. thermoglucosidasius* as a suitable candidate organism for lignocellulosic bioethanol production, due to its thermophilic nature, its genetic accessibility and its native ability to ferment a variety of carbohydrate sources to desirable products, including ethanol.

1.3 *Geobacillus thermoglucosidasius* and its Use for Industry

G. thermoglucosidasius is a facultatively anaerobic, Gram-positive, spore-forming bacillus capable of growth between 40°C and 70°C (Nazina *et al.* 2001). *G. thermoglucosidasius* is very flexible in its growth substrates being able to ferment pentose and hexose monomers, and oligomers. It naturally produces valuable products, predominantly lactate, but also small amounts of formate, acetate and ethanol (Hussein *et al.* 2015, Olson *et al.* 2015).

The establishment of a genetic tool kit and transformation protocols made this organism genetically tractable and allowed metabolic engineering through over-expression of genes on the plasmid pUCG18 or creating insertions and deletions using pTMO31 (Cripps *et al.* 2009, Taylor *et al.* 2008).

Growth under anaerobic conditions requires regeneration of NAD^+ from the NADH produced during glycolysis. In the wild type (wt) *G. thermoglucosidasius* NCIMB 11955 reduction of pyruvate to lactate is the dominant process to regenerate NAD^+. Therefore, in order to enhance ethanol production for industrial use, this organism has been modified by TMO Renewables Ltd. and partners, by knocking out lactate dehydrogenase (*ldh*) (Figure 1.3). This increased the ethanol yield from 20 to 47% of the theoretical maximum in fermentations at 60°C on glucose (Cripps *et al.* 2009). However, formate and acetate were still produced in small amounts and unlike fermentation in the wt organism, pyruvate accumulation was observed, indicating that expression of the natural *pdh* under oxygen-limited conditions was too weak to provide an adequate fermentation pathway.

The ethanol yield was further improved by upregulating the *pdh* (through regulation of its expression by an anaerobically-induced *ldh* promoter) and deleting *pfl*. In batch fermentations at 60°C the triple mutant (Δ*ldhA*, Δ*pflB*, pdh^{up}, Figure 1.3), known as TM242, could generate ethanol as the major product with yields of up to 68% on xylose, 82% on glucose and 92% of

the theoretical maximum on cellobiose, while reducing pyruvate accumulation to insignificant amounts (Cripps *et al.* 2009).

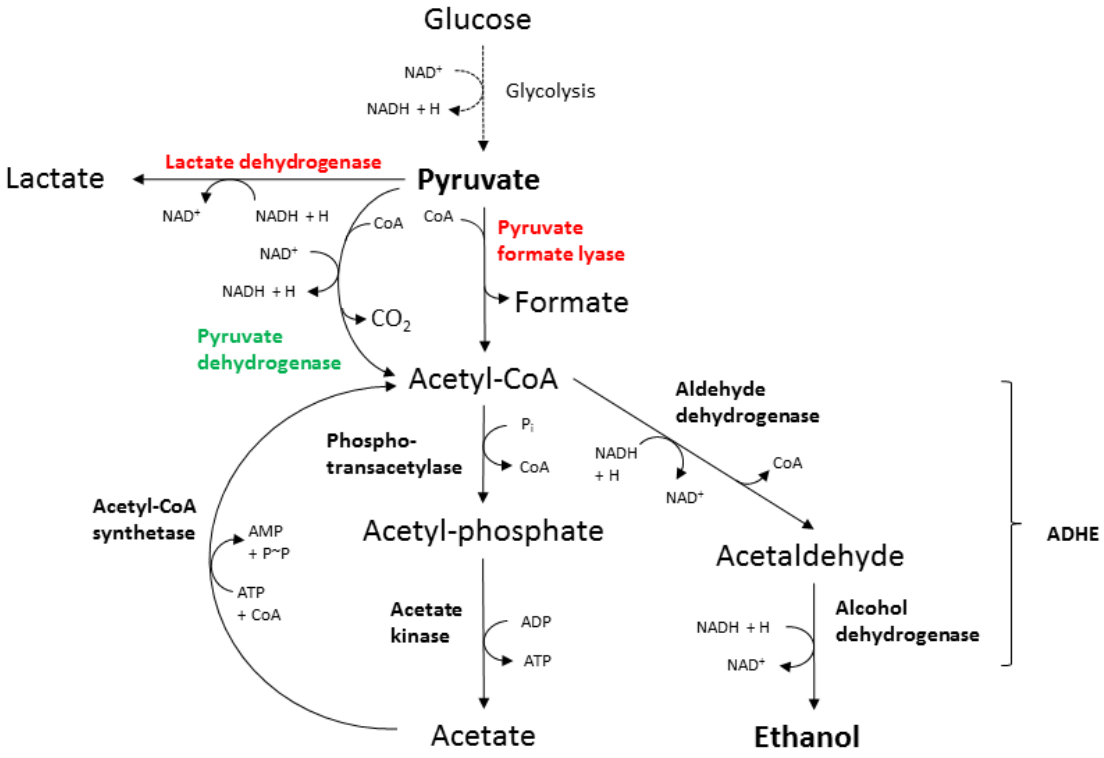


Figure 1.3 Schematic representation of the major fermentative metabolic pathways in *G. thermoglucosidasius*. In TM242 lactate dehydrogenase and pyruvate formate lyase (in red) have been deleted, while pyruvate dehydrogenase (in green) has been upregulated. Abbreviations are: ADHE, bifunctional aldehyde/alcohol dehydrogenase; P_i, orthophosphate; P~P, pyrophosphate.

Pyruvate is the major substrate for these fermentation pathways. In *G. thermoglucosidasius* it is generated through glycolysis via the Embden-Meyerhof-Parnas pathway (EMP) (Cooper 1978, TMO Renewables Ltd. unpublished data, Figure 1.4).

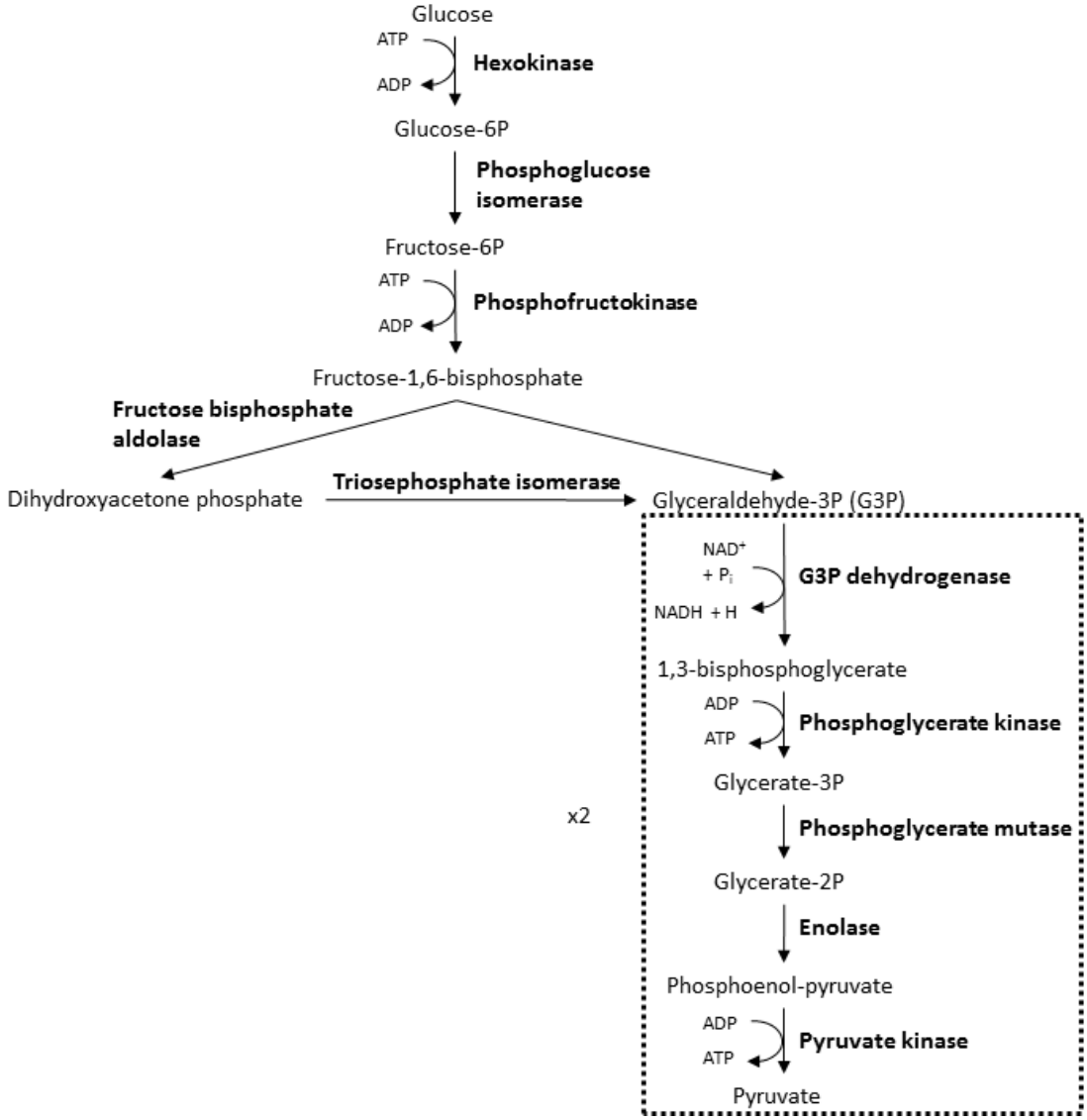


Figure 1.4 Schematic representation of the Embden-Meyerhof-Parnas pathway. The net reaction is:

$$\text{Glucose} + 2\ \text{ADP} + 2\ \text{NAD}^+ + 2\ \text{P}_i \rightarrow 2\ \text{pyruvate} + 2\text{ATP} + 2\text{NADH} + 2\ \text{H}^+$$

Dihydroxyacteone phosphate is converted to glyceraldehyde-3P, so the last five reactions occur twice for every molecule of glucose catabolised (highlighted by the dashed box). P is phosphate, P_i is orthophosphate.

For every molecule of glucose metabolised through the EMP pathway 2 molecules of pyruvate, 2 ATP and 2 NADH are generated. Further metabolism via PDH generates another NADH, and a molecule of acetyl-CoA from each molecule of pyruvate. Under anerobic conditions, this NADH must be reoxidsed to NAD^+ in order to restore the cellular redox balance; this requires the

reduction of acetyl-CoA to acetaldehyde by an AcDH. Then, taking into account the 2 NADH produced during glycolysis, to fully restore redox balance, further regeneration of NAD^+ through the ethanol yielding activity of ADH is required (Figure 1.5).

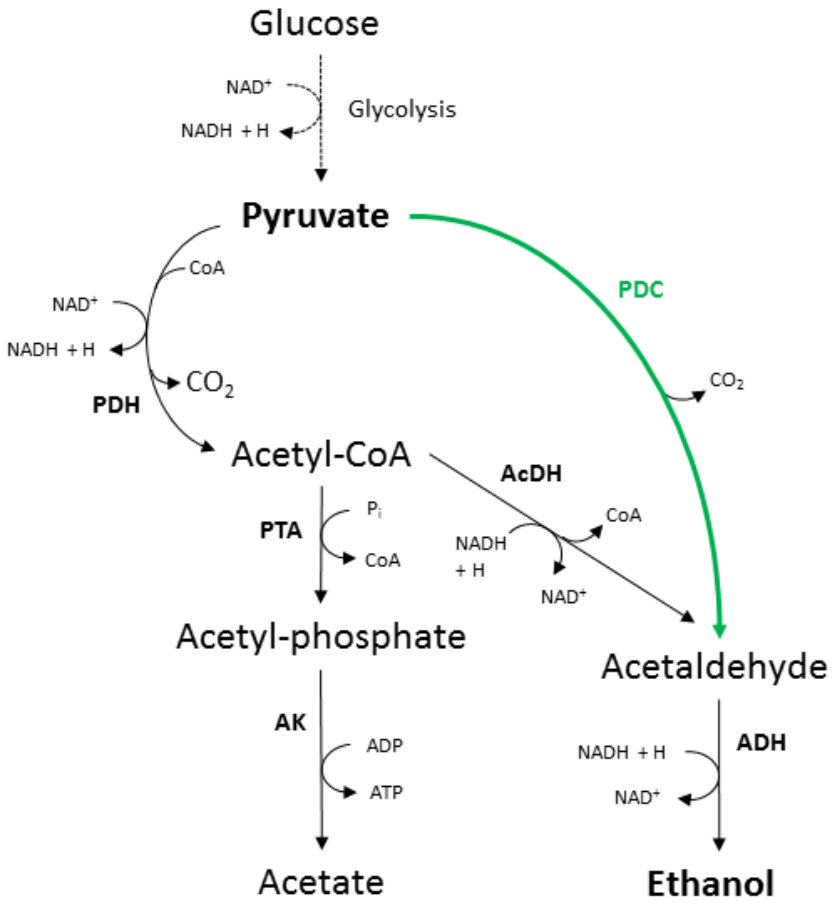


Figure 1.5 Schematic representation of various ethanol production strategies. Mixed acid producers such as *G. thermoglucosidasius* require an extensive metabolic network to restore the cellular redox balance through regeneration of NAD^+ during fermentation. Homoethanol producers, instead, use a more direct pathway via PDC (highlighted in green) and ADH to efficiently regenerate NAD^+. Enzyme abbreviations are: AcDH, acetaldehyde dehydrogenase; ADH, alcohol dehydrogenase; AK, acetate kinase; LDH, lactate dehydrogenase; PDC, pyruvate decarboxylase; PDH, pyruvate dehydrogenase; PFL, pyruvate-formate lyase; PTA, phosphotransacetylase.

In contrast to mixed acid producers, such as *G. thermoglucosidasius*, homoethanol producers use a more direct pathway via PDC and ADH to efficiently regenerate NAD^+ with high ethanol yields desirable for biotechnological applications such as industry-scale bioethanol production. While the extensive metabolic engineering by TMO Renewables Ltd. has yielded a strain with

greatly improved ethanol yield, unwanted by-products are still observed, in particular acetate (Cripps *et al.* 2009), which derives from acetyl-CoA. An alternative approach increasing carbon flux towards ethanol may be to introduce a PDC. The presence of a heterologous PDC in *G. thermoglucosidasius* would offer an alternative route for the production of acetaldehyde, being stoichiometrically equivalent to the PDH-AcDH pathway (Figure 1.5), but avoiding metabolism via acetyl-CoA.

Previous studies in mesophilic organisms indicate that this is a promising strategy. *Z. mobilis pdc* has been successfully used to produce mesophilic ethanologenic strains in *Escherichia coli* (Ingram *et al*. 1988, Ingram & Conway 1988), *Klebsiella oxytoca* (Zhou & Ingram 1999) and *Bacillus* spp. (Barbosa & Ingram 1994). Furthermore, *Zymobacter palmae pdc* has been functionally expressed in *Lactococcus lactis* (Liu *et al.* 2005).

Some attempts at expressing bacterial PDCs in *G. thermoglucosidasius* have been made, with limited success. The *pdc* from *Z. palmae* has been functionally expressed up to 48°C in *G. thermoglucosidasius* DL44 (DL33 Δ*ldh*) (Taylor *et al.* 2008), while Thompson *et al.* (2008) observed PDC activity up to 52°C when expressing the *Z. mobilis pdc* in *G. thermoglucosidasius* TN (LLD-R Δ*ldh*). More recently, van Zyl *et al.* (2014a) characterized a novel PDC from *Gluconobacter oxydans* and reported PDC activity up to 52°C with an ethanol yield of 68% of the theoretical maximum in *G. thermoglucosidasius* TM89 (NCIMB 11955 Δ*ldh*). While these attempts are promising for the application of a PDC homoethanologenic pathway in *G. thermoglucosidasius*, optimization to functionally and efficiently express a PDC at optimum fermentation temperature of 60°C is required.

The overall objective of this project was to functionally express a thermostable PDC in *G. thermoglucosidasius* to channel pyruvate through the PDC-ADH pathway and away from acetate production.

1.4 Bacterial Pyruvate Decarboxylase

Function

Pyruvate decarboxylase (PDC, EC 4.1.1.1) is a key enzyme in homo-fermentative metabolism where ethanol is the main fermentation product. Using magnesium (Mg^{2+}) and thiamine pyrophosphate (TPP) as cofactors, PDC catalyses the non-oxidative decarboxylation of pyruvate to acetaldehyde with the release of carbon dioxide. In a complete homo-

fermentative pathway, acetaldehyde is then reduced to ethanol by an alcohol dehydrogenase (ADH, EC 1.1.1.1).

PDC is widely represented in plants and fungi, but rare in bacteria (Raj *et al.* 2002). So far, only six bacterial PDCs have been described, including the one found in *Z. palmae* (ZpPDC, PDB entry 5EUJ, this study, Buddrus *et al.* 2016). The *Z. mobilis* enzyme (ZmPDC) has been extensively studied, with a variety of structural variants published (PDB entry 1ZPD, Dobritzsch *et al.* 1998; 2WVA, 2WVG, 2WHH, Pei *et al.* 2010; 3OEI, Meyer *et al.* 2010; 4ZP1, Wechsler *et al.* 2015). Other bacterial PDCs include those from *Acetobacter pasteurianus* (ApPDC, PDB entry 2VBI), *Gluconoacetobacter diazotrophicus* (GdPDC, PDB entry 4COK, van Zyl *et al.* 2014a), *G. oxydans* (GoPDC, van Zyl *et al.* 2014b), and the only known Gram-positive species possessing a PDC, *Sarcina ventriculi* (SvPDC, Lowe & Zeikus 1992). Table 1.1 summarises kinetic and temperature data of the known bacterial PDCs.

Unlike yeast PDCs, bacterial PDCs are not allosterically activated, with the exception of SvPDC (Raj *et al.* 2002). In the first step of the catalytic cycle, TPP is protonated at N-1' and deprotonated at 4'-NH_2. This imino tautomer in turn promotes the deprotonation at C2 on the thiazolium ring, thus creating the active ylid. The nucleophilic attack of the ylid on the carbonyl group of pyruvate generates a lactyl adduct (C2-α-lactylthiamine diphosphate intermediate), decarboxylation of which yields the enamine intermediate with concomitant carbon dioxide release. This intermediate is then protonated producing hydroxyethyl TPP, and finally the release of acetaldehyde regenerates the ylid (Pei *et al.* 2010, van Zyl *et al.* 2014a, see Chapter 3, Figure 3.20 for more detail).

Table 1.1 Characteristics of known bacterial PDCs.

	ZpPDC	ApPDC	ZmPDC	GoxPDC [g]	GdPDC [h]	SvPDC
Organism	*Zymobacter palmae* Gram-negative γ-Proteobacteria	*Acetobacter pasteurianus* Gram-negative α-Proteobacteria	*Zymomonas mobilis* Gram-negative α-Proteobacteria	*Gluconobacter oxydans* Gram-negative α-Proteobacteria	*Gluconoacetobacter diazotropicus* Gram-negative α-Proteobacteria	*Sarcina ventriculi* Gram-positive Firmicutes (Clostridium)
pH optimum	4.5 – 8, maximum at 7 [b] 5.5 – 6.0 [f]	3.5 – 6.5 [b] 5 – 5.5 [f]	5.5 – 8, maximum at 6 – 6.5 [b] 6.0 [f]	4.5 – 5.0	5 – 5.5	6.3 – 6.7 [c, f]
pH dependent stability half-life time	pH6.5/8: stable for several days [b] pH4/9: complete loss within 2h [b]	pH 5 – 7: no activity loss within 60 hours pH4: 2.3h [b]	NA	NA	NA	NA
Temperature optimum	55°C [b]	65°C [b]	60°C [b]	53°C	45 – 50°C	NA
Temperature dependent stability half-life time	30°C: 150h 40°C: 40h 50°C: 10h 60°C: 0.4h [b]	30°C: 144h 40°C: 34h 50°C: 12h 60°C: 2h 70°C: 0.4h [b]	50°C: 24h [b]	65°C: 0.16h	60°C: 0.3h	50°C: 0.5h [h]
Temperature dependence of activity retention (30 min exposure)	60°C: 100% 65°C: 80% 70°C: 0% [f]	50°C: 100% 60°C: 65% 65°C: 45% 70°C: 5% [f]	45°C: 85% 60°C: 65% 65°C: 45% 70°C: 0% [f]	55°C: 98% 60°C: 70% 65°C: 40%	NA	45°C: 95% 50°C: 0% [f]

	ZpPDC	ApPDC	ZmPDC	GoxPDC [g]	GdPDC [h]	SvPDC
Denaturing temperature	NA	NA	63°C [i]	NA	NA	NA
*V*max (U/mg)	116 ±2 (pH 6.5)[b] 130 (pH 6)[f] 140 (pH 7)[f]	110 ±1.9 (pH 6.5)[b] 97 (pH 5)[f] 79 (pH 7)[f]	121 (pH 6.5)[b] 100 (pH 6)[f] 78 (pH 7)[f] 120 [e] 181 [a]	57 (pH 5) 47 (pH 6) 125 (pH 7)	20 (pH 5) 39 (pH 6) 43 (pH 7)	103 [c] 45 (pH 6.5)[f] 35 (pH 7)[f]
K_M ($S_{0.5}$) (mM)	2.5 ±0.2 (pH 6.5)[b] 0.24 (pH 6)[f] 0.71 (pH 7)[f]	2.8 ±0.2 (pH 6.5)[b] 0.39 (pH 5)[f] 5.1 (pH 7)[f]	1.3 (pH 6.5)[b] 0.43 (pH 6)[f] 0.94 (pH 7)[f] 0.31 (pH 6)[d] 1.1[e] 0.4 (pH 6)[a]	0.12 (pH 5) 1.2 (pH 6) 2.8 (pH 7)	0.06 (pH 5) 0.6 (pH 6) 1.2 (pH 7)	13 [c] 5.7 (pH 6.5)[f] 4.0 (pH 7)[f]
k_{cat} (s^{-1})	341 – 508 [f]	341 – 508 [f]	486 (pH 6.5)[f] 150 (pH 6)[d]	57 (pH 5) 47 (pH 6) 125 (pH 7)	NA	412 [c]
k_{cat} / K_M ($M^{-1}.s^{-1}$)	1.4×10^6 (pH 6.0) (calculated based on [f]) [h]	1.3×10^6 (pH 5.0) (calculated based on [f]) [h]	1.9×10^6 (pH 6)[d] 4.4×10^5 (pH 6.5)[e] 1.79×10^6 (pH 6)[a]	1.9×10^6 (pH 5.0) 1.6×10^5 (pH 6.5) 1.8×10^5 (pH 7.0)	1.3×10^6 (pH 5.0) 2.6×10^5 (pH 6.0) 1.4×10^5 (pH 7.0)	3.2×10^4 [c] 0.87×10^4 (pH 7)[h]
Kinetics	Michaelis-Menten[b]	Michaelis-Menten[b]	Michaelis-Menten[b]	Michaelis-Menten	Michaelis-Menten[b]	Sigmoidal[f] (activated by pyruvate)
Activation energy	41 kJ mol^{-1} [b]	27.1 kJ mol^{-1} [b]	43 kJ mol^{-1} [b]	NA	46 kJ mol^{-1}	NA

	ZpPDC	ApPDC	ZmPDC	GoxPDC [g]	GdPDC [h]	SvPDC
Buffer used for the characterisation of decarboxylase activity and stability	50 mM potassium phosphate buffer pH 6.5, 2.5 mM $MgSO_4$, 0.1 mM TPP [b] 50 mM sodium citrate buffer pH 5, 1 mM $MgCl_2$, 1 mM TPP [f]	50 mM potassium phosphate buffer pH 6.5, 2.5 mM $MgSO_4$, 0.1 mM TPP [b] 50 mM sodium citrate buffer pH 5, 1 mM $MgCl_2$, 1 mM TPP [f]	50 mM potassium phosphate buffer pH 6.5, 2.5 mM $MgSO_4$, 0.1 mM TPP [b] 50 mM sodium citrate buffer pH 5, 1 mM $MgCl_2$, 1 mM TPP [f]	200 mM sodium citrate buffer pH 6.0, 5 mM $MgCl_2$, 0.1 mM TPP Irreversible denaturation in 50 mM MES buffer pH 6.5, 5 mM MgCl2, 0.1 mM TPP	200 mM sodium citrate buffer pH 6.0, 5 mM $MgCl_2$, 0.1 mM TPP	50 mM sodium citrate buffer pH 5, 1 mM $MgCl_2$, 1 mM TPP [f]
Accession numbers: GenBank (gene) GenBank (protein) PDB	AF474145.1 AAM49566.1 NA	AF368435.1 AAM21208.1 2VBI	M15393.2 AAA27696.2 1ZPD	KF650839.1 AHB37781.1 NA	KJ746104.1 AIG13066.1 4COK	AF354297.1 AAL18557.1 NA
Amino acid sequence identity to ZpPDC		73%	63%	67%	71%	31%

a. Bringer-Meyer *et al.* (1986), b. Gocke *et al.* (2009), c. Lowe & Zeikus (1992), d. Meyer *et al.* (2010), e. Siegert *et al.* (2005), f. Raj *et al.* (2002), g. van Zyl *et al.* (2014b), h. van Zyl *et al.* (2014a), i. Pohl *et al.* (1995)

SvPDC appears to be a special case amongst the bacterial PDCs. It is most similar to yeast PDCs. SvPDC's residues for TPP binding differ from other bacterial PDCs when compared to ZmPDC, and the residues involved in pyruvate binding to the allosteric site are conserved in yeast PDCs and SvPDC, but are not present in ZmPDC (Talarico *et al.* 2001).

A comparative analysis of a range of TPP-dependent enzymes revealed that they cluster into families based on specific enzyme function (Figure 1.6). SvPDC seems to be more closely related to eubacterial indole-pyruvate decarboxylases (IPD), while the PDCs of the Gram-negative bacteria group more closely with plant PDCs and an outgroup of fungal PDCs. Talarico *et al.* (2001) speculated that ZmPDC might have originated through horizontal transfer of a plant *pdc*, while SvPDC is more likely to have diverged earlier during evolution and shares the last common ancestor with eubacterial IPD and fungal PDCs.

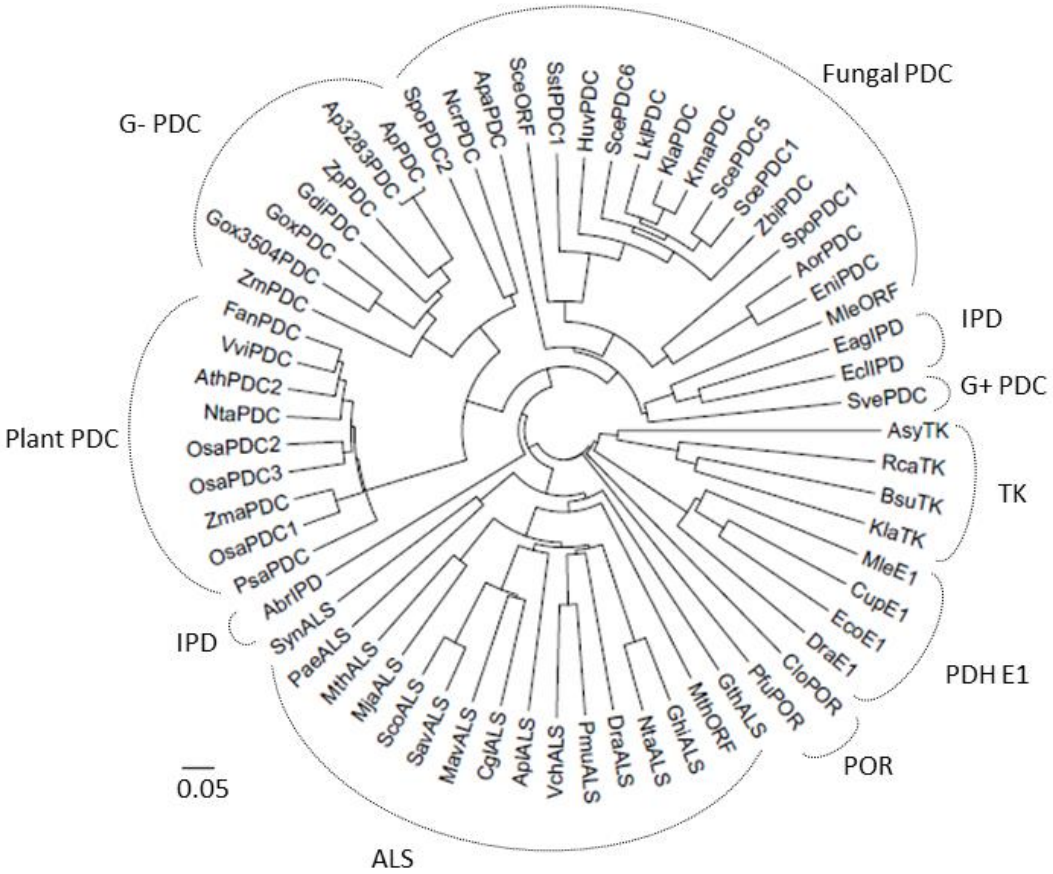


Figure 1.6 Unrooted tree depicting the relationships between selected TPP-dependent enzymes. The tree aligned protein sequences in ClustalW and displaying them in FigTree. Protein abbreviations are: ALS, acetolactate synthase; IPD, indole-pyruvate decarboxylase; PDC, pyruvate decarboxylase; PDH E1 or E1, the E1 component of pyruvate dehydrogenase;

POR, pyruvate ferredoxin oxidoreductase; TK, transketolase. Organism abbreviations and GenBank or SWISS-PROT accession numbers are: Ap, *Acetobacter pasteurianus* C7JF72 (NBRC 3283), H1URW8 (NBRC 106471); Abr, *Azospirillum brasilense* P51852, Aor, *Aspergillus oryzae* AAD16178; Apa, *Aspergillus parasiticus* P51844; Apl, *Arthrospira platensis* P27868; Asy, *Ascidia sydneiensis* BAA74730; Ath, *Arabidopsis thaliana* Q9FFT4; Bsu, *Bacillus subtilis* P45694; Cgl, *Corynebacterium glutamicum* P42463; Clo, *Clostridium* spp. L74 A0A0M1IXY2; Cup, *Cupriavidus necator* Q59097; Dra, *Deinococcus radiodurans* A75387 (ALS), A75541 (E1); Eag, *Enterobacter agglomerans* P71323; Ecl, *Enterobacter cloacae* P23234; Eco, *Escherichia coli* CAA24740; Ehe, *Erwinia herbicola* AAB06571; Eni, *Aspergillus (Emericella) nidulans* P87208; Fan, *Fragaria ananassa* AAG13131; Gdi, *Gluconacetobacter diazotrophicus* A0A075Q354; Ghi, *Gossypium hirsutum* S60056; Gox, *Gluconobacter oxydans* A0A067Z5Y9 (DSM 3504), K7SJD8 (H24); Gth, *Geobacillus thermoglucosidasius* A0A0M1QR06; Huv, *Hanseniaspora uvarum* P34734; Kla, *Kluyveromyces lactis* Q12629 (PDC), Q12630 (TK); Kma, *Kluyveromyces marxianus* P33149; Lkl, *Lachancea kluyveri* Q9P4E4; Mav, *Mycobacterium avium* Q59498; Mja, *Methanococcus jannaschii* Q57725; Mle, *Mycobacterium leprae* CAC31122 (ORF), CAC30602 (E1); Mth, *Methanobacterium thermoautotrophicum* A69081 (ORF), C69059 (ALS); Ncr, *Neurospora crassa* P33287; Nta, *Nicotiana tabacum* P51846 (PDC), P09342 (ALS); Osa, *Oryza sativa* P51847 (PDC1), P51848 (PDC2), P51849 (PDC3); Pae, *Pseudomonas aeruginosa* G83123; Pfu, *Pyrococcus furiosus* Q51799; Pmu, *Pasteurella multocida* AAK03712; Psa, *Pisum sativum* P51850; Rca, *Rhodobacter capsulatus* P0CZ16; Sav, *Streptomyces avermitilis* AAA93098; Sce, *Saccharomyces cerevisiae* P06169 (PDC1), P16467 (PDC5), P26263 (PDC6), Q07471 (ORF); Sco, *Streptomyces coelicolor* T35828; Spo, *Schizosaccharomyces pombe* Q09737 (PDC1), Q92345 (PDC2); Sst, *Scheffersomyces stipites* O43106; Sve, *Sarcina ventriculi* AF354297; Syn, *Synechocystis* sp. BAA17984; Vch, *Vibrio cholerae* A82375; Vvi, *Vitis vinifera* AAG22488; Zm, *Zymomonas mobilis* P06672; Zma, *Zea mays* P28516; Zbi, *Zygosaccharomyces bisporus* CAB65554; Zp, *Zymobacter palmae* Q8KTX6. G+ refers to Gram positive species. G- refers to Gram negative species. Scale bar represents 0.05 nucleotide substitutions per site.

Structure and Folding

In general, the PDC quaternary structure is a tetramer of four identical subunits that bind in a dimer-of-dimers fashion. Each monomer is composed of approximately 550 amino acids with a relative molecular mass of around 60 kDa. Each monomer contains an N-terminal pyrimidine binding (PYR) domain, a central regulatory (R) domain, and a C-terminal pyrophosphate binding (PP) domain.

Two monomers are tightly bound together in a dimer. Two TPP molecules bind across both subunits in each dimer, with the pyrimidine ring binding to the PYR domain from one subunit and the pyrophosphate group binding to the PP domain from the other, thus forming two active sites in the dimer (Figure 1.7). The Mg^{2+} ion anchors the TPP diphosphate group to the protein. Two dimers bind together to form a homotetramer. However, the dimer-dimer interaction is less tight than the interaction between the monomers within a dimer (Dobritzsch *et al.* 1998).

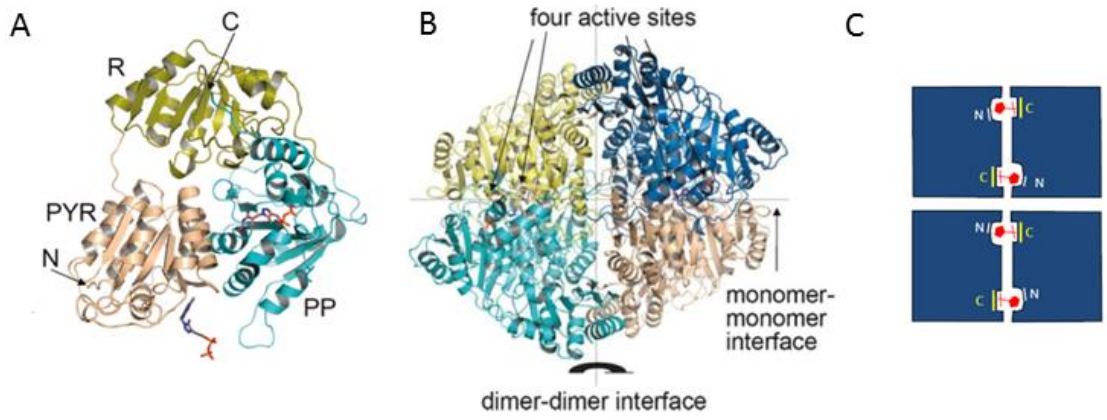


Figure 1.7 PDC structure and conceptual scheme of TPP binding to the active protein. (A) Cartoon representation of one monomer of ZmPDC. PYR-domain, PP-domain and R-domain coloured brown, cyan and yellow, respectively. Termini of the peptide are labelled N (PYR) and C (PP). (B) Overall structure of the homotetrameric ZmPDC. (C) Conceptual scheme of TPP binding to the active protein. PDC is effectively a dimer of dimers. Each dimeric monomer (blue) shares two TPP molecules (red) connecting the end-terminals as indicated. The orientation of TPP is highlighted by the pyrimidine ring. Adapted from Pei *et al.* (2010) and Waite (2010).

Denaturation and renaturation studies of the ZmPDC showed that its correct folding and activity strongly depends on the binding of the cofactors TPP and Mg^{2+} (Pohl *et al.* 1994). Cofactor-free renaturation results in an inactive enzyme. No activity is regained after cofactor-free refolding, even after incubation with excess cofactors. However, in the correctly folded protein TPP and Mg^{2+} are reversibly bound, and the holoenzyme has been shown to be greatly stabilised by cofactor binding. Once formed, the enzyme retains its tetramer formation even after dissociation from cofactors. The formation of catalytically active enzyme does, however, require cofactor binding.

1.5 Project Aims and Objectives

The overall objective of this project was to introduce a thermoactive PDC and acetaldehyde-reducing ADH into the host platform *G. thermoglucosidasius,* in order to channel pyruvate through the PDC-ADH pathway away from acetate production and towards higher ethanol yields. It was hypothesized that expression of a functional PDC-ADH pathway in *G. thermoglucosidasius* increases ethanol yields.

This involved:

- Exploring potentially suitable bacterial PDCs through ASR, which generated inferred ancestral PDCs, and through comparison of extant mesophilic bacterial PDCs.

- Expanding the knowledge on bacterial PDCs available by characterizing ZpPDC *in vitro*, including the crystal structure, and *in vivo* in *G. thermoglucosidasius.*

- Finding an appropriate ADH partner to complete the pathway and design a PET operon.

- Testing the PET operon under fermentative conditions in *G. thermoglucosidasius*.

Towards the overall objective, alternative approaches to cloning the wt Zp*pdc* were explored, especially focussing on codon harmonization and ancestral sequence reconstruction in search of a novel, potentially more thermoactive PDC.

1.6 Project Outline

Chapters 3 to 7 describe the characterization of ZpPDC, and the design and development of a producer of ethanol (PET) operon for expression in *G. thermoglucosidasius.* ZpPDC is one of the most thermostable bacterial PDCs currently known, with an *in vitro* thermoactivity of up to 65°C. Previous attempts to express this enzyme in *G. thermoglucosidasius* were promising with PDC activity observed up to a 50°C growth temperature (Taylor *et al.* 2008). Furthermore, unlike ApPDC and ZmPDC, there are currently no patent restrictions on ZpPDC, thus making it a great candidate for the optimization and evaluation of a PET pathway in *G. thermoglucosidasius* for biotechnological applications.

Chapter 3 describes the ZpPDC crystal structure and *in vitro* characterization, which gives vital information to further our understanding of bacterial PDCs. Please note that the data presented in this Thesis does not necessarily follow a chronological order. Unfortunately, the crystal structure data was not available for consideration in design approaches throughout this project.

Chapter 4 characterizes the expression of wt ZpPDC in aerobic cultures in *G. thermoglucosidasius*. The *in vitro* data suggest that purified ZpPDC is thermostable within the growth temperature range of *G. thermoglucosidasius*. Expression in this host at various growth temperatures resulted in detectable PDC activity in the unfractionated cell extract up to 65°C. However, with increasing growth temperature the activity rapidly decreased.

Codon harmonization was explored as one approach to improve recombinant expression of the ZpPDC in *G. thermoglucosidasius*. **Chapter 5** discusses a genome sequencing study to gather the information required for codon harmonization as described in **Chapter 6**. Applying this method did indeed improve detectable PDC activity in *G. thermoglucosidasius* cultures grown up to 65°C.

Chapter 7 describes the design and creation of a PET operon pairing the codon harmonized Zp*pdc* with the *G. thermoglucosidasius* ADH6. *In vitro* characterization of ADH6 kinetics and temperature optimum was performed by Dr. Luke Williams. All other work was performed by the author. *In vitro* coupled assays with these two enzymes were functional up to 70°C, indicating the potential for *in vivo* expression. The expression of the PET operon was successfully tested in *G. thermoglucosidasius* in aerobic cultures and tube fermentations up to 65°C. The discussion to this chapter also explores future work and several approaches to improve PDC activity in *Geobacillus* spp.

Chapter 8 presents ancestral sequence reconstruction as an alternative approach to finding a thermoactive bacterial PDC.

Chapter 9 is a general discussion and sums up the major findings of this Thesis.

2. General Methods

All chemicals were purchased from Sigma-Aldrich Company Ltd. (Sigma, Poole, UK), Merck Chemicals Ltd. (Nottingham, UK), BDH Laboratory Supplies Ltd. (Poole, UK), Melford (Ipswich, UK), Thermo Fisher Scientific (Waltham, MA, USA), Oxoid (Cambridge, UK), or Acros Organics (Loughborough, UK).

2.1 Bacterial Strains and Plasmids

Bacterial strains used in this project are listed in Table 2.1. Bacterial stocks were maintained in 20% (v/v) glycerol in cryogenic vials at -80°C. Bacteria were revived by streaking on appropriate media-agar plates. The plates were stored at 4°C for up to a month.

Table 2.1 Bacterial strains used in this study.

Strain	Genotype/Description	Growth conditions*	Source
***E.coli* BioBlue**	*end*A1, *rec*A1, *gyr*A96, *thi*-1, *hsd*R17 (r_{k-}, m_{k+}), *rel*A1, *sup*E44, *lac,* [F' *pro*AB *lacI*^{q}ZΔ15 Tn10 (Tet^{r})]	LB, 37°C	Bioline Reagents Ltd., London, UK
***E.coli* BL21 (DE3)**	*fhu*A2 [lon] *omp*T *gal* (λ DE3) [*dcm*] Δ*hsd*S λ DE3=λ sBamHIo ΔEcoRI-Bint :: (*lac*I :: PlacUV5 :: T7 gene1) i21 Δ*nin*5	LB, 37°C	NEB, Ipswich, MA, USA
***Geobacillus thermoglucosidasius* DL44**	Δ*ldh*, parent strain DL33	2TY, 2SPYNG, TSA, ASM 45-65°C	David Leak
***Geobacillus thermoglucosidasius* TM236**	Δ*ldh*, Δ*pfl,* parent strain NCIMB 11955	2TY, 2SPYNG, TSA, ASM 45-65°C	TMO Renewables Ltd., Guildford, UK
***Geobacillus thermoglucosidasius* TM400**	Δ*ldh*, Δ*pfl,* Δ*adh*E parent strain NCIMB 11955	2TY, 2SPYNG, TSA, ASM 45-65°C	TMO Renewables Ltd., Guildford, UK
***Zymobacter palmae* T109 (type strain, ATCC 51623)**	wild type	MYE, 30°C	NBRC, Japan

*See Table 2.3 for media compositions.

Table 2.2 gives a comprehensive list of the plasmids used in this study. Plasmid map manipulations were carried out using ApE (A plasmid Editor, version 2.0.30) and SnapGene Viewer (version 1.5.1).

Table 2.2 Plasmids used in this study.
Abbreviations are: ADH, alcohol dehydrogenase; amp^r, ampicillin resistance marker *bla*; Gk, *Geobacillus kaustophilus*; Gst, *Geobacillus stearothermophilus*; Gt, *Geobacillus thermoglucosidasius*; kan^r , kanamycin resistance marker; *ldh*, lactate dehydrogenase; MCS, multiple cloning site; Ori$_E$, *E. coli* origin of replication; Ori$_G$, *Geobacillus* spp. origin of replication; p, promoter; PDC, pyruvate decarboxylase; Zp, *Zymobacter palmae* T109 ATCC1623. GenBank accession numbers given in square brackets.

Name	Size (bp)	Description	Use	Source
pUCG18 [EU547236.2]	6331	MCS, Ori$_E$ pMB1 and amp^r marker from pUC18, Ori$_G$ repBST1 and kan^r TK101 from pBST22 (Liao & Kanikula 1990)	Thermostable (up to 68°C) *E. coli*/*Geobacillus* spp. shuttle vector	Taylor *et al.* (2008) [EU547236.2]
pGR002 (pUCG18pheB)	7490	pUCG18 backbone, contains Gst NCA1503 p*ldh* [D9042.1] – Gst *phe*B [DSMZ6285], [DQ146476.2]	Catechol 2,3-dioxygenase expression	Bartosiak-Jentys *et al.* (2012)
p778	9238	pUCG18 backbone, contains Gst NCA1503 p*ldh* - Zp wt *pdc* [AF474145.1] – Gst *adh*T [M19396.1]	PDC/ADH expression	Alex Pudney (unpublished)
p600 wt ZpPDC (p778*BamXba*OUT, pCJWGeo)	9238	p778 with site directed mutation in non-unique *Bam*HI and *Xba*I sites upstream of Gst p*ldh*	PDC/ADH expression	Chris Waite (2010)
p600 ZpPDC 2.0	9238	as p600 wt ZpPDC, but *pdc* codon harmonized	PDC/ADH expression	This study
pGR002 wt ZpPDC	8182	pUCG18 backbone, contains Gst NCA1503 p*ldh* – *phe*B RBS - Zp wt *pdc*	PDC expression	This study
pGR002 2.0	8182	as pGR002 wt ZpPDC, but *pdc* codon harmonized	PDC expression	This study
pUCGT PET	8543	pUCG4.8 (pUCG18 based, smaller, containing oriT for conjugation) carrying the PET operon (ZpPDC2.0 + Gt ADH6)	PET operon expression (PDC/ADH expression)	This study
pJET1.2	2974	High efficiency blunt-ended cloning vector, amp^r	*E. coli* cloning/ holding vector	Fermentas (UK) [EF694056.1]

Name	Size (bp)	Description	Use	Source
pET28a(+)	5369	T7 RNA polymerase-dependent expression plasmid, kan^r, containing an N-terminal His-tag/thrombin cleavage site/T7-tag and an optional C-terminal His-tag, *lac* operator/T7 promoter control	Protein tagging and purification	Novagen, Addgene, Cambridge, MA, USA
pET28ZpwtPDC	6902	pET28a – wt *pdc* (Zp) – C-terminal thrombin cleavage site/3xGlycine linker/6His-tag	Protein tagging and purification	Hernández Gómez (2011)
pET28 Node 27	6923	pET28a – Node 27 *pdc* – C-terminal thrombin cleavage site/3xGlycine linker/6His-tag	Protein tagging and purification	This study

2.2 Microbiological Techniques

2.2.1 Growth Media

Components used in the preparation of growth media used in this study are listed in Table 2.3. Media were prepared using deionized water (Elix, Millipore, Billerica, MA, USA) and sterilized by autoclaving at 121°C for 15 min. To prepare agar plates, 1.5% (w/v) agar (Melford) was added before autoclaving. If preparing media with antibiotics, the autoclaved medium was allowed to cool to approximately 50°C before the addition of appropriate antibiotic, previously filter sterilized (0.22 µm, Millipore), and subsequently poured into 90 mm petri dishes (Sterilin, Caerphilly, UK). Plates were stored at 4°C for up to 1 month. *E. coli* carrying a plasmid was generally grown on media containing ampicillin (100 µg/ml) or kanamycin (50 µg/ml), as required. *G. thermoglucosidasius* strains carrying a plasmid were grown on media containing kanamycin (12 µg/ml).

Table 2.3 Growth media and buffers used in this study.
Abbreviations are: DMSO, dimethyl sulfoxide; PEG, polyethylene glycol.

Medium	Components	Reference
Lysogeny broth (LB)	10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, pH 7 with NaOH	Sambrook *et al.* (1989)
2TY	16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl, pH 7 with NaOH	Sambrook *et al.* (1989)
2SPYNG	16 g/L soy peptone, 10 g/L yeast extract, 5 g/L NaCl, pH 7 with NaOH	Novacta method 3, TMO Renewables Ltd., Guildford, UK
TSA	30 g/L (17 g/L pancreatic digest of casein, 3 g/L enzymatic digest of soya bean, 5 g/L NaCl, 2.5 g/L K_2HPO_4, 2.5 g/L glucose, pH 7.3)	Powder: Oxoid, UK
MYE (NBRC No. 945)	20 g/L maltose, 10 g/L yeast extract, 2 g/L KH_2PO_4, 5 g/L NaCl, pH 6	NBRC, Japan
Overnight Express™ (Instant Terrific Broth medium)	60 g/L granules, 10 ml/L glycerol (microwave until boil)	Granules: Novagen, Darmstadt, Germany
Electroporation buffer	91 g/L sorbitol, 91 g/L mannitol, 10 ml/L glycerol	Novacta method 3, TMO Renewables Ltd., UK
TSS buffer	5 g/50 ml PEG 8000, 2.5 ml/50 ml DMSO, 0.30 g/50 ml magnesium chloride hexahydrate ($MgCl_2 \cdot 6H_2O$), LB to 50 ml	Chung *et al.* (1989)
ASM (modified rich ammonium salts medium with triple buffer)	3.12 g/L (20 mM final concentration) monosodium phosphate (NaH_2PO_4 $2H_2O$), 1.74 g/L (10 mM) potassium sulphate (K_2SO_4), 1.68 g/L (8 mM) citric acid, 1.23 g/L (5 mM) magnesium sulphate ($MgSO_4 \cdot 7H_2O$), 0.09 g/L (0.08 mM) calcium chloride ($CaCl_2$), 0.0004 g/L (1.65 µM) sodium molybdate ($Na_2MoO_4 \cdot 2H_2O$) (or 250 µl of 10 mM stock, 2.419 mg/ml), 3.3 g/L (25 mM) ammonium sulphate ($(NH_4)_2SO_4$), 10 g/L (1%) glucose, 5 g/L (0.5%) yeast extract, 0.00292 g/L (12 µM) biotin, 1.5 g/L (5 mM) thiamine, 5 ml/L sulphate trace elements solution (pH 7 with potassium hydroxide (KOH) at 60°C, make to 940 ml, and filter sterilize), add 20 ml of filter sterilized 1 M MOPS (3-(N-morpholino)propanesulfonic acid), 1 M HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), and 1 M Bis-Tris (2-[Bisamino]-2--1,3-propanediol) each to make 1 L	
Sulphate trace elements solution	5 ml/L concentrated sulphuric acid (H_2SO_4), 1.44 g/L (25 µM in final ASM medium) zinc sulphate heptahydrate ($Zn\ SO_4 \cdot 7H_2O$), 5.56 g/L (100 µM) iron sulphate heptahydrate ($Fe\ SO_4\ 7H_2O$), 1.69 g/L (50 µM) manganese sulphate heptahydrate ($Mn\ SO_4 \cdot 7H_2O$), 0.25 g/L (5 µM) copper sulphate heptahydrate ($Cu\ SO_4 \cdot 7H_2O$), 0.562 g/L (10 µM) cobalt sulphate heptahydrate ($Co\ SO_4 \cdot 7H_2O$), 0.886 g/L (16.85 µM) nickel sulphate heptahydrate ($Ni\ SO_4 \cdot 7H_2O$), 0.08 g/L boric acid (H3BO3)	

2.2.2 Quantification of Bacterial Cell Density

Cell culture samples were diluted up to 1/20 in growth media depending on the growth stage, added to a 96-well plate (BD Falcon, San Jose, CA, USA) and the optical density at 600 nm (OD_{600nm}) analysed in a Synergy™ HT plate reader (BioTek, Winooski, NH, USA). The reading was corrected against growth medium as the background. For *G. thermoglucosidasius* an indicative OD of 1 approximates 0.25 g/L dry weight (Taylor 2008).

2.2.3 General Growth Conditions

Constant temperature, mixing and aeration in liquid culture were controlled using an Innova44 shaking incubator (New Brunswick Scientific, Edison, NJ, USA) at 250 rpm. Plates were incubated in static incubators (Harrow Scientific, London, UK).

2.3 Molecular Biology Techniques

2.3.1 Primers

Primer pairs were designed to have a similar melting temperature, structures without hairpins and free 3' ends, ideally with a G or C at the 3' end, and a low probability of forming a primer homo- or hetero-dimer, assessed using mfold online-software (http://mfold.rna.albany.edu) and an online oligo analyzer (http://eu.idtdna.com). All primers were supplied in lyophilised form, salt-free, by Eurofins (Munich, Germany). Table 2.4 gives a comprehensive list of the sequencing primers used in this study. Cloning primers can be found in the relevant chapters.

Table 2.4 Sequencing primers used in this study.
Abbreviations are: F, forward; R, reverse; Zp, *Zymobacter palmae* T109 ATCC1623.

Name	Use	Sequence 5' → 3'
	Standard Primers	
T7F	Sequencing pET28a(+)	TAA TAC GAC TCA CTA TAG GG
T7R	Sequencing pET28a(+)	CTA GTT ATT GCT CAG CGG T
M13F (21 uni)	Sequencing pUCG18 based plasmids	TGT AAA ACG ACG GCC AGT
M13R (49)	Sequencing pUCG18 based plasmids	GAG CGG ATA ACA ATT TCA CAC AGG
pJET1.2F	Sequencing pJET1.2	CGA CTC ACT ATA GGG AGA GCG GC
pJET1.2R	Sequencing pJET1.2	AAG AAC ATC GAT TTT CCA TGG CAG

Name	Use	Sequence 5’ → 3’
	Gene Specific Primers	
wtZpPDC F1	Sequencing wt ZpPDC	CTG TTG CCC TAG CGG ACC
wtZpPDC F2	Sequencing wt ZpPDC	CCG ACG CTG ATC GAA TGT
2.0 F1	Sequencing harmonized ZpPDC	TTA GGC TGC GCT GTC ACG ATT ATG
2.0 F2	Sequencing harmonized ZpPDC	CGG TGC TAG GGT CGA ATT AG
Node 27 1F	Sequencing Node 27	GGA AGA TCA TCC GGG CTA TG
Node 27 1R	Sequencing Node 27	ATA GCC CGG ATG ATC TTC C

2.3.2 DNA Extraction and Purification

Plasmid DNA was extracted from overnight cultures using the QIAprep® Spin Miniprep Kit (Qiagen, Hilden, Germany, Product code: 27106) following the standard protocol with one exception; the initial cell harvest was done by centrifugation at room temperature (~21°C) for 10 min at 4,000 rpm. For *G. thermoglucosidasius* plasmid extractions an additional step was introduced. The cell pellet was resuspended in buffer P1 (from the kit) + 1 mg/ml lysozyme (Sigma) and incubated at 37°C for 30 min before continuing with step 2 of the supplier’s protocol.

2.3.3 DNA Quantification and Visualization

DNA was visualized by electrophoresis in 1% agarose gels (see Table 2.5) containing 0.005% SYBR® safe (Invitrogen, Thermo Fisher Scientific, Product code: S33102) at 90 to 110 V for 60 min. The gel was imaged using a G:BOX (Syngene, Cambridge, UK) and the associated software Genesnap (Syngene). Gene Ruler™ 1 kb DNA ladder (Fermentas, Thermo Fisher Scientific, Product code: SM0311) was routinely used for size estimations.

Genomic or plasmid DNA preparations were routinely quantified using NanoVue™ (GE Healthcare, Chalfont St Giles, UK).

Table 2.5 Agarose gel preparations.
Abbreviations are: TAE, Tris-acetic acid-EDTA buffer; EDTA, ethylenediaminetetraacetic acid.

	Components
50x TAE	242 g/L Tris, 57.1 ml/L acetic acid, 18.6 g/L EDTA
Agarose gel	1% (w/v) agarose, 1xTAE
6x Loading buffer	30% (v/v) glycerol, 0.25% (w/v) bromophenol bue, 0.25% (w/v) xylene cyanol FF

2.3.4 DNA AMPLIFICATION

Colony PCR uses target-specific primers which allow the identification of clones containing the target gene. Colony PCR was regularly performed to check colonies for successful transformation, using KAPATaq Ready Mix DNA polymerase (Kapabiosystems, Woburn, MA, USA, Product code: KK1024) according to the supplier's instructions (in a 20 µl reaction: 10 µl ready mix containing loading buffer, reaction buffer with Mg^{2+}, 0.5 U Taq polymerase and 0.4 mM of each dNTP, 1 µl 10 µM primer each, 8 µl MilliQ water). The DNA template was added to the reaction by picking individual colonies with a sterile pipette tip or toothpick. The reaction was carried out in a thermocycler (Mastercycler, Eppendorf, Hamburg, Germany).

PCR amplification of genes targeted for cloning was routinely carried out using KAPA HiFi HotStart (Kapabiosystems, Product code: KK2501) according to the supplier's instructions. A 25 µl reaction contained 5 µl 5x KAPA HiFi buffer, 0.75 µl KAPA dNTP mix (10 mM each dNTP), 0.75 µl 10 µM forward primer, 0.75 µl 10 µM reverse primer, template DNA (10-100 ng genomic DNA, 1-10 ng for plasmid DNA), 0.5 U KAPA HiFi HotStart DNA polymerase and MilliQ water to make up the final volume.

Alternatively, genes of interest were amplified using Phusion® High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Product code: F-530L) or Phusion® Hot Start II High Fidelity DNA Polymerase (Thermo Fisher Scientific, Product code: F-549L) according to the supplier's instructions. A 50 µl reaction contained 10 µl 5x Phusion® HF buffer, 1 µl dNTP mix (10 mM each dNTP), 1 µl 10 µM forward primer, 1 µl 10 µM reverse primer, template DNA (10-100 ng genomic DNA, 1-10 ng for plasmid DNA), 0.5 U Phusion® DNA polymerase and MilliQ water to make up the final volume.

This method was also used in site-directed mutagenesis, using primers phosphorylated at the 5' end for direct ligation of the PCR product.

Occasionally, PCR products were ligated into the pJET1.2/blunt cloning vector (CloneJET PCR Cloning Kit, Thermo Fisher Scientific, Product no: K1231), according to the manufacturer's directions. This allows sequencing before proceeding further, as well as a more efficient digest compared to PCR products.

2.3.5 Restriction Digest and Ligation

1 µg of DNA was digested with desired restriction enzymes according to the supplier's instructions (Fermentas or NEB); generally the digest was incubated for 3 h at 37°C in a heat block. Fermentas Fast Digest enzymes were incubated at 37°C for 10 min.

Restriction digest reaction products were separated on a 1% agarose gel and purified from it using Zymoclean™ Gel DNA recovery kit (Zymo Research, Irvine, CA, USA) following the supplier's instructions.

Linearized DNA was ligated in 20 µl reactions using T4 DNA ligase and an appropriate buffer (Fermentas) and incubation at room temperature (~21°C) for 10 to 30 min. 5-10 µl of the ligation reactions were transformed into high efficiency cloning strains, such as *E. coli* BioBlue.

The constructs were passaged through *E. coli* BioBlue to allow for DNA methylation prior to transformation into *G. thermoglucosidasius,* thus preventing endonuclease degradation in *G. thermoglucosidasius.*

2.3.6 Preparation of Competent Cells

G. thermoglucosidasius electrocompetent cells were prepared based on Novacta method 3 (Novacta, Biosystems Ltd., Herts, UK). A starter culture was set-up in 5 ml pre-warmed 2SPYNG by scraping cells off a spread plate with confluent growth, and incubated at 60°C, 250 rpm for 1-2 h. 1-2 ml of the starter culture were used to inoculate 50 ml pre-warmed 2SPYNG in a 250 ml baffled flask (starting OD_{600nm} ~0.1) and incubated at 60°C with shaking at 250 rpm. The OD_{600nm} was monitored to 1.4, at which point the flask was cooled on ice for 10 min. The culture was aliquoted into pre-chilled 1.5 ml microcentrifuge tubes. Cells were collected by centrifugation at 10,000 rpm at 4°C for 1 min and washed 3x with 1 ml ice-cold electroporation buffer. After the final wash step the cells were resuspended in 100 µl ice-cold electroporation buffer and used immediately for transformation or were stored at -80°C.

E. coli chemically competent cells were prepared based on Chung *et al.* (1989). An overnight culture was used to inoculate 50 ml LB in a 250 ml conical flask and grown to OD_{600nm} ~0.5 at 37°C with shaking at 250 rpm. When the culture had reached the desired OD the flask was cooled on ice for 10 min before harvesting the cells by centrifugation for 10 min at 3,000 rpm, 4°C. The supernatant was removed and the cell pellet was resuspended in ice-cold TSS buffer. Aliquots were added to pre-chilled microcentrifuge tubes and either transformed immediately or stored at -80°C until use.

2.3.7 TRANSFORMATION

Electrocompetent *G. thermoglucosidasius* were prepared as described and transformed using electroporation. Electrocompetent cells (50-100 µl) were mixed with 1-2 µl (100-500 ng) plasmid DNA using gentle stirring. This mixture was then added carefully to a pre-chilled electrocuvette (Cellprojects, EP-101, Harrietsham, UK) and kept on ice. The cuvette was tapped gently to ensure no air bubbles were trapped in the gap. The cuvette was wiped with a cloth to remove water and condensation before placing it into the electropod. An electric pulse was delivered using the exponential decay protocol, 2,500 V (25 kV/cm), 10 µF capacitance, 600 Ω resistance (Gene Pulser Xcell™, Bio-Rad, Hercules, CA, USA). Immediately after delivering the electric pulse, 1 ml pre-warmed 2TY was added to the cuvette and the suspension transferred to a 50 ml falcon tube and recovered at 60°C, 250 rpm for 2 h. After recovery 200 µl cells were plated out on agar plates (e.g., TSA containing 12 µg/ml kanamycin). Plates were incubated at 60°C for up to 48 h (covered in tinfoil to avoid evaporation).

Chemically competent *E. coli* was transformed using heat-shock (Chung *et al.* 1989). 1-2 µl (100-500 ng) plasmid DNA or ligation reactions were added to chemically competent *E. coli* (50-100 µl) and mixed gently. The mixture was incubated on ice for 5-30 min, heat-shocked at 42°C for 30-45 s and cooled on ice for a further 2 min. 1 ml LB broth was added and the cells were allowed to recover at 37°C for 1 h. After recovery, 200 µl cells were plated out on LB agar plates containing the appropriate antibiotic and incubated at 37°C for up to 24 h.

2.3.8 SEQUENCING

Samples of plasmid DNA were prepared according to required standards (50-100 ng/µl, 15 µl) and sent for sequencing by Eurofins. T7F and T7R were used for pET28 plasmids, M13F (21 uni)

and M13R (49) for all pUCG18 derivatives. If a non-standard gene- or vector-specific sequencing primer was required, this was added to the sample (2 µl of 100 pmol/µl).

2.3.9 RNA Extraction and cDNA Preparation

Cells were grown to the desired OD and RNAprotect® (Qiagen, Product code: 76506) added immediately to a culture sample to stabilise the RNA (OD_{600nm} 1.5 to 3, 6ml culture : 20 ml RNAprotect). This mix was incubated for 5 min at T_{room} before harvesting the cells by centrifugation at 4,000 rpm, for 10 min. The supernatant was discarded and the cell pellets stored at -80°C until use.

Cell lysis was performed by resuspending the cell pellet in 250 µl lysis buffer (30 mM Tris-HCl, 1 mM EDTA, pH 8, 15 mg/ml lysozyme (Sigma), 7.4% proteinase K (Qiagen, Product code: 19133)) and incubation at T_{room} for 10 min with regular vortexing (10 s every 2 min). 750 µl RTL + 1.5% ß-mercaptoethanol were added and the sample vortexed vigorously before adding 500 µl 96-100% ethanol. This solution was gently mixed by pipetting.

RNA was extracted using the RNeasy® kit (Qiagen, Product code: 74104), following the supplier's instructions for purification of total RNA from yeast, starting at step 2 and including the optional on-column DNase digestion step using the RNase-Free DNase Set (Qiagen, Product code: 79254).

cDNA was prepared using the High capacity cDNA reverse transcription kit (Applied Biosystems, Thermo Fisher Scientific, Product code: 4368814) following the supplier's instructions.

2.3.10 RT-qPCR

A list of RT-qPCR (reverse transcription quantitative PCR) primers used in this study can be found in Table 2.6. The RT-qPCR primers were designed using Primer3Plus (http://http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) based on the following criteria: pair towards 3' end, product size of 60-150 bp, GC content 50-60%, length of around 20 bp, melting temperature of around 60°C, max 3' self-complementary: 1, and max poly-x: 3.

Table 2.6 RT-qPCR primers used in this study.
Abbreviations are: F, forward; R, reverse; T_m, melting temperature.

Name	T_m (°C)	Product size (bp)	Sequence 5' → 3'
PDC4F	60.2	87	CGG CAG CAT TAG CTG AGA A
PDC4R	60.9		GCT CTG CCG CCT CTA TAC CT
ADH2F	60.7	192	CGA TAA CTG TGC CAC CTT CTG
ADH2R	60.6		GCA ATT TCC CTC TGG CAT C
RecN1F	61.5	156	CGT TGT CGG TTT CGT TTG AC
RecN1R	61.7		GCC CTT CTA TTT CCG CCT TT

20 µl RT-qPCR reactions were prepared using LuminoCt® SYBR®Green qPCR Ready Mix™ (Sigma, Product code: L6544) in 96-well plates (Thermo Fisher Scientific, Product code: AB1100/W) following the supplier's instructions, and run on a Chromo4™/DNA Engine® (Bio-Rad) with the accompanying software Opticon Monitor™.

The threshold cycle (C_T) data was analysed using relative quantification normalized to a reference gene and the Pfaffl method. Strains expressing pUCG18 (empty vector) were tested for the genes of interest and served as the "calibrator". *Rec*N codes for a DNA repair and recombination function, and serves as the housekeeping or "reference" gene. The following equation was used to give the expression ratio between the sample and calibrator:

$$Ratio = \frac{(E_{target})^{\Delta C_T, target\ (calibrator-test)}}{(E_{reference})^{\Delta C_T, reference\ (calibrator-test)}}$$

2.4 Protein Biochemistry

2.4.1 Protein Analysis Buffers and Preparations

Table 2.7 summarises buffers and solutions used for protein analysis in this study. Preparations were made using MilliQ deionized water.

Table 2.7 Protein analysis buffers and preparations.
Abbreviations are: APS, ammonium persulfate; EDTA, ethylenediaminetetraacetic acid; SDS, sodium dodecyl sulphate; TEMED, tetramethylethylenediamine; TPP, thiamine pyrophosphate; MES, 2-[N-Morpholino] ethanesulfonic acid.

	Components
Protein Gels:	
10x SDS running buffer	144 g/L glycine, 30 g/L Tris, 100 ml/L 10% (w/v) SDS
Coomassie Blue staining solution	0.25% (w/v) Coomassie Brillant Blue (R250, Sigma), 45% (v/v) methanol, 10% (v/v) acetic acid, 45% (v/v) water
12% Tris-glycine running gel (10 ml preparation for 1 gel, 1 mm gap)	3.3 ml water, 4 ml 30% acrylamide mix (National diagnostics, ProtoGel, Hessle, UK), 2.5 ml 1.5 M Tris-Cl pH 8.8 (Tris base, pH adjusted with HCl), 0.1 ml 10% SDS, 0.1 ml 10% (w/v) APS, 4 µl TEMED (National diagnostics)
Stacking gel (4 ml preparation for 1 gel)	2.7 ml water, 0.67 ml 30% acrylamide mix, 0.5 ml 1 M Tris pH 6.8, 0.04 ml 10% SDS, 0.04 ml 10% (w/v) APS, 4 µl TEMED
Destaining solution	30% (v/v) methanol, 10% (v/v) acetic acid
SDS loading buffer	63 mM Tris-HCl, 10% (w/v) glycerol, 2% (w/v) SDS, 0.0025% (v/v) Bromophenol Blue
Enzyme Assays:	
MES	1.2 g/L TPP (3 mM), 2.4 g/L magnesium sulphate ($MgSO_4 \cdot 7H_2O$) (20 mM), 9.76 g/L MES (50 mM), pH 6.5 with KOH
Sodium phosphate buffer	22.3 g/L tretra-sodium pyrophosphate (50 mM), pH 8.5
Citric acid buffer	9.6 g/L citric acid (50 mM), pH 6
Protein Purification:	
His-bind buffer	20 mM Tris pH 8.0, 300 mM NaCl, 20 mM imidazole
His-elute buffer	20 mM Tris pH 8.0, 300 mM NaCl, 1 M imidazole

2.4.2 Expression of Recombinant Protein

Seed cultures of *E. coli* strains carrying expression vectors were inoculated with a single colony from a fresh streak plate, and used to inoculate 50 ml Overnight Express™ + 100 µg/ml kanamycin (+ 5 mM thiamine if expressing a PDC) in a 250 ml conical flask (or 500 ml in a 2 L flask), 250 rpm, 30°C for 16 h. Cells were harvested by centrifugation at 4,000 rpm at 4°C for 20 min. Cell pellets were stored at -80°C until use. Thawed cells were resuspended in assay

buffer (10% of the culture volume) supplemented with a protease inhibitor cocktail (cOmplete™ Mini EDTA-free, Roche, Mannheim, Germany, Product code: 11836170001). Cells were lysed using 4 pulses of sonication at 12 µm for 15 s (exponential microprobe, Soniprep 150 plus, MSE, London, UK). The lysate was cleared of cell debris by centrifugation at 13,000 rpm at 4°C for 10 min. Supernatant containing soluble protein was aliquoted and used immediately for enzyme analysis or stored at -20°C until further use.

A starter culture of *G. thermoglucosidasius* strains carrying expression vectors was set-up in 5 ml pre-warmed 2TY + 12 µg/ml kanamycin by scraping cells off a spread plate with confluent growth, and incubated at 60°C, with shaking at 250 rpm for 1-2 h. This starter culture was then used to inoculate 50 ml 2TY + 12 µg/ml kanamycin + 5 mM thiamine in 250 ml baffled flasks, with shaking at 250 rpm, until an OD_{600nm} 1.5-2.5 was reached. This was considered to be "aerobic" growth conditions. Cells were harvested by centrifugation at 4,000 rpm, 4°C, for 20 min. Cell pellets were stored at -20°C until use. Thawed cells were resuspended in assay buffer (5% of the culture volume) supplemented with a protease inhibitor cocktail (cOmplete™ Mini EDTA-free). Cells were lysed using 4-6 pulses of sonication at 12 µm for 15-20 s (exponential microprobe, Soniprep 150 plus). The lysate was cleared of cell debris by centrifugation at 13,000 rpm, 4°C, for 30 min. The supernatant containing soluble protein was aliquoted and used immediately for enzyme analysis.

2.4.3 Protein Purification

Cell pellets were resuspended in His-bind buffer at a concentration of ~0.4 g/ml, sonicated with 4-6 x 20 s bursts at up to 16 µm (exponential microprobe, Soniprep 150 plus) and centrifuged for 10-20 min at 13,000 rpm and 4°C to obtain a soluble fraction.

For small-scale purifications His-tagged protein was purified using Talon® metal-affinity resin (cobalt charged, Clontech, Saint-Germain-en-Laye, France, Product code: 635502). 2 ml of resin suspension were used to provide 1 ml of column volume and 3 mg maximum protein binding capacity. The column was washed with 20 column volumes of MilliQ water and then 20 column volumes of His-bind buffer to equilibrate. The soluble fraction was applied to the column and the flow-through containing unbound protein was collected. The column was washed again with 20 column volumes of His-bind buffer to remove any unbound protein. The bound protein was then eluted using increasing concentrations of imidazole in 1.5 column volumes (see Table 2.8).

Table 2.8 Elution buffers for protein purification.

His-bind buffer (ml)	His-elute buffer (ml)	Final Imidazole Concentration (mM)
9.6	0.4	50
9.1	0.9	100
8.1	1.9	200
5.1	4.9	500
0.0	10.0	1000

The column was washed with 4 volumes of 20 mM MES pH 5.0, 300 mM NaCl to remove imidazole and protein, and stored in 20% (v/v) ethanol.

For large-scale protein expression experiments, the His-tagged protein was purified using a HisTrap™ HP 5 ml column on an ÄKTA explorer FPLC system (GE Healthcare) calibrated with His-bind buffer prior to loading the soluble protein fraction. Bound protein was eluted with increasing concentrations of imidazole. Eluted protein was monitored at 280 nm.

The eluted protein was further purified and the buffer exchanged by size-exclusion chromatography using a Superdex™ 200 10/300 GL gel filtration column on an ÄKTA explorer FPLC system, again monitoring elution at 280 nm. The column was equilibrated with MES buffer, pH 6.5, containing 20 mM $MgSO_4$ and 3 mM TPP.

2.4.4 Protein Quantification and Visualization

Total protein in cell lysates was quantified using a Bradford protein assay (Protein Assay, Bio-Rad, Product code: 500-0006), according to the supplier's instructions. Standards were made from bovine serum albumin (BSA standard, Thermo Fisher Scientific, Product code: 23210), 0-10 µg/ml. Samples were incubated at room temperature for 20 min. Absorbance at 595 nm was measured in a spectrophotometer (Varian Cary® 50 Bio, Agilent Technologies, Santa Clara, CA, USA) and protein concentrations extrapolated from the standard curve.

Protein fractions were visualized using SDS-PAGE (sodium dodecyl-polyacrylamide gel electrophoresis) according to Sambrook *et al.* (1989) using a 12% (v/v) running gel (see Table 2.7) and a Mini-PROTEAN electrophoresis system (Bio-Rad Laboratories GmbH, Munich, Germany). Unstained protein marker (Thermo Fisher Scientific, Product code: 26610) was run with the samples for size comparison. Samples and marker were denatured in SDS loading buffer at 95°C for 3 min, immediately after adding the loading buffer to prevent unwanted

proeteolysis. The gel was routinely run at 75 mA for ~25 min and stained with Coomassie Blue staining solution with gentle agitation (15-30 min, destain in destain solution for 30-90 min). Gel imaging was done using the G:BOX and accompanying software (Syngene) with a UV filter and set to 0-100 ms.

2.4.5 Enzyme Activity Assays

Enzyme activities were measured using a Varian Cary® 50 or Cary® 60 Bio UV/visible light spectrophotometer and a single-cell peltier temperature controller (Agilent Technologies), with the accompanying software CaryWinUV Kinetics Application 3.00.

It was confirmed that under all assay conditions the rate of reaction was directly proportional to the enzyme concentration.

2.4.5.1 PDC Activity Assay

The PDC activity assay is a coupled enzyme assay that follows the oxidation of NADH to NAD^+ spectrophotometrically at 340 nm (Pohl *et al.* 1994) (Figure 2.1). 1 mole of NADH is oxidised per mole acetaldehyde reduced, thus the rate of NADH oxidation is directly related to PDC activity.

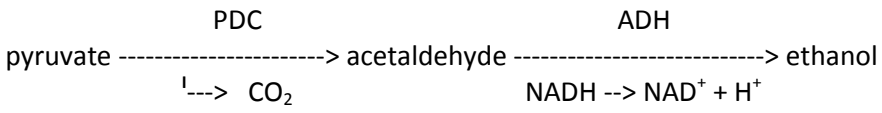


Figure 2.1 Schematic representation of the pyruvate decarboxylase coupled assay.
Acetaldehyde production by pyruvate decarboxylase (PDC) was coupled to the activities of alcohol dehydrogenase (ADH) and the rate of NADH oxidation was measured at 340 nm.

The standard assay was carried out at 30°C using commercially available ADH (*Saccharomyces cerevisiae* ADHs, Sigma). See Table 2.9 for volumes and concentrations used in the assay. The background rate was measured for 1 min before initiating the reaction with the addition of the substrate, sodium pyruvate. The reaction rate was measured over 2 min.

Table 2.9 PDC activity assay set-up. Substrates were dissolved in buffer.

1 ml final volume	Volume (µl)	Final Concentration
MES (pH 6.5)	875 (up to 1 ml)	
ßNADH (4.5 mM, 3.15 mg/ml)	34	0.15 mM
ADH (1,500 U/ml, 4.4 mg/ml)	7	10 U/ml
PDC/cell lysate	<50	
Sodium pyruvate (0.5 M, 110 mg/ml)	34	16.9 mM

Specific activity (µmol/min/mg) was calculated using the following equation:

$$\frac{\text{(reaction rate – background rate) * dilution factor}}{\text{total protein (mg/ml) * 6.22 mM}^{-1}\text{ cm}^{-1}\text{ (ε of NADH) * 1 cm (cuvette path length)}}$$

2.4.5.2 ADH Activity Assay

The assay was carried out using volumes and concentrations as specified in Table 2.10 below. The background rate was measured for 1 min before initiating the reaction with the addition of the substrate, acetaldehyde or ethanol. The reaction rate was measured over 2 min.

Table 2.10 ADH activity assay set-up.

1 ml final volume	Volume (µl)	Final Concentration
ADH activity assay using acetaldehyde (substrates dissolved in buffer)		
Citric acid (50 mM, pH 6)	850 (up to 1 ml)	48 mM
ßNADH (4 mM, 2.82 mg/ml)	50	0.2 mM
ADH/cell lysate	<50	
Acetaldehyde (4 M, 0.176 g/ml)	50	200 mM
ADH activity assay using ethanol (substrates dissolved in MilliQ water)		
MilliQ water	Up to 1 ml	
Sodium phosphate (50 mM, pH 8.5)	167	10 mM
NAD^+ (15 mM, 10.27 mg/ml)	33	0.5 mM
ADH/cell lysate	<50	
Ethanol (3 M)	33	100 mM

Specific activity (µmol/min/mg) was calculated using the following equation:

$$\frac{\text{(reaction rate – background rate) * dilution factor}}{\text{total protein (mg/ml) * 6.22 mM}^{-1}\text{ cm}^{-1}\text{ (ε of NADH) * 1 cm (cuvette path length)}}$$

2.4.5.3 Catechol Assay

A simple catechol conversion assay allows a quick test of expression of the reporter gene *phe*B under different growth conditions. The *phe*B gene encodes catechol 2,3-dioxygenase. This enzyme catalyzes the conversion of catechol into 2-hydroxymuconic semialdehyde, which has a vivid yellow colour (Bartosiak-Jentys *et al.* 2012). Colonies, pellets or lysates of cells expressing *phe*B were easily identified by a colour change after the addition of several drops of 100 mM catechol (dissolved in deionized water, Sigma).

2.4.6 Determining Kinetic Parameters

The enzyme kinetics module of SigmaPlot (Systat Software Inc, London, UK) was used to determine kinetic parameters. Where enzyme kinetics followed Michaelis-Menten kinetics, the parameters were determined using non-linear regression based on the Michaelis-Menten equation:

$$v = \frac{V_{max}[S]}{K_M + [S]}$$

2.4.7 Thermal Shift Assays

Thermal shift assays used dilute SYPRO®-Orange (90x final concentration) (Thermo Fisher Scientific, Product code: S6650) and a protein concentration of 0.1 mg/ml in a total volume of 25 µl in a desired buffer in optically clear Genie® II tubes (OptiGene, Horsham, UK). This reaction was cycled from 25-105°C, ramping at 0.05°C/s, and the fluorescence was measured at 470 nm excitation and 555 nm emission in a Genie® II real-time fluorescence detection instrument (OptiGene), kindly provided by Dr. Nick Morant (GeneSys Biotech Ltd., Camberely, UK).

SYPRO®-Orange is a fluorescent dye that specifically binds to hydrophobic residues of a protein. As the protein unfolds, more residues become available, which results in a significant increase in fluorescence emission (Figure 2.2). A denaturing temperature (T_d) can be assigned by using a derivative plot (rate of change in fluorescence with respect to change in temperature). The T_d is the temperature at which the rate of change in fluorescence with change in temperature is at its maximum, i.e., the temperature at the peak in the derivative plot.

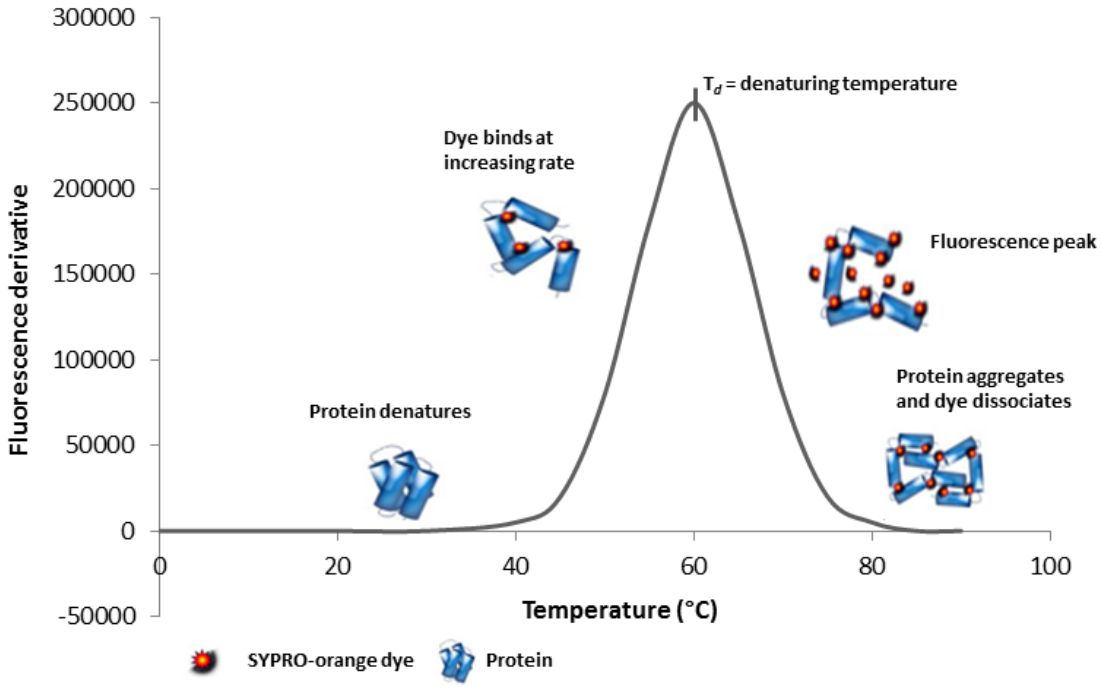


Figure 2.2 Schematic representation of the thermal shift assay showing the intercalation of SYPRO®-Orange. Shown is the derivative plot, i.e. the rate of change in fluorescence with respect to change in temperature. The T_d is the temperature at which the rate of change in fluorescence with change in temperature is at its maximum. Adapted from Morant (2014).

3. Characterization and Crystal Structure of the *Zymobacter palmae* Pyruvate Decarboxylase

3.1 Introduction

PDCs are relatively wide-spread in plants and fungi, but are rarely found in bacteria. At the time of writing, only six bacterial PDCs have been described, including the one found in *Zymobacter palmae* (ZpPDC). The *Zymomonas mobilis* enzyme (ZmPDC) has been extensively studied, with a variety of structural variants published (PDB entry 1ZPD, Dobritzsch *et al.* 1998; 2WVA, 2WVG, 2WVH, Pei *et al.* 2010; 3OEI, Meyer *et al.* 2010; 4ZP1, Wechsler *et al.* 2015). Other bacterial PDCs include those from *Acetobacter pasteurianus* (ApPDC, PDB entry 2VBI), *Gluconoacetobacter diazotrophicus* (GdPDC, PDB entry 4COK, van Zyl *et al.* 2014a), *Gluconobacter oxydans* (GoPDC, van Zyl *et al.* 2014b), and the only known Gram-positive species possessing a PDC, *Sarcina ventriculi* (SvPDC, Lowe & Zeikus 1992).

ZpPDC was first described by Raj *et al.* (2002). Recombinantly expressed ZpPDC displayed the highest specific activity (130 U/mg of protein) and lowest K_M for pyruvate (0.24 mM) when compared to other bacterial PDCs (Raj *et al.* 2002). Furthermore, in a comparison of irreversible denaturation effects with increasing temperature, ZpPDC exhibited the highest thermostability of these PDCs (see General Introduction section 1.4 for more detail). It retained 100% activity after 30 min incubation at 60°C and 80% activity after incubation at 65°C for 30 min.

ZpPDC being one of the most thermostable bacterial PDCs currently known prompted attempts to utilize ZpPDC to produce a homo-ethanologenic *Geobacillus* spp. In the first attempt Taylor *et al.* (2008) expressed ZpPDC aerobically in *G. thermoglucosidasius* grown at 45-50°C, and assayed the clarified cell extract for PDC activity by monitoring NADH depletion dependent decrease in absorbance at 340 nm (standard coupled assay). They found that with an increase in growth temperature from 45°C to 48°C PDC activity dropped sharply from 1067 to 89 nmol/min/mg total protein. No activity was detectable in cells grown at 50°C.

The disparity between high *in vitro* thermostability and limited *in vivo* expression of active PDC at higher temperatures is striking, and encouraged further investigation into ZpPDC thermostability and thermoactivity presented in this chapter, together with attempts to optimize expression in *G. thermoglucosidasius* (described in Chapter 4).

Furthermore, the ZpPDC crystal structure is of great interest as it gives vital information to further our understanding of bacterial PDCs. This may allow for the rational design of PDCs with improved thermostability and thermoactivty by improving the TPP-binding site efficiency at higher temperatures.

No published crystal structure was previously available for the ZpPDC. This chapter details the characterization of kinetic and thermal properties of the ZpPDC from *Z. palmae* ATCC 51623 recombinantly expressed and purified from *E.coli* and the determination of the crystal structure of the ZpPDC by X-ray crystallography to 2.15 Å resolution. The structure was submitted to the PDB (entry: 5EUJ, Buddrus *et al.* 2016). General concepts of X-ray crystallography are briefly introduced below.

This work is based on a collaboration between the University of Bath, UK, and the University of Waikato, New Zealand, supported by a Microbiology Society research visit grant (RVG14-10). The author particularly wishes to acknowledge support from Dr. Emma Andrews and Professor Vic Arcus at the University of Waikato, New Zealand, and Dr. Susan Crennell at the University of Bath.

Structure Determination by X-ray Crystallography

X-ray crystallography is a widely-used technique. The prerequisite for solving the 3-dimensional structure of a protein is a well-ordered crystal made of repeating protein units that will strongly diffract X-rays so that a diffraction pattern can be produced from which the structure can be determined.

Nucleation and crystal growth are affected by a number of environmental conditions, including protein and precipitant concentration, buffer type, pH, temperature, and the presence and concentration of metal ions or ligands. They must all allow the protein to be under supersaturated conditions. Failing to meet these conditions results in the protein remaining soluble or precipitating out of solution (Figure 3.1).

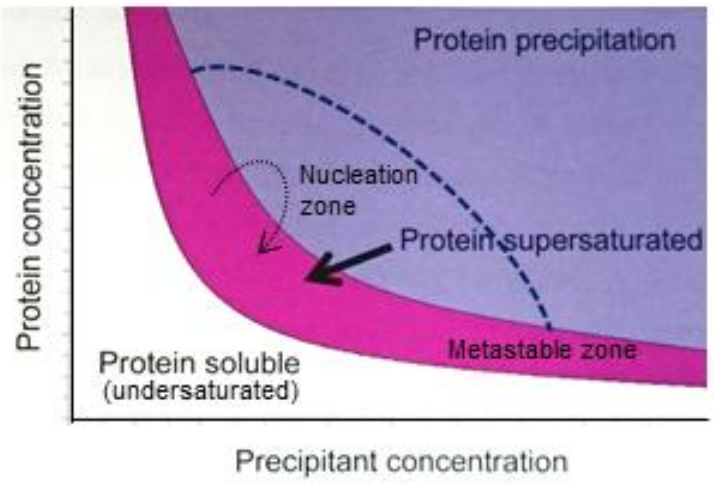


Figure 3.1 Crystallisation phase diagram. The diagram shows theoretical conditions for crystal nucleation and growth. A protein spontaneously precipitates under unfavourable conditions. However, in the nucleation zone a small number of crystal nuclei tend to form. If they re-enter the metastable zone, in which the protein is supersaturated, they can grow and form ordered crystals. During vapour-diffusion, water evaporates shifting the conditions to the nucleation zone allowing nuclei to form. This causes the protein concentration in the liquid phase to decrease and the system shifts back into the metastable zone. This is indicated by the trajectory of the dotted arrow. Adapted from Extance (2012) and Sherwood & Cooper (2011).

Vapour-diffusion uses a hanging or sitting drop of protein solution mixed with precipitant solution suspended above an undiluted precipitant solution reservoir. These are left to equilibrate during which water diffuses from the drop to the reservoir, gradually changing the conditions experienced by the protein. If successful, supersaturating conditions are eventually met and nucleation followed by crystal growth occurs.

For X-ray diffraction experiments a beam of parallel X-rays is directed onto the crystal. Heat and free radicals generated from the primary X-ray causes damage to the crystal. To minimize this, the crystal is cooled to 100 K using a stream of nitrogen gas. Before mounting the crystal, it is soaked in cryoprotectant, such as a glycerol-containing buffer. This replaces some of the water in the crystal and prevents the formation of ice that would destroy the crystal.

When the primary X-ray beam hits the crystal most of it passes straight through the crystal. However, some X-rays interact with the electrons on each atom causing them to oscillate. This oscillation creates new sources of X-rays that are scattered in all directions. Scattered X-rays that positively interfere with one another create a diffracted beam which can be recorded on a detector. The X-ray diffraction experiment aims to determine the intensity of the diffracted

beam. This information allows the deduction of the atom position within the crystal. The rules for diffraction are given by Bragg's law (Bragg & Bragg 1913).

The crystal is highly ordered and made up of repeating units (also known as unit cells). The unit cells are arranged such that each corner of each unit cell is on one of the planes in a set of parallel planes throughout the crystal (Figure 3.2). Bragg's law stipulates that X-rays reflected from adjacent planes (at the same angle) travel different distances and diffraction only occurs if the difference in distance is equal to (a multiple of) the wavelength of the X-ray beam. Only those scattered X-rays with the same diffraction angle that positively interfere with one another give rise to diffracted beams and thus distinct diffraction spots on the detector. Each diffraction spot position on the detector relates to a specific set of planes through the crystal.

Bragg's law describes the relation between the reflection angle (Θ), the distance between the planes (d), and the wavelength of the X-ray beam (λ) by $2d * sin\Theta = \lambda$. This can be used to determine the size of the unit cell (Branden & Tooze 1998).

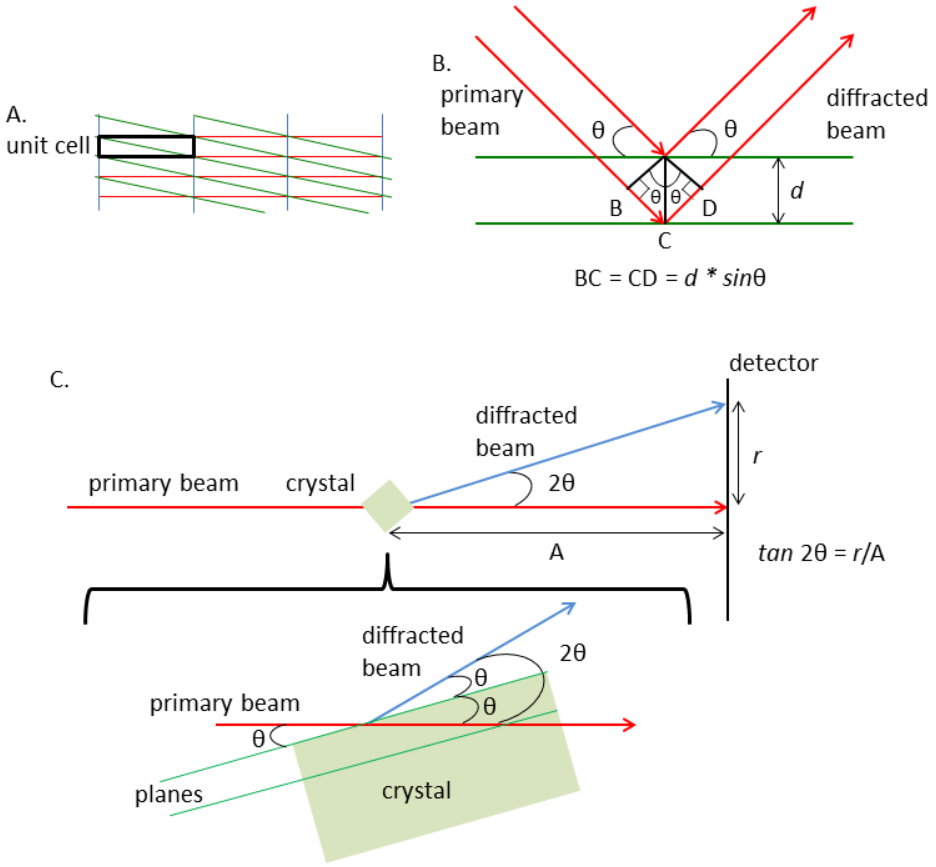


Figure 3.2 Diffraction of X-rays by a crystal. A. The diagram shows 3 simple sets of parallel planes (red, blue and green). Each corner of all unit cells is on one of the planes of the set. **B.** As the primary X-ray beam passes through the crystal, some of it is reflected by the atoms in the crystal. Depicted in green are 2 planes separated by the distance *d*. The primary beam hits the crystal at the angle Θ. The reflected beam leaves the crystal at the same angle, the reflection angle. X-rays being reflected from the lower plane will have to travel further than X-rays reflected from the upper plane to reach the detector. The difference in distance is given by BC + CD, which is equal to 2*d* * *sin*Θ. Bragg's law stipulates that diffraction only occurs if the difference in distance is equal to the wavelength of the X-ray beam (λ), i.e., 2*d* * *sin*Θ = λ. **C.** The reflection angle can be calculated from the distance between the diffraction spot and the location of the primary beam on the detector or film. This information can be used to calculate the unit cell size by measuring reflections obtained from a specific set of planes that is separated from any 2 adjacent planes by the length of one of the unit cell axes by using *d* = λ/(2**sin*Θ). This has to be repeated for the other 2 axes of the unit cell by reorienting the crystal and measuring further 2 sets of reflections. Adapted from Branden & Tooze (1998).

The information recorded as diffraction spots is actually the intensity of the diffracted beams, which is proportional to the square of the amplitudes, and given in terms of the structure factor *F(hkl)* (*h, k,* and *l* are known as the Miller indices denoting a set of crystal planes). To obtain all possible diffraction spots, the crystal is rotated in the beam so that the beam hits the crystal from many different directions. The diffractometer fitted with an area detector records

many 2-dimensional images of data, which are then computationally combined to give a complete diffraction pattern for the crystal.

Each diffracted beam recorded as a spot in the diffraction pattern has 3 properties: the amplitude, which is in proportion to the intensity of the spot as described above, the wavelength, which is determined by the X-ray source, and the phase. The phase is the relation of a beam's interference (positive or negative) to other beams, and is lost in the experimental data, thus creating the phase problem. The study presented here employed molecular replacement to solve the phase problem. This is a commonly used method that relies on comparing the Patterson map of the unknown structure and that calculated for a previously solved, similar structure to determine the orientation and positioning of molecules within the new unit cell (Sherwood & Cooper 2011).

From this initial phasing model, phases can be calculated, and together with the amplitude electron density maps can be derived. Using these data one can extract the information needed to assign the positions of atoms in the structure.

The initial model undergoes several rounds of manual building and computational refinement to ensure the best fit. This reduces the difference between experimentally obtained diffraction amplitudes and computationally derived diffraction amplitudes for a crystal containing the hypothetical model. The difference is given as the R factor, or residual agreement, given by the following equation:

$$R = \frac{\sum \big| |F_o| - |F_c| \big|}{\sum |F_o|}$$

F_o is $F_{observed}$; F_c is $F_{calculated}$, F being the structure factor.

The lower the R factor, the better the built model fits the experimental data.

3.2 METHODS

3.2.1 CLONING WT ZP *PDC* FOR EXPRESSION IN *E. COLI*

An *E. coli* expression construct for purification of ZpPDC was made by Mercedes Clara Hernández Gomes (Imperial College London). Sequencing of this construct and comparison to the GenBank entry AF474145.1 revealed that it contained a mutation changing amino acid 292 from valine to alanine. The mutation was corrected by site-directed mutagenesis using Phusion® High Fidelity DNA polymerase in a PCR following the supplier's instructions with an annealing temperature of 64.8°C and a 3 min extension at 72°C (see General Methods for details). Using primers ZpwtPDC A292V F (TAT GCT ACC GTT GGC TGG AAC) and R (GTC GTT GAA TAC CGG TGC), and pET28 wtPDC as the template, the nucleotide sequence for amino acid 292 was corrected from GCT (encoding alanine) to GTT (encoding valine).

Vector and gene specific primers (T7F and T7R, ZpwtPDC F1 and F2) were used in sequencing to confirm the sequence of pET28 ZpwtPDC was correct before transforming the plasmid containing the corrected gene sequence into *E. coli* BL21 (DE3) cells for protein expression.

3.2.2 ENZYME CHARACTERIZATION – KINETIC AND THERMAL PROPERTIES OF ZPPDC

The ZpPDC protein was expressed in *E. coli* BL21 (DE3) and purified through nickel-affinity chromatography, followed by further purification and buffer exchange through size-exclusion chromatography, as detailed in General Methods.

Kinetic properties of ZpPDC were determined at 30°C using a range of pyruvate concentrations following NADH depletion at 340 nm in the standard coupled assay with ADH (see General Methods).

To assess temperature-dependent PDC activity, assays were carried out in a 1 ml quartz cuvette at temperatures between 30 and 75°C following pyruvate depletion at 320 nm (based on Gocke *et al.* 2009).

The temperature at which the PDC was irreversibly denatured was assessed by incubating protein samples at 50°C, 55°C, 60°C, 65°C and 70°C for 30 min, then cooling the sample on ice and assaying using the standard coupled assay at 30°C.

Denaturing temperatures were determined by a fluorescence-based thermal shift assay using SYPRO®-Orange as described in General Methods.

3.2.3 Crystallisation of ZpPDC

Preliminary crystallisation screens were set up using a Mosquito liquid handler (TTP LabTech, Melbourn, UK) in the 96-well format using Index™, PEGRx HT™, SaltRx HT™ or Crystal Screen HT™ (Hampton Research, Aliso Viejo, CA, USA), and the sitting-drop vapour-diffusion method. Well solutions were tested in triplicate. The plate was sealed and incubated at 18°C. Crystal hits were achieved within minutes and were most promising in the PEGRx HT™ screen, in particular A5 (0.1 M BIS-TRIS pH 6.5, 25% (v/v) PEG 300), B6 (0.1 M HEPES pH 7.5, 30% (w/v) PEG 1,000), and C5 (0.1 M sodium citrate tribasic dehydrate pH 5.5, 18% (w/v) PEG 3,350).

From these initial hits further fine screens were set up using the hanging-drop vapour-diffusion method in 24-well XRL crystallisation plates with a 2 µl drop (1:1 protein:buffer) on siliconised glass cover slips above 400 µl reservoirs and sealed with vacuum grease. Optimization was generally based around buffer and PEG concentration.

Optimized crystallisation conditions:

Crystals of the purified ZpPDC were obtained using the hanging-drop vapour-diffusion method at 18°C (291 K). Purified ZpPDC (4.8 mg/ml) was incubated with 2 mM pyruvate for 30 min at room temperature (293 K), and then 2 µl drops of the enzyme/pyruvate mixed with crystallisation solution (0.15 M sodium citrate, pH 5.5, 14% (w/v) PEG 3350) in a 1:1 ratio were placed onto siliconised glass cover slips and were equilibrated against 400 µl reservoir solution. Crystals were looped out and soaked in cryoprotectant (10% (w/v) glycerol added to crystallisation buffer), before flash-cooling and storage in liquid nitrogen.

3.2.4 Data Collection and Processing

X-ray diffraction data were collected to 2.15 Å resolution on the Australian synchrotron beamline MX2 in Melbourne, Australia. *iMOSFLM* (Battye *et al.* 2011), *AIMLESS* (Evans & Murshudov 2013) and *BALBES* (Long *et al.* 2008) were used for data reduction, data scaling, and phasing, respectively.

3.2.5 Structure Solution and Refinement

The structure was solved by molecular replacement using ApPDC (PDB entry: 2VBI, 76% amino acid identity) as the starting model. The structure was refined by iterative cycles of manual building and modelling in *Coot* (Emsley *et al.* 2010) and refinement in *Refmac5* (*CCP4i* suite) (Winn *et al.* 2001; Potterton *et al.* 2003; Vagin *et al.* 2004). The quality of the final model was checked using *MolProbity* (http://molprobity.biochem.duke.edu/index.php, Chen *et al.* 2010).

3.3 Results

3.3.1 Cloning of wt ZpPDC for Expression in *E.coli*

The wt Z*ppdc* had previously been cloned into pET28a to generate the enzyme with a C-terminal hexa-histidine-tag (see plasmid map below, Figure 3.3).

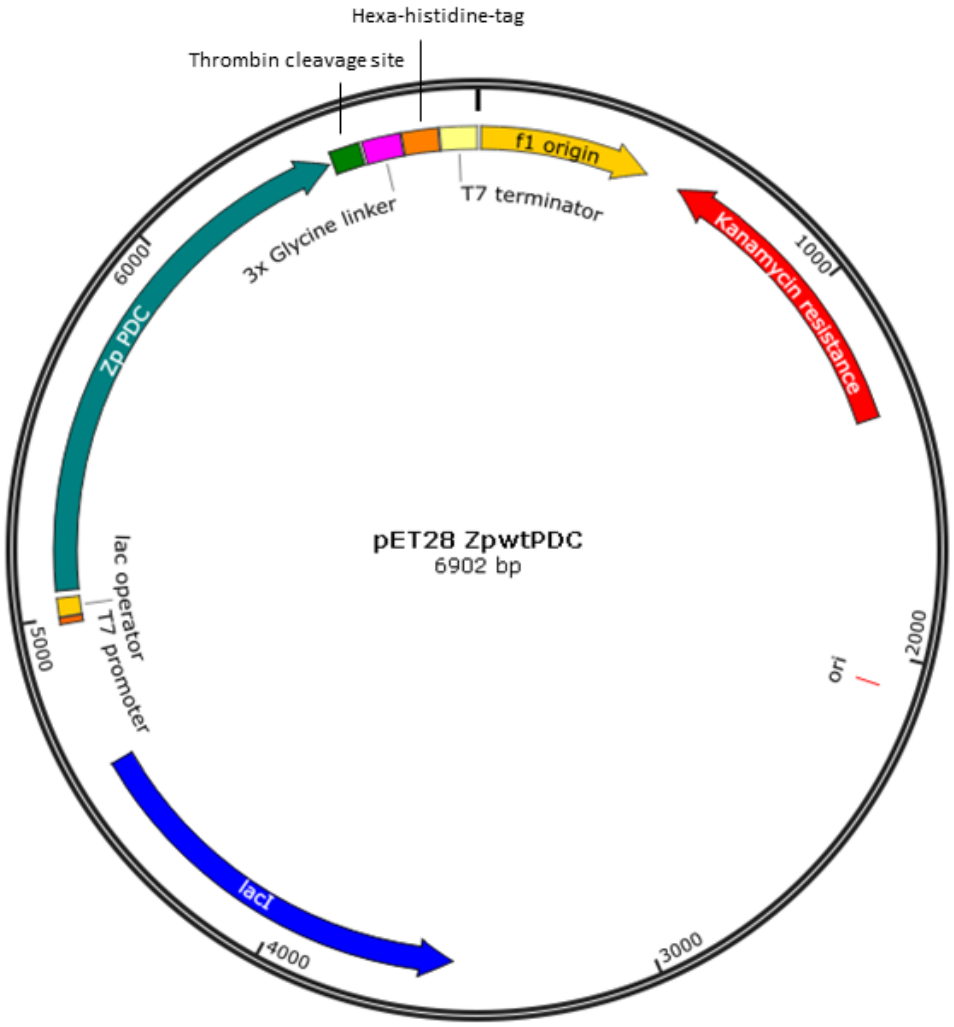


Figure 3.3 Plasmid map of pET28 ZpwtPDC. The neomycin phosphotransferase gene (labelled kanamycin resistance) confers resistance to kanamycin. In the presence of a T7 RNA polymerase, the Zp*pdc* is expressed from the T7 promoter under the control of the *lac* operator. *Lac*I encodes the *lac* operon repressor. The f1 origin is the origin of replication from an f1 phage.

Sequencing of this construct and comparision to the GenBank entry AF474145.1 showed that it was erroneous, and thus required correction by site-directed mutagenesis, changing the codon for alanine 292 to the correct valine (see Figure 3.4 for PCR results). Sequencing confirmed the presence of the correct nucleotide sequence.

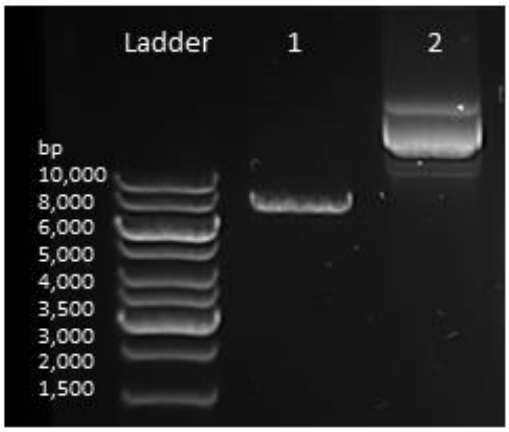


Figure 3.4 Agarose gel electrophoresis of the mutagenesis PCR. The PCR-amplified product was visualized alongside the GeneRuler™ 1 kb ladder (Thermo Fisher Scientific). Lane 1 contains the mutated pET28 ZpwtPDC A292V clearly linearized, while lane 2 contains unmodified, purified plasmid pET28 ZpwtPDC.

3.3.2 Recombinant Expression and Purification of ZpPDC

The Zp*pdc* was expressed in *E. coli* BL21 (DE3) from pET28 ZpwtPDC using the T7 expression system. The recombinant protein was tagged with a C-terminal hexa-histidine-tag, giving it a predicted monomer size of 61.8 kDa. His-tagged wt ZpPDC was purified by nickel-affinity chromatography, followed by a further purification and buffer exchange step through size-exclusion chromatography (see traces in Figure 3.5 and 3.6). PDC activity in the protein-containing fractions was confirmed by testing them in the standard assay and purity of the protein was assessed by SDS-PAGE (Figure 3.7). The pure protein was used in enzyme characterization studies.

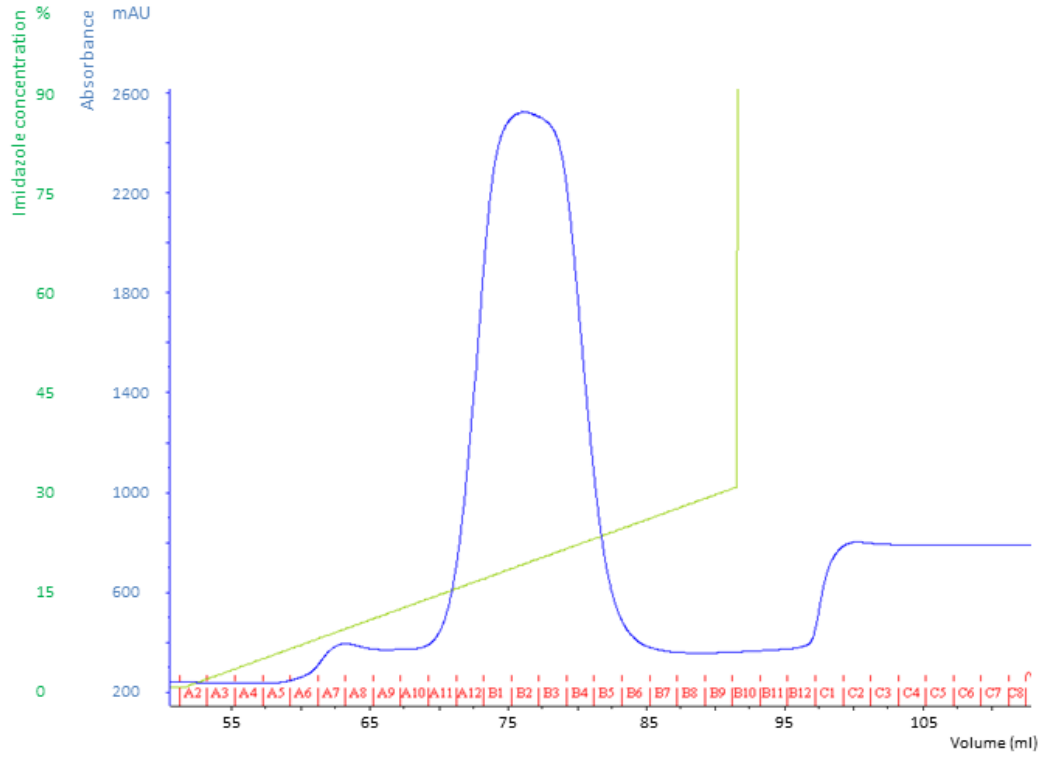


Figure 3.5 Nickel-affinity chromatogram for recombinant wt ZpPDC. The dark blue line is the absorbance at 280 nm (mAU). The green line indicates the imidazole concentration in the His-elute buffer. ZpPDC elutes at 120-200 mM imidazole.

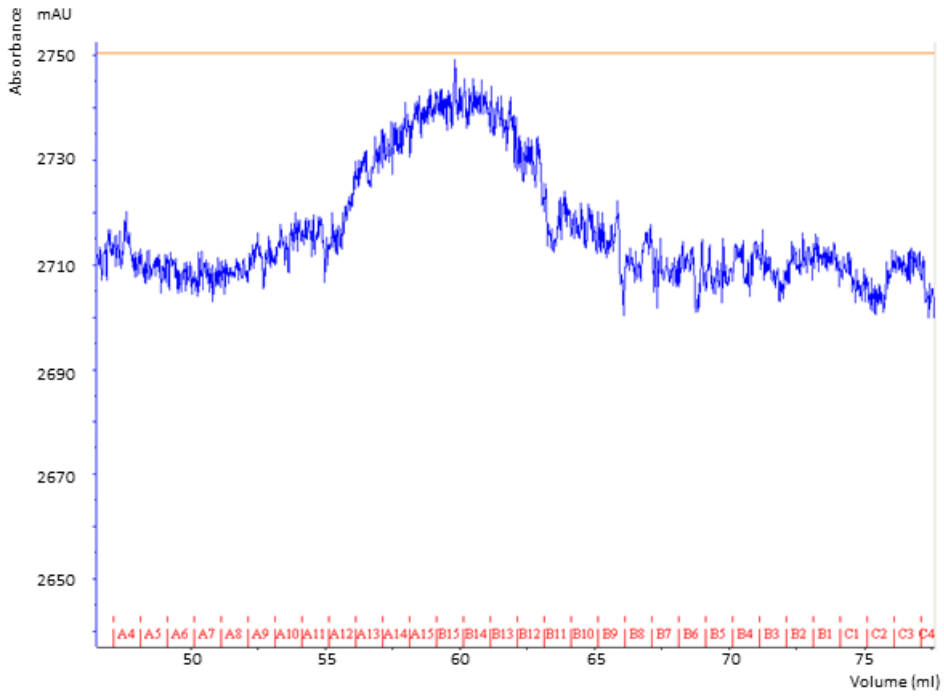


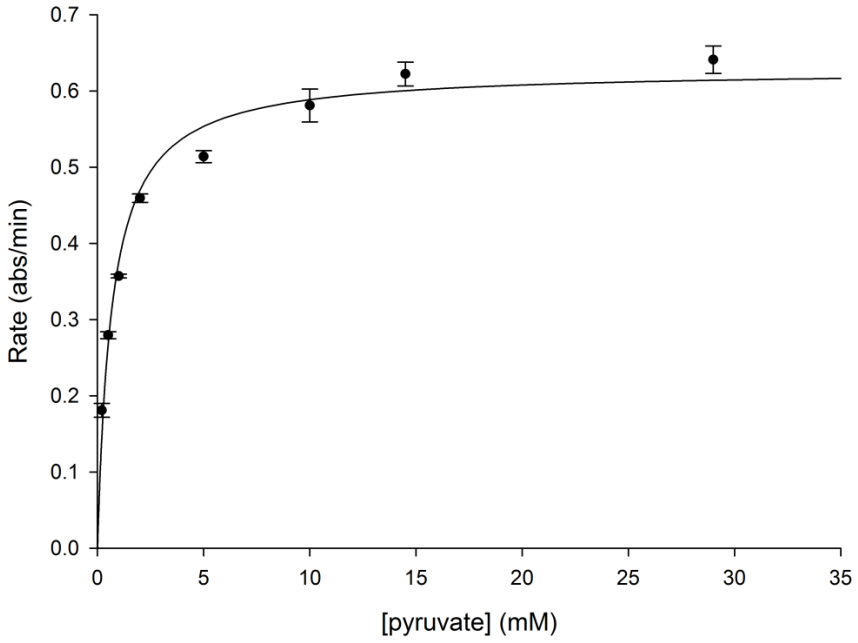
Figure 3.6 Size-exclusion chromatogram for recombinant wt ZpPDC. The dark blue line is the absorbance at 280 nm (mAU).

Figure 3.7 SDS-PAGE analysis of wt ZpPDC size-exclusion chromatography fractions. His-tagged wt ZpPDC (monomer size of 61.8 kDa) was purified by nickel-affinity chromatography followed by further purification and buffer exchange through size-exclusion chromatography. Lane M contains the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific).

3.3.3 Enzyme Characterization

Kinetic properties were analysed using the standard coupled assay at 30°C, pH 6.5. The data were analysed using the non-linear fit model from the enzyme kinetics module in SigmaPlot (Figure 3.8), resulting in a V_{max} of 165 ±3 µmol/min/mg and a K_M for pyruvate of 0.67 ±0.06 mM.

(A)



(B)

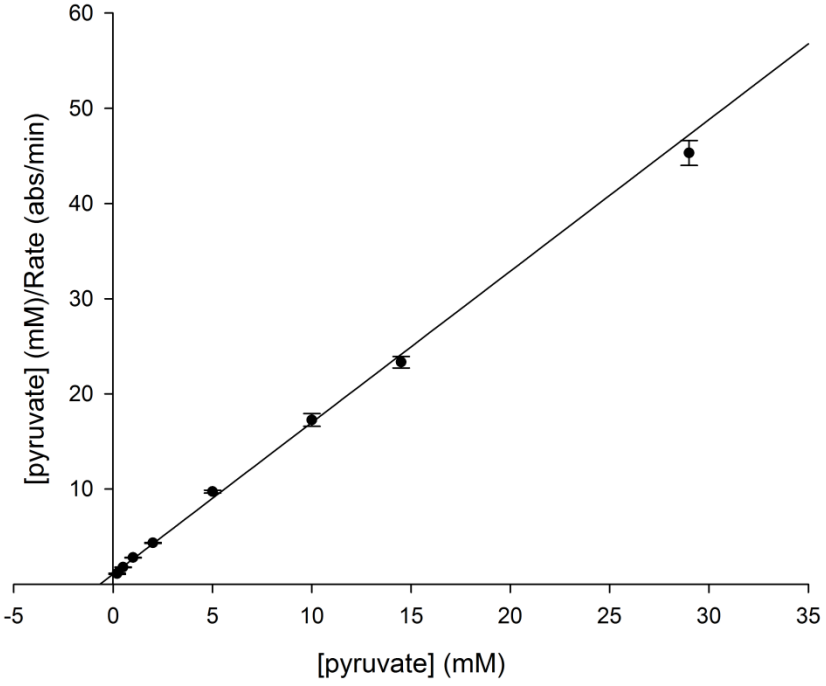


Figure 3.8 Dependence of wt ZpPDC activity on the concentration of pyruvate. The relationship between enzyme activity (abs/min) and pyruvate concentrations is displayed as (A) a Michaelis-Menten plot, and (B) a Hanes-Woolf plot. Error bars are standard errors based on 3 measurements. Kinetic parameters determined are: V_{max} = 0.63 ±0.01 abs/min (equates to 165 ±3 µmol/min/mg) and K_M = 0.67 ±0.06 mM.

The thermal properties of the enzyme were also investigated. The temperature optimum for activity *in vitro* was found to be 65°C (Figure 3.9). Irreversible denaturation analysis was carried out by incubation of the protein at 50 to 70°C for 30 min, and assaying for retained activity at 30°C. This showed that as previously reported, wt ZpPDC retained 100% activity at 60°C, 80% at 65°C, and completely lost activity at 70°C (Raj *et al.* 2002). Using thermal shift assays, the denaturing temperature was determined to be 70°C.

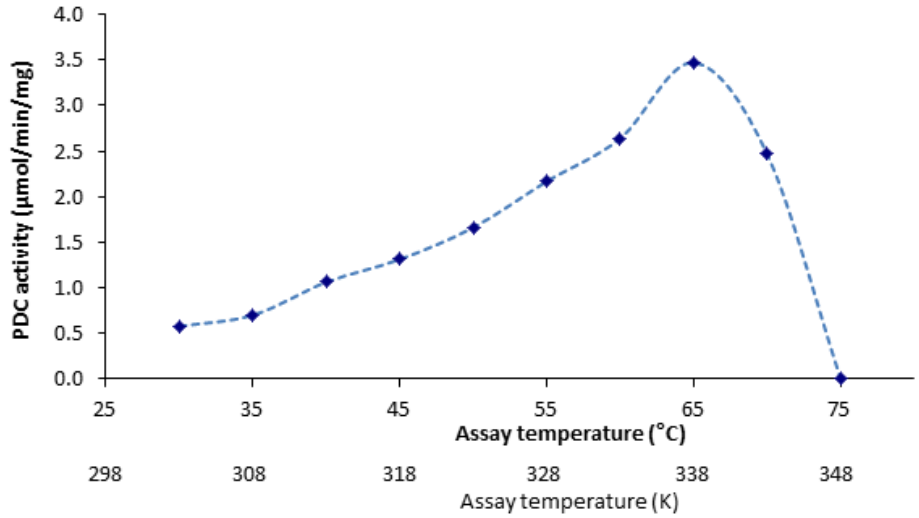


Figure 3.9 Relationship between temperature and wt ZpPDC activity. The relationship between PDC activity (rate in µmol/min/mg) and assay temperature was determined by monitoring the decrease in pyruvate-dependent absorbance at 320 nm.

3.3.4 Crystallisation

The purified His-tagged ZpPDC (4.8 mg/ml) was incubated with 2 mM pyruvate for 30 min at room temperature and crystallised in 0.15 M sodium citrate, pH 5.5, 14% (w/v) PEG 3350 using the hanging-drop vapour-diffusion method at 18°C (291 K). Figure 3.10 shows typical crystals.

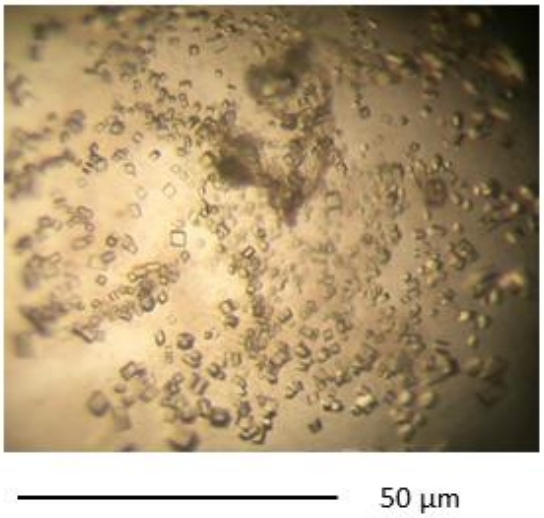


Figure 3.10 ZpPDC crystals. Microscope image of representative crystal forms.

3.3.5 Data Collection

X-ray diffraction data were collected to 2.15 Å resolution on the Australian synchrotron beamline MX2. The crystal belonged to space group $P2_1$ and contained six tetramers in the asymmetric unit. Data collection and processing statistics are summarised in Table 3.1.

Table 3.1 Data collection and processing.
Two data collections, at different crystal-detector distances and rotation ranges, were combined to form the final data set. Values for the outer resolution shell are given in parentheses.

Diffraction source	beamline MX2
Wavelength (Å)	0.9537
Temperature (K)	100.0
Detector	ADSC Q315*R* CCD
Crystal-detector distance (mm)	400 (low resolution set) / 300 (high resolution set)
Rotation range per image (°)	0.5 (low resolution set) / 0.25 (high resolution set)
Total rotation range (°)	180
Exposure time per image (s)	1
Space group	$P2_1$
a, *b*, *c* (Å)	204.56, 177.39, 244.55
α, β, γ (°)	90, 112.94, 90
Mosaicity (°)	0.3
Resolution range (Å)	75.470–2.150 (2.190–2.150)
Total No. of reflections	5108329 (153406)
No. of unique reflections	858032 (41628)
Completeness (%)	98.9 (96.9)
Redundancy	6.0 (3.7)
$\langle I/\sigma(I)\rangle$	19.9 (2.3)
$R_{r.i.m.}$†	0.175 (0.714)
Overall *B* factor from Wilson plot (Å^2)	19.9

† Estimated $R_{r.i.m.} = R_{merge}[N/(N-1)]^{1/2}$, where N = data multiplicity.

3.3.6 Structure Solution and Refinement

ApPDC (PDB entry: 2VBI, 76% amino acid identity) was used as the starting model for molecular replacement. The correct solution was supported by the presence of positive electron density for TPP, which was not part of the search model.

The structure was refined by iterative cycles of manual building and modelling in *Coot* and refinement in *Refmac5* (*CCP4i* suite). Initially, the non-crystallographic symmetry was used to minimize the rebuilding task; however, in the later stages this was not used, all 24 monomers being refined independently. The final structure was quality checked using *MolProbity*. The final model was refined to an *R* factor of 18.6%, and an R_{free} value of 22.3%, using data between 75.40 – 2.15 Å. Table 3.2 summarises the structure solution and refinement statistics.

Table 3.2 Structure solution and refinement.
Values for the outer shell are given in parentheses.

Resolution range (Å)	225.21–2.15 (2.207–2.151)
Completeness (%)	98.8 (97.6)
σ cutoff	$F > 0.000\sigma(F)$
No. of reflections, working set	814954 (59398)
No. of reflections, test set	42940 (3176)
Final R_{cryst}	0.186 (0.271)
Final R_{free}	0.220 (0.300)
Cruickshank DPI	0.2148
No. of non-H atoms:	
Protein	102016
Ligand	648
Solvent	4316
Total	107004
R.m.s. deviations:	
Bonds (Å)	0.008
Angles (°)	1.307
Average B-factors (Å^2):	
Protein	26.935
Ligand	24.626
Ramachandran plot:	
Most favoured (%)	98.05
Allowed (%)	1.91

98.05% of the residues are located in favoured regions of the Ramachandran plot. In chains G, J, K, M and O, Ser73 is located in the disallowed region (Figure 3.11), but is well defined in its electron density and seems to be located in a tight bend. This has been noted previously in ZmPDC (Dobritzsch *et al.* 1998). The structure has been deposited in the PDB and was given the code 5EUJ (Buddrus *et al.* 2016).

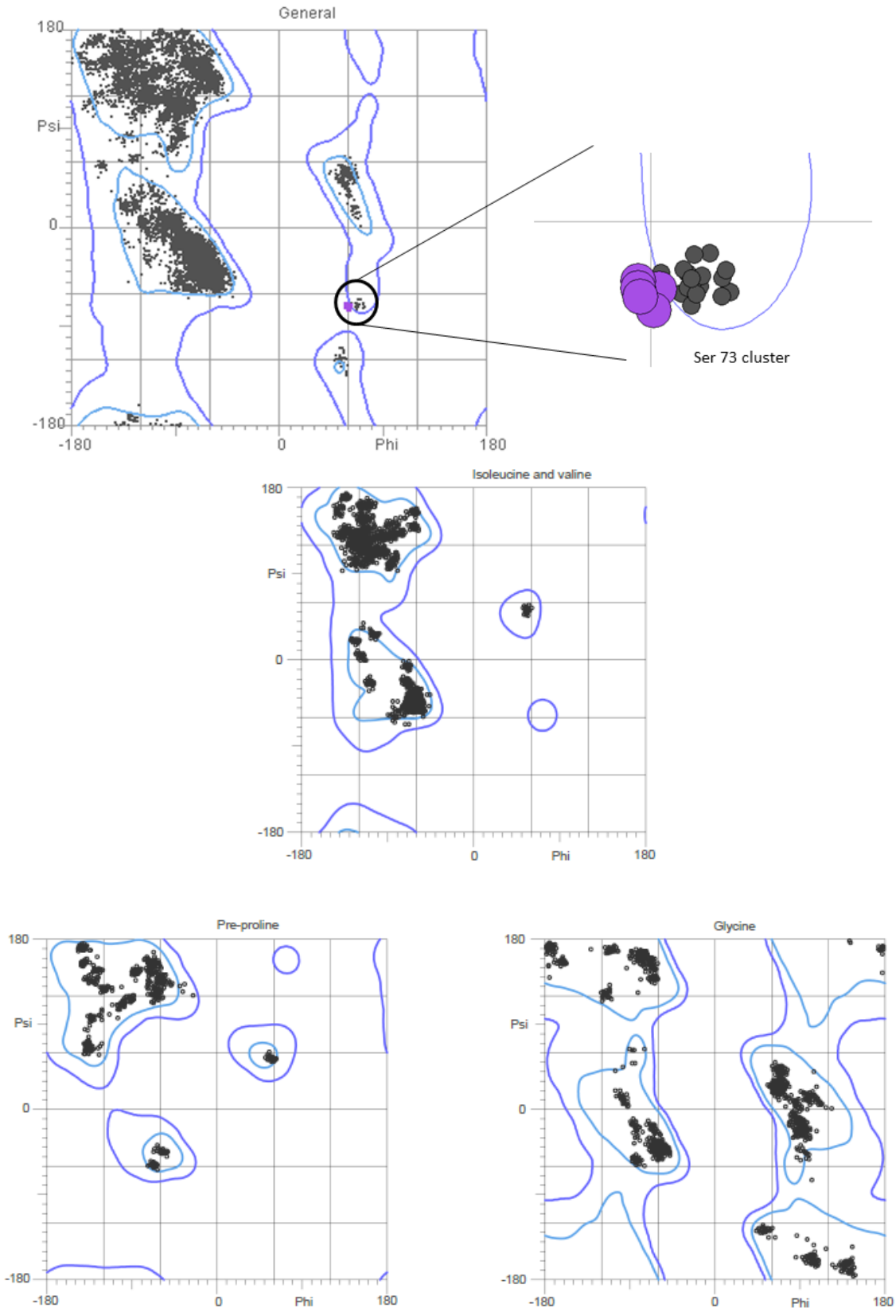


Figure 3.11 MolProbity Ramachandran analysis of ZpPDC. The light blue line surrounds favoured regions; the purple line marks the allowed boundary.

3.3.7 Overall Structure of ZpPDC

As described for other bacterial PDCs, the quaternary structure of the ZpPDC is a homo-tetramer, or dimer-of-dimers. Each subunit consists of 556 amino acids with a molecular mass of 59.4 kDa. The domains in each monomer can be assigned as follows: amino acids 1 – 190 PYR (pyrimidine binding, with 176 – 190 linker), 191 – 355 R (regulatory, with 345 – 355 linker), and 356 – 555 PP (pyrophosphate binding). The TPP molecules bind across both subunits in each dimer, with the pyrophosphate group binding to the PP domain from one subunit and the pyrimidine ring binding to the PYR domain from the second subunit, thus forming two active sites in the dimer (Figure 3.12).

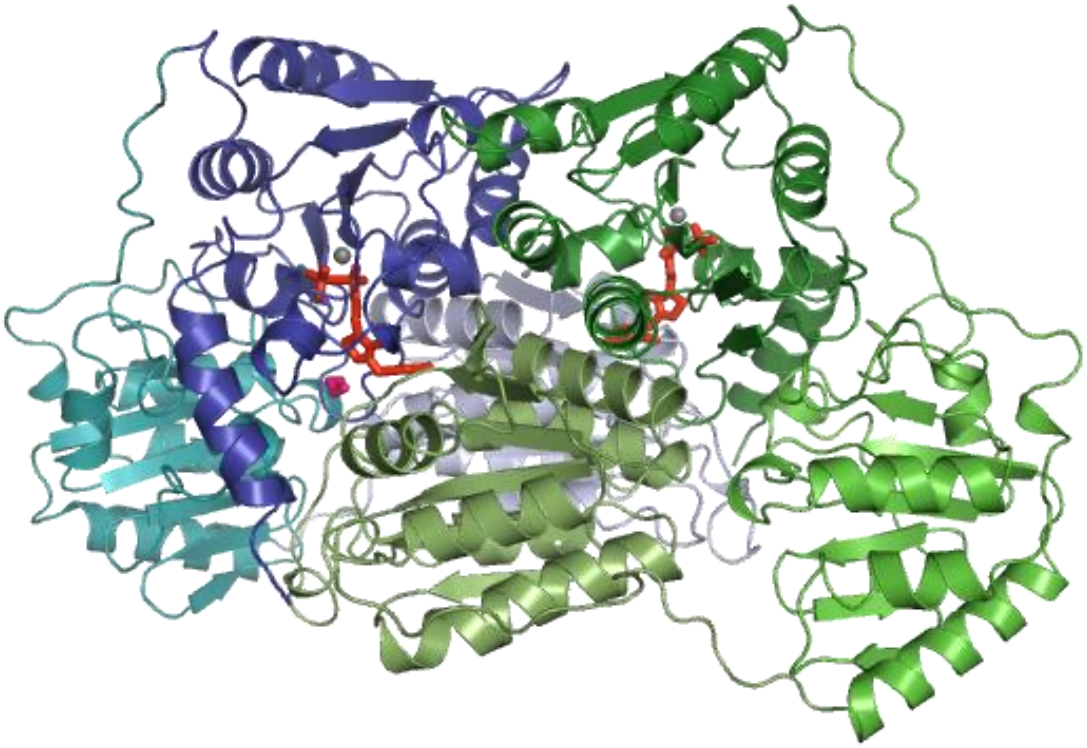


Figure 3.12 Cartoon representation of the ZpPDC dimer. One monomer is coloured blue, with the PYR domain (residues 1 to 190) in pale blue, the R domain (residues 191 to 355) in teal, and the PP domain (residues 356 to 556) in dark blue. The second monomer is coloured green, with the PYR domain pale green, the R domain bright green, and the PP domain dark green. The active site magnesium ions are represented as grey spheres and an 1,2-ethanediol molecule bound in the active site of the blue monomer in pink. Two thiamine pyrophophate (TPP) molecules bound between the PYR domain of one monomer and the PP domain of the other are represented as orange sticks.

The final model contains six tetramers in the asymmetric unit, each with a surface area of 65,000 Å^2. It comprises 24 amino acid chains of 555 amino acids (labelled A to X), each with a TPP and a Mg(II) ion, and 4,147 water molecules. Despite crystallisation in the presence of pyruvate, no pyruvate molecules were visible anywhere in the structure, probably due to catalytic turnover of the substrate to acetaldehyde and subsequent loss of this volatile product. However, six 1,2-ethanediol (EDO) molecules are present, four of which can be found in the active sites of chains I, N, Q and R.

Overall, the electron density map was of good quality, apart from chains U, V, W and X, which seem exceptionally flexible (Figure 3.13). The average B factor for all tetramers is 27 Å^2; the average B factor for UVWX is 42 Å^2. Other common flexible areas include the exposed N-terminal Met, the exposed C-terminus, and the 340 – 355 loop, which is the exposed linker between the R and PP domains.

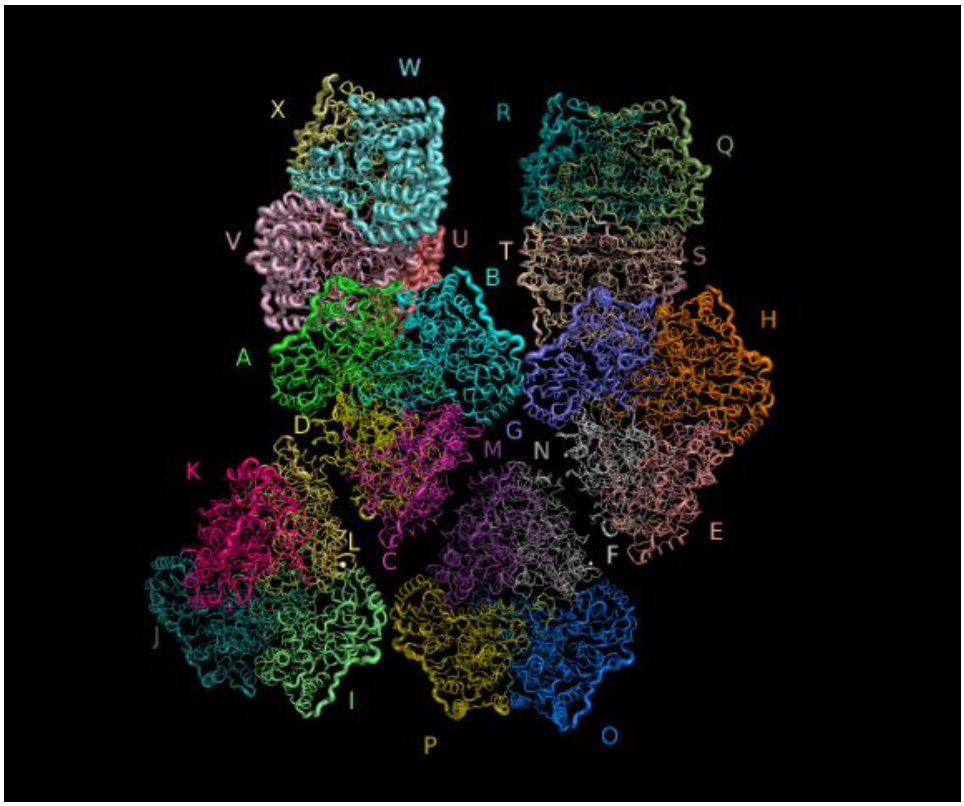


Figure 3.13 C-alpha representation of the asymmetric unit. Representation containing the six tetramers coloured by chain, labelled with the chain name, with the thickness of the trace determined by the B-factors. The UVWX tetramer has anomalously high B-factors and correspondingly poor electron density.

3.4 Structural Analysis and Discussion

3.4.1 Kinetic and Thermal Properties of the *Zymobacter palmae* PDC

The kinetic parameters found here (V_{max} of 165 ±3 U/mg, K_M for pyruvate 0.67 ±0.06 mM, at pH 6.5) are comparable to previously reported values: at pH 6, V_{max} 130 U/mg, K_M 0.24 mM, and at pH 7, V_{max} of 140 U/mg, K_M 0.71 mM (Raj *et al.* 2002).

Irreversible denaturation data were in agreement with those reported by Raj *et al.* (2002). However, the temperature optimum found in this study is 10°C higher than previously reported (55°C, Gocke *et al.* 2009). This may be explained by the fact that Gocke *et al.* (2009) used a potassium phosphate buffer and low concentrations of the cofactors (0.1 mM $MgSO_4$, 0.1 mM TPP). The study presented here used a MES-based buffer, which does not chelate magnesium ions, and 20 mM $MgSO_4$ and 3 mM TPP. Perhaps the presence of TPP and magnesium in excess balances out cofactor dissociation effects as the temperature increases.

Thermal shift assays determined the denaturing temperature to be 70°C (previously unreported). Together with the irreversible temperature data and the high temperature optimum, these data would suggest the ZpPDC was well suited for high temperature expression.

3.4.2 Comparison of Known Bacterial PDCs

Most bacterial PDCs studied so far are very similar in their amino acid sequence and kinetic parameters, with the exception of SvPDC. SvPDC seems to be more closely related to fungal PDCs, whereas the other bacterial PDCs are more similar to plant PDCs (Raj *et al.* 2002, Talarico *et al.* 2001). SvPDC is the only PDC identified from a Gram-positive bacterium, shares only 31% amino acid identity with ZpPDC, and shows sigmoidal rather than Michaelis-Menten kinetics.

ZpPDC shows a high temperature optimum at 65°C, and retains 80% activity after incubation at 65°C for 30 min (Raj *et al.* 2002; confirmed in this study), making it one of the most thermostable bacterial PDCs currently known. See Table 3.3 for a summary of the thermostability properties of known PDCs.

Based on the root-mean-square deviation (RMSD) values, the tetramers in 5EUJ are more similar to each other than to other known bacterial PDCs. The MNOP tetramer is the most representative of the 6 tetramers in 5EUJ, based on low RMSD values and B-factors when

compared to all other chains in the structure (see Table 3.4), and was used to compare the ZpPDC structure to other known bacterial PDC structures (see Table 3.3).

Table 3.3 Properties of known bacterial PDCs in order of decreasing thermostability. RMSD and Q scores were calculated in PDBeFOLD using the ZpPDC tetramer MNOP against the indicated PDB entries. These are averages of chains M,N,O and P against all chains in the relative PDB entries.

	ZpPDC	ApPDC	ZmPDC	GdPDC	GoPDC	SvPDC
Amino acid identity (%)	Reference	73	63	71	67	31
Temperature optimum	65°C	65°C [a]	60°C [a]	45-50°C [d]	53°C [c]	NA
Temperature dependence of activity retention	60°C:100% 65°C:80% [a] 70°C: 0%	50°C: 100% 60°C: 65% 65°C: 45% 70°C: 5% [b]	45°C: 85% 60°C: 65% 65°C: 45% 70°C: 0% [b]	NA (half-life, 60°C 0.3h [d])	55°C: 98% 60°C: 70% 65°C: 40% [c]	45°C: 95% 50°C: 0%[b]
PDB entry	5EUJ	2VBI	1ZPD	4COK	NA	NA
Genbank gene/ protein	AF474145.1 AAM49566	AF368435.1 AAM21208.1	M15393.2 AAA27696.2	KJ746104.1 AIG13066.1	KF650839.1 AHB37781.1	AAL18557.1 AF354297.1
RMSD (Å)		0.71 ±0.03	0.70 ±0.05	0.61 ±0.02	NA	NA
Q Scores		0.93 ±0.01	0.89 ±0.01	0.94	NA	NA
Matched residues		551 ±1	544 ±2	545 ±1	NA	NA

a. Gocke *et al.* (2009) b. Raj *et al.* (2002) c. van Zyl *et al.* (2014b) d. van Zyl *et al.* (2014a)

Table 3.4 RMSD value comparison of all chains in ZpPDC (5EUJ). RMSD (Å) values calculated by PDBeFOLD over all C-alpha in each chain. Green fields contain the lowest scores; these rise through yellow, orange, red, purple and blue. The average B-factor (Å^2) for each chain is given below the RMSD table, and was calculated using CCP4i Baverage.

RMSD (Å) calculated by PDBeFOLD	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
A	0	0.17	0.21	0.15	0.14	0.17	0.16	0.14	0.15	0.12	0.14	0.16	0.16	0.16	0.12	0.15	0.14	0.14	0.13	0.16	0.26	0.19	0.25	0.18
B	0.17	0	0.2	0.19	0.16	0.15	0.19	0.16	0.19	0.15	0.16	0.17	0.18	0.18	0.15	0.16	0.17	0.15	0.17	0.17	0.3	0.2	0.25	0.18
C	0.21	0.2	0	0.2	0.18	0.18	0.2	0.18	0.23	0.2	0.19	0.24	0.15	0.18	0.19	0.21	0.22	0.18	0.22	0.2	0.25	0.23	0.25	0.2
D	0.15	0.19	0.2	0	0.14	0.18	0.14	0.14	0.16	0.13	0.16	0.17	0.17	0.15	0.14	0.15	0.16	0.16	0.15	0.17	0.27	0.19	0.26	0.21
E	0.14	0.16	0.18	0.14	0	0.15	0.16	0.14	0.15	0.12	0.13	0.15	0.14	0.14	0.12	0.13	0.15	0.13	0.14	0.16	0.27	0.19	0.27	0.18
F	0.17	0.15	0.18	0.18	0.15	0	0.18	0.16	0.18	0.16	0.15	0.17	0.16	0.12	0.15	0.16	0.17	0.15	0.17	0.12	0.28	0.2	0.25	0.19
G	0.16	0.19	0.2	0.14	0.16	0.18	0	0.15	0.17	0.15	0.17	0.2	0.18	0.16	0.15	0.15	0.16	0.15	0.17	0.17	0.26	0.2	0.25	0.22
H	0.14	0.16	0.18	0.14	0.14	0.16	0.15	0	0.17	0.13	0.16	0.19	0.17	0.15	0.14	0.16	0.16	0.14	0.16	0.17	0.26	0.19	0.24	0.19
I	0.15	0.19	0.23	0.16	0.15	0.18	0.17	0.17	0	0.15	0.16	0.15	0.16	0.18	0.14	0.15	0.15	0.16	0.14	0.16	0.3	0.22	0.28	0.2
J	0.12	0.15	0.2	0.13	0.12	0.16	0.15	0.13	0.15	0	0.13	0.14	0.15	0.15	0.12	0.13	0.13	0.13	0.13	0.16	0.27	0.18	0.26	0.17
K	0.14	0.16	0.19	0.16	0.13	0.15	0.17	0.16	0.16	0.13	0	0.15	0.15	0.16	0.12	0.15	0.15	0.13	0.14	0.16	0.28	0.19	0.26	0.18
L	0.16	0.17	0.24	0.17	0.15	0.17	0.2	0.19	0.15	0.14	0.15	0	0.17	0.19	0.15	0.15	0.15	0.17	0.14	0.15	0.31	0.21	0.29	0.19
M	0.16	0.18	0.15	0.17	0.14	0.16	0.18	0.17	0.16	0.15	0.15	0.17	0	0.16	0.14	0.16	0.17	0.16	0.15	0.16	0.28	0.22	0.27	0.16
N	0.16	0.18	0.18	0.15	0.14	0.12	0.16	0.15	0.18	0.15	0.16	0.19	0.16	0	0.13	0.16	0.18	0.14	0.17	0.13	0.27	0.21	0.26	0.21
O	0.12	0.15	0.19	0.14	0.12	0.15	0.15	0.14	0.14	0.12	0.12	0.15	0.14	0.13	0	0.13	0.14	0.12	0.12	0.14	0.27	0.18	0.25	0.17
P	0.15	0.16	0.21	0.15	0.13	0.16	0.15	0.16	0.15	0.13	0.15	0.15	0.16	0.16	0.13	0	0.15	0.15	0.15	0.15	0.29	0.2	0.27	0.2
Q	0.14	0.17	0.22	0.16	0.15	0.17	0.16	0.16	0.15	0.13	0.15	0.15	0.17	0.18	0.14	0.15	0	0.15	0.13	0.16	0.29	0.18	0.27	0.18
R	0.14	0.15	0.18	0.16	0.13	0.15	0.15	0.14	0.16	0.13	0.13	0.17	0.16	0.14	0.12	0.15	0.15	0	0.15	0.16	0.26	0.17	0.23	0.19
S	0.13	0.17	0.22	0.15	0.14	0.17	0.17	0.16	0.14	0.13	0.14	0.14	0.15	0.17	0.12	0.15	0.13	0.15	0	0.15	0.29	0.19	0.26	0.17
T	0.16	0.17	0.2	0.17	0.16	0.12	0.17	0.17	0.16	0.16	0.16	0.15	0.16	0.13	0.14	0.15	0.16	0.16	0.15	0	0.28	0.2	0.26	0.2
U	0.26	0.3	0.25	0.27	0.27	0.28	0.26	0.26	0.3	0.27	0.28	0.31	0.28	0.27	0.27	0.29	0.29	0.26	0.29	0.28	0	0.26	0.26	0.31
V	0.19	0.2	0.23	0.19	0.19	0.2	0.2	0.19	0.22	0.18	0.19	0.21	0.22	0.21	0.18	0.2	0.18	0.17	0.19	0.2	0.26	0	0.24	0.22
W	0.25	0.25	0.25	0.26	0.27	0.25	0.25	0.24	0.28	0.26	0.26	0.29	0.27	0.26	0.25	0.27	0.27	0.23	0.26	0.26	0.26	0.24	0	0.28
X	0.18	0.18	0.2	0.21	0.18	0.19	0.22	0.19	0.2	0.17	0.18	0.19	0.16	0.21	0.17	0.2	0.18	0.19	0.17	0.2	0.31	0.22	0.28	0
Average RMSD of each monomer	0.16	0.17	0.20	0.16	0.15	0.16	0.17	0.16	0.17	0.15	0.16	0.17	0.17	0.16	0.15	0.16	0.16	0.15	0.16	0.16	0.27	0.19	0.25	0.19
Average RMSD overall	0.173																							
Average B-factor (Å²) (CCP4i Baverage)	28.24	27.38	21.30	22.28	21.25	18.55	25.64	27.30	25.19	22.73	24.38	23.33	18.05	18.72	24.32	23.33	27.59	27.29	25.58	24.99	44.65	43.49	47.92	32.93
Average B-factor overall	26.93																							
Average B-factor without UVWX	23.87																							
Average B-factor MNOP alone	21.11																							
Average B-factor UVWX alone	42.25																							

The residues involved in interactions on interfaces (i.e., forming hydrogen bonds or salt bridges) are well conserved between all bacterial PDCs analysed. Figure 3.14 indicates the positioning of the various interfaces. Table 3.5 summarises the interface areas, and Table 3.6 the number of interactions made between different interfaces. A trend of increasing interface area between two functional dimers and increasing number of salt bridges can be observed as the thermoactivity and thermostability of the PDCs increase.

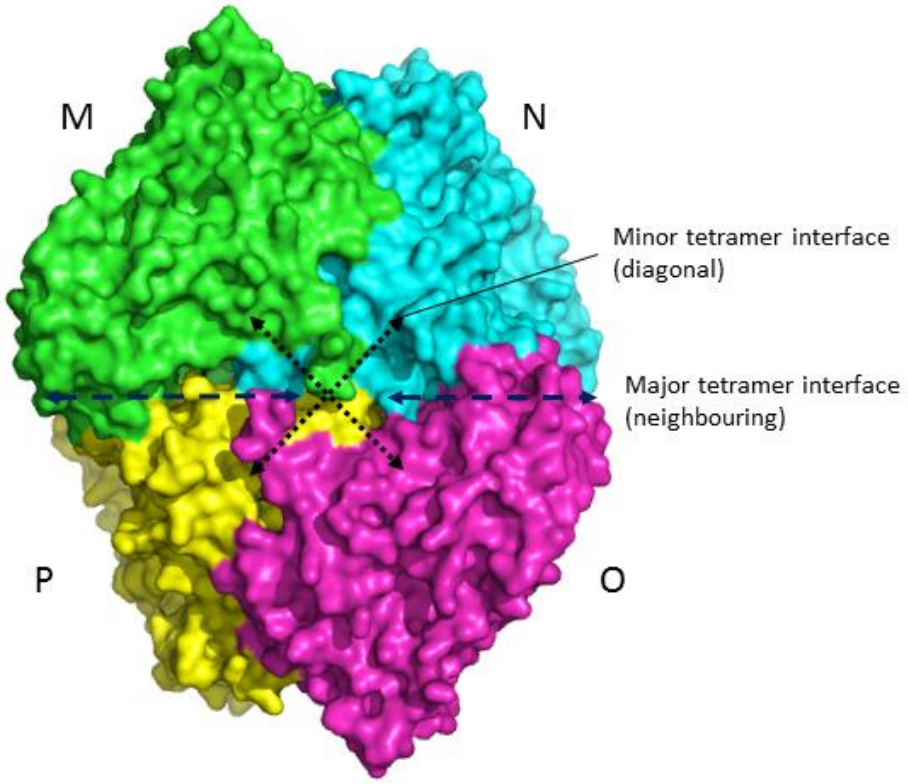


Figure 3.14 Surface mesh representation of the ZpPDC MNOP tetramer. This representation shows the interfaces between the monomers within a functional dimer (dimer interface), i.e., M (green) with N (cyan), and O (magenta) with P (yellow). The major tetramer interface is the interaction between neighbouring monomers such as M with P or N with O on the interfacing area between the two dimers, and is indicated by the dashed dark blue lines. The minor tetramer interface is the interaction between monomers diagonally across the tetrameric centre, such as M with O or N with P, as indicated by the black dotted lines.

Table 3.5 Comparison of interface areas of known bacterial PDCs in order of decreasing thermostability; determined using PDBePISA.

	ZpPDC (5EUJ)	ApPDC (2VBI)	ZmPDC (1ZPD)	GdPDC (4COK)
Interface area between monomers within a functional dimer (Å^2)	3813.13	3761.3	4144.5 (*4387)	3749.8
Percentage of total surface of the monomer	17.08%	16.95%	18.39% (*19.4%)	17.83%
Interaction area between two functional dimers to form a tetramer (Å^2)	2912.16	2840	2489.2 (*4405)	1851.4
Tetramer interface as a percentage of total surface of one dimer	13.06%	12.78%	11.03% (*12.1%)	8.79%

* Dobritzsch *et al.* (1998)

Table 3.6 Comparison of interactions within interfaces of bacterial PDC structures in order of decreasing thermostability; interactions determined using PDBePISA.

Interactions on Interfaces	ZpPDC (5EUJ)		ApPDC (2VBI)		ZmPDC (1ZPD)		GdPDC (4COK)	
	Hydrogen Bonds	Salt bridges	Hydrogen Bonds	Salt bridges	Hydrogen Bonds	Salt bridges	Hydrogen Bonds	Salt bridges
Dimer interface (between monomers within a functional dimer)	73	12	61	16	76 (66*)	14 (7*)	63	13
Major tetramer interface (neighbour)	31	24	34	14	29 (64*)	8 (25*)	17	9
Minor tetramer interface (diagonal)	2	0	4	0	6	2	4	3
TPP pyrimidine ring	11	0	10	0	12	0	10	0

* Dobritzsch *et al.* (1998)

Neighbour: interactions between neighbouring monomers; diagonal: interactions between monomers diagonally across tetrameric centre, as shown in Figure 7.7.

Using PROMALS3D (Pei *et al.* 2008) to generate a structure-based alignment, conserved regions were analysed (Figure 3.15). The PYR and PP domains are well conserved; R and the linker regions are much less so. 1ZPD contains some extra residues in the R and PP domains, as discussed in van Zyl *et al.* (2014a). In particular, they mention 2 regions, both of which are

linker regions and are extended in 1ZPD (Figure 3.16). In 4COK these regions are shorter or less ordered. However, none of these factors seem to correlate with the thermostability and thermoactivity differences observed between these PDCs.

Figure 3.15 PROMALS3D structural sequence alignment (below). refmac30.pdb is ZpPDC 5EUJ chain M. Two black boxes highlight position 134 and 245, which are incorrect in the ZpPDC GenBank entry. The GenBank entry states 134R and 245E; both should be A, which is well conserved among the bacterial PDCs. Each representative sequence has a magenta name and is colored according to PSIPRED secondary structure predictions (red: alpha-helix, blue: beta-strand). The last two lines show consensus amino acid sequence (Consensus_aa) and consensus predicted secondary structures (Consensus_ss). Consensus predicted secondary structure symbols: alpha-helix: h; beta-strand: e. Consensus amino acid symbols are: conserved amino acids are in **bold** and uppercase letters; aliphatic (I, V, L): *l*; aromatic (Y, H, W, F): @; hydrophobic (W, F, Y, M, L, I, V, A, C, T, H): *h*; alcohol (S, T): o; polar residues (D, E, H, K, N, Q, R, S, T): p; tiny (A, G, C, S): t; small (A, G, C, S, V, N, D, T, P): s; bulky residues (E, F, I, K, L, M, Q, R, W, Y): b; positively charged (K, R, H): +; negatively charged (D, E): -; charged (D, E, K, R, H): c.


Conservation:
refmac30.pdb_chainM_s001 1 -MYTVGMYLAERLAQIGLKHHFAVAGDYNLVLLDQLLLNKDMEQVYCCNELNCGFSAEGYARARGAAAAI 69
2vbi_chainA_p002 1 -TYTVGMYLAERLVQIGLKHHFAVAGDYNLVLLDQLLLNKDMKQIYCCNELNCGFSAEGYARSNGAAAAV 69
4cok_chainA_p004 1 MTYTVGRYLADRLAQIGLKHHFAVAGDYNLVLLDQLLLNTDMQQIYCSNELNCGFSAEGYARANGAAAAI 70
1zpd_chainA_p003 1 -SYTVGTYLAERLVQIGLKHHFAVAGDYNLVLLDNLLLNKNMEQVYCCNELNCGFSAEGYARAKGAAAAV 69
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 70 VTFSVGAISAMNAIGGAYAENLPVILISGSPNTNDYGTGHILHHTIGTTDYNYQLEMVKHVTCAAESIVS 139
2vbi_chainA_p002 70 VTFSVGAISAMNALGGAYAENLPVILISGAPNSNDQGTGHILHHTIGKTDYSYQLEMARQVTCAAESITD 139
4cok_chainA_p004 71 VTFSVGALSAFNALGGAYAENLPVILISGAPNANDHGTGHILHHTLGTTDYGYQLEMARHITCAAESIV- 139
1zpd_chainA_p003 70 VTYSVGALSAFDAIGGAYAENLPVILISGAPNNNDHAAGHVLHHALGKTDYHYQLEMAKNITAAAEAIYT 139
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 140 AEEAPAKIDHVIRTALRERKPAYLEIACNVAGAECVRPGPINSLLRELEVDQTSVTAAVDAAVEWLQDRQ 209
2vbi_chainA_p002 140 AHSAPAKIDHVIRTALRERKPAYLDIACNIASEPCVRPGPVSSLLSEPEIDHTSLKAAVDATVALLEKSA 209
4cok_chainA_p004 140 AADAPAKIDHVIRTALREKKPAYLEIACNVAGAPCVRPGGIDA-------DEASLKAAVDAALAFIEQRG 202
1zpd_chainA_p003 140 PEEAPAKIDHVIKTALREKKPVYLEIACNIASMPCAAPGPASALFNDEASDEASLNAAVDETLKFIANRD 209
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 210 NVVMLVGSKLRAAAAEKQAVALADRLGCAVTIMAAAKGFFPEDHPNFRGLYWGEVSSEGAQELVENADAI 279
2vbi_chainA_p002 210 SPVMLLGSKLRAANALAATETLADKLQCAVTIMAAAKGFFPEDHAGFRGLYWGEVSNPGVQELVETSDAL 279
4cok_chainA_p004 203 SVTMLVGSRIRAAGAQAQAVALADALGCAVTTMAAAKSFFPEDHPGYRGHYWGEVSSPGAQQAVEGADGV 272
1zpd_chainA_p003 210 KVAVLVGSKLRAAGAEEAAVKFTDALGGAVATMAAAKSFFPEENALYIGTSWGEVSYPGVEKTMKEADAV 279
Consensus_aa:
Consensus_ss:


Conservation:
refmac30.pdb_chainM_s001 280 LCLAPVFNDYATVGWNSWPKGDNVMVMDTDRVTFAGQSFEGLSLSTFAAALAEKAPSRPATT----QGTQ 345
2vbi_chainA_p002 280 LCIAPVFNDYSTVGWSAWPKGPNVILAEPDRVTVDGRAYDGFTLRAFLQALAEKAPARPASA----QKSS 345
4cok_chainA_p004 273 ICLAPVFNDYATVGWSAWPKGDNVMLVERHAVTVGGVAYAGIDMRDFLTRLAAHTVRRDATA----RGGA 338
1zpd_chainA_p003 280 IALAPVFNDYSTTGWTDIPDPKKLVLAEPRSVVVNGIRFPSVHLKDYLTRLAQKVSKKTGSLDFFKSLNA 349
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 346 APVLGIEAAEPNAPLTNDEMTRQIQSLITSDTTLTAETGDSWFNASRMPIPGGARVELEMQWGHIGWSVP 415
2vbi_chainA_p002 346 VPTCSLTATSDEAGLTNDEIVRHINALLTSNTTLVAETGDSWFNAMRMTLPRGARVELEMQWGHIGWSVP 415
4cok_chainA_p004 339 YVTPQTPAAAPTAPLNNAEMARQIGALLTPRTTLTAETGDSWFNAVRMKLPHGARVELEMQWGHIGWSVP 408
1zpd_chainA_p003 350 GELKKAAPADPSAPLVNAEIARQVEALLTPNTTVIAETGDSWFNAQRMKLPNGARVEYEMQWGHIGWSVP 419
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 416 SAFGNAVGSPERRHIMMVGDGSFQLTAQEVAQMIRYEIPVIIFLINNRGYVIEIAIHDGPYNYIKNWNYA 485
2vbi_chainA_p002 416 SAFGNAMGSQDRQHVVMVGDGSFQLTAQEVAQMVRYELPVIIFLINNRGYVIEIAIHDGPYNYIKNWDYA 485
4cok_chainA_p004 409 AAFGNALAAPERQHVLMVGDGSFQLTAQEVAQMIRHDLPVIIFLINNHGYTIEVMIHDGPYNNVKNWDYA 478
1zpd_chainA_p003 420 AAFGYAVGAPERRNILMVGDGSFQLTAQEVAQMVRLKLPVIIFLINNYGYTIEVMIHDGPYNNIKNWDYA 489
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 486 GLIDVFNDEDGH----GLGLKASTGAELEGAIKKALDNRRGPTLIECNIAQDDCTETLIAWGKRVAATNS 551
2vbi_chainA_p002 486 GLMEVFNAGEGH----GLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDMLVQWGRKVASTNA 551
4cok_chainA_p004 479 GLMEVFNAGEGN----GLGLRARTGGELAAAIEQARANRNGPTLIECTLDRDDCTQELVTWGKRVAAANA 544
1zpd_chainA_p003 490 GLMEVFNGNGGYDSGAAKGLKAKTGGELAEAIKVALANTDGPTLIECFIGREDCTEELVKWGKRVAAANS 559
Consensus_aa:
Consensus_ss:
Conservation:
refmac30.pdb_chainM_s001 552 RKPQ-- 555
2vbi_chainA_p002 552 RKT--- 554
4cok_chainA_p004 545 RPP--- 547
1zpd_chainA_p003 560 RKPVNK 565
Consensus_aa:
Consensus_ss:

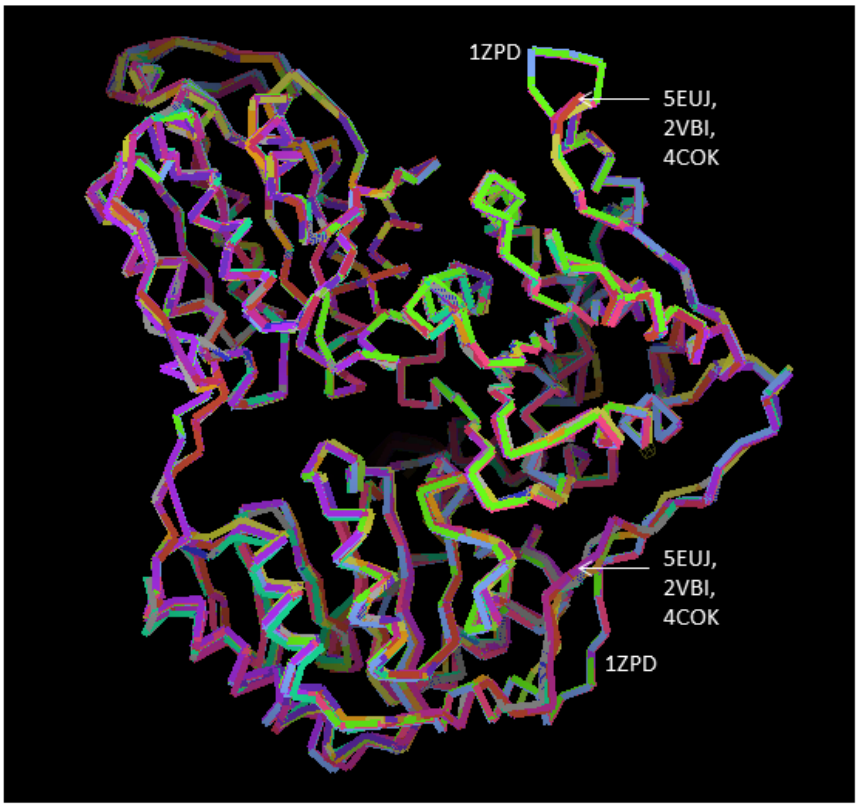


Figure 3.16 PROMALS3D structural sequence alignment of 5EUJ tetramer MNOP, 1ZPD, 4COK and 2VBI, visualized in *Coot* highlighting the 2 most variable linker regions.

3.4.3 The Active Site and TPP Binding

There is good quality electron density evidence for a TPP molecule in each of the 24 chains of 5EUJ. However, as noted previously (Dobritzsch *et al.* 1998), the bound TPP appears chemically modified at the C2. The electron density evidence suggests that the thiazolium ring has been opened and the C2 atom has been lost as seen in 1ZPD (Figure 3.17). This was suggested to be most likely due to partial degradation of the TPP during crystallisation (Dobritzsch *et al.* 1998) or possibly radiation damage during data collection since the break is in close proximity to a sulphur atom.

Through a pseudo 222 symmetry, the four monomers in a tetramer create four cofactor and substrate binding sites, located in narrow clefts on the interfaces between the PYR domain of one monomer and the PP domain of a second monomer. The TPP molecule binds as indicated

by the domain nomenclature (Figure 3.17), with the pyrophosphate group binding to the PP domain of one monomer and the pyrimidine ring binding to the PYR domain of a second monomer in the dimer.

Ile410 from the first monomer holds the TPP in a V-shape (Figure 3.17). The backbone amide groups of Ile467 and Val466 form hydrogen bonds with the oxygen atoms of the pyrophosphate. Glu49 (from the second monomer) forms a hydrogen bond to the N1 atom on the pyrimidine ring. This is essential for the C2 deprotonation mechanism (Pei *et al.* 2010). Other residues involved in TPP binding are Tyr465 and Glu468 from one monomer around the pyrophosphate, and Thr 71 from the second monomer around the pyrimidine ring.

A water molecule in the second monomer supports the active site arrangement, interacting with Asp26, Thr71, and His113, and is present even in the absence of substrate. This is thought to play a pivotal role in the organisation of the substrate complex and of the hydrogen bond network (Pei *et al.* 2010).

The pyrophosphate group of the TPP is anchored to the protein by a Mg(II) ion. It forms an octahedral coordination sphere with the two oxygen atoms on the diphosphate group of the TPP, the side-chain oxygen atom of Asp435 and Asn462, the main-chain oxygen atom of Gly464, and a water molecule (Figure 3.17).

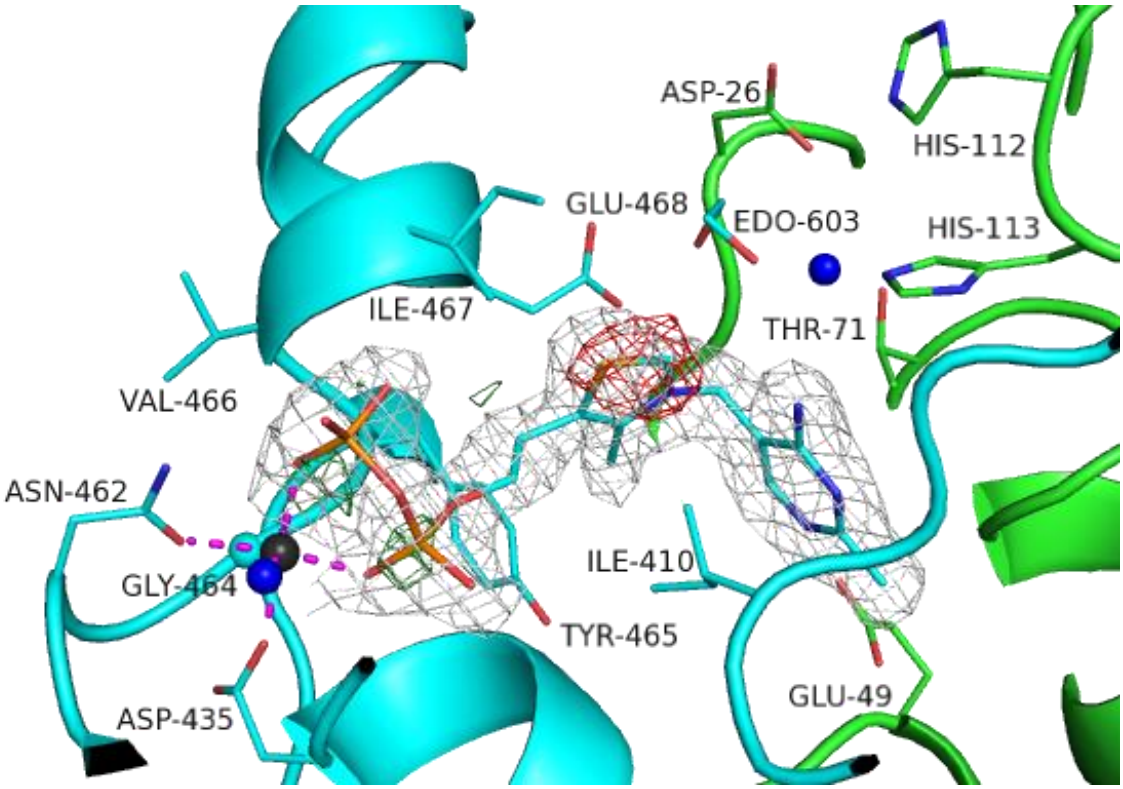


Figure 3.17 Cartoon and stick depiction of TPP binding in the active site. Residues of one monomer are coloured in cyan; residues of the other monomer are coloured in green. The magnesium (dark grey) and water (blue) molecules are represented as spheres. The cyan chain's 1,2-ethanediol (EDO) and TPP are shown as stick models and coloured by atom. The $2F_oF_c$ density (grey) surrounding the TPP is shown contoured at 1σ. The F_oF_c density is shown contoured at 3σ (dark green) and at -3.4σ (red).

The active sites appear to be connected by a water tunnel (Figure 3.18). This has been noted in PDCs and other TPP-binding enzymes (Frank *et al.* 2004, Pei *et al.* 2010). Pei *et al.* (2010) suggested it may play a role as a form of communication system between the active sites, perhaps as a proton relay system. ZpPDC 5EUJ shows the same pattern of water molecules as ZmPDC 2WVA. Similarly, the residues lining the tunnel seem to be well conserved and include Glu49, Asn48, Leu50, and His409. Frank *et al.* (2004) described a water tunnel linking two active sites in E1 of the PDH complex in *G. stearothermophilus*, which is a prerequisite for the ping-pong kinetics displayed by this enzyme and many other TPP-dependent enzymes. Frank *et al.* proposed a proton-wire model stating that the first TPP is activated by binding, while the other is activated by the decarboxylation and consequent proton relay from the first active site. This coordinates the uptake of substrates and release of products. However, the water

tunnel described by Frank *et al.* is much richer in acidic residues than the water tunnel found in the ZpPDC.

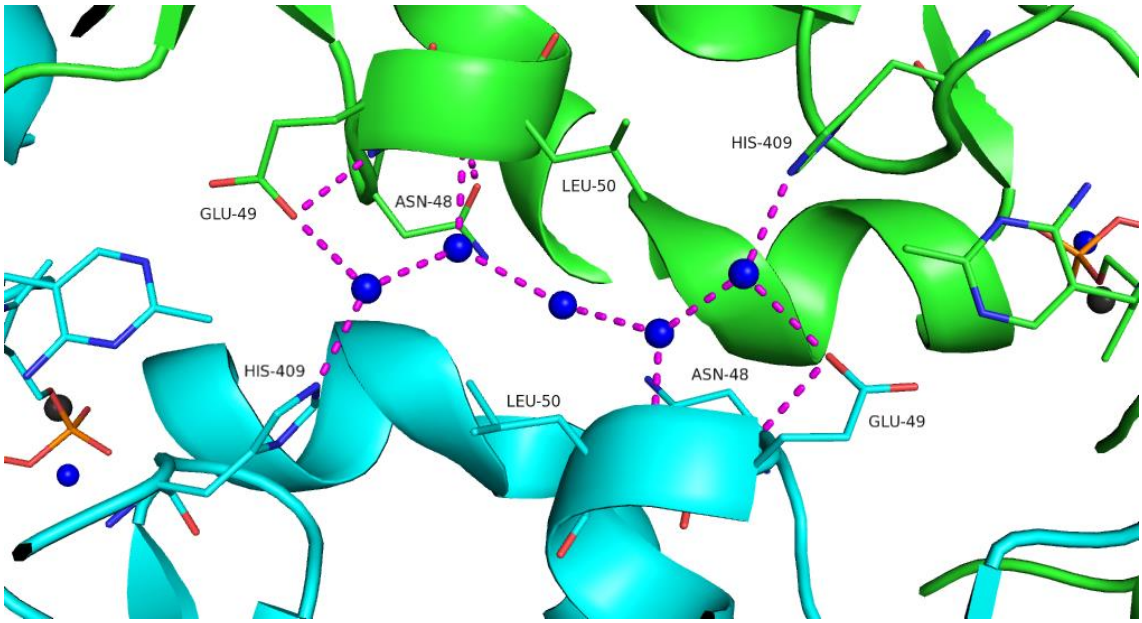


Figure 3.18 Model of the water tunnel connecting the two active sites in the ZpPDC dimer. Water molecules are shown as blue spheres; magnesium ions as dark grey spheres. Residues and the TPP from one monomer have carbon atoms and cartoon coloured cyan, while the other monomer in the dimer is coloured green.

A comparison of ZmPDC apo- and holo-structures by Pei *et al.* (2010) revealed that TPP binding induces a conformational change involving the loop and adjacent α-helix between Asn467-Tyr481 (Asn462-Tyr476 in 5EUJ). This region is structured in the apo-enzyme, and thus requires major conformational changes to accommodate TPP binding (Pei *et al.* 2010). Similar changes would presumably occur in ZpPDC as this region is structurally highly-conserved in bacterial PDCs.

3.4.4 Substrate Binding and the Catalysis Mechanism

As mentioned above, no pyruvate was found in the crystal structure presented here. However, EDO was bound in the active site of some chains (Figure 3.17), acting as a mimic of part of a pyruvate molecule as can be seen by the very good alignment with the positioning of pyruvate when a PDC structure containing pyruvate, such as the ZmPDC structure 2WVA chain F, is

superposed (Figure 3.19). This allowed the observation of interactions that might be made with pyruvate and comparison with those reported by Pei *et al.* (2010) and Dobritzsch *et al.* (1998).

Glu468 and Tyr289 from one monomer, and Asp26, His112 and His113 from a second monomer, interact with EDO through an extensive hydrogen bond network (Figure 3.19).

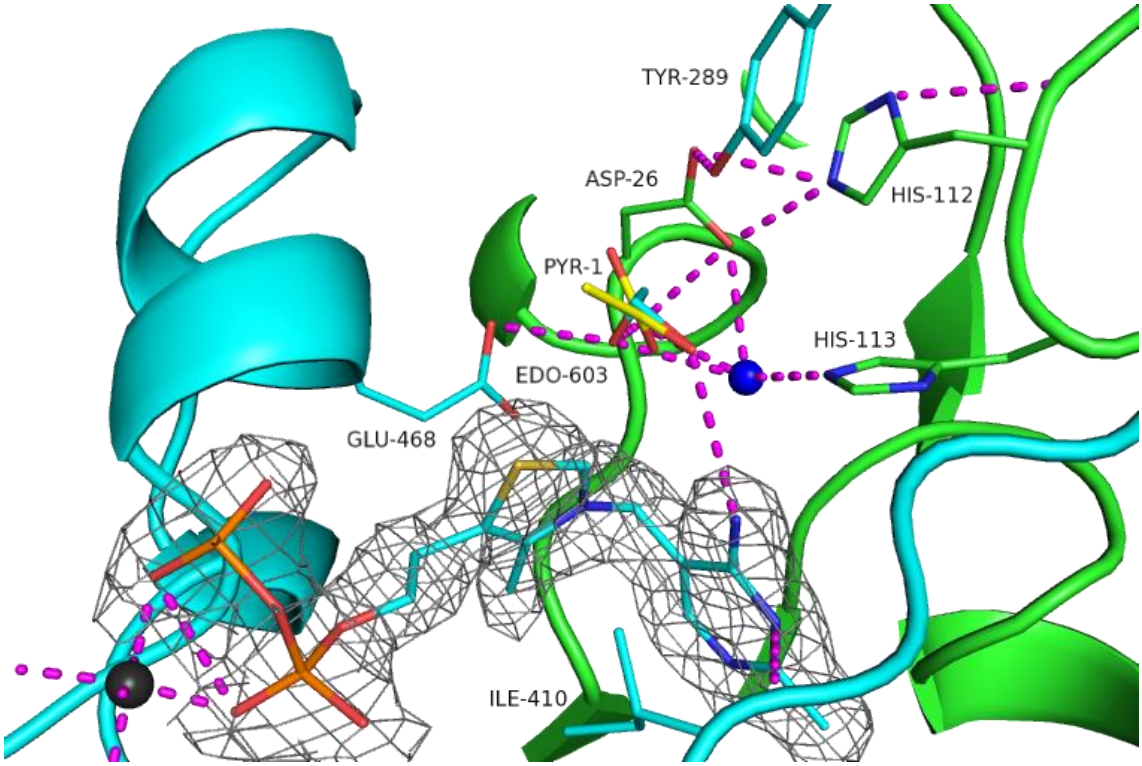


Figure 3.19 Cartoon and stick depiction of the active site. Residues of one monomer are coloured in cyan; residues of the other monomer are coloured in green. The magnesium (dark grey) and water (blue) molecules are represented as spheres. The cyan chain's 1,2-ethanediol (EDO) and TPP are shown as stick models and coloured by atom. Pyruvate (PYR, yellow) has been overlaid following a superposition of *Z. mobilis* PDB entry 2WVA chain F and lies on top of the EDO. Selected hydrogen bond interactions within the active site are shown as dotted magenta lines. The $2F_oF_c$ density (grey) surrounding the TPP is shown contoured at 1σ.

Glu468 is very likely to have a key role in catalysis, and is thought to be the proton acceptor from the 4'-NH_2 group of TPP, thus deprotonating C2 to form the active ylid (Kern *et al.* 1997, Pei *et al.* 2010) (Figure 3.20, step 3). It is further thought to form a stabilising hydrogen bond

interaction with the dianion formed after the nucleophilic attack of the thiazolium ring carbanion on pyruvate (step 4, Dobritzsch *et al.* 1998). In the lactyl-TPP intermediate, the five membered ring is positively-charged, facilitating a reverse proton transfer from Glu468 to 4'-N (step 4). The now negatively-charged Glu468 destabilises the adjacent carboxylate group on the pyruvate and thus facilitates subsequent decarboxylation (step 5, Pei *et al.* 2010). Tittman *et al.* (2003) and Meyer *et al.* (2010) investigated the role of this residue further, and through a variety of mutants found it to be crucial in substrate binding and catalysis. His112 may not be directly involved in catalysis, but plays an important role in maintaining the active site environment, in particular in retaining His113 uncharged. This in turn is essential to allow proton abstraction from the C2 in the first step of catalysis (step 3, Dobriztsch *et al.* 1998). Furthermore, His112 is likely to be involved in holding the carboxylate group of Asp26 in the correct state and position, supported by Tyr289 (Pei *et al.* 2010). Asp26 may also be involved in acetaldehyde release (step 6, Pei *et al.* 2010). His113 interacts with pyruvate (O3) and TPP (N4') (Dobritzsch *et al.* 1998).

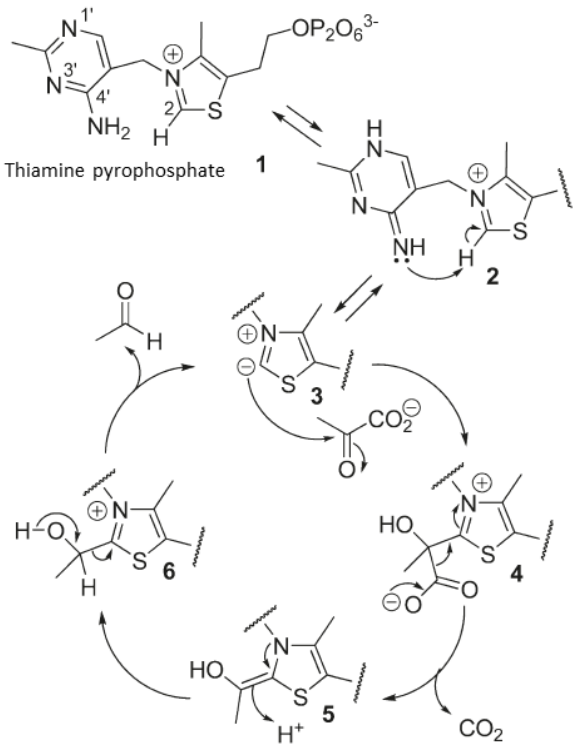


Figure 3.20 Representation of the catalytic cycle of pyruvate decarboxylation by PDC. Protonation at N-1' and deprotonation at 4'-NH_2 of the TPP (1) give rise to the amino tautomer (2). This in turn promotes the deprotonation of C2 on the thiazolium ring to form the active ylid (3). The ylid attacks C2 of pyruvate, thus creating the lactyl adduct (4). Decarboxylation then leads to the enamine intermediate (5) and protonation creates hydroxyethyl-TPP. The release of the acetaldehyde produced regenerates the ylid (3). From Pei *et al.* (2010).

Dobritzsch *et al.* (1998) remarked that large conformational changes upon substrate binding are unlikely due to the extensive interface regions. Instead, it is thought that the C-terminal helix swings out of the way to allow access to the active site, and closes upon substrate binding to create a hydrophobic active site environment. This helix is exposed in ZpPDC as well, so it is likely that a similar mechanism applies here.

3.5 Final Remark

In summary, this chapter presented the crystal structure of the *Z. palmae* PDC and a functional and structural comparison to known bacterial PDCs. Bacterial PDCs are structurally well conserved, which may allow in-depth studies carried out on ZmPDC of the mechanism of folding as presented by Pohl *et al.* (1994) and the mechanism of catalysis as described by Dobritzsch *et al.* (1998), Pei *et al.* (2010) and Meyer *et al.* (2010), to be applied to ZpPDC.

Structural analysis suggests that the different thermostability and thermoactivity displayed by these PDCs may be correlated to increased oligomeric interfaces and salt bridges as has been seen in many other protein families (Sterner & Liebl 2001). This Thesis hereby adds to the structural knowledge of bacterial PDCs, generating information that has the potential to be very useful in design approaches for enzyme engineering and biotechnology applications. Unfortunately, rational design approaches utilizing the crystal structure data were beyond the scope of this project as the crystal structure only became available at the end of this Thesis project.

However, the high *in vitro* thermostability and thermoactivity of ZpPDC, with a denaturation temperature of 70°C and an optimum temperature of 65°C, would suggest the ZpPDC was well suited for high temperature expression. As mentioned, initial attempts to express the ZpPDC in *Geobacillus* spp. were not very successful, with a complete loss of activity being observed in cell extracts of *G. thermoglucosidasius* cultures grown at 50°C (Taylor *et al.* 2008).

Chapter 4 describes the improved expression of the ZpPDC in *G. thermoglucosidasius* up to 65°C, using an optimized expression system.

4. EXPRESSING *ZYMOBACTER PALMAE* PDC IN *GEOBACILLUS THERMOGLUCOSIDASIUS*

4.1 INTRODUCTION

In vitro ZpPDC shows a high temperature optimum at 65°C, and retains 80% activity after incubation at 65°C for 30 min (Raj *et al.* 2002; confirmed in this study, Chapter 3), making it one of the most thermostable bacterial PDCs currently known.

In the first attempt to utilize ZpPDC in *G. thermoglucosidasius*, Taylor *et al.* (2008) expressed ZpPDC aerobically in *G. thermoglucosidasius* grown at 45-50°C, and assayed the clarified cell extract for PDC activity by monitoring NADH depletion dependent decrease in absorbance at 340 nm (standard coupled assay). With an increase in growth temperature from 45°C to 48°C a sharp drop in PDC activity from 1067 to 89 nmol/min/mg total protein was observed. No activity was detectable in cells grown at 50°C.

Given the high *in vitro* thermostability of ZpPDC, an optimized expression system should allow PDC activity to be observed at *G. thermoglucosidasius* growth temperatures of 50-65°C.

This chapter reports the improved aerobic expression of ZpPDC from *Z. palmae* ATCC 51623 in *G. thermoglucosidasius* up to 65°C.

4.2 METHODS

4.2.1 CLONING WT ZP *PDC* FOR CHARACTERIZATION IN *GEOBACILLUS THERMOGLUCOSIDASIUS*

Expression of ZpPDC from p600 wtZpPDC (Waite 2010) could not be reliably replicated. Analysis of the promoter sequence failed to identify a functional ribosome binding site (RBS), and therefore, prompted the redesign of the expression vector with particular attention to the promoter-RBS combination. The *ldh* promoter-*pheB* RBS combination has been shown to work very effectively with GFP and *phe*B (personal communication with Ben Reeve; Bartosiak-Jentys *et al.* 2012). Thus, the *pdc* sequence was spliced into pGR002, which already contains this promoter-RBS combination.

The wt *pdc* gene was amplified from p600 wtZpPDC using the following primers: wtpGR002F (CTA GTC TAG ATA AGG AGT GAT TCG AAT GTA TAC CGT TGG TAT GTA CTT GGC) and wtpGR002R (**GAG CTC** GCA AAA AAA CGC CCC CTT TCG GGG CGC GAT TAC GCT TGT GGT TTG

CGA GAG), and Phusion® Hot Start II polymerase (see General Methods) with an annealing temperature of 67.7°C and 1:30 min extension at 72°C. This PCR added an *Xba*I recognition site (underlined) and the *pheB* RBS (in blue) to the upstream region of the gene, and the *pheB* downstream region (in green) containing a transcriptional terminator region and a *Sac*I recognition site (bold, underlined) to the downstream region of the *pdc*.

The PCR product was cloned into the pJET1.2 cloning vector (see General Methods), which was then transformed into *E. coli* BioBlue, reisolated and sequenced with pJET1.2 F/R and wtZpPDC F1/F2. A clone containing an insert with the correct sequence was then chosen for digestion with *Xba*I and *Sac*I, as was the pGR002 vector. The released fragment and the linearized vector were agarose gel purified, ligated (using 20 ng linearized vector and 17 ng insert) into pGR002, and the product transformed into *E. coli* BioBlue.

An *E. coli* BioBlue colony carrying the correct clone of pGR002 wtZpPDC, as confirmed by sequencing with M13 F/R and wtZpPDC F1/F2, was used to propagate the plasmid, which was purified and transformed into *G. thermoglucosidasius* DL44 and TM236 (see General Methods). The resulting strains were characterized for PDC expression in aerobic conditions at 50°C, 60°C and 65°C.

4.3 Results

4.3.1 Cloning for Expression of ZpPDC in *Geobacillus thermoglucosidasius*

For expression of wt Zp*pdc* in *G. thermoglucosidasius*, an *E. coli-Geobacillus* spp. shuttle vector based on pUCG18 (Taylor *et al.* 2008) carrying the wt Zp*pdc* was made through PCR amplification of the *pdc* from an available construct (Waite 2010) (see Figure 4.1 for PCR results), and cloning of the gene into pGR002 using the *Sac*I and *Xba*I restriction sites (Bartosiak-Jentys *et al.* 2012) (Figure 4.2), putting the *pdc* under the control of the *G. stearothermophilus* NCA1503 *ldhA* promoter combined with the highly efficient RBS from *G. stearothermophilus* DSMZ6285 *pheB*. pGR002 wtZpPDC plasmid was then transformed by electroporation into *G. thermoglucosidasius* DL44 (DL33 Δ*ldh*) and TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*).

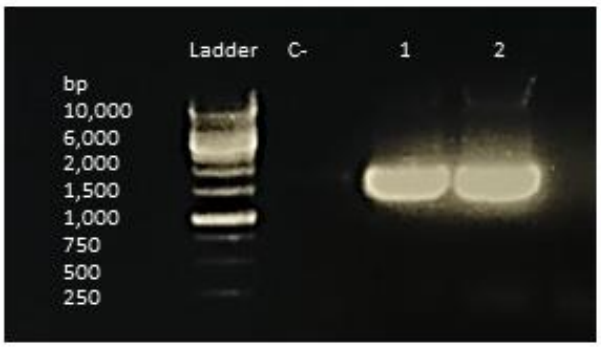


Figure 4.1 Agarose gel electrophoresis of the *pdc* amplification product. The 1,727 bp PCR-amplified product was visualized alongside the GeneRuler™ 1kb ladder (Thermo Fisher Scientific). C- is a PCR negative control containing no template. Lanes 1 and 2 are from replicate PCR reactions.

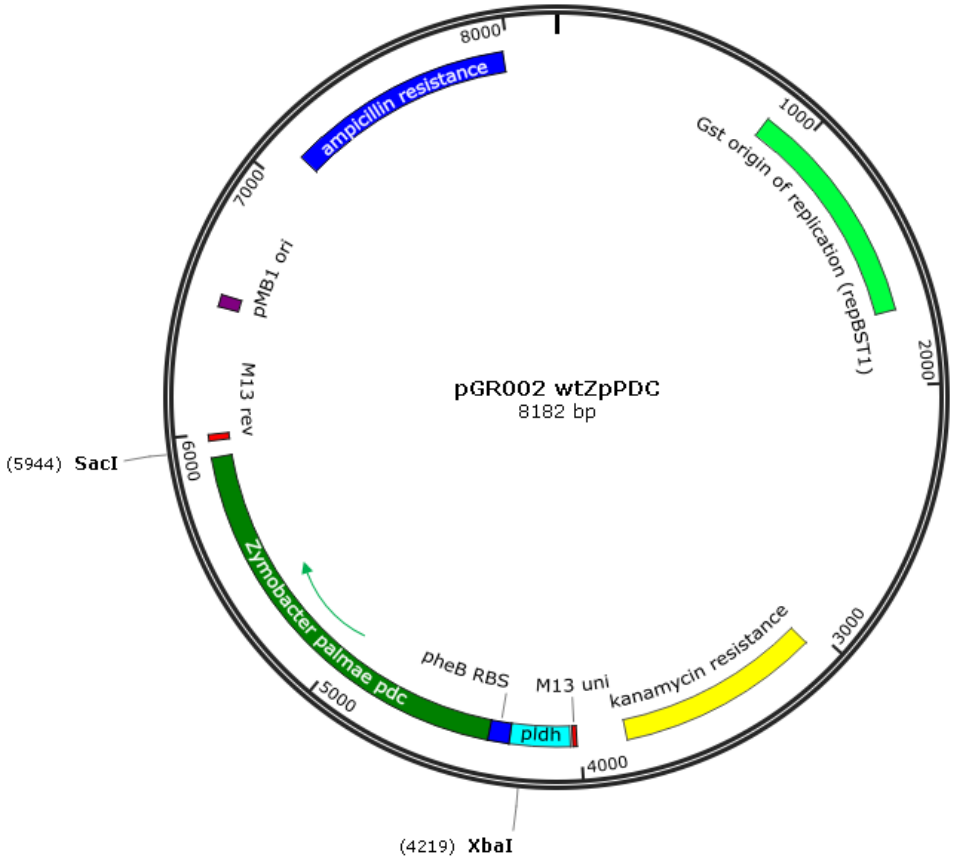


Figure 4.2 Plasmid map of pGR002 wtZpPDC. pGR002 wtZpPDC is based on the *E. coli-Geobacillus* spp. shuttle vector pUCG18, containing features for expression in both hosts. RepBST1 and the kanamycin resistance gene are *G. stearothermophilus* derived and thus allow replication and selection of the plasmid at higher growth temperatures, whereas pMB1 ori and the ampicillin resistance gene are *E. coli* derived. Pldh denotes the *G. stearothermophilus* NCA1503 *ldhA* promoter, while *phe*B RBS marks the position of the ribosome binding site from the *G. stearothermophilus* DSMZ6285 *pheB* gene. M13 uni/rev are sequencing primer complimentary sequences.

4.3.2 Aerobic Expression of ZpPDC in *Geobacillus thermoglucosidasius*

G. thermoglucosidasius DL44 (DL33 Δ*ldh*) and TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) carrying pUCG18 (empty plasmid backbone) or pGR002 wtZpPDC were grown aerobically in 50 ml 2TY + 12 µg/ml kanamycin + 5 mM thiamine in 250 ml baffled flasks, with shaking at 250 rpm, at 50, 60 and 65°C to an OD_{600nm} of 1.5 to 2.5. After sonication, the clarified cell extract from these cells was assayed for PDC activity (see Table 4.1).

As noted in Table 4.1, the PDC activities observed were generally higher in the TM236 expression background. Moreover, the activity rapidly decreased with increasing growth temperature. Very little activity was detected in cells grown at 65°C, at which point it approached the detection limits of the assay. No over-expressed protein band could be identified by SDS-PAGE analysis (see example gel image below, Figure 4.3); the PDC monomers would be expected to appear at a relative molecular mass (Mr) of 59.4 kDa.

Table 4.1 PDC activities in cell extracts of *G. thermoglucosidasius* DL44 and TM236 carrying pGR002 wtZpPDC grown aerobically at various temperatures.

Growth temperature	Specific activity (µmol/min/mg total protein)		
	DL44/TM236 pUCG18	DL44 pGR002 wtZpPDC	TM236 pGR002 wtZpPDC
50°C	0	No data	1.06 ±0.33
60°C	0	0.081 ±0.018	0.53 ±0.34
65°C	0	0	0.16 ±0.03

Specific activity was determined from cell extracts by monitoring the decrease in NADH-dependent absorbance at 340 nm in standard coupled assays at 30°C. At least two independent expression experiments were carried out, and each was assayed three times. The error is standard error.

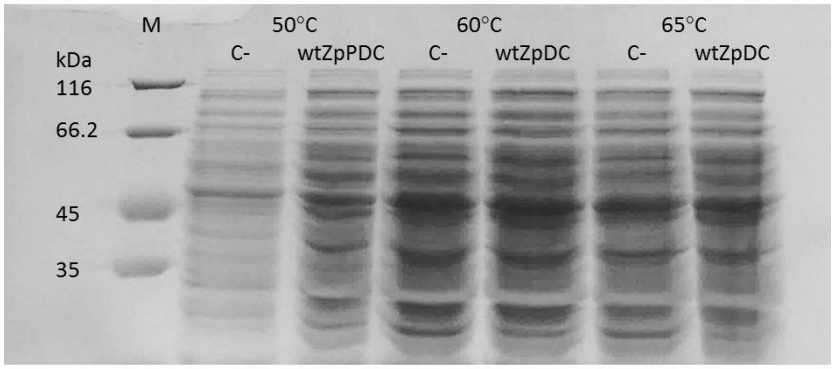


Figure 4.3 SDS-PAGE analysis of *G. thermoglucosidasius* TM236 pGR002 wtZpPDC following expression under aerobic conditions. TM236 was grown at 50, 60 and 65°C; cells were lysed and the extract clarified by centrifugation. The soluble fraction (~10 µg total protein) was run on a 12% SDS-PAGE gel. C- is TM236 carrying an empty pUCG18 plasmid. WtZpPDC is TM236 carrying the pGR002 wtZpPDC expression vector. M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific). Wt ZpPDC would be expected at 59.4 kDa.

G. thermoglucosidasius DL44 carrying pGR002 *phe*B (Bartosiak-Jentys *et al.* 2012) served as an expression control, and were positive for the catechol conversion assay at all growth temperatures (Figure 4.4). This suggested that growth conditions were suitable to trigger expression from the pGR002 p*ldh*.

Figure 4.4 Catechol test on culture samples. *G. thermoglucosidasius* DL44 carrying pGR002 *pheB* was grown under the same conditions as DL44 carrying pGR002 wtZpPDC. Shown here is a culture grown at 60°C to OD_{600nm} 2. *Phe*B encodes catechol 2,3-dioxygenase, which converts catechol into 2-hydroxymuconic semialdehyde, which has a vivid yellow colour. Cells expressing *phe*B can be easily identified by covering them in a 100 mM catechol solution and observing the colour change. Left: DL44 pGR002 *phe*B, yellow indicates a positive test. Right: DL44 pGR002 wtZpPDC, white indicates a negative test.

4.4 Discussion

In a first attempt at expressing wt Zp*pdc* in *G. thermoglucosidasius* DL44 (DL33 Δ*ldh*) Taylor *et al.* (2008) reported PDC activities of 1067 nmol/min/mg total protein from cells grown at 45°C and 89 nmol/min/mg total protein from cells grown at 48°C. No activity was detectable in cells grown at 50°C. With a change in promoter-RBS combination to p*ldh*-RBS*ldh*-RBS*phe*B we were able to considerably improve expression at higher temperatures, and detected PDC activity of 81 ±18 nmol/min/mg total protein at 60°C growth temperature (although no activity in cells grown at 65°C).

Expression seemed to be further improved by changing the expression background to *G. thermoglucosidasius* TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*), with PDC activity being detected up to a growth temperature of 65°C. PDC activity was 510 ±340 nmol/min/mg total protein from cells grown at 60°C and 160 ±30 nmol/min/mg total protein from cells grown at 65°C.

Nevertheless, a drastic decrease in detected PDC activity at increasing growth temperatures was observed. Considering the high *in vitro* thermostability and thermoactivity of the pre-folded protein and the limited *in vivo* expression at higher temperatures, this disparity points towards a problem in expression at high temperatures rather than thermostability *per se*. One limiting factor may be impaired translation and co-translational folding around TPP.

Codon harmonization has been found to be a successful strategy for optimizing translation in a recombinant host (Angov *et al.* 2008). Unlike codon optimization, which uses high-frequency codons based on the recombinant host for high-titre expression, codon harmonization aims to recreate the pattern of codon usage frequency along the protein as found in the native host with the objective that this allows appropriate co-translational folding in the recombinant host.

However, very little data are available on *Z. palmae* as an expression host, so expanding the available information by genome sequencing is the focus of Chapter 5, and is a prerequisite for an attempt at applying codon harmonization to Zp*pdc* to improve expression in *G. thermoglucosidasius* presented in Chapter 6.

5. *ZYMOBACTER PALMAE* GENOME SEQUENCING

5.1 INTRODUCTION

Z. palmae was first isolated from palm tree sap in Okinawa Prefecture, Japan (Okamoto *et al.* 1993). *Z. palmae* is a Gram-negative, facultatively-anaerobic, non-sporeforming, motile rod. *Z. palmae* is a mesophilic organism and optimum growth occurs at 30°C.

Z. palmae is part of the *Halomonadaceae* family (de la Haba *et al.* 2010), but is the only member that utilizes PDC in a homo-fermentative metabolism. It produces ethanol as its major fermentation product from a variety of hexose sugars and oligo-saccharides (Horn *et al.* 2000, Okamoto *et al.* 1993).

Engineering efforts have been undertaken to increase the use of *Z. palmae* for commercial ethanol production (Yanase *et al.* 2005 & 2007). Furthermore, the *pdc* gene from *Z. palmae* ATCC 51623 has been used to engineer a variety of ethanologenic strains, including *G. thermoglucosidasius* (Taylor *et al.* 2008) and *Lactococcus lactis* (Liu *et al.* 2005).

Expression of ZpPDC in *G. thermoglucosidasius* at 50°C, 60°C and 65°C showed a drastic decrease in detected PDC activity with increasing growth temperatures (Chapter 4). Considering the high *in vitro* thermostability and thermoactivity of the ZpPDC, the limited *in vivo* expression at higher temperatures may point towards impaired translation and co-translational folding around the co-factor TPP. Codon harmonization has been shown to be a successful strategy for optimizing translation in a recombinant host (Angov *et al.* 2008). Codon harmonization strategies require data on codon usage frequency and t-RNA availability of the native and recombinant host. The *G. thermoglucosidasius* C56-YS93 genome had been sequenced previously (Genbank: NC_015660, NC_015661, NC_015665). However, at the start of this work no genome data were available for any of the *Z. palmae* strains, so the data available for codon harmonization were extremely limited.

The *Z. palmae* type strain T109/ ATCC 51623 was acquired from the NBRC, Japan, and genomic DNA was sent for sequencing using Ion Torrent technology. This chapter describes the manipulation of the sequencing data into a draft genome. The data obtained were a valuable addition to allow codon harmonization to be applied to the Zp*pdc* with the potential of improving expression in *G. thermoglucosidasius* (described in Chapter 6).

5.2 Methods

5.2.1 Taxonomic Studies of Species Phylogenetically Related to *Z. palmae*

Taxonomic studies used 16S rRNA sequences (Table 5.1) aligned using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/, Sievers *et al.* 2011). The alignment was then fed into ClustalW Phylogeny (http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/, Larkin *et al.* 2007) with the default paramters (not excluding gaps, not correcting for distance, clustering by neighbour-joining). The resulting phylogenetic tree was visualized using EvolView (http://evolgenius.info/evolview.html, Zhang *et al.* 2012).

For some of the organisms no separate 16S rRNA entry was found in the GenBank database (http://www.ncbi.nlm.nih.gov/genbank/), so these were identified using the BLAST function in the CLC Genomics Workbench (CLCBio, Aarhus, Denmark) searching against the *Z. palmae* 16S rRNA (GenBank: D14555.1).

Table 5.1 Summary of strain details.

Organism	Strain	NCBI Taxonomy ID	Genome accession number	16S rRNA sequence accession	Identity (%)
Zymobacter palmae	ATCC51623	33074	NA	D14555.1	100
Zymomonas mobilis	ATCC10988	555217	NC_017262.1	AF281031.1	78.9
Acetobacter pasteurianus	386B	1266844	NC_021991	BLAST	41.8
Halomonas elongata	ATCC33173	768066	NC_014532.1	BLAST	90.3
Chromohalobacter salexigens	ATCCBAA-138	290398	NC_007963.1	BLAST	90.9

5.2.2 DNA Extraction and Purification

Genomic DNA was extracted from overnight cultures using DNeasy® Blood and Tissue Kit (Qiagen, Product code: 69504) following the supplier's protocol for Gram-negative and Gram-positive bacteria. The incubation period for step 2 of the animal tissue spin-column protocol was 30 min and the optional step of adding 4 µl 100 mg/ml RNase A (Sigma) to remove residual RNA was included.

5.2.3 Genome Sequencing and *De Novo* Assembly

Genomic DNA was extracted from *Z. palmae* T109 (ATCC 51623) as described above. The quality of the preparation was checked by agarose gel electrophoresis and spectrophotometric quantification using a NanoVue (GE Healthcare). Two samples were sent to the University of Bristol Genomics Facility according to their specifications (2-3 µg per sample). An Ion Xpress Plus gDNA library was constructed and sequenced using an Ion Torrent technology 316 chip.

The sequencing reads were assembled into contigs using the CLC Genomics Workbench 6.0.5 (CLCbio); see Figure 5.1 for the workflow. Reads were trimmed with the quality score set to 0.05 and a maximum of 2 ambiguous nucleotides allowed in read ends. The reads were assembled *de novo* using de Bruijn graphs with a words size of 21 nucleotides and a bubble size of 175 bp. Input reads were mapped back onto the assembled contigs using the default parameters: mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.5 and similarity fraction 0.8.

The contigs were further analysed and assembled using the CLC Microbial Genome Finishing Tool (CLCbio). The first round of automatically joining the contigs used the input contigs as references as well as the genome sequences of *Z. mobilis* (NC_017262.1), *A. pasteurianus* (NC_021991), *H. elongata* (NC_014532.1) and *C. salexigens* (NC_007963.1) with a reference genome weight factor of 1 and a contig overlap weight factor of 6. Minimum BLAST word size was 20, minimum e-value 0.1 and minimum match size of 5 bp. The second round of automatically joining the contigs used the same parameter settings, but no reference genomes. The third round of automatically joining the contigs used a BLAST word size of 30 bp. The resulting contigs were further joined manually.

The trimmed reads were mapped back against the resulting 54 contigs using mismatch cost 1, insertion cost 1, deletion cost 1, length fraction 0.1, similarity fraction 0.5. Contigs were extended using the read consensus sequences and manually joined. The trimmed reads were mapped again using mismatch cost 2, insertion cost 2, deletion cost 3, length fraction 0.5, similarity fraction 0.9, and the resulting consensus sequence was extracted as a FASTA file.

Further analysis of the genome sequence included identification and annotation of predicted coding regions using PROKKA (http://www.vicbioinformatics.com). The *Z. palmae* genome-wide CUF data were obtained by analysing the genome using DAMBE (v.4.0.36) (Xia & Xie 2001). A preliminary metabolic model was constructed using RAST (Aziz *et al.* 2008) and used in the analysis of carbohydrate metabolism and fermentation pathways.

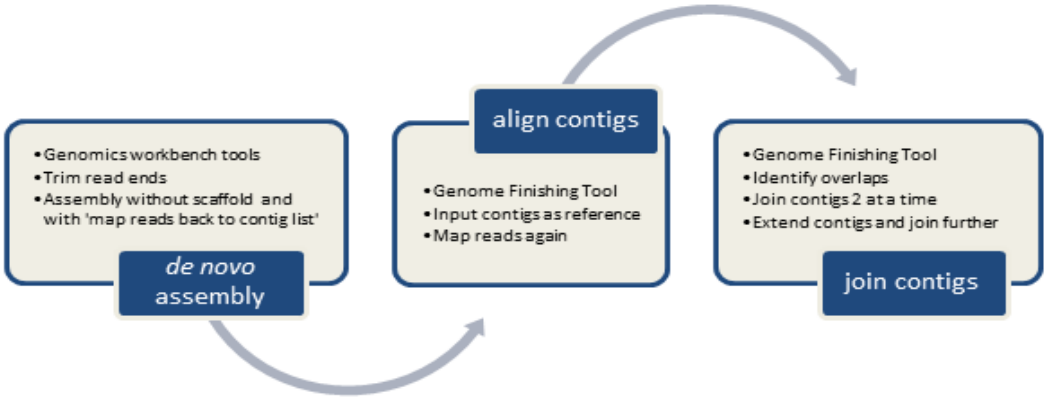


Figure 5.1 Workflow of genome assembly. *De novo* assembly was carried out using the CLC Genomics Workbench 6.0.5. Contigs were further aligned and joined using the CLC Microbial Genome Finishing Tool.

5.3 Results

5.3.1 Assessing Species Phylogenetically Related to *Z. palmae*

In association with the genome sequencing project presented here phylogenetic relations of *Z. palmae* were investigated. *Z. palmae* is part of the *Halomonadaceae* family (de la Haba *et al.* 2010). Published genome sequences of organisms in the same family include *Halomonas elongata* ATCC33173/DSM2581 and *Chromohalobacter salexigens* ATCCBAA-138/DSM3043. The *Z. palmae pdc* is similar to *pdc* genes found in *Z. mobilis* ATCC10988 and *A. pasteurianus* 386B, so these organisms were included in the analysis (see Table 5.1 for a summary of the strain details). Comparing 16S rRNA produced the phylogenetic tree shown in Figure 5.2.

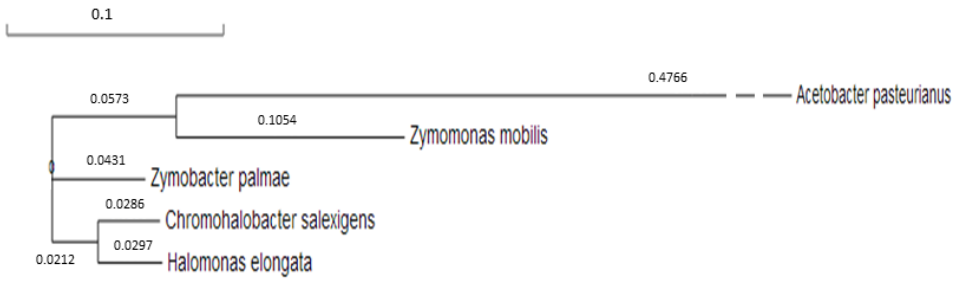


Figure 5.2 Unrooted phylogenetic tree. Neighbour-joining phylogentic tree based on 16S rRNA gene sequence comparison. The bar indicates 0.1 substitutions per nucleotide position, also given on each branch.

This suggested that aligning the genome sequencing reads against *C. salexigens* and *H. elongata* would be a useful tool in assembling the *Z. palmae* genome. Furthermore, using *Z .mobilis* and *A. pasteurianus* in the assembly may be useful as they contain ethanol fermentation pathways not found in *C. salexigens* and *H. elongata*.

5.3.2 Processing of the Genome Sequencing Data into Contigs

The genome sequencing run resulted in 3,121,811 reads with a mean read length of 176 bp and a total nucleotide sum of 550 Mb. The sequencing reads were assembled into contigs using the CLC Genomics Workbench 6.0.5.

In the first step the reads were trimmed in order to remove low quality reads (quality score limit set to 0.05) and ambiguous read ends (maximum 2 nucleotides allowed). Sequencing data is generally annotated with a quality score, which indicates the probability of base-calling error. Trimming annotates low quality or ambiguous regions in the reads, which are then ignored during assembly. This removed 50 low quality reads and annotated 497,736 reads bringing the mean read length to 175 bp.

Next, the reads were assembled *de novo* using de Bruijn graphs. The appropriate size of words (unique sub-sequences of a certain length) and bubbles (a bifurcation in the Bruijn graph caused by a sequencing error in some of the overlapping reads) was automatically determined based on the amount of input data (the more data, the larger the word size). In this case the software used a word size of 21 bp and a bubble size of 175 bp. In summary, the CLC assembler went through the following general procedures: (1) creation of a table of "words" observed in the sequence data, (2) creation of a de Bruijn graph from the word table, (3) utilization of the input reads to resolve small repeats or errors, (4) output of the resulting contigs. Contigs of a minimum size of 2,000 bp were generated. This resulted in 100 contigs of an average length of 29,365 bp (maximum 114,421, minimum 2,065) and of 2,936,480 bp in total (Figure 5.3 for contig length distribution). Half the contigs taken together account for 2.5 Mb, with a total accumulated contig length of 3 Mb. N75, N50 and N25 were 29,398, 52,752 and 71,632 bp, respectively. These values are calculated by summing the lengths of the contigs from longest to smallest until 25, 50 or 75% of the total contig length is reached. See Table 5.2 for nucleotide distribution data. This indicated a GC content of 56.2%, which is in agreement with the experimentally determined GC content of 55.8 ±0.4 mol% (Okamoto *et al.* 1993).

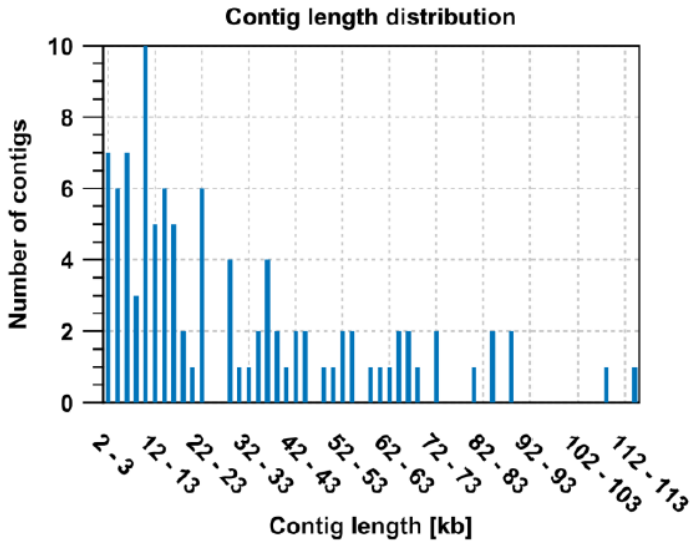


Figure 5.3 Contig length distribution. Shown here are the number of contigs of various lengths.

Table 5.2 Nucleotide distribution. Summary of *Zymobacter palmae* genome nucleotide distribution.

Nucleotide	Count	Frequency (%)
Adenine (A)	641,443	21.8
Cytosine (C)	821,014	28.0
Guanine (G)	828,172	28.2
Thymine (T)	645,851	22.0

After *de novo* assembly, the assembler mapped the input reads back to the reference (i.e., the contigs generated) using the following default parameters: mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.5 and similarity fraction 0.8. A match is always scored 1. The various costs determine how the reads should be aligned to the reference. The mismatch cost refers to a mismatch between read and reference sequence. The insertion cost refers to an insertion in the read sequence, causing a gap in the reference sequence, and the deletion cost refers to a gap in the read sequence. A length fraction of 0.5 means that at least half of the read has to match the reference sequence, in order for that read to be included in the final mapping. A similarity fraction of 0.8 means that the read has to have at least 80% sequence identity to the reference sequence across the length fraction (here half of the read length) in order for that read to be included in the final mapping. The mapping tool of the assembler then uses local alignment to map the reads onto the reference. This allows read ends to

remain unaligned if they do not match the reference sequence. This is desirable, as the sequencing quality often drops towards the end of a read. If no reads map back to the reference sequence, it is assumed that there is no data to support this sequence. Therefore, areas of zero coverage will be deleted from the final contigs. This resulted in 97.6% of the reads being mapped (2.4% or 75,094 unmapped reads, average length between 1 and 144 bp).

The contigs were further analysed and assembled using the CLC Microbial Genome Finishing Tool (CLCbio). A first round of automatically joining the contigs used the input contigs as references as well as the genome sequences of *Z. mobilis* (NC_017262.1), *A. pasteurianus* (NC_021991), *H. elongata* (NC_014532.1) and *C. salexigens* (NC_007963.1) with a reference genome weight factor of 1 and a contig overlap weight factor of 6. Minimum BLAST word size was 20, minimum e-value 0.1 and minimum match size of 5 bp. Here, the software used the "align contigs" tool to visually inspect how the contigs align to a reference or the contigs themselves and performed pairwise joins on contigs with at least 80% weighted overlap. Of the 100 contigs 28 were joined. The resulting 86 contigs were put through a second round of automatically joining the contigs using the same parameter settings, but no reference genomes. The resulting 82 contigs went through a third round using a BLAST word size of 30 bp. The resulting 81 joined contigs were then further joined manually. Running the contigs through the aligner tool highlights overlaps, which can then be joined in a pairwise manner.

Joining contig overlaps manually resulted in 54 contigs. The trimmed reads were mapped against these again using mismatch cost 1, insertion cost 1, deletion cost 1, length fraction 0.1 and similarity fraction 0.5. This resulted in 97.92% of the reads being mapped (2.08% unmapped reads, average length between 1 and 144 bp).

The contigs of the original *de novo* assembly contig list were extended using read consensus sequences. Reads often continue outside the contig end, e.g, when the assembler failed to connect a repeat sequence to the contig. The consensus sequence of the reads in this area can then be used to extend the contig. Contigs not joined previously were merged with the joined contigs. This contig list was then put through another round of automatically joining contigs, which resulted in 47 contigs. These were again aligned and further joined manually.

The trimmed reads were again mapped against the resulting 38 contigs (93% of reads mapped, mismatch cost 2, insertion cost 2, deletion cost 3, length fraction 0.5, similarity fraction 0.9) and the consensus sequence for each contig was extracted as a FASTA file. The average contig length was 73,829 bases, with the smallest being 2,716 bases (contig 27) and the largest being

372,699 bases (contig 35). The contigs accumulate to an overall genome length of 2,805,510 bp, but it is evident that there are still gaps in the sequencing.

The average coverage of the final 38 contigs was 1.9 kb to 140 bases. Only very few ends (2 to 31 nucleotides at the 5' or 3' end of the contig) had a low coverage (<8), including contigs 14, 17, 21, 23, 24, 26, 27, and 37.

5.3.3 Genome Annotation

The 38 contigs were analysed in PROKKA (courtesy of Dr. Andrew Preston, University of Bath), which identified 2,531 coding regions without a pseudo or pseudogene qualifier. They are on average 911 bases long and the density is 0.902 genes per kb. This accounts for 2,307,933 bases (82%) of the genome. The GC content of the coding regions is 57.53%, which is very similar to the whole genome (56.2%).

Some genes of interest are located on the following contigs: pyruvate decarboxylase *pdc* 23, alcohol dehydrogenases *adh*I 28 and *adh*II 7, 16S rRNA 32. Furthermore, 51 tRNA genes have been identified (Figure 5.4, Appendix II). Analysis of the draft genome using DAMBE enabled the collection genome-wide CUF data, which are summarized in Figure 5.4.

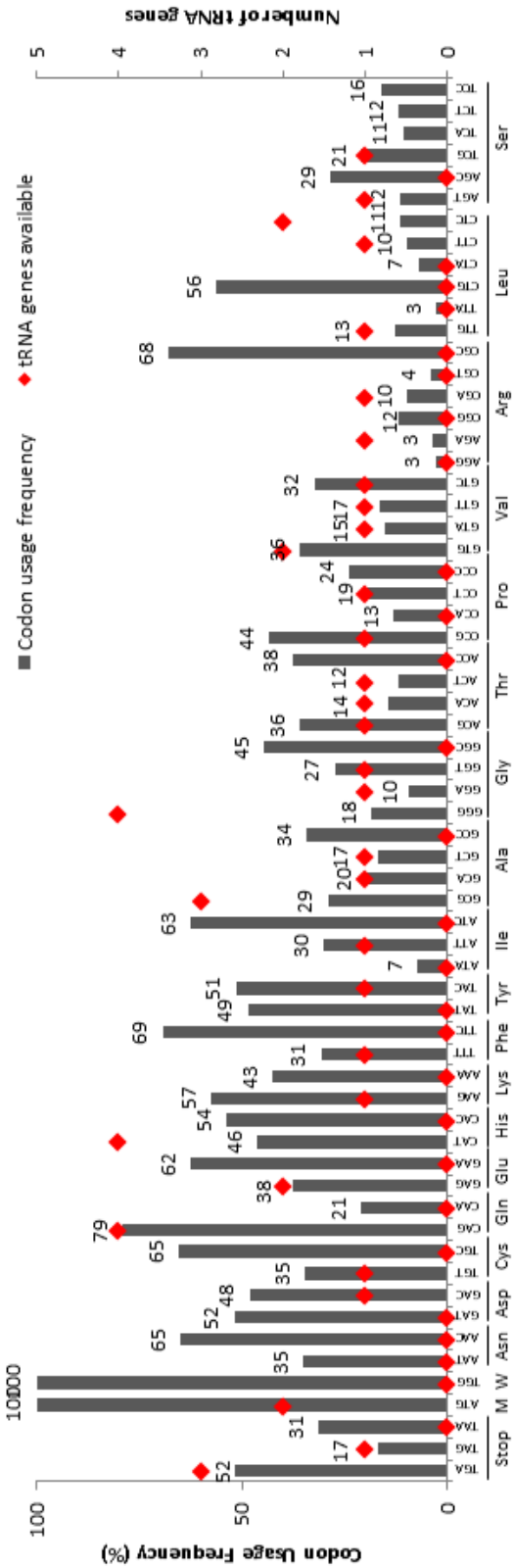


Figure 5.4 *Z. palmae* genome-wide codon usage frequency and tRNA data. Analysis of the *Z. palmae* draft genome using PROKKA identified 51 tRNA genes, displayed as red markers for each codon. Codon usage frequency (CUF) of all open reading frames found in the draft genome was analysed using DAMBE. CUF data is displayed as relative frequency in % in a grey bar chart.

Additionally, a preliminary metabolic model was constructed using RAST, which allowed the analysis of carbohydrate metabolism. Glucose metabolism appears to be directed through the Entner-Doudoroff pathway (Entner & Doudoroff 1952) (Figure 5.5). Genes for pentose-phosphate pathway enzymes are also present, including transketolase, ribose-phosphate pyrophosphokinase, 6-phosphogluconolactonase, transaldolase, riobose 5-phosphate isomerase, ribulose-phosphate 3-epimerase and glucose-6-phosphate 1-dehydrogenase.

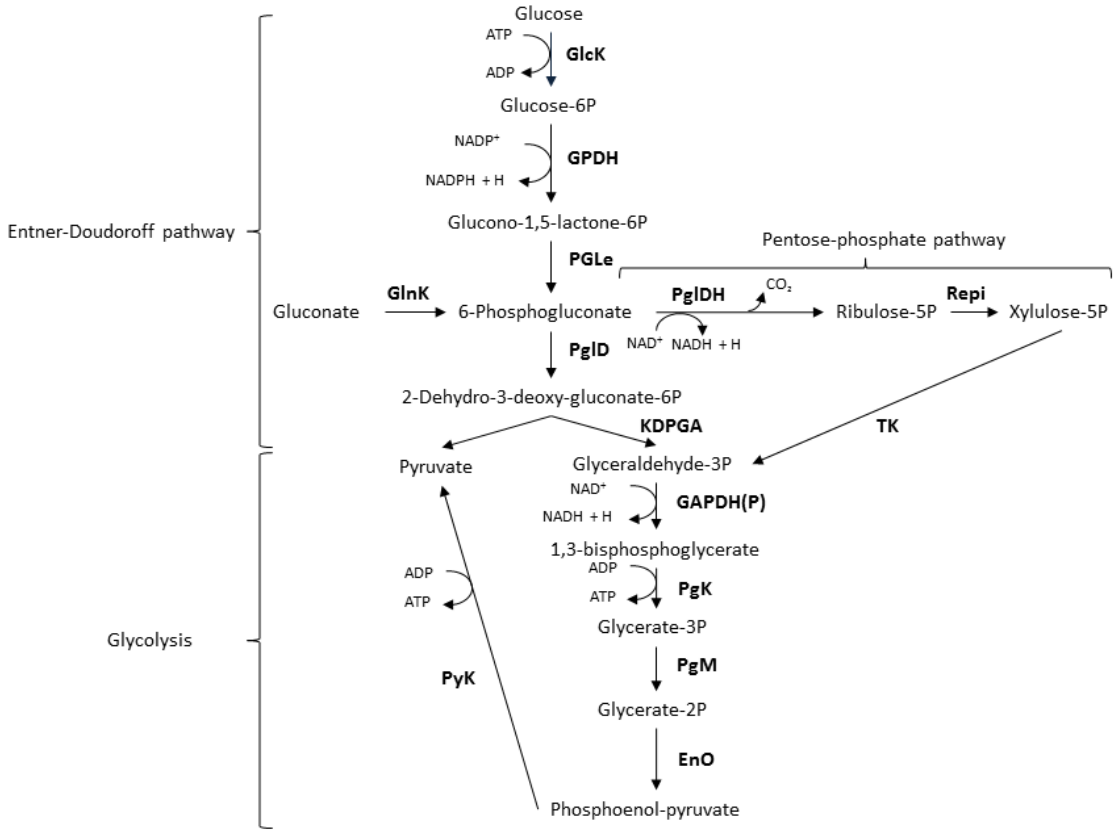


Figure 5.5 Major pathways of glucose metabolism in *Z. palmae*. Abbreviations are: P, phosphate; GlcK, Glucokinase; GPDH, Glucose-6P 1-dehydrogenase; GlnK, Gluconokinase; PGLe, 6-Phosphogluconolactonase; PglD, Phosphogluconate dehydrogenase; KDPGA, 2-dehydro-3-deoxyphosphogluconate aldolase; GAPDH(P), NAP(P)-dependent gylceraldehyde-3P-dehydrogenase; PgK, Phosphoglycerate kinase; PgM, Phosphoglycerate mutase; EnO, Enolase; PyK, Pyruvate kinase; PglDH, 6-Phosphogluconate dehaydrogenase; Repi, Ribulose-phosphate 3-epimerase; TK, transketolase.

5.3.4 Identifying the *Z. palmae pdc* Gene

The GenBank entry from the Zp*pdc* (AF474145.1) was BLAST-searched against the genome sequence. This identified the probable *pdc* gene but with 3 nucleotide differences (C399G, G400C, A733C) from the sequence in GenBank. The *pdc* was cloned from genomic Zp DNA as described below and resequenced in order to confirm the most likely nucleotide sequence for this gene. Wild type Zp*pdc* was amplified from genomic DNA using primers gZpPDC F (AAT CAG CAC ATA GGG TCT AAG AGG CAC G) and R (GCG CCG TAA GAG GGG CTA TGT GG), and KAPA HiFi polymerase (see General Methods for details, using 10 ng genomic DNA, annealing temperature 60°C, extension time 1:30 min). The PCR product (1671 nts in length) was inserted into the blunt-end cloning vector pJET1.2 and transformed into *E. coli* BioBlue. The resulting construct was isolated and sequenced using pJET1.2 F/R and wtZpPDC F1/F2. This provided further evidence that the "changes" were due to an erroneous GenBank entry.

These nucleotide changes result in 2 amino acid changes (R134A, E245A). Alanine is conserved in both positions when compared to other bacterial PDCs (see Chapter 3, Figure 3.15).The differences had previously been noted in ZpPDC constructs in the lab, but had originally been considered to be mutations in the constructs. This genome sequence confirms that in fact these "changes" are not due to mutation but that the GenBank entry for Zp*pdc* is erroneous.

5.4 Discussion

The genome sequencing data resulted in 38 contigs. This is not unusual, as difficulties in assembling repeat regions lead to ambiguity in the genome reconstruction and hence, to fragmentation. Finishing of the genome was outside of the scope of this project and not necessary for the purpose of acquiring data for codon harmonization. However, if one wanted to complete the genome sequence, finishing would have 2 objectives: closing gaps and validating the sequence.

As this sequencing data does not contain paired-reads, it is not easy to determine pairs of adjacent contigs; the order of the contigs and the size of the gaps are unknown. Primer walking can be used to sequence short, consecutive regions starting with the last ~20 bases of the known sequence. A short strand of DNA with an unknown sequence is synthesized and sequenced. As sequence information for a region becomes available, a new primer can be

designed to synthesize and sequence the next region and so on, until the gap is closed. Combinatorial PCR experiments use a combination of primers binding to unique sites in the genome, thus, simultaneously amplifying several regions. The short, unique fragments synthesized in the PCR then become themselves primers for another round of amplification. The synthesized DNA fragments can be sequenced by multiplex sequencing. This reduces the number of steps required to close multiple gaps (Nagarajan *et al.* 2010). Other approaches may include resequencing using paired-read libraries or PacBio technology, which gives longer read lengths often desirable for more complete genome coverage.

Despite being an incomplete genome, the analysis of this draft has yielded some interesting results. The experimentally determined GC content of 55.8 ±0.4 mol% (Okamoto *et al.* 1993) was confirmed, as the data indicated a GC content of 56.2%.

An analysis of the coding regions (2,531 in total) confirmed errors in the GenBank entry for Z*ppdc* and gave some insight into putative genes present. Among the protein coding regions, genes for ubiquinone, rod-shape determining proteins, sporulation inhibitor proteins, flagella and pili associated proteins, and two catalase enzymes were found, but no cytochrome c oxidase gene. These findings are in agreement with previous phenotype descriptions of *Z. palmae* (Okamoto *et al.* 1993).

Furthermore, analysis of the draft sequence has identified a repertoire of genes involved in carbohydrate metabolism, supporting previous observations that *Z. palmae* is able to utilize a variety of carbohydrate substrates, including, but not limited to, glucose, sucrose, galactose, arabinose, melibiose, maltose, mannitol, and sorbitol (Okamoto *et al.* 1993). Anaerobic conversion of glucose to ethanol and carbon dioxide appears to proceed through the Entner-Doudoroff pathway (Figure 5.6), similarly to the homoethanol fermenter *Z. mobilis* (Dawes *et al.* 1966). The non-ethanol producing species of the *Halomonadaceae* family *H. elongata* and *C. salexigens* use the Entner-Doudoroff pathway in glucose metabolism as well (according to the analysis of their genome annotation on KEGG, http://www.genome.jp/kegg/, Kyoto Encyclopedia of Gene and Genomes).

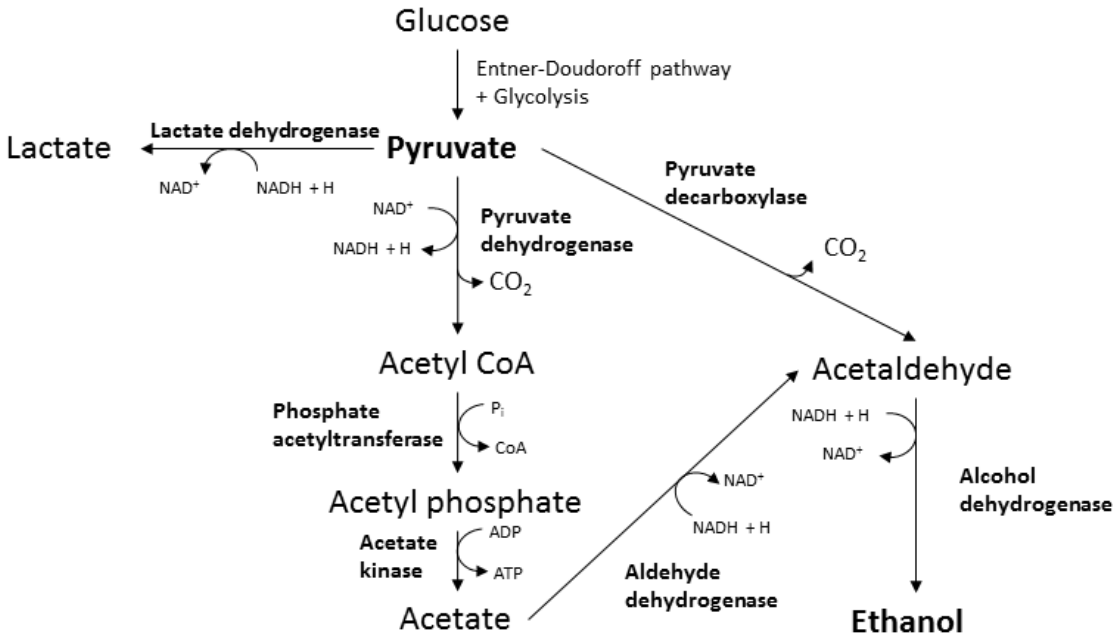


Figure 5.6 Fermentative metabolism in *Z. palmae*. P_i is orthophosphate.

With the codon usage frequency and tRNA availability data now available, a codon harmonization strategy was applied to the Zp*pdc* as described in Chapter 6.

6. Codon Harmonization of the *Zymobacter palmae pdc*

6.1 Introduction

The process of translation is complex and influenced by a variety of factors, affecting translational speed and co-translational folding of the nascent polypeptide chain. Some of these include direct interactions with the ribosome tunnel which may lead to stalling translation, interactions with chaperones, and the rate of tRNA binding (Angov 2011). Synonymous substitutions of a codon may not be as "silent" as they were once thought to be. A change in codon usage frequency (CUF) has been shown to affect protein expression, structure and in some cases function (Angov 2011).

Translation does not proceed in a uniform manner, but rather in pulses, creating patterns of different translational speed along the length of a protein. Conventionally, codons which are found at a high frequency in a genome are associated with "fast" translation (Figure 6.1, rabbit) and highly ordered structural elements of the protein, such as α helices, whereas low frequency (rare) codons are associated with slow translation speed or pauses (Figure 6.1, tortoise). Clusters of slow translating codons are commonly found in domain boundaries, i.e., link- or end-elements that separate higher order structural elements (Angov 2011). It is thought that low frequency codons slow down translation and, thus, allow proper co-translational binding of enzymatic co-factors and/or proper folding around these, at least partially (Angov *et al.* 2008).

Figure 6.1 Translation speed pattern along the mRNA.
The rabbit symbolizes codons that are translated more rapidly than slow codons indicated by the tortoise. The translation initiation complex is positioned on the translation initiation site. Translation slows down (tortoise) at putative pause sites. This is thought to allow for co-translational binding of co-factors and co-translational folding of the nascent polypeptide. From Angov (2011).

Studies in *Z. mobilis* PDC have shown proper binding and folding around the co-factors magnesium and TPP is essential for correct folding into an active enzyme (Pohl *et al.* 1994). The difference in codon bias in various organisms affects translational speeds of the same codon, therefore, affecting heterologous expression of a protein. Improper co-factor binding and folding due to these differences in codon bias may be alleviated through codon harmonization.

Unlike codon optimization, where codons in the target sequence are substituted by codons classed as high frequency in the expression host to allow rapid translation, codon harmonization aims to recreate the pattern of CUF along the protein in the recombinant expression host as found in the native host. Absolute CUF describes the occurrence of a codon throughout all coding regions known for an organism. Relative CUF is the ratio between absolute CUF and the sum of absolute CUF of a synonymous group, i.e., codons coding for the same amino acid.

Codon harmonization based on relative CUF has been applied to a mesophilic PDC for expression in *G. thermoglucosidasius* with some success (van Zyl *et al.* 2014b). Harmonizing the *G. oxydans pdc*, van Zyl *et al.* were able to improve expression of this PDC in *G. thermoglucosidasius* TM89 (NCIMB 11955 Δ*ldh*) such that it was fully functional up to 45°C (reported 0.22 U/mg at 45°C), whereas the wt was not. This is still below the optimum growth temperature for *G. thermoglucosidasius* at 60°C, but a promising attempt, especially considering that the *G. oxydans* PDC has lower thermoactivity and thermostability than the *Z. palmae* PDC (see General Introduction for details).

Conventionally, it was thought that codon bias (usage frequency) and tRNA iso-acceptor availability co-evolved, so that high frequency codons correlate with an abundance of the corresponding tRNA isoacceptor (Bulmer 1987). However, this is not necessarily true. When comparing CUF and tRNA availability of a given organism, the idea of co-evolution between codon bias and tRNA iso-acceptor abundance holds true for most codons. Yet there are exceptions, where high frequency codons do not have a tRNA gene available for strict Watson-Crick base pairing (Zhang *et al.* 2009). Research has shown the importance of tRNA availability and decoding type for translation speed, i.e., decoding of the codon by Watson-Crick base pairing in all three positions of the codon:anticodon interaction or non-Watson-Crick base pairing (“wobble” pairing) in the third position of the codon. There is evidence that suggests that wobble type decoding is slower than strict WC pairing (Spencer & Barral 2012, Spencer *et al.* 2012).

Taking this into consideration, a combined approach was employed here. With a preliminary *Z. palmae* genome sequence now available, harmonization was performed based on relative CUF and tRNA availability with respect to both the native and recombinant host. After harmonization, codons along the entire Zp*pdc* gene should follow similar patterns in CUF and tRNA availability, thus recreating the translational pattern as found in *Z. palmae* as closely as possible and improving recombinant expression of the ZpPDC in *G. thermoglucosidasius*.

6.2 Methods

6.2.1 Assessing Codon Usage Frequencies and t-RNA Availability

The *Z. palmae* genome-wide CUF data were obtained by running the genome through DAMBE (v.4.0.36) (Xia & Xie 2001). The *Z. palmae* genome-wide CUF data were compared to the codon usage table of *G. thermoglucosidasius* as found in the online database http://www.kazusa.or.jp/codon/, using the Graphical Codon Usage Analyzer (GCUA, http://gcua.schoedl.de/), in particular the codon vs. usage table option and the "frequency" output rather than "relative adaptiveness". The output data were then brought together in a spreadsheet and compared between the organisms for each codon along the Zp*pdc* gene.

Using the PROKKA annotation of the *Z. palmae* genome, tRNA genes were assessed for the native host and compared to data for *G. thermoglucosidasius* C56-YS93 (Genbank: NC_015660, NC_015661, NC_015665) as found through the tRNAscanSE software (v.1.3.1) (http://lowelab.ucsc.edu/tRNAscan-SE/, Lowe & Eddy 1997). This output was combined with the CUF data, and codon usage was adjusted following these design rules: as similar as possible and keep patterns in clusters. The preservation of the amino acid sequence was checked using Blastp (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

6.2.2 Cloning of the Harmonized Zp *pdc* for Expression in *Geobacillus thermoglucosidasius*

The harmonized gene was synthesized as 2 GeneArt® Strings™ DNA fragments (Thermo Fisher Scientific), which were combined together using a *Bsa*I site in the mid-gene region and Golden Gate DNA Assembly.

Golden Gate DNA assembly used 100 ng linearized vector backbone, equimolar amounts of the additional assembly pieces, 1.5 µl 10x T4 buffer (NEB), 0.15 µl 100x BSA (NEB, supplied with T4 ligase), 1 µl *Bsa*I (NEB, Product code: R3535S), 1 µl T4 ligase (NEB, Product code: M0202T, high concentration ligase, 2 million units/ml), and MilliQ water to make up 15 µl. The assembly reaction was performed in a thermocycler (Eppendorf, Mastercycler) using the following programme: 25 cycles of 3 min at 30°C and 4 min at 16°C, followed by 5 min at 50°C and 5 min at 80°C. 5-10 µl of the ligation reactions were transformed into high efficiency cloning strains, such as *E. coli* BioBlue.

The gene is flanked by *Bam*HI and *Xho*I recognition sites, which allowed digestion and ligation into the desired expression vector, initially pUCG18 under the control of the *ldh* promoter, thus creating p600 ZpPDC 2.0.

Note that this work was carried out in parallel to the work on the wt Zp*pdc* described in Chapter 4. Expression from p600 ZpPDC 2.0 could not be reliably replicated and analysis of the promoter sequence failed to identify a functional RBS in p600. This prompted the redesign of the expression vector with particular attention to promoter-RBS combination. As mentioned in Chapter 4 the *ldh* promoter-*phe*B RBS combination has been shown to work very effectively and thus the genes were moved into pGR002, which already contains this combination.

The harmonized *pdc* gene was amplified from p600 ZpPDC 2.0 using the following primers: 2.0pGR002F (CTA GTC TAG ATA AGG AGT GAT TCG AAT GTA CAC GGT TGG TAT GTA TCT AGC AG) and 2.0pGR002R (**GAG CTC** GCA AAA AAA CGC CCC CTT TCG GGG CGC GAT TAG GCC TGT GGT TTG CG), and Phusion® Hot Start II polymerase (as described in General Methods) with an annealing temperature of 65°C and 1:30 min extension at 72°C. This PCR added an *Xba*I recognition site (underlined) and the *pheB* RBS (in blue) to the upstream region of the gene, and the *pheB* downstream region (in green) containing a transcriptional terminator region and a *Sac*I recognition site (bold, underlined) to the downstream region of the *pdc*.

The PCR product was cloned into the pJET1.2 cloning vector (see General Methods), which was then transformed into *E. coli* BioBlue, reisolated and sequenced with pJET1.2 F/R and 2.0 F1/F2. A clone containing an insert with the correct sequence was then chosen for digestion

with *Xba*I and *Sac*I, as was the pGR002 vector. The released fragment and the linearized vector were agarose gel purified, ligated (using 20 ng linearized vector and 17 ng insert) into pGR002, and the product transformed into *E. coli* BioBlue.

An *E.coli* BioBlue colony carrying the correct clone of pGR002 2.0, as confirmed by sequencing with M13 F/R and 2.0 F1/F2, was used to propagate the plasmid, which was then purified and transformed into *G. thermoglucosidasius* DL44 and TM236 (see General Methods).

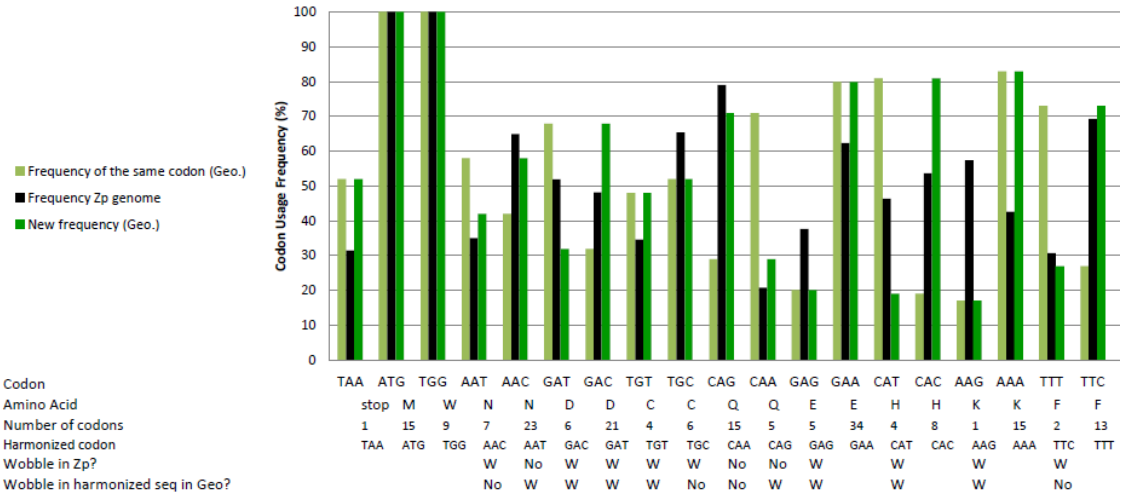
The resulting strains were characterized for *pdc* expression in aerobic conditions at various growth temperatures.

6.3 Results

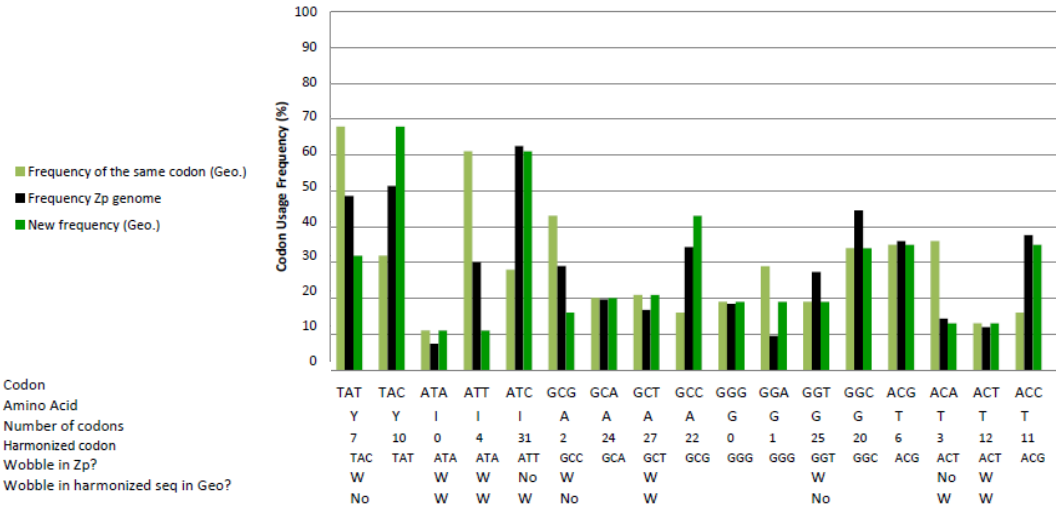
6.3.1 Assessing Codon Usage and Harmonizing the Zp *pdc* Gene Sequence

With a preliminary *Z. palmae* genome sequence now available, harmonization was performed based on relative CUF and tRNA availability with respect to the expression hosts *G. thermoglucosidasius* and *E. coli*. 308 "silent" nucleotide substitutions were made, leading Zp*pdc* 2.0 to be 18 % different in nucleotide sequence compared to the wt Zp*pdc*; the amino acid sequence remained unchanged. Figure 6.2 below summarises the harmonization strategy employed for each codon.

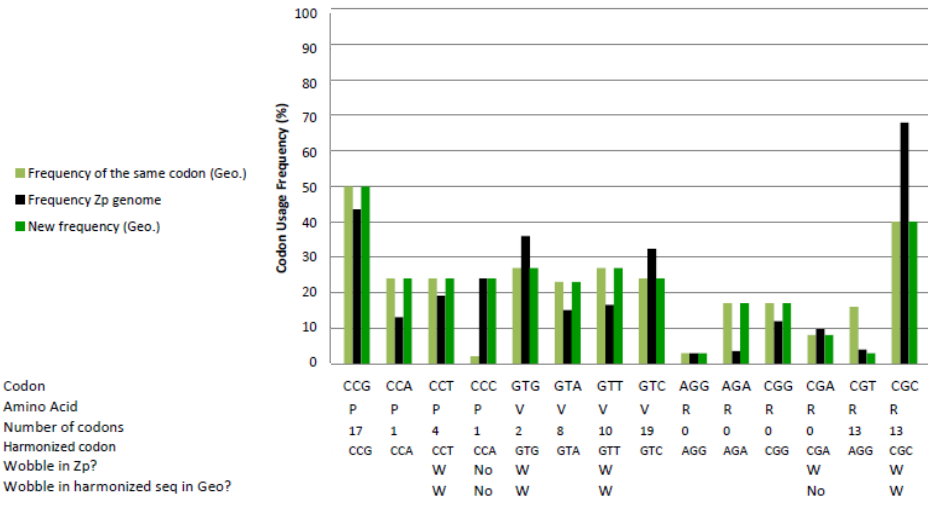
Figure 6.2 Codon harmonization of the *Z. palmae pdc* (below). Illustrated here is the harmonization strategy using genome-scale relative codon usage frequency (CUF, in %) and tRNA availability ("W" denotes wobble coding) summarised by codon type. These were manually applied along the entire gene. The bars are coloured as follows: light green denotes the CUF for the indicated codon in *G. thermoglucosidasius*; black denotes the CUF for the indicated codon in *Z. palmae*; dark green denotes the CUF for a codon in *G. thermoglucosidasius*, which encodes the same amino acid but with a more similar CUF to the codon used in *Z. palmae*. Each set of bars contains the following annotations on the x-axis: the codon; the number of amino acids present in the Zp *pdc* encoded by the particular codon; the new/harmonized codon to be used instead of the native codon; an annotation of wobble coding of the native codon in *Z. palmae*; and an annotation of wobble coding of the new/harmonized codon when used in *G. thermoglucosidasius.*



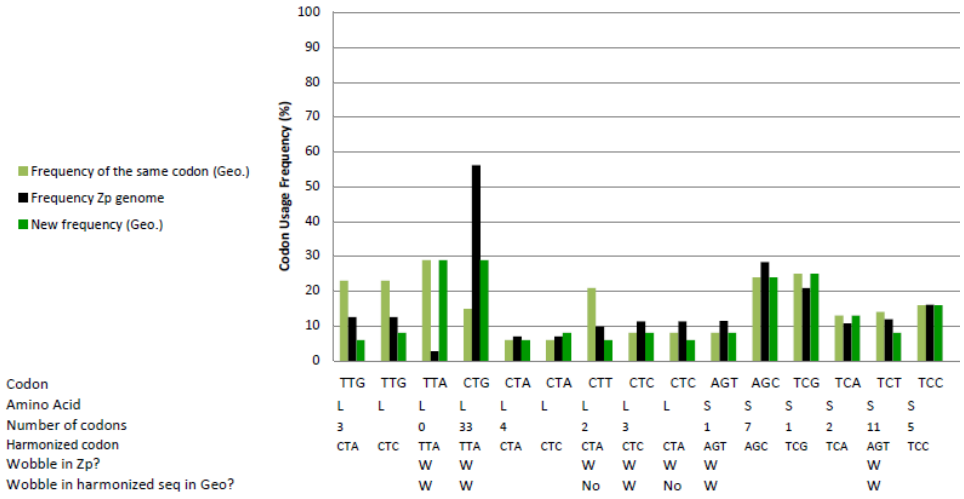
Codon	TAA	ATG	TGG	AAT	AAC	GAT	GAC	TGT	TGC	CAG	CAA	GAG	GAA	CAT	CAC	AAG	AAA	TTT	TTC
Amino Acid	stop	M	W	N	N	D	D	C	C	Q	Q	E	E	H	H	K	K	F	F
Number of codons	1	15	9	7	23	6	21	4	6	15	5	5	34	4	8	1	15	2	13
Harmonized codon	TAA	ATG	TGG	AAC	AAT	GAC	GAT	TGT	TGC	CAA	CAG	GAG	GAA	CAT	CAC	AAG	AAA	TTC	TTT
Wobble in Zp?				W	No	W	W	W	W	No	No	W		W		W		W	
Wobble in harmonized seq in Geo?				No	W	W	W	W	No	No	W	W		W		W		No	



Codon	TAT	TAC	ATA	ATT	ATC	GCG	GCA	GCT	GCC	GGG	GGA	GGT	GGC	ACG	ACA	ACT	ACC
Amino Acid	Y	Y	I	I	I	A	A	A	A	G	G	G	G	T	T	T	T
Number of codons	7	10	0	4	31	2	24	27	22	0	1	25	20	6	3	12	11
Harmonized codon	TAC	TAT	ATA	ATA	ATT	GCC	GCA	GCT	GCG	GGG	GGG	GGT	GGC	ACG	ACT	ACT	ACG
Wobble in Zp?	W		W	W	No	W		W				W			No	W	
Wobble in harmonized seq in Geo?	No		W	W	W	No		W				No			W	W	



Codon	CCG	CCA	CCT	CCC	GTG	GTA	GTT	GTC	AGG	AGA	CGG	CGA	CGT	CGC
Amino Acid	P	P	P	P	V	V	V	V	R	R	R	R	R	R
Number of codons	17	1	4	1	2	8	10	19	0	0	0	0	13	13
Harmonized codon	CCG	CCA	CCT	CCA	GTG	GTA	GTT	GTC	AGG	AGA	CGG	CGA	AGG	CGC
Wobble in Zp?			W	No	W		W					W		W
Wobble in harmonized seq in Geo?			W	No	W		W					No		W



6.3.2 Cloning the Harmonized Zp *pdc* Gene for Expression in *G. thermoglucosidasius*

To make the final expression construct, the gene was amplified adding appropriate up- and down-stream regions and sub-cloned into pGR002 (see Figure 6.3 for PCR results), thus creating pGR002 2.0 (Figure 6.4). Plasmid DNA was purified from *E. coli* and sequencing confirmed the correct clone, which was then transformed into *G. thermoglucosidasius* DL44 (DL33 Δ*ldh*) and TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*).

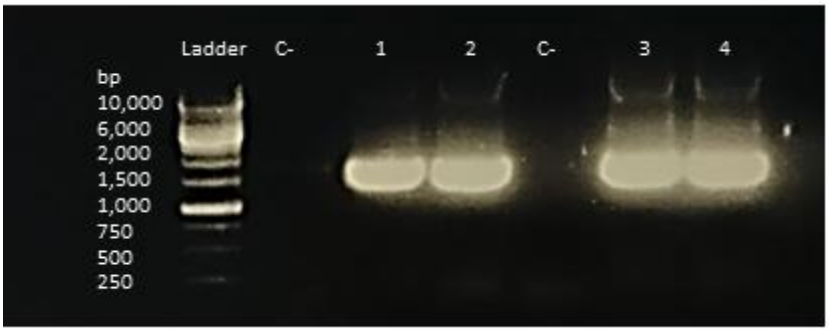


Figure 6.3 Agarose gel electrophoresis of the *pdc* amplification product. The 1,727 bp PCR-amplified product was visualized alongside the GeneRuler™ 1 kb ladder (Thermo Fisher Scientific). C- is a PCR negative control containing no template. Lanes 1 and 2 are from replicate PCR reactions of wt Zp*pdc*; lanes 3 and 4 are replicate PCR reactions of wt Zp*pdc* 2.0.

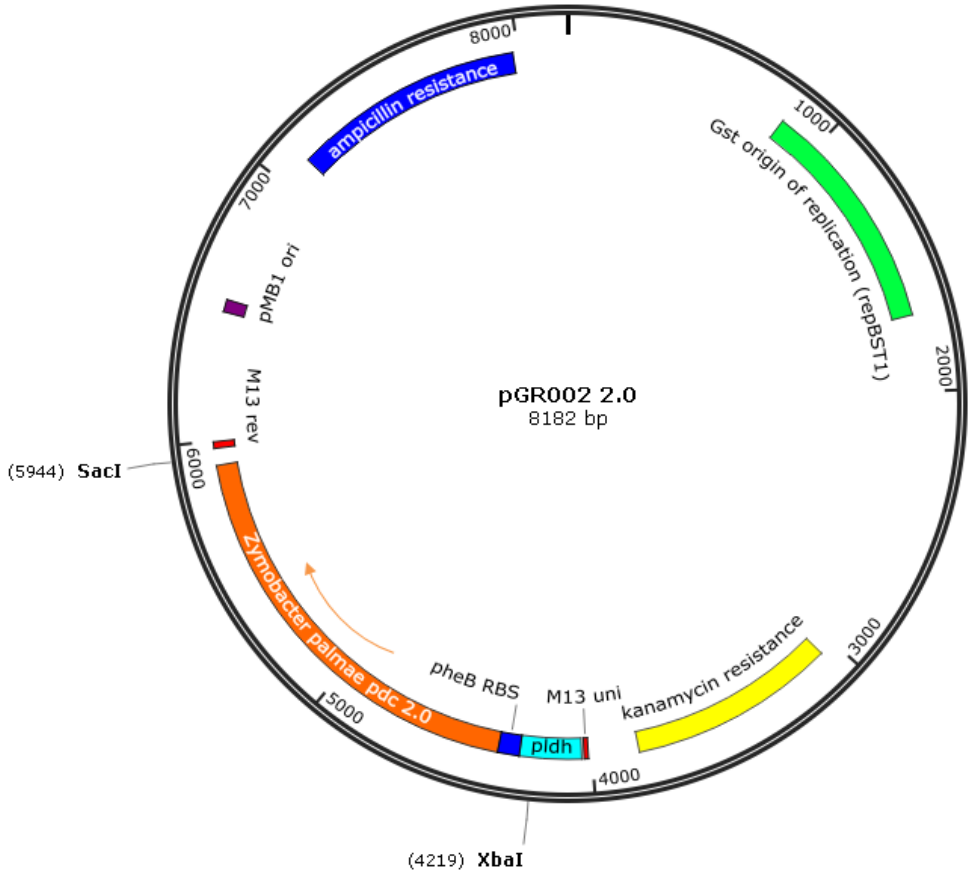


Figure 6.4 Plasmid map of pGR002 2.0. pGR002 2.0 is based on the *E. coli-Geobacillus* spp. shuttle vector pUCG18, containing features for expression in both hosts. RepBST1 and the kanamycin resistance gene are *G. sterothermophilus* derived and thus allow replication and selection of the plasmid at higher growth temperatures, whereas pMB1 ori and the ampicillin resistance gene are *E. coli* derived. Pldh denotes the *G. stearothermophilus* NCA1503 *ldhA* promoter, while *phe*B RBS marks the position of the ribosome binding site from the *G. stearothermophilus* DSMZ6285 *pheB* gene. M13 uni/rev are sequencing primer complimentary sequences.

6.3.3 Aerobic Expression of the Codon Harmonized *Zp pdc* 2.0 in *G. thermoglucosidasius*

G. thermoglucosidasius DL44 (DL33 Δ*ldh*) and TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) carrying pUCG18 (empty plasmid backbone), pGR002 wtZpPDC (see Chapter 4) or pGR002 2.0 were grown aerobically in 50 ml 2TY + 12 µg/ml kanamycin + 5 mM thiamine in a 250 ml baffled flask, with shaking at 250 rpm, at 50, 60 and 65°C to an OD_{600nm} of 1.5 to 2.5. After sonication, the clarified cell extract was assayed for PDC activity.

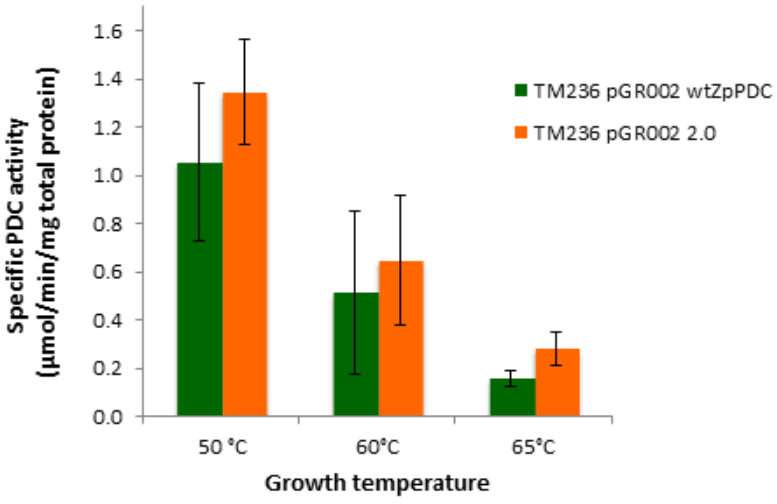
As observed with the wtZpPDC, there are differences in PDC activity measured between DL44 and TM236, with TM236 seemingly being a favourable strain for expression at higher

temperatures as it produces increased PDC activities at growth temperatures of 60 and 65°C (see Table 6.1 and Figure 6.5 below).

Table 6.1 PDC activities in cell extracts of *G. thermoglucosidasius* DL44 carrying pGR002 wtZPPDC or pGR002 2.0 grown aerobically at various temperatures.

Growth temperature	Specific activity (µmol/min/mg total protein)		
	DL44 pUCG18	DL44 pGR002 wtZpPDC	DL44 pGR002 2.0
60°C	0	0.081 ±0.018	0.081 ±0.034
65°C	0	0	0.045 ±0.017

Specific activity was determined from cell extracts by monitoring the decrease in NADH-dependent absorbance at 340 nm in standard coupled assays at 30°C. At least two independent expression experiments were carried out, and each was assayed three times. The error is standard error.



TM236 pGR002 wtZpPDC	1.06 ±0.33	0.52 ±0.34	0.16 ±0.03
TM236 pGR002 2.0	1.34 ±0.22	0.65 ±0.27	0.28 ±0.07

Figure 6.5 Specific PDC activities of *G. thermoglucosidasius* TM236 variants grown aerobically at different temperatures. Specific activity was determined in cell extracts by monitoring the decrease in the NADH-dependent absorbance at 340 nm in standard coupled assays at 30°C. These are averages of 3 independent expression experiments with 3 assays each; error is standard error. Note that no activity was detected for the TM236 strain carrying the empty pUCG18.

Codon harmonization did improve PDC activity in *G. thermoglucosidasius* TM236 by 21% from cells grown at 50°C, 20% from cells grown at 60°C, and 42% from cells grown at 65°C in comparison to the wt Zp*pdc* (based on the average PDC activity). However, similarly to the wt ZpPDC, activity drastically decreased with increasing growth temperature. The activity detected for the wt ZpPDC from cells grown at 65°C is only at 15% of the activity level detected in cells grown at 50°C, and 20% for the codon harmonized *pdc* 2.0.

As noted with wt ZpPDC expression, no over-expressed protein band could be identified by SDS-PAGE analysis indicating that expression levels were low (see example gel image below, Figure 6.6).

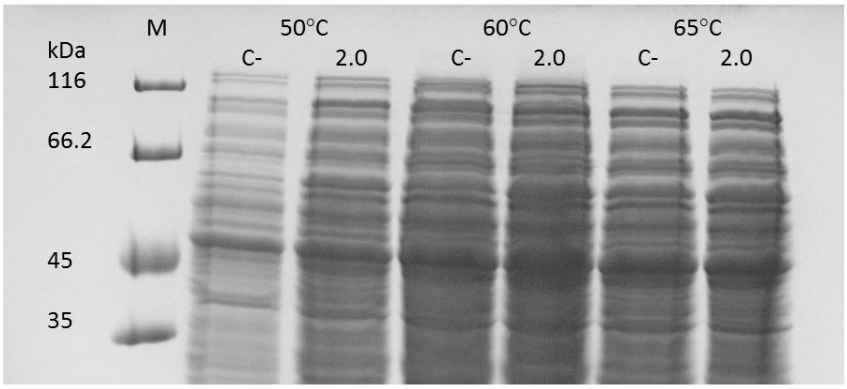


Figure 6.6 SDS-PAGE analysis of *G. thermoglucosidasius* TM236 pGR002 2.0 expression under aerobic conditions. TM236 was grown at 50, 60 and 65°C, cells were lysed, and the extract was clarified by centrifugation. The soluble fraction (~10 µg total protein) was run on a 12% SDS-PAGE gel. C- is TM236 carrying an empty pUCG18 plasmid. 2.0 is TM236 carrying the pGR002 2.0 expression vector. M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific). ZpPDC would be expected at 59.4 kDa.

6.3.4 Analysis of Gene Expression by RT-qPCR

Aerobic cultures of *G. thermoglucosidasius* TM236 pGR002 2.0 grown at 60 and 65°C were also analysed for gene expression using RT-qPCR as described in General Methods. Gene expression analysis showed the same trend as observed in PDC activity measurements. Using TM236 pUCG18 as the calibration strain, *pdc* expression in TM236 pGR002 2.0 was upregulated 80-fold at 60°C but only 28-fold at 65°C.

6.4 Discussion

Codon harmonizing the wt Zp*pdc* improved observed PDC activity in *G. thermoglucosidasius* across the growth temperature range. DL44 (DL33 Δ*ldh*) seems to be a less favourable expression strain (as noted with wt Zp*pdc* expression in Chapter 4). However, harmonized Zp*pdc* showed PDC activity up to a growth temperature of 65°C (0.045 ±0.017 µmol/min/mg total protein) in this strain, whereas there was no detectable PDC activity in the cultures expressing the wt ZpPDC at this growth temperature. Specific PDC activity measured in cells grown at 60°C was very similar for both the wt ZpPDC (0.081 ±0.018 µmol/min/mg total protein) and the harmonized version (0.081 ±0.034 µmol/min/mg total protein) expressed in DL44. This is a vast improvement to Taylor *et al.* (2008) who reported PDC activities of 0.089 µmol/min/mg total protein from cells grown at 48°C, and detected no PDC activity at 50°C. This is most likely due to the optimization of the expression system by changing the promoter-RBS combination to p*ldh*-RBS*ldh*-RBS*phe*B.

In TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*), higher activities of the wt ZpPDC were noted at all growth temperatures compared to DL44. This was further increased by codon harmonization. At a growth temperature of 65°C detected PDC activity was increased by 42%, from 0.16 ±0.03 µmol/min/mg total protein of the wt to 0.28 ±0.07 µmol/min/mg total protein of the harmonized ZpPDC.

This is a major improvement for high temperature recombinant expression of a bacterial PDC, and surpasses any previously reported attempts. Van Zyl *et al.* (2014b) reported a specific activity of 0.22 U/mg from cells grown at 45°C with their codon harmonized *G. oxydans pdc* (GoxPDC), but detected no PDC activity at growth temperatures above that. This improvement may be due to the ZpPDC being more thermostable and more thermoactive than GoxPDC, with a temperature optimum of 65°C and 53°C, respectively, and retained activity after exposure to 65°C for 30 min of 80% and 40%, respectively, thus making ZpPDC a better candidate for high temperature expression.

Despite improved high temperature expression after codon harmonization, the detected PDC activity drastically decreased with increases in growth temperature, following a similar trend to the wt ZpPDC. The lack of an overexpressed protein band in the SDS-PAGE analysis points towards a deficiency in this expression system. Gene expression analysis by RT-qPCR confirmed the reduced expression of Zp*pdc* with increasing growth temperature, which may explain the reduced PDC activity detected in cells grown at higher temperatures. Gene expression levels at 65°C growth temperature were 35% of the level detected at 60°C growth temperature.

This decrease in gene expression levels may be explained by the decrease of expression from the *ldh* promoter as oxygen becomes limited in the cultures grown at higher temperatures. Bartosiak-Jentys *et al.* (2012) showed that expression from the *ldh* promoter decreases as oxygen becomes increasingly limited, and it is well known that with increasing temperature gas solubility decreases, which means that with increasing growth temperatures the culture becomes more oxygen limited.

Nonetheless, these results are promising, and may be further improved by the development of a producer of ethanol (PET) operon, combing the harmonized Zp*pdc* with a suitable ADH, as described in the next chapter. This will then also allow further testing of the functionality under fermentative conditions.

7. Designing a PET Operon For Expression in *G. thermoglucosidasius*

7.1 Introduction

So far in this study the Zp*pdc* has been heterologously expressed on its own in *G. thermoglucosidasius*. In order for the strains expressing a PDC to become effective ethanol producers an appropriate thermoactive ADH should be paired with PDC to convert the acetaldehyde produced into ethanol; this would prevent loss of the volatile acetaldehyde, or its build-up to toxic levels that can compromise cell growth and function.

Liu *et al.* (2005) showed that the overexpression of *Z. palmae pdc* in *L. lactis* led to an increase in detectable acetaldehyde levels (up to 8-fold), but also noted that the endogenous ADH activity was not sufficient to efficiently reduce the acetaldehyde produced to ethanol. This had also been noted by Ingram & Conway (1988) when expressing *Z. mobilis pdc* in *E. coli.* However, the expression of both the *Z. mobilis pdc* and *adh*B led to the successful creation of an ethanologenic *E. coli* (Ingram & Conway 1988, Ingram *et al.* 1987). This PET (producer of ethanol) operon was developed further and shown to restore redox balance in an *E. coli* (Δ*ldh*, Δ*pfl*) strain deficient in anaerobic growth, enabling anaerobic growth and high ethanol production (Hespell *et al.* 1996). These studies suggest that the co-expression of a suitable ADH would be beneficial in engineering an efficient ethanol-producing *Geobacillus* strain.

Unfortunately the *Z. palmae* ADHI and II are not very thermostable (characterization in Appendix III), and thus were not appropriate candidates for expression in *G. thermoglucosidasius.* Dr. Luke Williams (University of Bath) characterized the suite of *G. thermoglucosidasius* ADHs, and identified GtADH6(E) as a good candidate for acetaldehyde to ethanol conversion with a V_{max} of 398 µmol/min/mg and a K_M of 1.5 mM for acetaldehyde (at 50°C) (Williams 2015).

This chapter describes the *in vitro* characterization of the ZpPDC-GtADH6 pathway. The results from these initial studies were very promising and prompted the design of a PET operon for expression in *G. thermoglucosidasius*. The chapter further describes the design and aerobic expression of the PET operon in *G. thermoglucosidasius*, as well as expression under fermentative conditions in tube fermentations.

Some future work and alternative approaches to improve PDC activity in *Geobacillus* spp. are explored in the Discussion of this chapter.

7.2 Methods

7.2.1 Expression of PDC and ADH for *in vitro* Characterization of the PDC-ADH Pathway

The ZpPDC was expressed and purified from pET28 ZpwtPDC in *E. coli* BL21 (DE3) as described in Chapter 3.

The *G. thermoglucosidasius adh*6 gene (RTMO03537, RAST1773, GtADH6) was cloned into pET28 ADH6 as described by Williams (2015), and expressed and purified from *E. coli* BL21 (DE3) as described in General Methods (with the addition of 34 µg/ml chloramphenicol in the growth medium) using a small-scale Talon® metal-affinity resin column.

7.2.2 *In vitro* Characterization of the PDC-ADH Pathway

To assess whether ADH6 would make an efficient coupling partner for ZpPDC, the coupled assay (see General Methods) was performed in 1 ml quartz cuvettes using ADH6 as the coupling enzyme, at temperatures up to 70°C with both enzymes purified from *E. coli*.

Furthermore, the stoichiometric balance of the reaction was assessed by performing these coupled assays in 1 ml sealed high-performance liquid chromatography (HPLC) glass vials incubated at 50, 60 and 65°C for 30 min, using 4-8 mM NADH and 10 mM pyruvate. Assay conditions were previously tested and optimized by following the reaction spectrophotometrically, monitoring the NADH depletion through the decrease in absorbance at 340 nm at 50, 60 or 65°C. The enzyme concentrations used were adjusted to ensure an excess of ADH6 (molar ratio per active site, ADH6:PDC is 2.2:1). Immediately following incubation at the desired reaction temperature, the vials were placed in boiling water for 10 min, cooled on ice and the products analysed by HPLC (see section 7.2.5 below).

7.2.3 Cloning the PET Operon for Expression in *Geobacillus thermoglucosidasius*

The initial PET operon was cloned using Gibson assembly. 1.2 ml of Gibson master mix were made up as follows, and stored in 15 µl aliquots at -20°C: 320 µl 5x ISO buffer, 0.64 µl 10 U/µl T5 exonuclease, 20 µl 2 U/µl Phusion® polymerase, 160 µl 40 U/µl Taq ligase and MilliQ water to 1.2 ml (all from NEB). 5x ISO buffer was prepared in 6 ml batches and stored in 320 µl aliquots at -20°C. It contains 3 ml 1 M Tris-HCl pH 7.5, 150 µl 2 M $MgCl_2$, 240 µl 100 mM dNTP

mix (25 mM each, dGTP, dCTP, dATP, dTTP, NEB), 300 µl 1 M DTT, 1.5 g PEG 8,000, 300 µl 100 mM NAD^+, and MilliQ water to 6 ml.

Gibson assembly fragments were designed with a 20 bp overlap using the NEBuilder® Assembly Tool (v1.8.1). The linearized vector was prepared by enzyme digest of pGR002 2.0 with *Xba*I and *Sac*I. The gene fragments were prepared by PCR using Phusion® Hot Start II (Thermo Fisher Scientific, see General Methods) and primers pUCG-PDC2.0NLF (cat tat att gag gga gga ttT CTA GAT AAG GAG TGA TTC GAA TG, *Xba*I recognition site underlined, upper case - gene specific, lower case - overhang) and ADH6-PDC2.0NLR (ctt act cga gTT AGG CCT GTG GTT TGC G, *Xho*I recognition site underlined) with an annealing temperature of 60.6°C and 30 s extension at 72°C to amplify the PDC-containing fragment from pGR002 2.0, and primers PDC2.0-ADH6NLF (aca ggc cta act cga gta agg agt gat tcg aat gta cac ggt tgg tat gAA TAC ATT CTT CTT GAA ACC AAA AAT C, *Xho*I recognition site underlined) and pUCG18-ADH6R (cat gat tac gaa ttc gag ctc gca aaa aaa cgc ccc ctt tcg ggg cgc gaT TAT CCG TTA TAT GCC CAT T, *Sac*I recognition site underlined) with an annealing temperature of 60°C and 30 s extension at 72°C to amplify the ADH6-containing fragment from pET28 ADH6. The PCR fragments and linearized pGR002 2.0 were gel purified as described in General Methods.

100 ng of the linearized vector and equimolar amounts of the gene fragments were prepared in a volume of 5 µl. This was added to 15 µl Gibson master mix (prepared as described above). This reaction was incubated at 50°C for 60 min. 5 µl of this reaction were transformed into *E. coli* BioBlue.

Sequencing highlighted an erroneous region upstream of the *pdc* lacking the desired *Xba*I recognition site. This was corrected by amplifying the *pdc* from pGR002 2.0 using 2.0HindIIIF (CCC AAG CTT GCA TGC CTG, *Hind*III recognition site underlined) and 2.0XhoIR (CCG CTC GAG TTA GGC CTG TG, *Xho*I recognition site underlined) and an annealing temperature of 68°C and 1 min extension at 72°C with Phusion® Hot Start II (see General Methods). The PCR fragment and pGR002 PET were digested with *Hind*III and *Xho*I, and ligated as described in General Methods.

An *E. coli* BioBlue colony carrying the correct clone of pGR002 PET, as confirmed by sequencing with M13 F/R, 2.0 F1/F2, was used to propagate the plasmid.

Attempts to transform pGR002 PET into TM236 failed, possibly because the increased size of the vector reduced transformation efficiency. This prompted cloning of the PET operon into the smaller pUCGT using enzyme digestion with *Hind*III and *Sac*I. The ligation was set up with

the gel purified fragments (20 ng linearized pUCGT, 34 ng PET operon). *E. coli* BioBlue was transformed with the ligation reaction mix and a colony carrying the correct clone of pUCGT PET, as confirmed by sequencing as above, was used to propagate the plasmid. Purified pUCGT PET was successfully transformed (by electroporation) into *G. thermoglucosidasius* TM236 (see General Methods). This vector also offered the possibility to transform by conjugation if electroporation had not been successful.

7.2.4 Tube Fermentations

Characterization of strains for fermentation products was carried out in tube fermentations. In a 15 ml Falcon tube, 10 ml of sterile ASM + 12 µg/ml kanamycin were inoculated with 1 ml of an aerobically grown seed culture in exponential phase. Cultures were incubated at the desired growth temperature (50, 60 or 65°C), shaking at 250 rpm for at least 48 h to allow the transition from aerobic to micro-aerobic growth.

7.2.5 Analysis of Fermentation Products by HPLC

Fermentation products in clarified culture samples were quantified by HPLC on a Hewlett Packard 1100 system detecting organic acids by UV absorption at 215 nm and glucose and ethanol by refractive index. Samples were analysed using an Agilent Technologies 1200 series LC system with an 300 x 7.8mm Rezex ROA-Organic Acid H+ (8%) column (Phenomenex, Cheshire, UK), a flow rate of 0.6 ml/min at 65°C, and the mobile phase being 5 mM sulphuric acid, with a 5 µl injection and a run time of 25 min.

7.3 Results

7.3.1 *In vitro* Characterization of the PDC-ADH Pathway

GtADH6 was expressed and purified from *E. coli* BL21 (DE3); see SDS-PAGE gel below (Figure 7.1). Using a thermal shift assay, the denaturing temperature was determined to be 81.5°C.

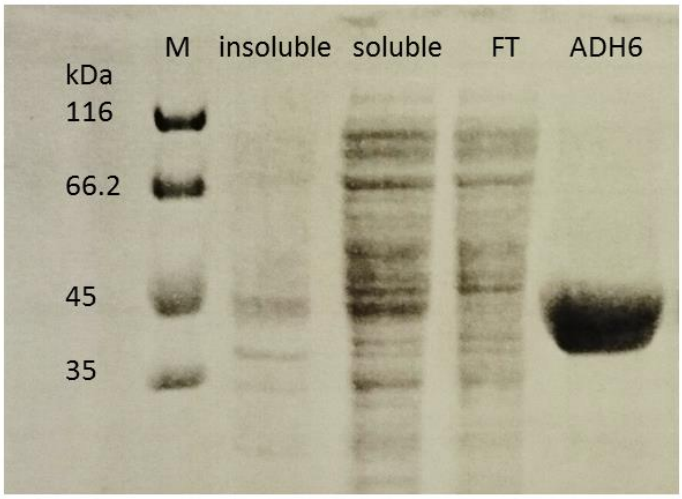


Figure 7.1 SDS-PAGE analysis of GtADH6. His-tagged GtADH6 (monomer size of 44.8 kDa) was purified by metal-affinity chromatography (MAC). Lane M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific). Insoluble is the insoluble protein fraction, soluble the soluble protein fraction, FT the MAC column flow-through and ADH6 the fraction eluted with 100 mM imidazole in His-elute buffer (20 mM Tris, 300 mM NaCl, pH 8).

ZpPDC was purified from *E. coli* as described in Chapter 3, and the purified enzymes were used together in the coupled assay. Enzyme concentrations and ratios were not optimized. Nonetheless, the coupled assay worked efficiently across a wide temperature range up to 70°C (Figure 7.2).

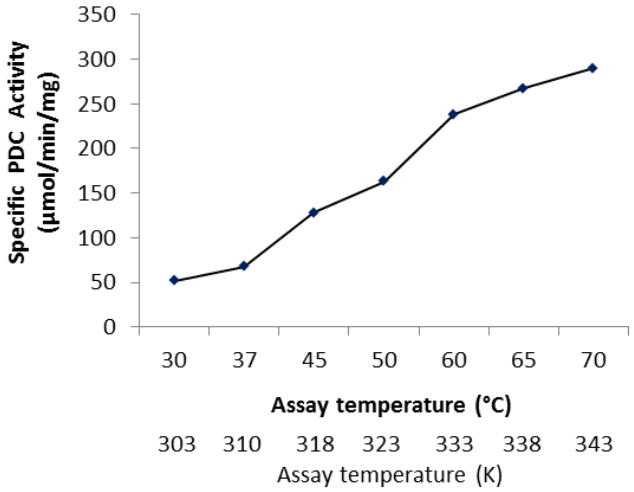


Figure 7.2 Relationship between temperature and wt ZpPDC activity using GtADH6 in the coupled assay. Specific PDC activity was determined in coupled assay using purified ZpPDC and GtADH6 by monitoring the decrease in NADH-dependent absorbance at 340 nm.

To further assess whether GtADH6 was an appropriate partner for ZpPDC, the products of coupled assays run at 60 and 65°C were analysed by HPLC (Figure 7.3 below). Pyruvate was present in excess, while NADH concentration was varied and limiting the amount of ethanol that could be produced. The data showed that for all assays pyruvate was fully or almost fully consumed. With 8 and 4 mM NADH, at 60°C less ethanol than expected was produced. However, in all other assays, including those run at 65°C, expected amounts of ethanol were produced suggesting that GtADH6 efficiently converts the acetaldehyde produced by the wt ZpPDC into ethanol.

These results indicated that the combination of ZpPDC and GtADH6 was functional up to 65°C, which is very promising for the expression of a PET operon in *G. thermoglucosidasius*.

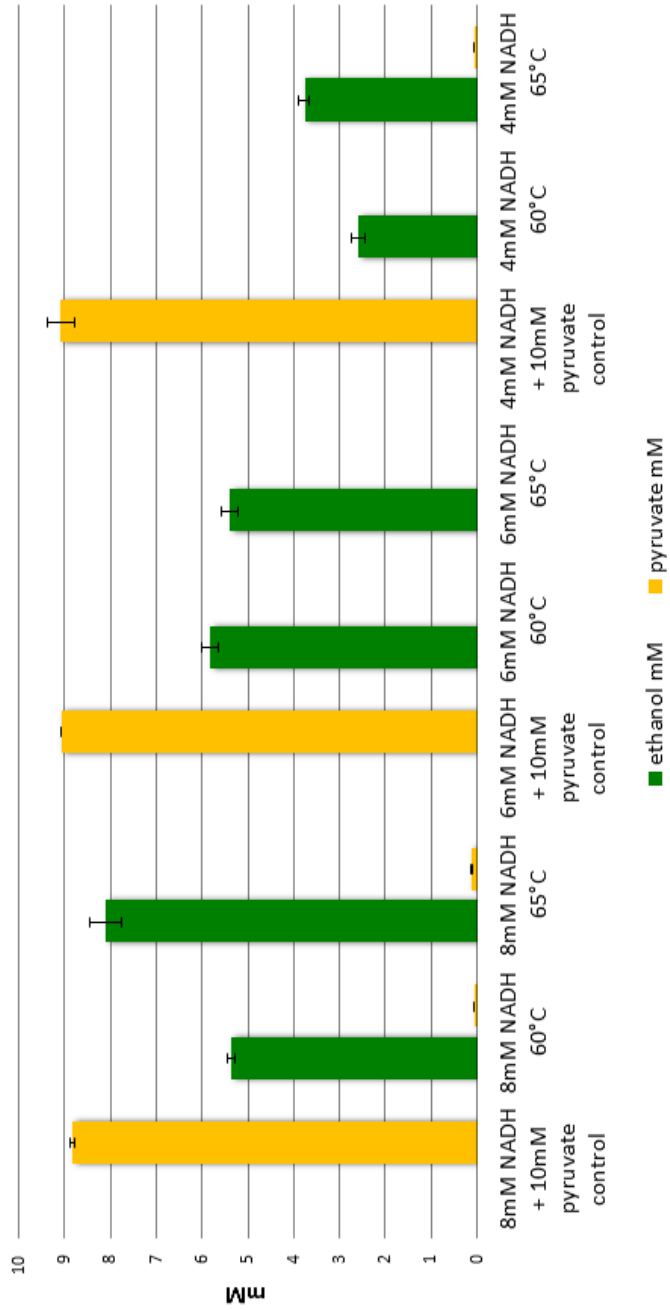


Figure 7.3 Relationship between NADH concentration and ethanol production in ZpPDC-GtADH6 coupled assay. Coupled assays used recombinantly-expressed (in *E. coli*) and purified wt ZpPDC and GtADH6 (molar ratio per active site is 1:2.2) with pyruvate in excess and a variety of NADH concentrations limiting the amount of ethanol that could be converted by GtADH6 from the acetaldehyde produced by ZpPDC. Assays were run at 60 or 65°C for 10 min before product analysis by HPLC. The control is a no-enzyme control. These are averages of 3 independent experiments with 3 measures; error is standard error.

7.3.2 Cloning the PET Operon for Expression in *Geobacillus thermoglucosidasius*

The PET operon was designed so that both genes were under the control of the *ldh* promoter as in previous pGR002 constructs, but was each preceded by the *phe*B RBS. Cloning developed into a complex task as described in the methods above. The final construct was pUCGT PET shown in Figure 7.4, which was transformed into *G. thermoglucosidasius* TM236 (NCIMB Δ*ldh*, Δ*pfl*).

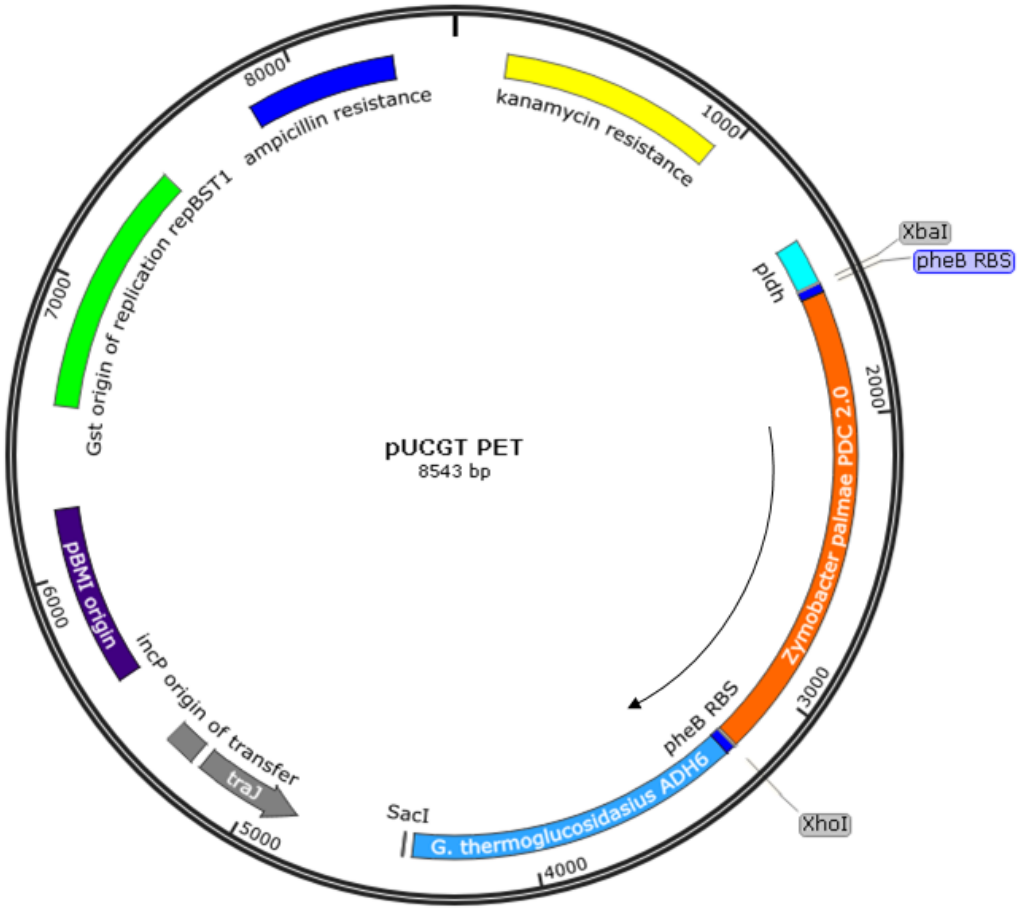
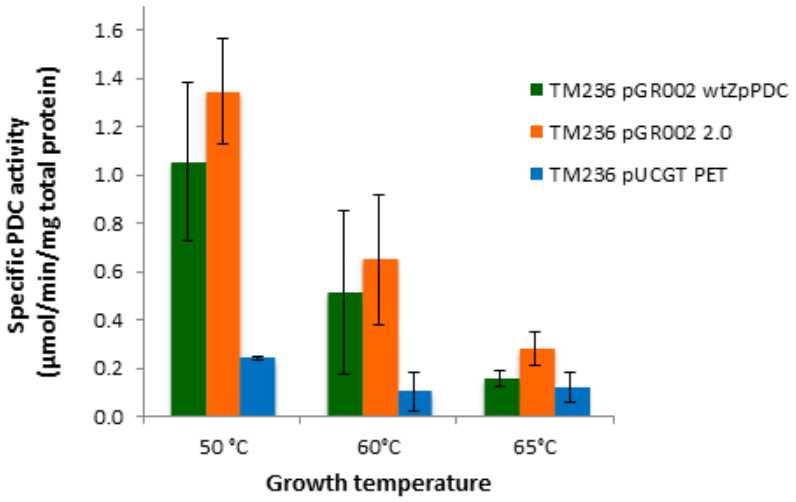


Figure 7.4 Plasmid map of pUCGT PET. pUCGT is based on the *E. coli-Geobacillus* spp. shuttle vector pUCG18, containing features for expression in both hosts. RepBST1 and the kanamycin resistance gene are *G. sterothermophilus* derived and thus allow replication and selection of the plasmid at higher growth temperatures, whereas pMB1 ori and the ampicillin resistance gene are *E. coli* derived. Pldh denotes the *G. stearothermophilus* NCA1503 *ldhA* promoter, while *phe*B RBS marks the position of the ribosome binding site from the *G. stearothermophilus* DSMZ6285 *pheB* gene. This plasmid further contains oriT (incP), an origin for conjugation and *tra*J, an oriT recognizing protein. M13 uni/rev are sequencing primer complimentary sequences.

7.3.3 Aerobic Expression of the PET Operon in *Geobacillus thermoglucosidasius*

G. thermoglucosidasius TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) carrying pUCGT PET was tested in aerobic cultures as described for pGR002 wtZPPDC (see Chapter 4) and pGR002 2.0 (see Chapter 6), with TM236 carrying pUCG18 (empty plasmid backbone) as the negative control. The cultures were grown aerobically in 50 ml 2TY + 12 µg/ml kanamycin + 5 mM thiamine in a 250 ml baffled flask, with shaking at 250 rpm, at 50, 60 and 65°C to an OD_{600nm} of 1.5 to 2.5 and, after sonication, the clarified cell extract was assayed for PDC activity. At the standard assay conditions (at 30°C with added *Saccharomyces cerevisiae* ADH, see General Methods), the PDC detected in pUCGT PET samples was lower than previously observed for pGR002 2.0 (Figure 7.5). SDS-PAGE analysis showed a potentially overexpressed band for the ZpPDC, but no such band for the GtADH6 (Figure 7.6).



TM236 pGR002 wtZpPDC	1.06 ±0.33	0.52 ±0.34	0.16 ±0.03
TM236 pGR002 2.0	1.34 ±0.22	0.65 ±0.27	0.28 ±0.07
TM236 pUCGT PET	0.24 ± 0.007	0.11 ±0.08	0.12 ±0.06

Figure 7.5 Specific PDC activities of *G. thermoglucosidasius* TM236 variants grown aerobically at different temperatures. Specific activity was determined in cell extracts by monitoring the decrease in NADH-dependent absorbance at 340 nm in standard coupled assays at 30°C. These are averages of at least 2 independent expression experiments with 3 assays; error is standard error. Note that no activity was detected for the TM236 strain carrying the empty pUCG18.

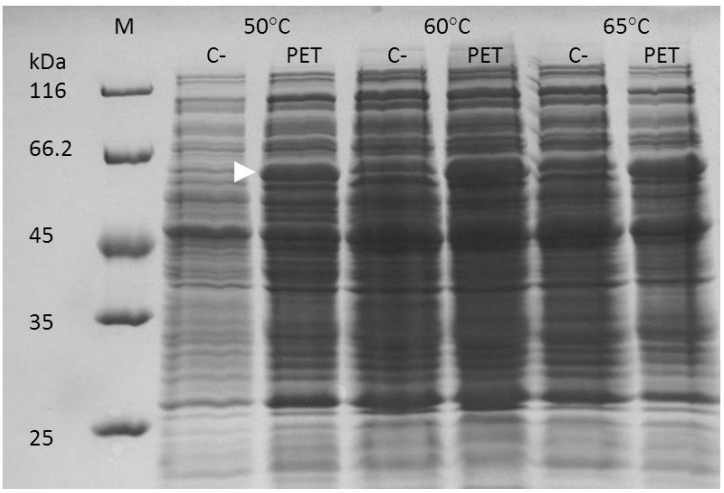


Figure 7.6 SDS-PAGE analysis of *G. thermoglucosidasius* TM236 pUCGT PET expression under aerobic conditions. TM236 was grown at 50, 60 and 65°C, cells lysed, and the extract clarified by centrifugation. The soluble fraction (~10 µg total protein) was run on a 12% SDS-PAGE gel. C- is TM236 carrying an empty pUCG18 plasmid. PET is TM236 carrying the pUCGT PET expression vector. M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific). The white arrow indicates a potential overexpressed band for the ZpPDC at 59.4 kDa. GtADH6 would be expected at 42.5 kDa.

Furthermore, the same culture samples were assayed for ADH activity using acetaldehyde as the substrate at 60°C (see General Methods for assay details, no added ADH). Assays on TM236 pGR002 2.0 were used to give the background ADH activity. This was subtracted from the pUCGT PET data to give the final ADH activity as shown in Figure 7.7 below. Specific ADH activity was increased by the expression of GtADH6 from the PET operon, as expected, but also seemed to follow the PDC trend of decreasing activity being detected with increasing growth temperature of the culture.

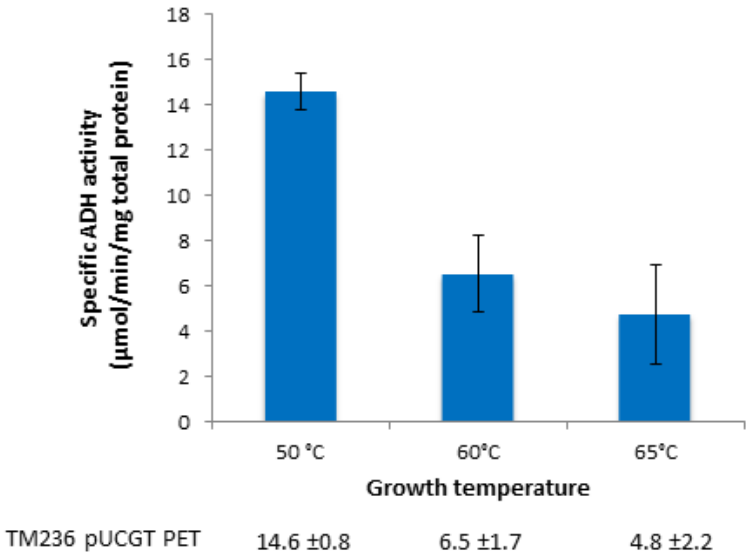


Figure 7.7 Specific ADH activities of *G. thermoglucosidasius* TM236 pUCGT PET grown aerobically at different temperatures. Specific activity was determined in cell extracts by monitoring the decrease in NADH-dependent absorbance at 340 nm at 60°C. These are averages of at least 2 independent expression experiments; error is standard error. The background was measured in TM236 pGR002 2.0 and removed to give the final ADH activity.

Additionally, gene expression was tested by RT-qPCR for these cultures as described in General Methods, using TM236 pUCG18 as the calibration strain. Both *adh*6 and *pdc* were found to be expressed as expected (see Table 7.1). However, Zp*pdc* was much less expressed than observed for TM236 pGR002 2.0, which is in agreement with the PDC activity assay data. Nonetheless, the data showed that both the ZpPDC and GtADH6 were upregulated in the PET operon expressing strain under the aerobic growth conditions used.

Table 7.1 Gene expression levels in *G. thermoglucosidasius* TM236 carrying pGR002 2.0 or pUCGT PET grown aerobically at various temperatures.

Growth temperature	Ratio of the target gene expression in the expression strain vs calibration strain		
	TM236 pGR002 2.0	TM236 pUCGT PET	
	Zp *pdc* 2.0	Zp *pdc* 2.0	Gt *adh*6
50°C	No data	12	23
60°C	80	0.12	0.26
65°C	28	1.78	0.65

7.3.4 Fermentation through the PET Operon in *Geobacillus thermoglucosidasius*

Following on from the promising *in vitro* and aerobic culture results suggesting that this PET operon is functioning at high temperatures and that expression of both enzymes is upregulated even under aerobic conditions (50 ml culture volume in 250 ml baffled flask, grown with shaking at 250 rpm) , the PET operon functionality was tested under fermentative conditions using tube fermentations. These are small-scale fermentations that allow the culture to naturally transition from aerobic growth to micro-aerobic conditions as the headspace is limited to 25% (v/v), the tubes have poor oxygen transfer and no additional air is supplied over the 48 h culture time. They were set-up as described in the methods above using rich ASM (ammonium salts medium + 1% glucose +0.5% yeast extract, see General Methods) and incubated at 50, 60 and 65°C. The fermentation products in the supernatant were analysed using HPLC. Figure 7.8 shows all tested fermentation products and residual glucose; Figure 7.9 shows the ethanol yield (g ethanol produced per g glucose consumed).

It is worth noting that *G. thermoglucosidasius* TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) can naturally produce ethanol through the PDH-ADHE pathway. Therefore, a TM236 strain carrying the empty pUCG18 vector was included as a background control in the fermentation experiments.

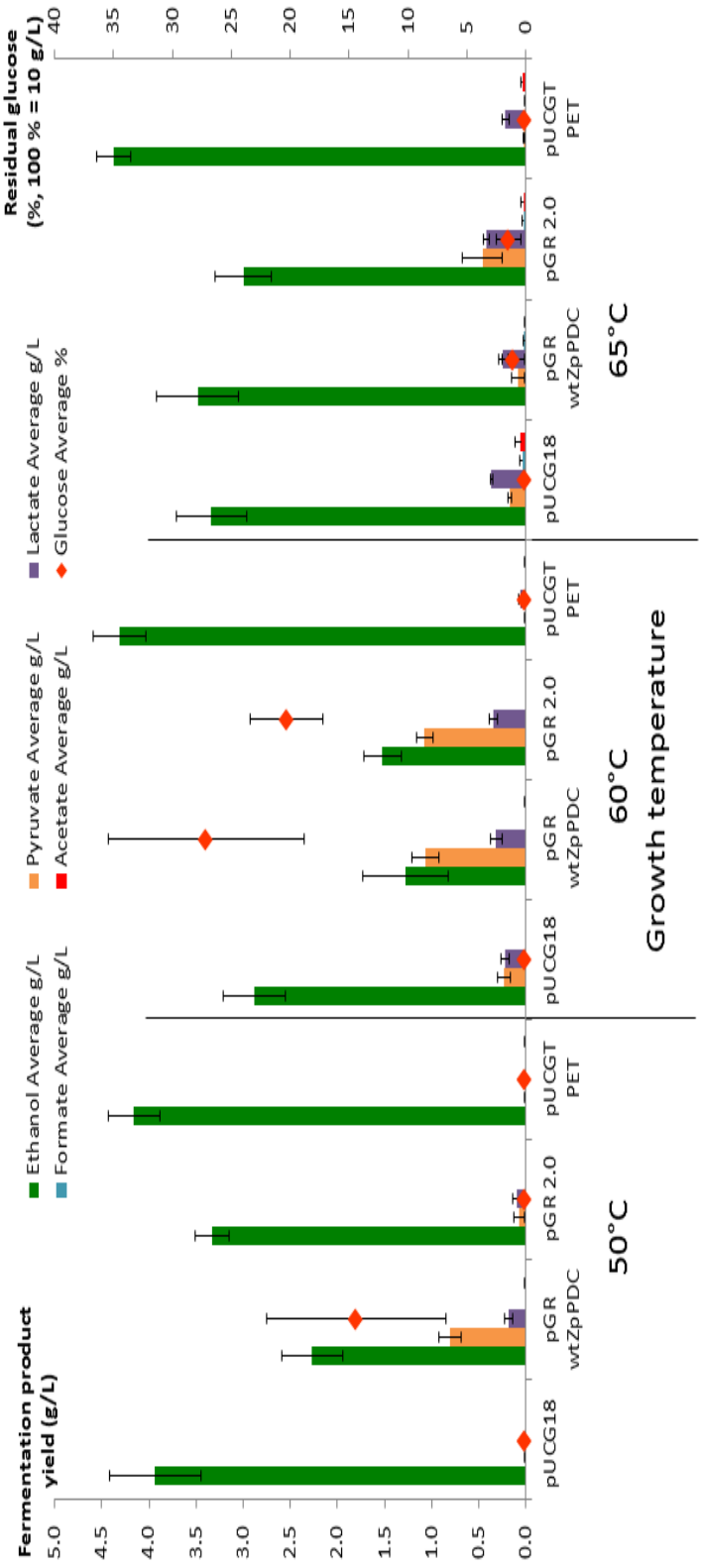


Figure 7.8 Fermentation products and residual glucose in tube fermentations of *G. thermoglucosidasius* TM236 expressing the PET operon and various controls at different growth temperatures. Fermentation product data are shown in g/L, including ethanol, pyruvate, lactate, formate, and acetate; glucose is in % remaining. Bars are labelled with the plasmid carried in TM236. pUCG18 is the empty vector; pGR is pGR002 wtZpPDC; pGR 2.0 is pGR002 2.0 (the codon harmonized *Zppdc*); pUCGT PET (codon harmonized *Zppdc* + GtADH6). These are averages of 3 independent expression experiments each performed in triplicate; error is standard error.

As expected, in all strains across the growth temperature range, ethanol was the major fermentation product. At 50°C growth temperature, TM236 pUCG18 and pUCGT PET had a very similar fermentation product profile with high ethanol production (around 4 g/L) and very few by-products, while TM236 expressing the ZpPDC alone, wt or 2.0 (codon harmonized), accumulated some pyruvate and lactate, and produced less ethanol. At 60°C these differences were enhanced, with higher glucose remaining and more pyruvate and lactate accumulation in the pGR002 wtZpPDC and 2.0 strains. TM236 pUCG18 also accumulated some pyruvate at 60°C fermentation. TM236 expressing the PET operon, however, consumed all the glucose and produced the highest amount of ethanol, on average 4.3 g/L. At 65°C, some accumulation of formate and acetate (0.03 to 0.05 g/L) was observed in all strains.

As the amount of glucose utilized varies between the different strains and fermentation temperatures, the amount of ethanol produced (g/L) is not necessarily a good measure for comparison. However, ethanol yield (g ethanol produced/g glucose consumed) takes those variations into account and thus is a more appropriate measure for comparison of ethanol production. The ethanol yield data is presented in Figure 7.9.

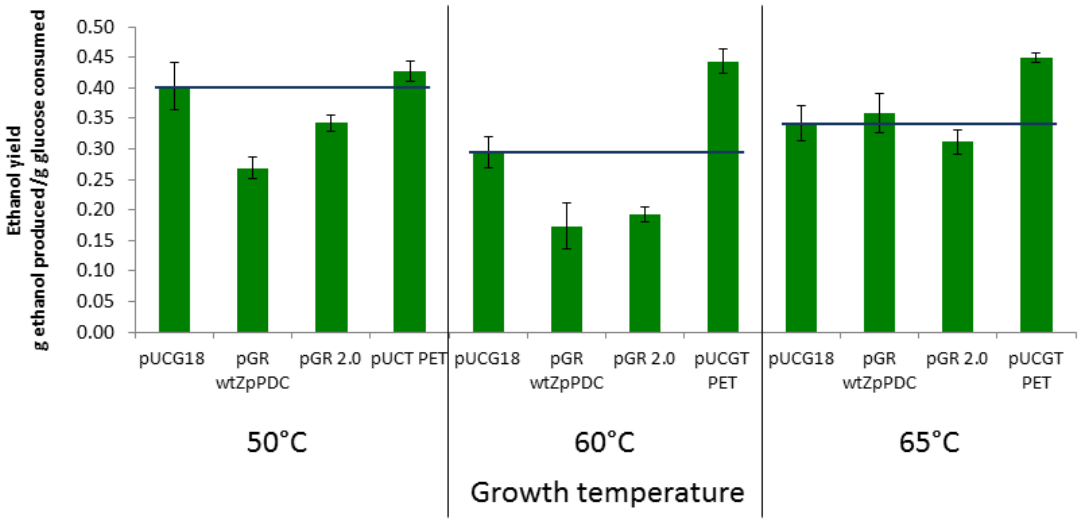


Figure 7.9 Ethanol yield in tube fermentations of *G. thermoglucosidasius* TM236 expressing the PET operon and various controls at different growth temperatures. Ethanol yield is expressed in g ethanol produced per g glucose consumed. Bars are labelled with the plasmid carried in TM236. pUCG18 is the empty vector; pGR is pGR002 wtZpPDC; pGR 2.0 is pGR002 2.0 (the codon harmonized Zp*pdc*); pUCGT PET (codon harmonized Zp*pdc* + GtADH6). These are averages of 3 independent expression experiments each performed in triplicate; error is standard error. The blue lines mark the background level (TM236 pUCG18) at each growth temperature.

The highest and most stable average yields across the temperature range were observed for TM236 expressing the PET operon (between 0.42 and 0.45 g/g). In strains expressing the PDC alone, wt or 2.0, ethanol yields dropped below background levels, apart from TM236 pGR002 wtZpPDC at 65°C.

At 60 and 65°C ethanol yields of the backgound TM236 pUCG18 and the PET operon expressing strain were significantly different (t-Test, $p_{60°C} = 0.01$, $p_{65°C} = 0.03$).

The unfractionated cell extract was assayed for PDC activity with slight modifications to the standard assay (see General methods). In an attempt to remove the influence of NADH-dependent respiratory chain enzymes, such as NADH: quinone oxidoreductoase or complex I, on the assay background, the cell lysate was centrifuged at 100,000 g for 30 min, or the assay was sparged with helium before starting the reaction by the addition of pyruvate to the sealed cuvette. However, no PDC activity could be detected.

Furthermore, fermentation samples of TM236 pGR002 2.0 and pUGT PET were assayed for ADH activity as described for aerobic expression trials above. However, no increased ADH activity was detected for pUCGT PET from 50 and 60°C fermentation samples, and only 0.9 µmol/min/mg total protein were measured in the extract from the fermentation grown at 65°C.

7.3.5 Expression of the PET Operon in *Geobacillus thermoglucosidasius* Δ*adh*E

G. thermoglucosdidasius TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) is able to naturally produce ethanol through the PDH-ADHE pathway. To really test the efficiency of the PET operon in balancing the redox potential in the cell and producing high ethanol yields, attempts were made to express the PET operon in a *G. thermoglucosdidasius* Δ*adh*E strain. *G. thermoglucosdidasius* TM400 (NCIMB 11955 Δ*ldh*, Δ*pfl*, Δ*adh*E, pdh^{up}) does not grow under fully anaerobic conditions and does not produce ethanol (Extance *et al.* 2013). A fully functioning PDC-ADH pathway should be able to restore the redox balance and hence "rescue" growth under anaerobic conditions. Expression of pUCGT PET was not successful in "rescuing" TM400 under fully anaerobic conditions on agar plates or in tube fermentations. In fermentations at 50 and 60°C no ethanol was produced and only 20% of the glucose were utilized, while pyruvate (0.02-0.04 g/L), lactate (0.01-0.05 g/L) and acetate (0.01 g/L) accumulated.

7.4 Discussion

7.4.1 Design and Characterization of the PET Operon

In vitro Characterization of the PET Operon

When purified from *E. coli* BL21 (DE3) and used together in *in vitro* coupled assays, ZpPDC and GtADH6 are functional and stoichiometrically tightly coupled at assay temperatures up to 65°C, suggesting that despite the volatility of acetaldehyde it was not escaping from the reaction in significant amounts.

Aerobic Expression of the PET Operon in *G. thermoglucosidasius*

In crude cell extracts from aerobically grown *G. thermoglucosidasius* TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) cultures expressing the PET operon, PDC activity was following a similar trend compared to cultures expressing the wt or codon harmonized Zp*pdc* on its own, although measurements were unexpectedly low. At growth temperatures of 50 and 60°C the PDC activity measured from cells expressing the PET operon was at 15-20% of the level detected in cell extracts from cultures expressing the wt or codon harmonized Zp*pdc* on its own. However, at a growth temperature of 65°C, PDC activity measured from cells expressing the PET operon was similar to the wt Zp*pdc* expressing strain with 0.12 ±0.06 µmol/min/mg total protein and 0.16 ±0.03 µmol/min/mg total protein, respectively. Compared to the cells expressing the codon harmonized Zp*pdc* on its own, PDC activity in the PET operon expressing cells was at 42% in cultures grown at 65°C. Perhaps this reduction in PDC activity can be explained with the added metabolic burden of over-expressing the GtADH6. When analysing the cell extract on SDS-PAGE, a potential band for an over-expressed PDC could be identified, unlike analyses of cell extracts from cultures expressing the wt or codon harmonized Zp*pdc* on its own. This stands in contrast to the PDC activity measured in cell extracts. Perhaps this is inactive protein.

ADH activities measured in cell extracts from aerobically grown *G. thermoglucosidasius* TM236 cultures expressing the PET operon were increased compared to TM236 carrying the empty plasmid backbone, but no band of over-expressed protein could be identified in SDS-PAGE analysis. However, similar to the PDC data, ADH activity decreased with increasing growth temperature.

Both codon harmonized Zp*pdc* and GtADH6 gene expression levels were analysed by RT-qPCR and showed a 100-fold decrease with increasing growth temperature from 50°C to 60°C.

Peculiarly, gene expression levels for both genes were slightly higher at 65°C than 60°C. Perhaps this was due to experimental error.

It is likely that the decrease in PDC and ADH activity and expression levels observed from cells grown at increasing temperatures is at least partially due to the decrease in expression from the *ldh* promoter as oxygen became increasingly limited in these cultures (Bartosiak-Jentys *et al.* 2012).

Nonetheless, these data were promising for a high temperature PDC-ADH pathway functionally expressed in *G. thermoglucosidasius*.

Fermentation through the PET Operon in *G. thermoglucosidasius*

Further analysis of the PET operon under fermentative conditions showed that the products of the operon were functional up to 65°C. Unlike TM236 strains expressing a PDC on its own, TM236 pUGT PET accumulated very few by-products. Interestingly, TM236 pGR002 wtZpPDC and 2.0 seemed to suffer from some metabolic burden compared to both TM236 carrying the empty pUCG18 vector or expressing the PET operon accumulating more pyruvate. This has also been noted by van Zyl *et al.* (2014b) and suggests that expression of PDC somehow disrupts the metabolism of pyruvate through the PDH pathway. Perhaps this is due to acetaldehyde accumulation and toxicity effects. Acetaldehyde has been shown to inhibit lysine-dependent enzymes by binding to the lysine at the catalytic site (Mauch *et al.* 1986). Peculiarly, 60°C seemed to be the least favourable condition for these strains; large amounts of glucose remained unused and pyruvate accumulated to relatively high levels (around 1 g/L), whereas at 50 and 65°C both strains accumulated more ethanol and less pyruvate and lactate.

With regards to ethanol yields (g ethanol produced per g glucose consumed) TM236 pUGT PET had the highest and most stable yields across the temperature range. It outperformed TM236 pUCG18 by 30% at 60 and 65°C, which is a significant difference. Accumulation of pyruvate, lactate, formate and acetate in TM236 pUCG18 fermentations at 60 and 65°C, perhaps indicate that the flux through PDC is becoming limited at 60°C. Accumulation of lactate from this Δ*ldh* strain was unexpected, but it may be D-lactate which may have been produced in the reversal of the catabolic pathway or some other route triggered by the pyruvate accumulation.

The theoretical maximum yield is 0.51 g/g (Cripps *et al.* 2009). According to Cripps *et al.* (2009) the current production strain TM242 (NCIMB 11955 Δ*ldh*, Δ*pfl*, *pdh*^{up}, developed by TMO Renewables Ltd.) achieves about 0.42 g/g (82% of the theoretical maximum yield) on glucose at 60°C (0.47 g/g, 92% on cellobiose). Under similar conditions TM236 expressing the PET operon achieved 0.44 g/g on gucose (87% of the theoretical maximum yield). That suggests a previously unreported improvement of 5% in ethanol yield on glucose.

Of course these data are preliminary and only comparable to a certain extend as tube fermentation conditions and bioreactor batch fermentations are not necessarily fully equivalent. However, in this study TM236 pUCG18 achieved an ethanol yield of 0.29 g/g (57% of the theoretical maximum) at 60°C in tube fermentations, whereas Cripps *et al.* (2009) reported 0.28 g/g for TM236 in bioreactor batch fermentations at 60°C, on glucose, thus supporting the comparability of the data.

7.4.2 Future Work and Alternative Approaches

In the future, the data presented here should be supported by further analysis of the PET operon under fermentative conditions in the controlled environment of bioreactors, as well as direct comparisons to strains of interest, such as TM242. Furthermore, it would be valuable for industrial purposes to explore further carbon sources, such as cellobiose.

The state of the overexpressed PDC and ADH6 at the protein level remained unclear as there was no detectable activity in the cell extracts from tube fermentations, despite the high ethanol yields observed. Additional investigations into potential protein expression bottlenecks may be useful when engineering these expression strains and constructs further.

Strain Engineering

Consideration should also be given to the strain background. TM236 (NCIMB 11955 Δ*ldh*, Δ*plf*) was certainly a good starting point for the initial design and testing of a functional PET operon. However, the development of a production strain might want to channel carbon flow further through the PDC-ADH pathway by reducing flow through by-products such as acetate. Dr. Chris Hills (2014) has developed *G. thermoglucosidasius* strains with various knock-outs in the acetate production pathway, in particular phosphotransacetylase, phosphotransbutyrylase and acetate kinase. These may be beneficial to the functionality of the PET operon as shown by

Solem *et al.* (2013); they reported ethanol as the sole fermentation product in a *L. lactis* Δ[3]*ldh*, Δ*pta*, Δ*adh*E strain expressing the codon optimized *Z. mobilis* PDC and ADHB.

In the study presented here, some attempts were made to express the PET operon in a *G. thermoglucosdidasius* Δ*adh*E strain, with limited success. *G. thermoglucosdidasius* TM400 (NCIMB 11955 Δ*ldh*, Δ*pfl*, Δ*adh*E, *pdh*^{up}) does not grow under fully anaerobic conditions and does not produce ethanol (Extance *et al.* 2013). Expression of pUCGT PET was not successful in restoring the redox balance and hence "rescuing" TM400 under fully anaerobic conditions on agar plates, or in tube fermentations (no ethanol was produced, and only 20% of the glucose were utilized).

Perhaps the overexpressed PDH outcompetes the PDC. The K_M for pyruvate at 30°C is 0.0011 mM for *G. stearothermophilus* PDH (Fries *et al.* 2003), a closely-related enzyme to the *G. thermoglucosidasius* PDH with 87% amino acid identity. The *Z. palmae* PDC K_M for pyruvate at 30°C is substantially higher at 0.67 mM.

Expression of the PET operon in a suitable Δ*adh*E strain would also create a good basis for a selection strategy in forced evolution experiments. Forced evolution can be a fantastic tool for altering pathways and improving enzyme function under the desired conditions, such as increased thermostability or thermoactivity (Couñago *et al.* 2006, Liao *et al.* 1986, Suzuki *et al.* 2015).

Targeting the PDC-ADH Pathway into Bacterial Microcompartments

In addition to strain engineering approaches to reduce competition for pyruvate and channel carbon flow through the PDC-ADH pathway, compartmentalisation of the reaction may be beneficial and may improve reaction efficiency and product output. Compartmentalisation reduces competition for the substrate by shielding the reaction from endogenous pathways. It increases the proximity between substrate and enzyme and thus enhances pathway efficiency, while also retaining toxic intermediates and thus reducing the toxic effect of acetaldehyde accumulation on the cell. Bacterial microcompartments (BMCs) are organelles that encapsulate reactions in a selectively permeable protein shell (Kerfeld & Erbilgin 2015). Such reactions include the breakdown of carbon compounds via an aldehyde to its cognate alcohol. Proteins appear to be targeted into the BMC by conserved N-terminal peptide sequences (Chessher *et al.* 2015).

Lawerence *et al.* (2014) demonstrated that BMCs can be manipulated and repurposed by targeting the *Z. mobilis* PDC and ADH into the empty propanediol utilization BMC of *Citrobacter freudii* expressed in *E. coli*, thereby creating "ethanol bioreactors". The ethanol yield in strains expressing the PDC and ADH targeted to the microcompartment was higher than the yield in strains expressing cytoplasmically located enzymes.

Hybrid PDCs

Considering that PDC activity data throughout this study showed a drastic decrease with increasing growth temperature of the culture, there is still room for improvement of PDC expression and function at higher temperatures. Unfortunately, the limited resources of this project did not allow for an in-depth investigation into the fate of the PDC during expression at high temperatures. One possible bottle-neck could be the co-translational binding of TPP. If the binding-efficiency of the ZpPDC TPP-binding sites decreases with increasing temperature, this could drastically affect expression of active protein. Perhaps a way to overcome this limitation is to replace the mesophilic TPP-binding sites with sites from thermophilic enzymes. Unlike PDCs, TPP-dependent enzymes, such as acetolactate synthase (ALS), are not limited to mesophiles, but are common in thermophilic organisms, including *Geobacillus* spp. Attempts at creating hybrid PDCs using *G. kaustophilus* ALS TPP-binding sites had previously been made with very limited success (Waite 2010). However, with the ZpPDC crystal structure now available (see Chapter 3), a more rational approach may be possible for this hybrid design.

Reverse engineering Acetolactate Synthase

Another alternative may be reverse engineering a *Geobacillus* spp. ALS to perform the PDC function. As mentioned above, ALS is a common enzyme in thermophilic bacteria. ALS is also a TPP-dependent enzyme and catalyses the carboligation between two pyruvate molecules to form acetolactate and carbon dioxide. Despite differing functions, ALS and PDC share a high sequence similarity. Furthermore, the conversion of pyruvate proceeds via a common reaction intermediate, 2-hydroxyethyl-TPP, in both enzymes. ALS catalyses the carboligation between 2-hydroxyethyl-TPP and a second pyruvate molecule to form the reaction product acetolactate, while in the PDC 2-hydroxyethyl-TPP is protonated to form acetaldehyde.

Cheng *et al.* (2016) produced a mutant library of the thermophilic ALS from *Thermus thermophilus* and through screening identified a quadruple mutant with improved acetaldehyde production (3.1 fold). This enzyme is very stable (retains 100% activity after incubation at 60°C for 60 min), and although not yet a perfect acetaldehyde producer with some acetolactate activity still remaining, it is a promising attempt in creating a thermoactive PDC for biotechnological applications. However, whether this modified enzyme would function well *in vivo* remains to be seen.

Ferredoxin Oxidoreductases

Some bacteria and archaea use an unusual variation of mixed acid fermentation pathways to convert pyruvate into aldehydes and acetyl-CoA (Figure 7.10). Instead of using NAD^+/NADH-dependent enzymes, they use ferredoxin oxidoreductases (iron-sulphur proteins). Several ferredoxin oxidoreductases have been studied in-depth and some have been found to be bifunctional enzymes with PDC often as a secondary function.

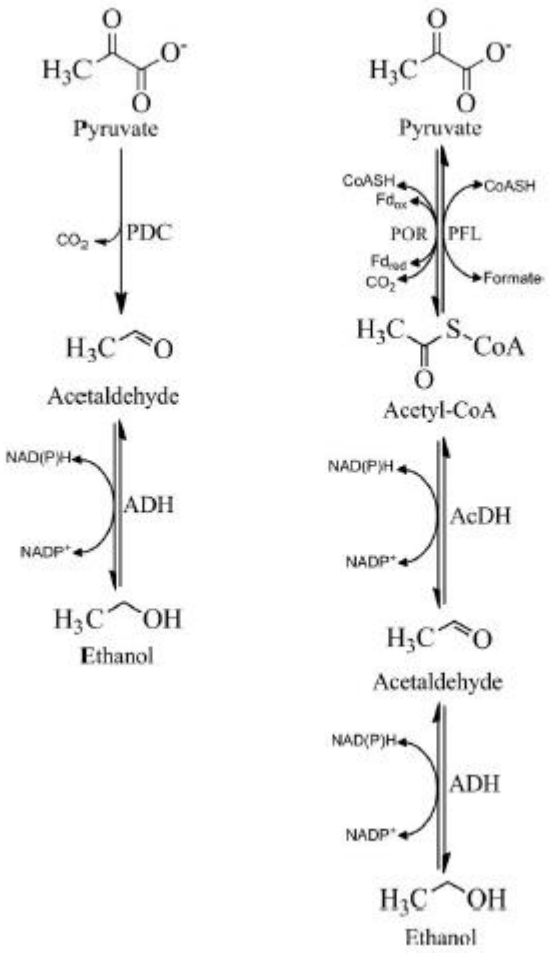


Figure 7.10 Two variations of ethanol production pathways from pyruvate. Abbreviations are: Fd_{ox}, oxidized ferredoxin; Fd_{red}, reduced ferredoxin; POR, pyruvate ferredoxin oxidoreductase; CoASH, coenzyme A; PDC, pyruvate decarboxylase; PFL, pyruvate formate lyase; AcDH, acetaldehyde dehydrogenase; ADH, alcohol dehydrogenase. From Eram & Ma (2013).

Ma *et al.* (1997) described an α-keto acid ferredoxin oxidoreductase (KAOR), or more specifically a pyruvate ferredoxin oxidoreductase (POR or PFOR), from the hyperthermophilic archaeon *Pyrococcus furiosus*, a key enzyme in its fermentation pathway (Figure 7.11). In the presence of CoA, this POR catalyses the anaerobic oxidation of pyruvate to acetyl-CoA using ferredoxin (Fd) as the electron acceptor. However, as a second function, it also catalyses Co-A-dependent, Fd-independent, anaerobic, non-oxidative decarboxylation of pyruvate to acetaldehyde. The reaction in *P. furiosus* is, however, highly oxygen-sensitive, and POR has a T_{opt} for acetaldehyde production of about 90°C. Which function is carried out seems to depend on the cellular redox potential.

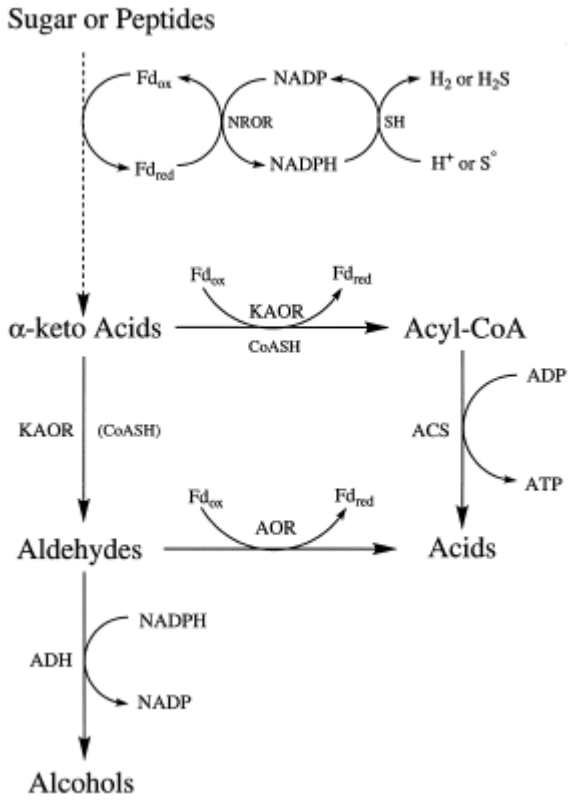


Figure 7.11 Fermentation pathway using ferredoxin oxidoreductases based on proposed metabolic roles in *Pyrococcus furiosus.* Abbreviations are: Fd_{ox}, oxidized ferredoxin; Fd_{red}, reduced ferredoxin; KAOR, α-keto acid ferredoxin oxidoreductase; CoASH, coenzyme A; AOR, aldehyde ferredoxin oxidoreductase; ACS, acetyl-CoA synthases; FNOR, ferredoxin NADP oxidoreductase; SH, sulfhydrogenase. From Ma & Adams (1999).

Bifunctional PORs have also been described from the hyperthermophilic bacteria *Thermotoga maritima* and *Thermotoga hypogea*. Again, the PDC activity is CoA-dependent and highly oxygen sensitive (Eram *et al.* 2015), which makes them unsuitable for simple cloning into *G. thermoglucosidasius*.

The POR from the mesophilic, anaerobic sulphate-reducing bacterium *Desulfovibrio africanus* is stable towards oxygen and has reportedly been expressed in anaerobically grown *E. coli* (Piuelle *et al.* 1997). Several crystal structures are available on PDB. Whether this mesophilic enzyme would be suitable for expression in a thermophile is not known.

Bioinformatics searches in *G. thermoglucosidasius* found two genes (RTMO 04514 and 04513) with a ferredoxin oxidoreductase annotation. RTMO 04514 seems to belong to the POR superfamily, but is specific for 2-oxoglutarate/succinyl-CoA. RTMO 04513 is likely another

subunit. No genes were found annotated with the PDC function. A BLAST search against known PORs with PDC function did not yield any results in *G. thermoglucosidasius.*

With increasing information on PORs it may be possible to find an enzyme suitable for cloning and expression in *G. thermoglucosidasius*.

An alternative to cloning and modifying modern PDCs or enzymes with PDC function is ancestral sequence reconstruction which is a computational molecular evolution strategy with the potential to yield enzymes with high thermostability and thermoactivity, as discussed in Chapter 8.

8. Ancestral Sequence Reconstruction of Bacterial PDCs

8.1 Introduction

Our current knowledge on bacterial PDCs is extremely limited, and no PDC has yet been identified in a thermophile. One approach to open up new opportunities for bacterial PDCs adapted to a variety of environments, potentially including high temperature conditions, is to exploit the diversity provided by evolution. Ancestral sequence reconstruction (ASR), first envisaged by Pauling and Zuckerkandl (1963), is a method of computational molecular evolution to infer extinct ancestral protein sequences, which can then be synthesized and experimentally characterized. ASR explores a sequence space that has already been screened over evolutionary time spans, thus reducing the non-functional space that would otherwise be included in protein libraries generated by random mutagenesis, for example. Thus, ASR has the advantage over random or purely computational approaches in that it limits the "design-space" to proteins that are properly folded and have a demonstrable activity (Cole & Gaucher 2011, Hobbs *et al.* 2012). Of course, one has to be cautious of bias and inaccuracies in the results, which may lead the inferred ancestor to exhibit an inaccurate phenotype, especially when trying to accurately predict ancestor function or relating features of ancestral proteins to environmental conditions at the time these may have existed (Gaucher *et al.* 2008, Wheeler *et al.* 2016).

What is ASR?

ASR uses evolutionary information stored in extant sequences of proteins from extant organisms and their phylogenetic relationships to infer the sequences of their extinct ancestors. It is a 6-step process:

1. Retrieve extant protein sequences (amino acid and nucleotide sequences) and align them.
2. Generate a phylogenetic tree to illustrate the evolutionary relationship between the sequences.
3. Determine the statistical model of evolution for nucleotide and amino acid substitution.
4. Apply an algorithm containing the model of evolution to the sequence alignments and the phylogenetic tree to infer the ancestral sequences for each node in the phylogenetic tree using three inference methods: amino acid, codon and nucleotide.

5. Infer gaps in the ancestral sequences.
6. Generate a consensus sequence for nodes of interest.

The model of evolution characterizes the evolutionary processes responsible for trait evolution, while the phylogenetic tree specifies the route through which the ancestral sequences evolved into extant sequences and gives the measure of evolutionary time that has past as determined by the branch length in the phylogenetic tree. The inference algorithm considers each site in the sequence alignment one at a time, examines which residues are present at the site in the extant sequences and decides which residue was most likely at that site in the ancestral sequence, given the length of evolutionary time that has past.

The inference method applied in the work described in this Thesis used maximum likelihood (ML) probability. ML provides statistical confidence in each reconstructed ancestral state as it considers three pieces of information, namely the alignment, the tree topology and branch length, and the model of evolution (amino acid or nucleotide substitution rates) (Thornton 2004). The likelihood of each possible amino acid occupying a particular site in the ancestral sequence is calculated for each position in the sequence at each internal node of the phylogenetic tree. Furthermore, the likelihood is the probability of observing all the extant states, given the predicted ancestor state, the tree and the model (Harms & Thornton 2010, Thornton 2004). However, ML does not account for sampling errors when estimating parameters (personal communication with Dr. Andrews).

ASR has two major limitations: (1) it is limited by the availability of extant sequences in a variety of species or domains of life, and (2) its difficulty handling insertions and deletions, or sequence regions that cannot be correctly aligned (Arenas & Posada 2010, Cole & Gaucher 2011). These limitations may be overcome with the ever increasing sequence data base and our increasing understanding of the evolution of insertion and deletion events (Cole & Gaucher 2011).

Investigating Thermophily using ASR

It is thought that the thermostability of proteins in microorganisms is dictated by the environmental temperature, i.e., the hotter the environmental conditions, the more thermostable the enzyme will have to be in order to be functional, and this correlation works

well across the entire temperature range (Gromiha *et al.* 1999). However, it remains unclear which general mechanism underlies increased thermostability of enzymes (Yano & Poulos 2003). It seems that a balance between better packing and increased solubility of the protein at higher temperatures increases its thermostability (Gromiha *et al.* 1999). With this limited understanding of thermophily, a design approach may not always lead to the desired results.

It has been suggested that ASR may be most useful when investigating proteins with an interest in increased thermostability (Cole & Gaucher 2011). It has been hypothesised that thermophily is a primitive trait exhibited by the oldest forms of life on Earth (Di Giulio 2003, Pace 1991). This hypothesis is based on the fact that thermophilic organisms are found on branches close to the root of the tree of life (Di Giulio 2003, Stetter 1996) (Figure 8.1).

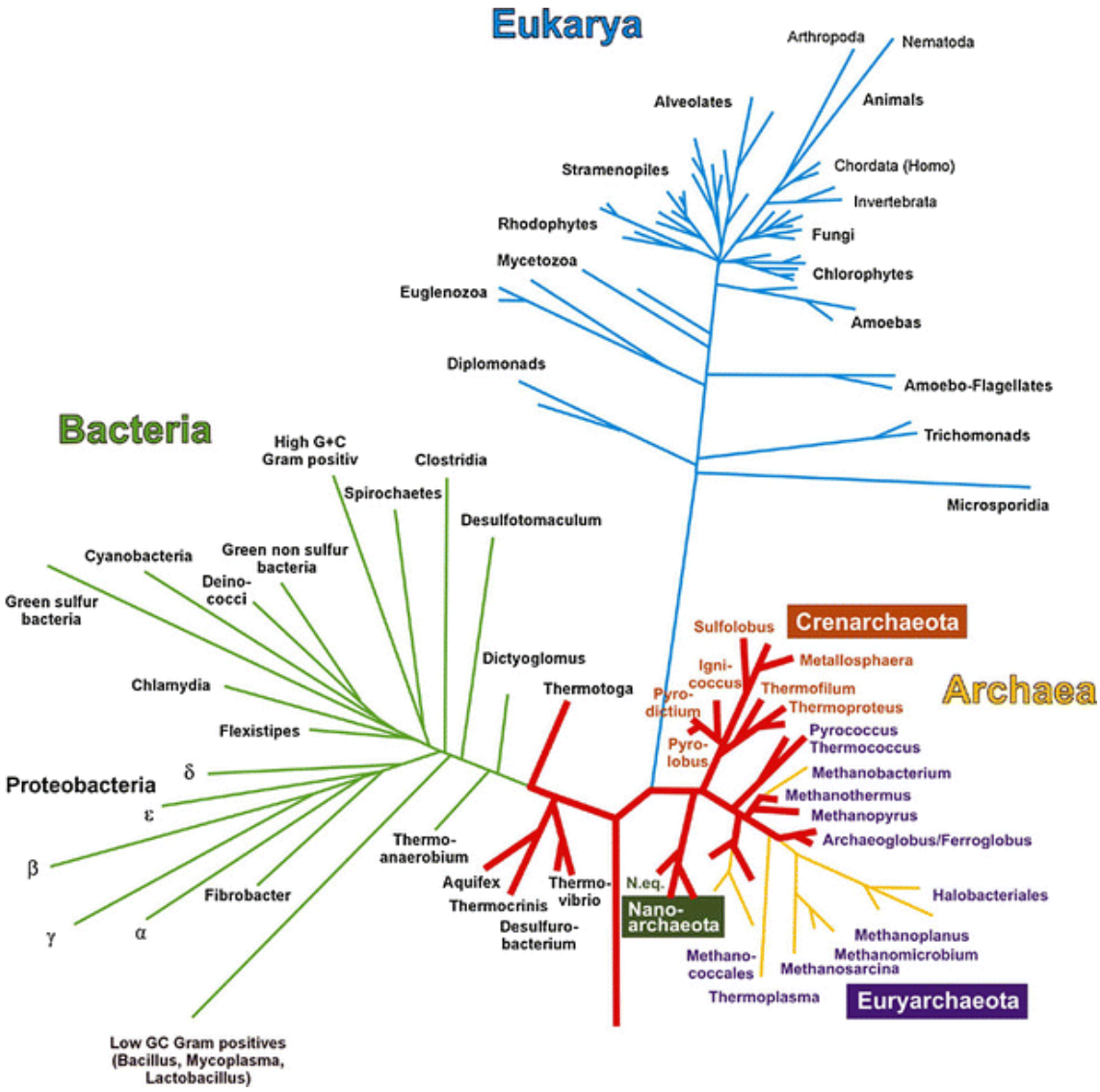


Figure 8.1 The tree of life. Phylogenetic tree based on 16S ribosomal RNA comparisons. The red, bulky lineages represent hyperthermophiles. From Stetter (2006).

There have been several studies using ASR to evaluate the hypothesis that thermophily is a primitive trait. Exploiting the correlation between environmental temperatures and thermostability of a protein, Gaucher *et al.* (2008) showed that environmental conditions between 3.5 and 0.5 billion years ago progressively cooled by 30°C. Using ASR Gaucher *et al.* computationally reconstructed several ancestral sequences of the bacterial elongation factor TU and experimentally characterized these proteins. They found that the older the inferred ancestor, the higher the thermostability, e.g., the inferred last proteobacterial ancestor of the protein was thermostable up to 57°C and the last inferred bacterial common ancestor up to 73°C, which was 30°C more thermostable than the extant *E. coli* TU.

Akanuma *et al.* (2013) observed a similar effect when using nucleoside diphosphate kinases in their ASR study of the last bacterial ancestor. Similarly, Risso *et al.* (2013) observed inferred, hyperstable β-lactamase ancestors from around 1 to 3 billion years ago, which were up to 35°C more thermostable than extant enzymes. These studies suggest that the older the inferred ancestor is, the more likely it is that the inferred protein is more thermostable.

However, some recent studies using ASR to characterize inferred Precambrian ancestors of extant mesophilic and thermophilic species have revealed fluctuating trends in thermaladaptation. Using LeuB (3-isopropylmalate dehydrogenase) from various *Bacillus* spp. Hobbs *et al.* (2012) reconstructed several functional inferred Precambrian enzymes. Despite the lack of deeply branching, thermophilic bacteria such as *Thermotoga* spp. in the inference, the ancestral LeuBs move progressively closer to extant *Thermotoga* spp. LeuB in sequence identity. Moreover, Hobbs *et al.* (2012) observed a fluctuating trend in thermal evolution: with a thermostable oldest inferred ancestor, temporary adaptation towards mesophily and a more recent return to thermophily. Structural analysis suggested that thermophily arose twice, independently.

Using ASR to trace the divergence of ribonuclease RNHI from the common ancestor towards an *E. coli* or *Thermus thermophilus* lineage, Hart *et al.* (2014) observed opposite trends in thermostability. On the one hand, following the thermophilic lineage towards the extant *T. thermophilus* RNHI, the denaturing temperature of the protein increased as it became more modern. On the other hand, following the mesophilic lineage, the older the inferred ancestor, the more thermostable it appeared to be, with the oldest inferred ancestor exhibiting an 11°C increase in denaturing temperature. However, the mechanisms of thermostability fluctuated within both lineages.

These studies demonstrate that thermophily may not be an exclusively primitive trait, but is a rather complex evolutionary trait and that predictions about thermostability of ancestral proteins must be made with caution.

Whichever trend these studies have discovered, they all show that ASR has great potential use in the development of enzymes with high thermostability and thermoactivity for biotechnology.

The work discussed in this chapter was completed in collaboration with Dr. Emma Andrews, Konny Shim and Professor Vic Arcus at the University of Waikato, New Zealand, who provided extensive technical support regarding ASR. This collaboration was supported by a Microbiology Society research visit grant (RVG14-10) and the Biochemical Society.

8.2 Methods

8.2.1 Retrieving Pyruvate Decarboxylase Sequences

PDC amino acid and nucleotide sequences were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/genbank). The input for this reconstruction was 25 bacterial PDC sequences. Confirmed PDCs included *A. pasteurianus* (Gocke *et al.* 2009), *G. oxydans* (van Zyl *et al.* 2014b), *S. ventriculi* (Lowe & Zeikus 1992), *Z. mobilis* (Dobritzsch *et al.* 1998) and *Z. palmae* PDC (Raj *et al.* 2002).

The other 20 PDC sequences (details in Table 8.1) were identified through a BLAST search based on >50% amino acid sequence identity over >90% coverage to the reference *Z. palmae* PDC (see Appendix I for sequence).

Table 8.1 Sequences used in ASR. *Z. palmae* PDC is the reference protein.

Strain	GenBank accession number amino acid/nucleotide	% amino acid identity to reference
Acetobacter aceti	KDE19620.1/JEOA01000009.1	76
Acetobacter malorum	KFL91998.1/JOJU01000008.1	63
Acetobacter nitrogenifigens	WP_026398870.1/NZ_AUBI01000020.1	68
Acteobacter papaya	WP_025859940.1/NZ_BAIN01000005.1	62
Acetobacter pasteurianus	AF368435.1	76
Acetobacter persici	WP_025828656.1/NZ_BAJW01000064.1	69
Acetobacter pomorum	WP_006115789.1/NZ_AEUP01000018.1	76
Acetobacter tropicalis	WP_006559524.1/NZ_BABS01000096.1	62
Acidomonas methanolica	GAJ29946.1/BAND01000085.1	71
Beijerinckia indica	WP_012386294.1/NC_010581.1	65
Commensalibacter sp MX01	EUK18520.1/ATSX01000001.1	56
Gluconacetobacter diazotrophicus	AIG13066.1/KJ746104.1	71
Gluconobacter frateurii	WP_010505240.1/NZ_BADZ01000024.1	70
Gluconobacter morbifer	WP_008852112.1/NZ_AGQV01000006.1	69
Gluconobacter oxydans	KF650839.1	67
Gluconobacter thailandicus	WP_007283613.1/NZ_BAON01000015.1	68
Komagataeibacter europaeus	WP_019085054.1/NZ_CADP01000006.1	65
Komagataeibacter hansenii	WP_003622049.1/NZ_CM000920.1	64
Komagataeibacter medellinensis	WP_014105323.1/NC_016027.1	64
Komagataeibacter oboediens	WP_010515737.1/NZ_CADT01000033.1	66
Kozakia baliensis	WP_029604671.1/NZ_JNAB01000024.1	78
Ktedonobacter racemifer	WP_007922190.1/NZ_ADVG01000005.1	52
Sarcina ventriculi	AF354297.1	31
Zymomonas mobilis	AAA27696.2/M15393.2	63

8.2.2 Ancestral Sequence Reconstruction

Amino acid sequences were aligned using Geneious (version R7 7.1.7, Kearse *et al.* 2012) MUSCLE alignment option (8 iterations). The alignment was exported as a FASTA file and refined using Gblocks (Castresana 2000). Gblocks trims variable regions from the MUSCLE alignment by selecting conserved blocks, anchoring them with positions that can be aligned with high confidence and trimming non-conserved positions around gaps.

Using the Gblocks alignment, ProtTest (Abascal *et al.* 2005) was then run to determine the most appropriate model of amino acid evolution according to Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The model determined was the Whelan and Goldman model (WAG) with the base frequency +F.

The Gblocks alignment generated and the best model of amino acid evolution as determined by ProtTest (WAG) were used in PhyML (version 3.0, Guindon *et al.* 2010) to build a phylogenetic guide tree based on ML phylogenies.

This guide tree, together with the MUSCLE alignment described above, was then used in PRANK (available at http://www.ebi.ac.uk/goldman-srv/prank/prank/) to generate an amino acid alignment based on phylogenetic information.

The resulting amino acid alignment was analysed using GARLI 2.0 (Zwickl 2010) while implementing the WAG model of evolution to find the best ML tree, based on tree topology (the more consistent between several iterations, the better) and log likelihood scores (the smaller the score, the "better" the tree). GARLI was also used to bootstrap the tree using 1,024 replicates. This tree was rooted to the outgroup (PDC sequences with <60% amino acid identity to the *Z. palmae* reference PDC, in this case *S. ventriculi* with 31% and *K. racemifer* with 52%) in Geneious.

The *pdc* nucleotide sequences were collated in Geneious and manually curated to include gaps as determined by the PRANK amino acid sequence alignment.

JModelTest (version 2.1.5, Posada 2008) was performed to determine the best model of nucleotide evolution under the AIC and BIC. The model determined was the general time-reversible model (GTR, aka REV) with the rate variation +I and +G.

Three different methods of ancestral inference (amino acid, nucleotide and codon inference) were performed using PAML software (version 4.3, Yang 2007) under the ML criterion. For nucleotide sequence inference in BASEML and codon inference in CODEML, the GTR substitution model was employed. For amino acid inference in CODEML the model WAG was used. The PAML algorithm uses the sequence alignment and the best ML GARLI tree to consider each site in an alignment, one at a time. It determines which residues/bases are present at that site in the extant sequences and decides which residue/base was most likely at that site in the ancestral sequence, given the length of evolutionary time that has passed based on the tree branch lengths.

Ancestral gaps were inferred using PRANK based on the PRANK amino acid alignment and the best ML GARLI tree.

Taken together the inferred sequences for nodes of interest were compiled using Awk through the Cygwin terminal (from http://cygwin.com/install.html), thus generating a consensus sequence from all methods of inference.

Any ambiguities in the resulting consensus amino acid sequence were resolved taking the following into consideration: (1) physiochemical properties, (2) structural environment,

(3) most common residues present in extant sequences, (4) residue as predicted by codon inference method (considered to be the most robust method, Hobbs *et al.* 2012).

8.2.3 Node Age Estimates

The best ML GARLI tree was aged using r8s (version 1.8, Sanderson 2003, available at http://loco.biosci.arizona.edu/r8s/). The point of divergence of Proteobacteria and Firmicutes (3.19 billion years ago) was used as the calibration point, based on the robust prokaryotic phylogeny study in Battistuzzi *et al.* (2004). Battistuzzi *et al.* (2004) used 32 protein sequences and molecular divergence times estimated from geological calibration points in their phylogenetic study. The resulting tree was visualized in FigTree (version 1.4.2, available at http://tree.bio.ed.ac.uk/software/figtree/).

8.2.4 Gene Synthesis and Cloning for Expression in *E. coli*

The resulting amino acid sequence for Node 27 was backtranslated in Geneious using *Geobacillus stearothermophilus* codon usage as a guide. The genes were synthesized by Eurofins (Germany) and delivered in pEXK (a holding vector). *Xba*I and *Xho*I were used to excise the gene, as well as cutting the target pET28a. The released fragment and the linearized vector were agarose gel purified, ligated (using 20 ng linearized vector and 17 ng insert), and the product transformed into *E. coli* BioBlue as described in General Methods.

An *E.coli* BioBlue colony carrying the correct clone of pET28 Node 27, as confirmed by sequencing with T7 F/R and Node 27 F/R, was used to propagate the plasmid, which was purified and transformed into *E. coli* BL21 (DE3) cells for protein expression (see General Methods).

8.2.5 Enzyme characterization – Kinetic and Thermal Properties

The protein was expressed in *E. coli* BL21 (DE3) and purified through nickel-affinity chromatography and buffer exchange through size-exclusion purification, as detailed in General Methods.

Kinetic and thermal properties were determined as described in Chapter 3 (see section 3.2.2 in particular).

8.3 Results

8.3.1 Ancestral Sequence Reconstruction

Node 27 (with an inferred age of 1,248 million years, Figure 8.2) was reconstructed using ASR. The three methods of inference (amino acid, nucleotide and codon inference) resulted in <3% ambiguous sites, which were resolved as described in section 8.2.1 (sequence in Appendix I). The average posterior probability at each site was 0.6946 (0.89443, 0.43704, 0.77432, confidence scores for amino acid, codon and nucleotide inference respectively). The final amino acid sequence showed an amino acid identity of <79% compared to all extant sequences used in the inference (Table 8.2).

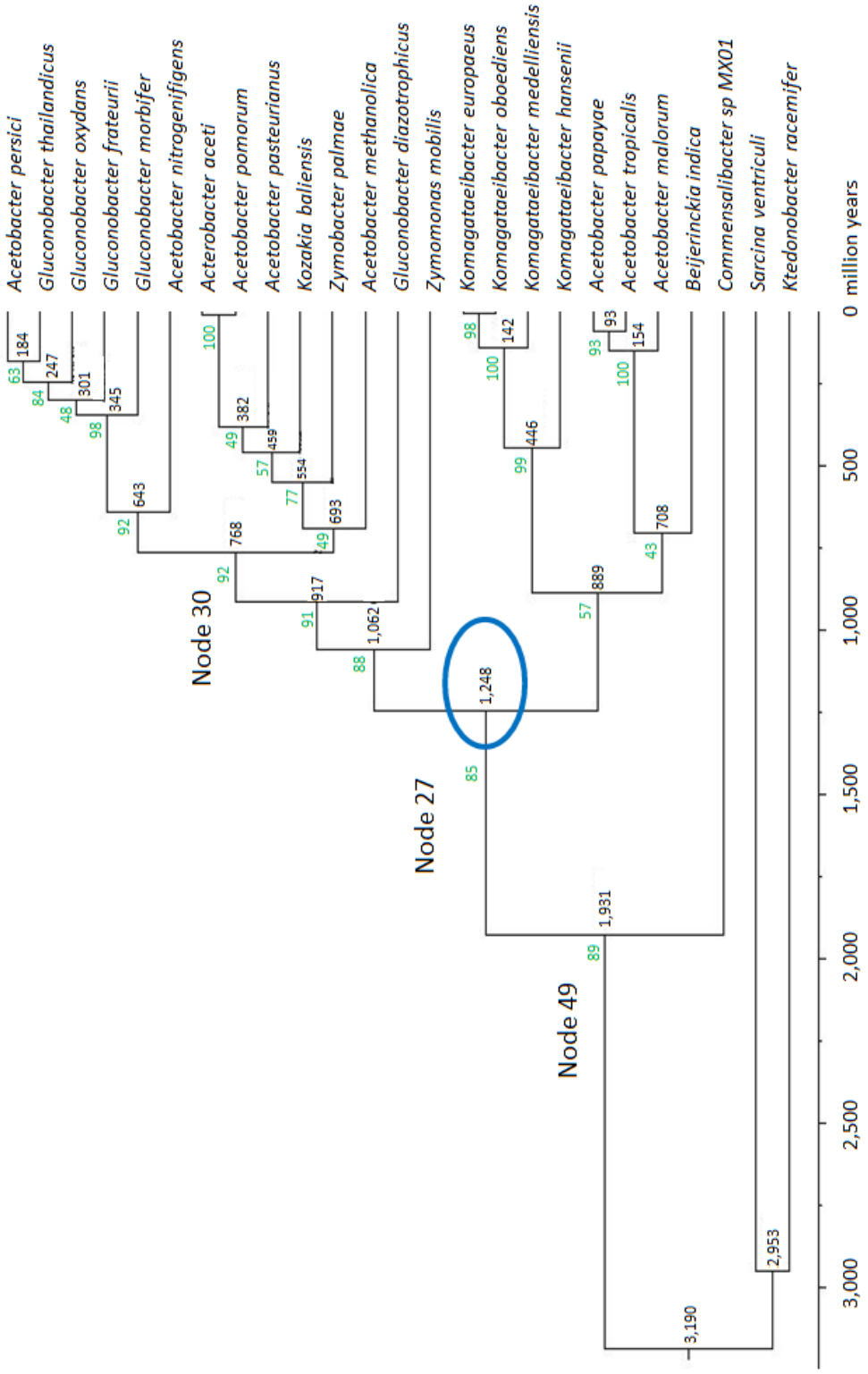


Figure 8.2 Maximum likelihood chronogram based on PDC amino acid sequences. Node 27 (1,248 million years old) is highlighted in blue and was reconstructed for characterization. Inferred node ages are indicated in black. Bootstrap percentages were assessed by 1,024 bootstrap replicates and are given in green. *S. ventriculi* and *K. racemifer* represent the outgroup. The point of divergence of Proteobacteria and Firmicutes (3.19 billion years ago) was used as the calibration point for node age estimates.

Table 8.2 Sequence comparison of Node 27 PDC to all PDCs used as ASR input.
This information was calculated in Geneious using ClustalW alignments. 27 refers to Node 27. Full organism names can be found in Figure 8.2. Mol… is molecule type, AA is amino acid.

Name ▲		... ...	Sequence Length	Topology	Mol...	# Sequences	% Pairwise Identity	% Identical Sites	Max Sequence Length	Alignment method	Min Sequence Length
27Aaceti		... -	563	linear	-	2	68.4%	68.4%	563	ClustalW Alignment	558
27Amalorum		... -	566	linear	-	2	74.0%	74.0%	563	ClustalW Alignment	563
27Amethanol		... -	563	linear	-	2	67.5%	67.5%	563	ClustalW Alignment	558
27Anitrogen		... -	566	linear	-	2	64.7%	64.7%	563	ClustalW Alignment	561
27Apapayae		... -	566	linear	-	2	73.1%	73.1%	563	ClustalW Alignment	563
27Apasteur		... -	563	linear	-	2	67.0%	67.0%	563	ClustalW Alignment	557
27Apersici		... -	569	linear	-	2	66.3%	66.3%	563	ClustalW Alignment	561
27Apormorum		... -	563	linear	-	2	68.2%	68.2%	563	ClustalW Alignment	558
27Atropicalis		... -	566	linear	-	2	73.1%	73.1%	563	ClustalW Alignment	563
27Bindica		... -	565	linear	-	2	77.0%	77.0%	563	ClustalW Alignment	562
27CspMX01		... -	568	linear	-	2	67.1%	67.1%	566	ClustalW Alignment	563
27Gdiazotrop		... -	563	linear	-	2	74.4%	74.4%	563	ClustalW Alignment	558
27Gfrateurii		... -	569	linear	-	2	67.0%	67.0%	563	ClustalW Alignment	563
27Gmorbifer		... -	568	linear	-	2	67.8%	67.8%	563	ClustalW Alignment	562
27Goxydans		... -	569	linear	-	2	66.1%	66.1%	563	ClustalW Alignment	563
27Gthailand		... -	569	linear	-	2	66.7%	66.7%	563	ClustalW Alignment	561
27Kbaliensis		... -	563	linear	-	2	70.0%	70.0%	563	ClustalW Alignment	558
27Keuropaeus		... -	566	linear	-	2	76.3%	76.3%	564	ClustalW Alignment	563
27Khansenii		... -	566	linear	-	2	77.7%	77.7%	564	ClustalW Alignment	563
27Kmedell		... -	566	linear	-	2	78.1%	78.1%	564	ClustalW Alignment	563
27Koboediens		... -	566	linear	-	2	76.7%	76.7%	564	ClustalW Alignment	563
27Kracemifer		... -	582	linear	-	2	60.9%	60.9%	582	ClustalW Alignment	563
27Sventri		... -	566	linear	-	2	33.0%	33.0%	563	ClustalW Alignment	552
27Zmobilis		... -	569	linear	-	2	71.1%	71.1%	568	ClustalW Alignment	563
27Zpalmae		... -	563	linear	-	2	68.8%	68.8%	563	ClustalW Alignment	556
28 node27(23) cons ...		... -	563	linear	AA	-	-	-	-	-	-

8.3.2 Cloning for Expression in *E.coli*

The Node 27 *pdc* was cloned into pET28a, to generate the enzyme with a C-terminal hexa-histidine-tag (see plasmid map below, Figure 8.3).

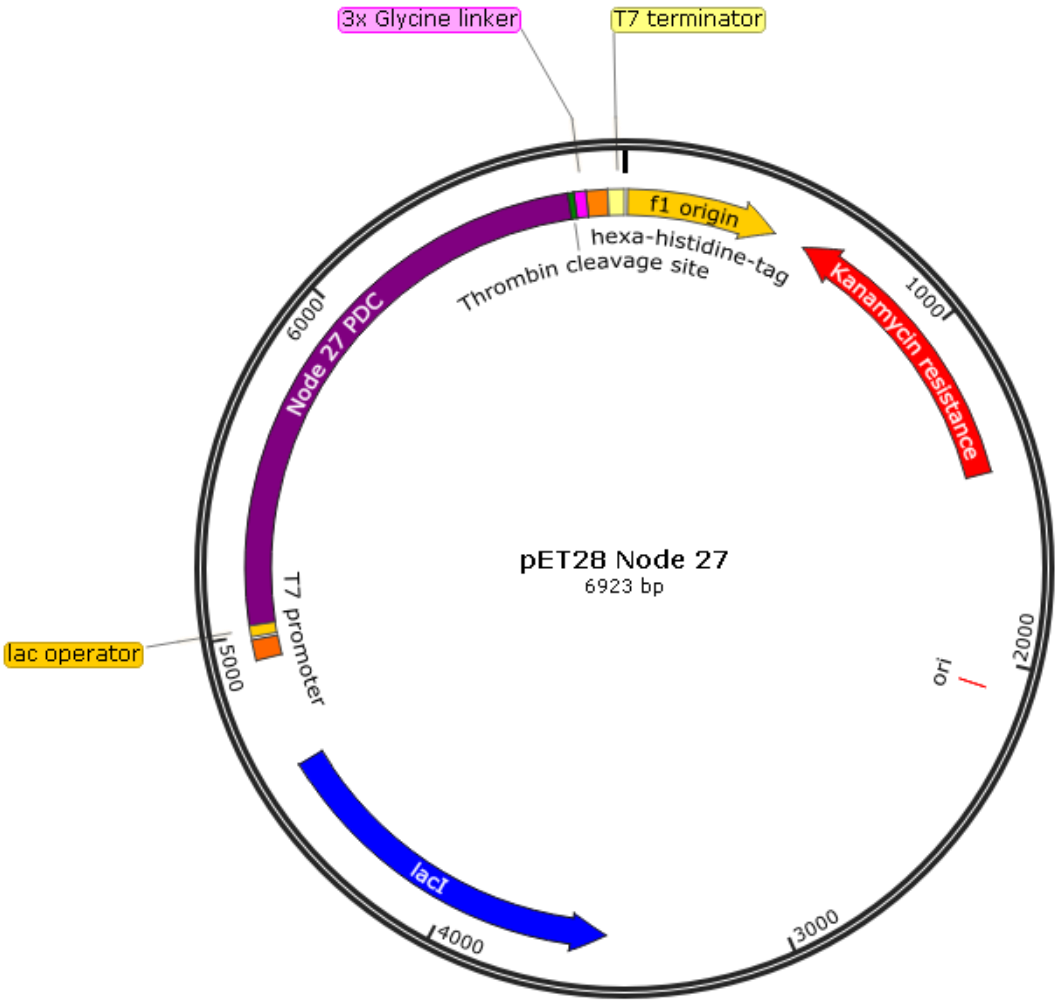


Figure 8.3 Plasmid map of pET28 Node 27 PDC. The neomycin phosphotransferase gene (labelled kanamycin resistance) confers resistance to kanamycin. In the presence of a T7 RNA polymerase, the Node 27 *pdc* is expressed from the T7 promoter under the control of the *lac* operator. *Lac*I encodes the *lac* operon repressor. The f1 origin is the origin of replication from an f1 phage.

8.3.3 Recombinant Expression and Purification

The Node 27 *pdc* was expressed in *E. coli* BL21 (DE3) from pET28 Node 27 using the T7 expression system. The recombinant protein is tagged with a C-terminal hexa-histidine-tag, with a predicted size of 62.8 kDa, and was purified by nickel-affinity chromatography, followed by a further purification and buffer exchange step through size-exclusion chromatography (see traces in Figure 8.4 and 8.5 below). PDC activity in the protein-containing fractions was confirmed and purity of the protein was assessed by SDS-PAGE (Figure 8.6). The pure protein was used in enzyme characterization studies.

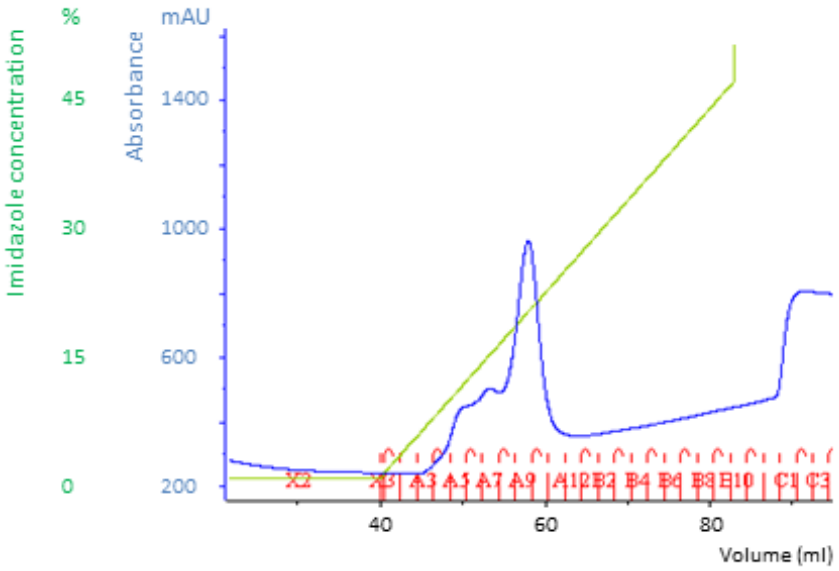


Figure 8.4 Nickel-affinity chromatogram for recombinant Node 27 PDC. The dark blue line is the absorbance at 280 nm. The green line indicates the imidazole concentration in the His-elute buffer. Node 27 PDC elutes at about 100-200 mM imidazole.

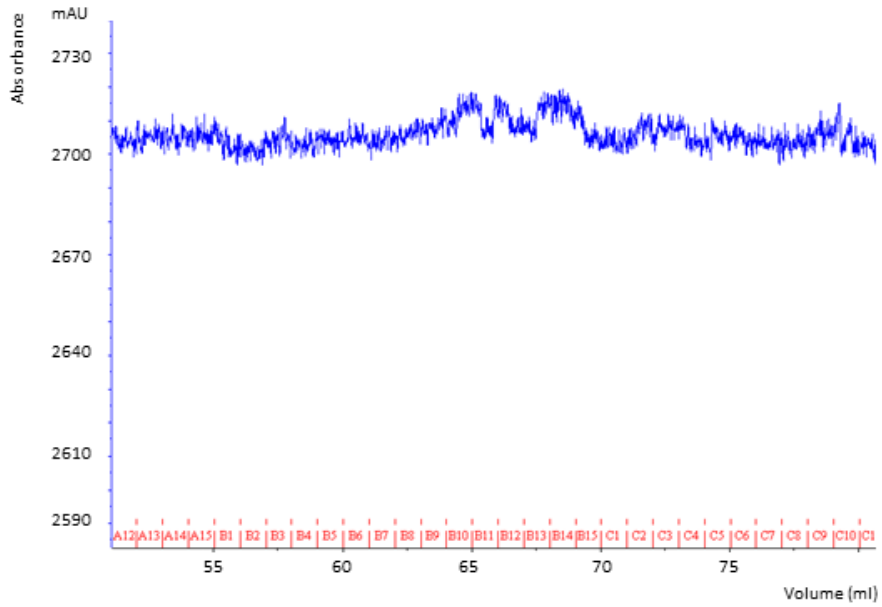


Figure 8.5 Size-exclusion chromatogram for recombinant Node 27 PDC. The dark blue line is the absorbance at 280 nm.

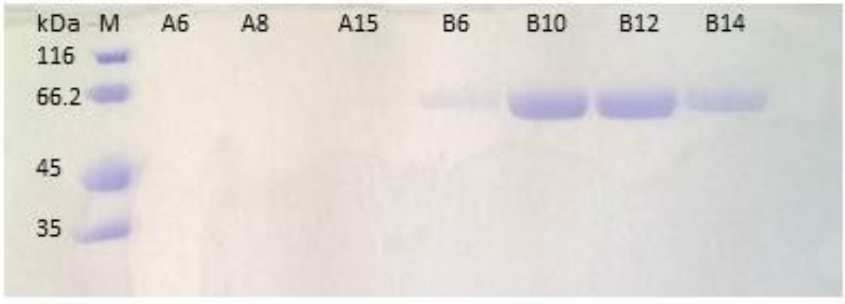
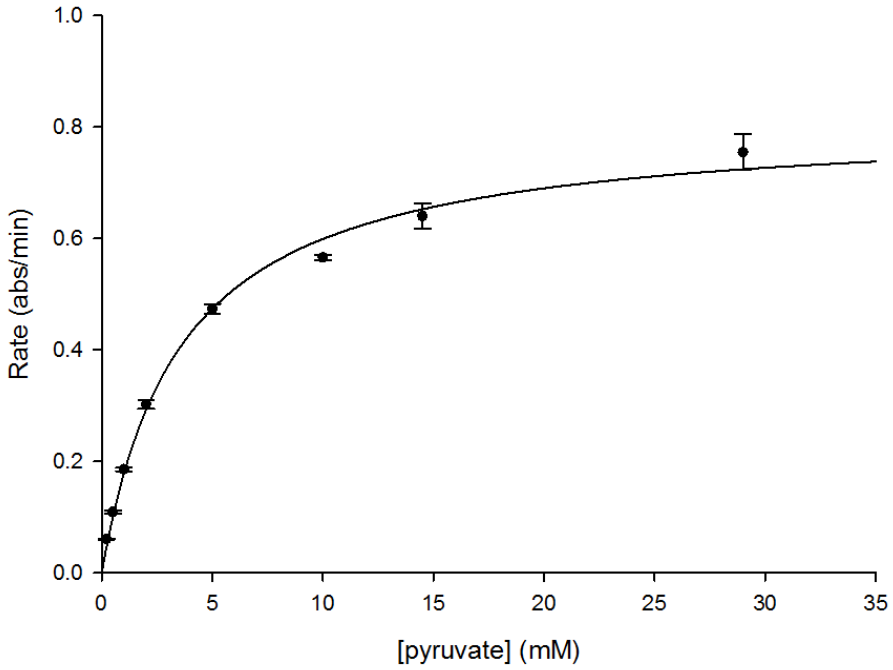


Figure 8.6 SDS-PAGE analysis of Node 27 PDC size-exclusion chromatography fractions. His-tagged Node 27 PDC (monomer size of 62.8 kDa) was purified by nickel-affinity chromatography followed by further purification and buffer exchange through size-exclusion chromatography. Lane M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific).

8.3.4 Enzyme Characterization

Kinetic properties were analysed using the standard coupled assay at 30°C, pH 6.5. The data were analysed using the non-linear fit model from the enzyme kinetics module in SigmaPlot (Figure 8.7), resulting in a V_{max} of 536 ±13 µmol/min/mg and a K_M for pyruvate of 3.6 ±0.3 mM.

(A)



(B)

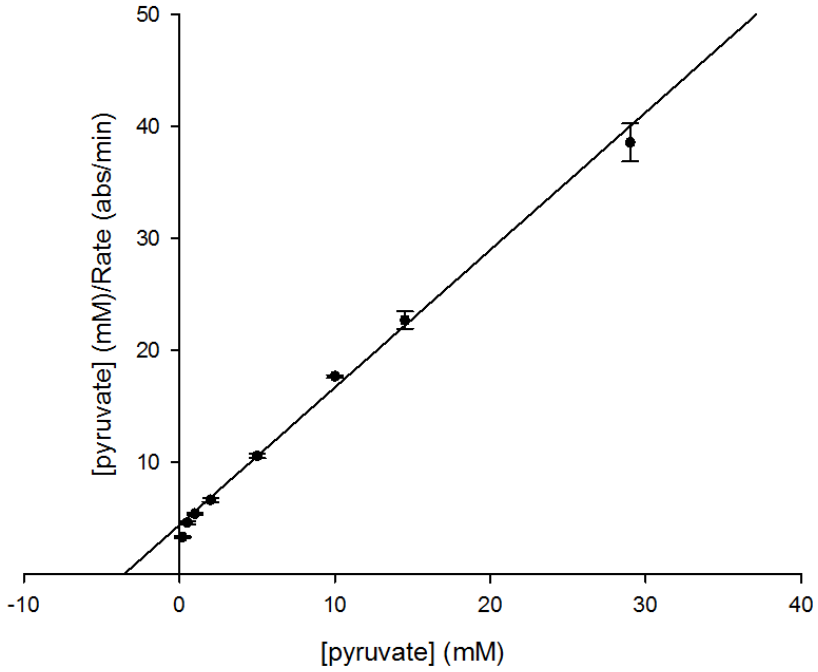


Figure 8.7 Dependence of Node 27 activity on the concentration of pyruvate. The relationship between enzyme activity (abs/min) and pyruvate concentrations is displayed as (A) a Michaelis-Menten plot, and (B) a Hanes-Woolf plot. Error bars are standard errors based on 3 measurements. Kinetic parameters determined are: V_{max} = 0.81 ±0.02 abs/min (equates to 536 ±13 µmol/min/mg) and K_M = 3.6 ±0.3 mM.

The thermal properties of the enzyme were also investigated. The temperature optimum was found to be approximately 50°C (Figure 8.8). Irreversible denaturation analysis was carried out by incubation of the protein at 50 to 70°C for 30 min, and assaying for retained activity at 30°C using the standard coupled assay. This showed that Node 27 PDC retains 90% activity at 50°C, 38% at 55°C, and completely lost activity at 60°C. Using thermal shift assays, the denaturing temperature was determined to be 62°C.

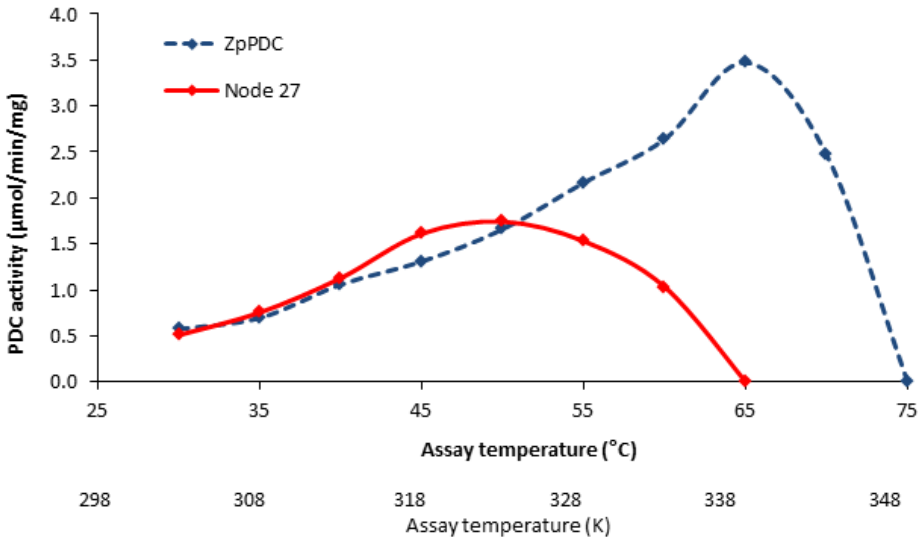


Figure 8.8 Relationship between temperature and PDC activity. The relationship between PDC activity (rate in µmol/min/mg) and assay temperature was determined by monitoring the decrease in pyruvate-dependent absorbance at 320 nm. The wild type *Zymobacter palmae* PDC (ZpPDC) activity is displayed as a dark blue dashed line, and showed a sharp peak at 65°C. Node 27 PDC activity is displayed as a solid red line, peaking at approximately 50°C. Node 27 showed a broader profile with lower maximum rates compared to ZpPDC.

8.4 Discussion

Through ASR a novel and functional PDC was generated, dubbed Node 27. Node 27 shares only 78% amino acid sequence identity with its closest extant homologue, yet it is fully functional. This is not only a previously unreported PDC, but also a novel application for ASR, in the sense that this enzyme family has not previously been tested and that this ASR study only used proteins from mesophilic organisms in the inference input.

The kinetic parameters for Node 27 were determined in standard conditions (at 30°C, in 50mM MES, pH 6.5, with 3 mM TPP, 20 mM $MgSO_4$). Node 27 has a V_{max} of 536 ±13 µmol/min/mg and a K_M for pyruvate of 3.6 ±0.3 mM. The K_M is comparable to data from previously characterized, extant bacterial PDCs. However, the V_{max} is higher than for any reported bacterial PDC (see General Introduction for more detail). The highest previously reported V_{max} was 181 U/mg determined for *Z. mobilis* PDC (at pH 6, 30°C, Bringer-Meyer *et al.* 1986).

Thermal properties of Node 27 PDC were also investigated. The temperature optimum for Node 27 PDC activity was found to be approximately 50°C. Irreversible denaturation experiments showed that this PDC is less thermostable than some of the modern PDCs, including ZpPDC characterized in Chapter 3. Node 27 retained 90% activity at 50°C, only 38% at 55°C, and completely lost activity after incubation at 60°C for 30 min. Thermal shift assays determined the denaturing temperature to be 62°C.

Considering these data, Node 27 PDC is one of the least thermostable and thermoactive bacterial PDCs currently known. Node 27 is more thermostable than the *S. ventriculi* PDC, which retained 95% activity after incubation at 45°C for 30 min, but lost all activity after incubation at 50°C. It is perhaps similar to *G. diazotropicus* PDC which has a temperature optimum of 45-50°C (van Zyl *et al.* 2014a); it is a close second to *G. oxydans*, which has a temperature optimum of 53°C, but is more thermostable retaining 70% activity after incubation at 60°C for 30 min (van Zyl *et al.* 2014b).

Node 27 is 33%, 74% and 66% identical in amino acid sequence to *S. ventriculi*, *G. diazotropicus* and *G. oxydans*, respectively. Similarly, it is 67%, 71% and 69% identical in amino acid sequence to the more thermostable and thermoactive PDCs of *A. pasteurianus*, *Z. mobilis* and *Z. palmae*, respectively.

An investigation into the crystal structure and comparisons to the known bacterial PDCs and other TPP-binding enzymes might shed some light unto the mysteries of PDC thermostability

and thermoactivity determining features. Unfortunately, the crystal structure investigation of Node 27 went beyond the limited resources of this Thesis project.

Furthermore, there are another 20 nodes that arose from the ASR process. Attempts to characterize 2 of them were made (Node 30 and Node 49, Figure 8.2); however, gene synthesis and expression of soluble protein proved to be a major issue with both of them and could not be resolved during the time on this project. Reducing the GC-content through using a different codon guidance in the backtranslation may perhaps be an option to overcome the gene synthesis problem. This was at the time not possible due to the limited capabilities of the production company. Expression of soluble protein may be improved through the use of different expression strains or growth temperatures. Several of these were tested for Node 49, including *E.coli* Rosetta, C41 and C43, between 16-45°C, to no avail. Perhaps, expression of insoluble protein should be optimized, so that the PDC may be extracted from inclusion bodies and resolubilized. Node 27 is 83% and 79% identical in amino acid sequence to Node 30 and Node 49, respectively.

As mentioned above, this ASR study was a novel approach as it used only mesophilic species as the inference input to reconstruct a complex bacterial enzyme. Other ASR studies described in the literature use a broader range of species with a wider growth temperature range.

In a study using plant peroxidases, the ancestor inferred by Ryan *et al.* (2008) and further characterized by Loughran *et al.* (2014) was less thermostable than some of the extant peroxidases, despite the influence of highly thermostable input sequences. However, it was comparatively young (only 110 million years old) and has increased oxidative stability.

Perhaps Node 27 is simply too young. Under close examination of the study conducted by Gaucher *et al.* (2008) and their description of the progressive cooling of the environmental conditions correlating to a progressive decrease in denaturing temperature of the bacterial elongation factor TU between 3.5 and 0.5 billion years ago, Node 27's denaturing temperature of 62°C fits well into their data. Their data suggest that in ancestors from 1,000 to 2,000 million years ago, a denaturing temperature of around 50°C is likely to be observed. However, earlier life (3-4 billion years ago) is likely to have experienced conditions similar to today's hot springs (at 60-80°C) (Gaucher *et al.* 2008).

Unfortunately, the oldest inferred ancestor, Node 49, could not be characterized during this study. Perhaps the ASR would have benefitted from expanding input sequences to thermophilic species with functionally-related TPP-dependent enzymes.

9. General Discussion

Geobacillus thermoglucodsidasius is naturally able to ferment a wide variety of substrates and produces lactate, formate, acetate and ethanol. This makes *G. thermoglucosidasius* a good candidate for the production of bioethanol from lignocellulosic feedstocks, as a second-generation biofuel. The engineered strain TM242 (NCIMB 11955 Δ*ldh*, Δ*pfl*, *pdh*^{up}) produces ethanol as its major fermentation product with yields of up to 92% of the theoretical maximum (at 60°C, on cellobiose) (Cripps *et al.* 2009). However, *G. thermoglucosidasius* TM242 still produces unwanted by-products, in particular acetate. The work presented in this Thesis explored an alternative fermentation pathway in which carbon flow would be channelled away from acetate and towards ethanol production through the introduction of a thermoactive homoethanologenic pathway.

Yeast and a limited range of mesophilic bacteria use a homoethanologenic fermentation pathway, instead of the mixed acid fermentation pathway found in *G. thermoglucosidasius.* This homoethanol pathway employs pyruvate decarboxylase (PDC), in conjunction with alcohol dehydrogenase (ADH), to convert pyruvate to ethanol. Despite extensive screening efforts, no enzyme with a PDC function could be identified in a thermophilic bacterium.

This Thesis aimed at identifying a suitable PDC-ADH pathway to introduce into *G. thermoglucosidasius* and develop a functional producer of ethanol (PET) operon for *in vivo* expression. It was hypothesized that expression of a functional PDC-ADH pathway in *G. thermoglucosidasius* increases ethanol yields.

This involved:

- Exploring potentially suitable bacterial PDCs through ASR, which generated inferred ancestral PDCs, and through comparison of extant mesophilic bacterial PDCs.

- Expanding the knowledge on bacterial PDCs available by characterizing ZpPDC *in vitro*, including the crystal structure, and *in vivo* in *G. thermoglucosidasius.*

- Finding an appropriate ADH partner to complete the pathway and design a PET operon.

- Testing the PET operon under fermentative conditions in *G. thermoglucosidasius*.

TOWARDS EXPLORING SUITABLE BACTERIAL PDCS AND EXPANDING THE KNOWLEDGE ON BACTERIAL PDCS

One alternative to engineering modern bacterial PDCs was explored in Chapter 8, using ancestral sequence reconstruction (ASR) to open up opportunities for bacterial PDCs beyond the known modern, mesophilic enzymes. ASR did indeed yield at least one functional PDC. However, this particular one was not highly thermostable or thermoactive, and thus not suitable for work in *G. thermoglucosidasius*. Nevertheless, that is not to say that none of the ancestral PDCs would be. The limited resources of this project did not allow the use of ASR to its full potential. It would have perhaps been interesting not only to test further predicted PDCs, but also to include investigations into the evolutionary relationship of PDC, POR, ALS, and other thermostable TPP-dependent enzymes.

Chapter 3 described the *in vitro* characterization of the most thermostable bacterial PDC currently known: the *Zymobacter palmae* PDC (ZpPDC). *In vitro* ZpPDC is most active at 65°C with a denaturing temperature of 70°C, when sourced from a recombinant mesophile.

Chapter 4 described the *in vivo* characterization of ZpPDC being expressed in *G. thermoglucosdidasius.* When expressed aerobically in *G. thermoglucosdidasius* TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) PDC activity was detectable in the crude cell extract up to 65°C growth temperature. That is a vast improvement on any previously reported data.

Taylor *et al.* (2008) reported no detectable activity at 52°C when expressing the ZpPDC in DL44 (DL33 Δ*ldh*). With an improved expression construct using the *phe*B RBS rather than the *ldh*A RBS, PDC activity was detectable up to 60°C growth temperature in DL44, almost to levels they had reported for 48°C (81 nmol/min/mg total protein, while Taylor *et al.* reported 89 nmol/min/ mg total protein).

A change in expression strain background to TM236 (NCIMB 11955 Δ*ldh*, Δ*pfl*) further improved detectable PDC activity in cell extracts from cultures grown up to 65°C. However, there was a noticeable decrease in PDC activity measured across the growth temperature range (50°C: 1.06 ±0.33 µmol/min/mg total protein, 60°C: 0.52 ±0.34 µmol/min/mg total protein, 65°C: 0.16 ±0.03 µmol/min/mg total protein).

Perhaps, the disparity between the high *in vitro* thermostability and limited *in vivo* expression of ZpPDC was due to impaired translation and co-translational folding around TPP in the thermophilic host.

Codon harmonization has been known to improve recombinant expression of proteins. Unlike codon optimization, which aims to overexpress the recombinant protein, codon harmonization aims to create the pattern of translation as found in the native host, and thus increases yields of correctly folded and functionally active protein in the recombinant expression host. Chapter 6 explored this approach for the ZpPDC using the newly acquired data from the first draft genome of *Z. palmae* at the time (described in Chapter 5).

Codon harmonization resulted in 308 "silent" substitutions, which changed 18% of the nucleotide sequence. The amino acid sequence remained unchanged. *In vivo* characterization experiments showed that codon harmonization did improve PDC activity in aerobic *G. thermoglucosidasius* TM236 cultures by 22% in cells grown at 50°C (to 1.34 ±0.22 µmol/min/mg total protein), 20% in cells grown at 60°C (to 0.65 ±0.27 µmol/min/mg total protein), and 42% in cells grown at 65°C (to 0.28 ±0.07 µmol/min/mg total protein) in comparison to TM236 expressing the wt Zp*pdc*. Similar to the wt ZpPDC, detectable activity noticeably decreased with increasing growth temperature. Perhaps, one limiting factor is co-translational binding and folding around the TPP. It is not clear whether TPP binding may be affected by the increased growth temperature as no studies have been done to investigate this.

Furthermore, RT-qPCR analysis of the gene expression levels also showed a decrease with increasing growth temperature. This may be due to expression from the *ldh* promoter decreasing as the cultures were increasingly oxygen limited at higher growth temperatures.

Nonetheless, these results showed a major improvement for high temperature recombinant expression of a bacterial PDC, and surpassed any previously reported attempts. Van Zyl *et al.* (2014b) reported a specific activity of 0.22 µmol/min/mg at 45°C with their codon harmonized *Gluconobacter oxydans pdc*, and detected no PDC activity at growth temperatures above that. Any activity at the *G. thermoglucosidasius* optimum growth temperature of 60°C outside the work of this Thesis has yet to be reported.

Additionally, towards expanding the knowledge on bacterial PDCs, the crystal structure of the ZpPDC was solved to 2.15 Å and was presented in Chapter 3. Structural comparison to other known bacterial PDCs suggested that the variations in thermostability and thermoactivity displayed by these PDCs may be correlated to increased oligomeric interfaces and salt bridges with increased thermal adaptation. Further analysis and comparison to thermophilic TPP-

containing enzymes may provide invaluable information on TPP-binding site design and for rational design approaches in future PDC engineering.

TOWARDS FINDING AN APPROPRIATE ADH PARTNER TO COMPLETE THE PATHWAY

Genome sequencing of *Z. palmae* did allow for the identification of the previously unreported ZpADHI and ADHII. Unfortunately, these were not suitably thermostable or thermoactive for use in *G. thermoglucosidasius.* However, work previously carried out by Dr. Luke Williams (University of Bath) identified a potentially suitable ADH in *G. thermoglucosidasius*, ADH6. The functionality of ZpPDC paired with ADH6 was analysed in Chapter 7, *in vitro* and *in vivo*.

Using both enzymes recombinantly expressed and purified from *E.coli* in the standard coupled assay worked efficiently across a wide temperature range up to 70°C. Furthermore, analysis of the assay products by HPLC showed that the ZpPDC and GtADH6 activities were stoichiometrically tightly coupled at assay temperatures up to 65°C (70°C was not tested here), suggesting that despite the volatility of acetaldehyde it was not escaping from the reaction in significant amounts.

These data were promising and the codon harmonized Zp*pdc* and GtADH6 were paired as a PET (producer of ethanol) operon on an expression construct for *in vivo* testing in *G. thermoglucosidasius* TM236*.* Aerobic cultures grown at 50, 60 and 65°C showed reduced detectable PDC activity in ufractionated cell extracts compared to data from ZpPDC 2.0 expressed on its own (50°C: 0.24 ±0.007 µmol/min/mg total protein, 60°C: 0.11 ±0.08 µmol/min/mg total protein, 65°C: 0.12 ±0.06 µmol/min/mg total protein). However, ADH activity was upregulated. Perhaps the overexpression of a second protein inflicted an increased metabolic burden that led to the decreased PDC expression and hence decreased detectable PDC activity.

Gene expression was analysed by RT-qPCR and showed that both ZpPDC 2.0 and GtADH6 genes were induced in the PET operon expressing strain, but that levels decreased with an increase in growth temperature. This decrease may be explained by expression from the *ldh* promoter decreasing as oxygen availability became limited at higher growth temperatures.

TESTING THE PET OPERON UNDER FERMENTATIVE CONDITIONS IN *G. THERMOGLUCOSIDASIUS*

Furthermore, Chapter 7 described the testing of the PET operon in *G. thermoglucosidasius* TM236 under fermentative conditions using tube fermentations. At a growth temperature of 50°C, TM236 pUCG18 and TM236 pUCGT PET showed a very similar fermentation product profile with high ethanol production (around 4 g/L) and very few by-products. At a 60°C growth temperature, TM236 pUCG18 also accumulated some pyruvate, while TM236 expressing the PET operon produced the highest amount of ethanol, on average 4.3 g/L, with very few by-products. At 65°C all strains showed some accumulation of formate, lactate and acetate.

TM236 expressing wt ZpPDC or ZpPDC 2.0 on their own accumulated some pyruvate and lactate, and produced less ethanol. The ethanol yield was often less than the TM236 pUCG18 background. This may be due to metabolic strain inflicted by misfolded PDC, or perhaps the accumulation of acetaldehyde to toxic levels in the cell impairing pyruvate metabolism.

Pairing the codon harmonized Zp*pdc* with GtADH6 generated a PET operon functional up to 65°C with ethanol yields of 87% of the theoretical maximum on glucose. This strain consistently showed the highest and most stable ethanol yield across the temperature range. At 60 and 65°C, these ethanol yields were significantly higher in the PET operon expressing strain than in the backgound TM236 pUCG18, which can naturally produce ethanol through the PDH-ADHE pathway (t-Test, $p_{60°C} = 0.01$, $p_{65°C} = 0.03$).

These preliminary data also indicated a 5% increase in ethanol yields from 60°C fermentations on glucose compared to data from the current production strain TM242 (Cripps *et al.* 2009). Further optimization will be required to produce a potential production strain. However, this increase in yield is very promising and could make a significant difference for industrial-scale production.

In summary, the work presented in this Thesis added to our knowledge of *Z. palmae* by the addition of a draft genome sequence, and of ZpPDC with the addition of further *in vitro* data, a crystal structure and characterization *in vivo* in *G. thermoglucosidasius.* It further added to our understanding of modern PDCs by the exploration of evolutionary relationships through ASR. Finally, the work conducted in this Thesis fulfilled the over-arching aim of identifying a suitable PDC-ADH pathway to introduce into *G. thermoglucosidasius* and developing a functional PET operon for *in vivo* expression. Expression data surpassed any previous reports on cloning PDCs

into *G. thermoglucosidasius*, and pairing the codon harmonized Zp*pdc* with GtADH6 produced a functional PET operon with ethanol yields 5% higher than the current production strain, at temperatures of up to 15°C higher than previously reported for any bacterial PDC expressed in a thermophilic host. Thus, the hypothesis that expressing a functional PDC-ADH pathway in *G. thermoglucosidasius* increases ethanol yields was validated.

REFERENCES

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

Agrawal M, Wang Y, Chen RR. 2012. Engineering efficient xylose metabolism into an acetic acid-tolerant *Zymomonas mobilis* strain by introducing adaptation-induced mutations. *Biotechnology Letters* 34: 1825-1832.

Akanuma S, Nakajima Y, Yokobori S, Kimura M, Nemoto N, Mase T, Miyazono K, Tanokura M, Yamagishi A. 2013. Experimental evidence for the thermophilicity of ancestral life. *PNAS* 110: 11067-11072.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2008. Molecular Biology of the Cell. 5th ed. New York, USA: Garland Science.

Angov E. 2011. Codon usage: nature's roadmap to expression and folding of proteins. *Biotechnology Journal* 6(6): 650-659.

Angov E, Hillier CJ, Kincaid RL, a Lyon J. 2008. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PloS ONE* 3(5): e2189.

Antoni D, Zverlov VV, Schwarz WH. 2007. Biofuels from microbes. *Applied Microbiology and Biotechnology* 77:23-35.

Arenas M & Posada D. 2010. The effect of recombination on the reconstruction of ancestral sequences. *Genetics* 184: 1133-1139.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9: 75.

Barbosa M de FS & Ingram LO. 1994. Expression of the *Zymomonas mobilis* alcohol dehydrogenase II (*adhB*) and pyruvate decarboxylase (*pdc*) genes in *Bacillus*. *Current Opinion in Microbiology* 28: 279-282.

Barnard D, Casanueva A, Tuffin M, Cowan D. 2010. Extremophiles in biofuel synthesis. *Environmental Technology* 31(8-9): 871-888.

Bartosiak-Jentys J, Eley K, Leak DJ. 2012. Application of *phe*B as a reporter gene for *Geobacillus* spp., enabling qualitative colony screening and quantitative analysis of promoter strength. *Applied and Environmental Microbiology* 78(16): 5945-5947.

Basen M, Schut GJ, Nguyen DM, Lipscomb GL, Benn RA, Prybol CJ, Vaccaro BJ, Poole FL, Kelly RM, Adams MWW. 2014. Single gene insertion drives bioalcohol production by a thermophilic archaeon. *PNAS* 111(49): 17618-17623.

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 Evolutionary Biology* 4: 44.

Battye TGG, Kontogiannis L, Johnson O, Powell HR, Leslie AGW. 2011. iMOSFLM: A new graphical interface for diffraction image processing with MOSFLM. *Acta Crystallographica D* 67: 271-281.

Bettiga M, Bengtsson O, Hahn-Hagerdal B, Gorwa-Grauslund MF. 2009. Arabinose and xylose fermentation by recombinant *Saccharomyces cerevisiae* expressing a fungal pentose utilization pathway. *Microbial Cell Factories* 8: 40.

Bioconversion Technologies Limited. 2007. Enhancement of microbial ethanol production. US20090226992 A1. Filed 26. March 2007.

Bragg WH & Bragg WL. 1913. The refrexion of x-rays by crystals. Proceedings of the Royal Society of London: A 88(605): 428-438.

Brat D, Boles E, Wiedemann B. 2009. Functional expression of a bacterial xylose isomerase in *Saccharomyces cerevisiae*. *Applied and Environmental Microbiology* 75(8): 2304-2311.

Brenden CI & Tooze J. 1998. Introduction to Protein Structure. 2nd ed. New York, USA: Garland Publishing Inc.

Bringer-Meyer S, Schimz KL, Sahm H. 1986. Pyruvate decarboxylase from *Zymomonas mobilis*. Isolation and partial characterization. *Archives Microbiology* 146: 105-110.

Buddrus L, Andrews ES, Leak DJ, Danson MJ, Arcus VL, Crennell SJ. 2016. Crystal structure of pyruvate decarboxylase from *Zymobacter palmae*. *Acta Crystallographica F* 72(9): 700-706. (see Appendix IV)

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. *Nature* 325: 728-730.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540-552.

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica D* 66: 12-21.

Cheng M, Yoshiyasu H, Okano K, Ohtake H, Honda K. 2016. Redirection of the reaction specificity of a thermophilic acetolactate synthase toward acetaldehyde formation. *PLoS ONE* 11(1): e0146146.

Chessher A, Breitling R, Takano E. 2015. Bacterial microcompartments: biomaterials for synthetic biology-based compartmentalization strategies. *ACS Biomaterials Science & Engineering* 1: 345-351.

Chung D, Cha M, Guss AM, Westpheling J. 2014. Direct conversion of plant biomass to ethanol by engineered *Caldicellulosiruptor bescii*. *PNAS* 111: 8931-8936.

Chung CT, Niemela SL, Miller RH. 1989. One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution. *PNAS* 86(7): 2172-2175.

Claassen PAM, van Lier JB, Contreras AML, van Niel EWJ, Sijtsma L, Stams AJM, de Vries SS, Weusthuis RA. 1999. Utilisation of biomass for the supply of energy carriers. *Applied Microbiology and Biotechnology* 52(6): 741-755.

Cole MF & Gaucher EA. 2011. Utilizing natural diversity to evolve protein function: applications towards thermostability. *Current Opinion in Chemical Biology* 2011, 15: 399-406.

Cooper RA. 1978. Intermediary metabolism of monosaccharides by bacteria. University Park Press: Baltimore, MD, USA.

Couñago R, Chen S, Shamoo Y. 2006. *In vivo* molecular evolution reveals biophysical origins of organismal fitness. *Molecular Cell* 22: 441-449.

Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, Boakes S, Martin S, Atkinson T. 2009. Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. *Metabolic Engineering* 11(6): 398-408.

Dawes EA, Ribbons DW, Large PJ. 1966. The route of ethanol formation in *Zymomonas mobilis*. *Biochemistry Journal* 98(3): 795-803.

de la Haba RR, Arahal DR, Márquez MC, Ventosa A. 2010. Phylogenetic relationships within the family *Halomonadaceae* based on comparative 23S and 16S rRNA gene sequence analysis. *International Journal of Systematic and Evolutionary Microbiology* 60: 737-748.

Di Giulio M. 2003. The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence. *Journal of Theoretical Biology* 221: 425-436.

Dobritzsch D, König S, Schneider G, Lu G. 1998. High resolution crystal structure of pyruvate decarboxylase from *Zymomonas mobilis* – implications for substrate activation in pyruvate decarboxylases. *Journal of Biological Chemistry* 273: 20196-20204.

Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of Coot. *Acta Crystallographica D* 66: 486-501.

Entner N & Doudoroff M. 1952. Glucose and gluconic acid oxidation of *Pseudomonas saccharophila*. *Journal of Biological Chemistry* 196(2): 853-862.

Eram MS & Ma K. 2013. Decarboxylation of pyruvate to acetaldehyde for ethanol production by hyperthermophiles. *Biomolecules* 3: 578-596.

Eram MS, Wong A, Oduaran E, Ma K. 2015. Molecular and biochemical characterization of bifunctional pyruvate decarboxylases and pyruvate ferredoxin oxidoreductases from *Thermotoga maritima* and *Thermotoga hypogea*. *Journal of Biochemistry* 158(6): 469-466.

Evans PR & Murshudov GN. 2013. How good are my data and what is the resolution? *Acta Crystallographica D* 69: 1204-1214.

Extance JP. 2012. Bioethanol production: characterization of a bifunctional alcohol dehydrogenase from *Geobacillus thermoglucosidasius*. PhD Thesis, University of Bath, Bath, UK.

Extance JP, Crennell SJ, Eley K, Cripps R, Hough DW, Danson MJ. 2013. Structure of a bifunctional alcohol dehydrogenase involved in bioethanol generation in *Geobacillus thermoglucosidasius*. *Acta Crystallographica Section D* 69: 2104-2115.

Fong JCN, Svenson CJ, Nakasugi K, Leong CTC, Bowman JP, Chen B, Glenn DR, Neilan BA, Rogers PL. 2006. Isolation and characterization of two novel ethanol-tolerant facultative-anaerobic thermophilic bacteria strains from waste compost. *Extremophiles* 10(5): 363-372.

Frank RA, Titman CM, Pratap JV, Luisi BF & Perham RN. 2004. A molecular switch and proton wire synchronize the active sites in thiamine enzymes. *Science* 306: 872-876.

Fries M, Chauhan HJ, Domingo GJ, Jung HI, Perham RN. 2003. Site-directed mutagenesis of a loop at the active site of E1 ($\alpha_2\beta_2$) of the pyruvate dehydrogenase complex. *European Journal of Biochemistry* 270: 861-870.

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451(7): 704-708.

Georgieva TI, Mikkelsen MJ, Ahring BK. 2007. High ethanol tolerance of the thermophilic anaerobic ethanol producer *Thermoanaerobacter* BG1L1. *Central European Journal of Biology* 2(3): 364-377.

Gocke D, Graf T, Brosi H, Frindi-Wosch I, Walter L, Müller M, Pohl M. 2009. Comparative characterization of thiamin diphosphate-dependent decarboxylases. *Journal of Molecular Catalysis B: Enzymatic* 61: 30-35.

Grabek-Lejko D, Kurylenko OO, Sibirny VA, Ubiyvovk VM, Penninckx M, Sibirny AA. 2011. Alcoholic fermentation by wild-type *Hansenula polymorpha* and *Saccharomyces cerevisiae* versus recombinant strains with an elevated level of intracellular glutathione. *Journal of Industrial Microbiology and Biotechnology* 38: 1853-1859.

Gromiha MM, Oobatake M, Sarai A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemistry* 82: 51-67.

Gronenberg LS, Marcheschi RJ, Liao JC. 2013. Next generation biofuel engineering in prokaryotes. *Current Opinion in Chemical Biology* 17: 462-471.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59(3): 307-321.

Harms MJ & Thornton JW. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology* 20: 360-366.

Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW, Marqusee S. 2014. Thermodynamic system drift in protein evolution. *PLoS Biology* 12: e1001994.

Hernández Gómez MC. 2011. Characterisation of hybrid pyruvate decarboxylases for application in thermophiles. MRes Report, Imperial College, London, UK.

Hespell RB, Wyckoff H, Dien BS, Bothast RJ. 1996. Stabilization of *pet* operon plasmids and ethanol production in *Escherichia coli* strains lacking lactate dehydrogenase and pyruvate formate-lyase activities. *Applied and Environmental Microbiology* 62(12): 4594-4597.

Hills CA. 2014. Acetate metabolism in *Geobacillus thermoglucosidasius* and strain engineering for enhanced bioethanol production. PhD Thesis, University of Bath, Bath, UK.

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: 825-835.

Hood EE. 2016. Plant-based biofuels. *F1000Research* 5 (F1000 Faculty Rev): 185.

Horn SJ, Aasen AM, Østgaard K. 2000. Production of ethanol from mannitol by *Zymobacter palmae*. *Journal of Industrial Microbiology and Biotechnology* 24: 51-57.

Hosaka M, Komuro Y, Yamanaka M, Yamaura T, Nakata H, Sakai T. 1998. Fermentability of sake moromi prepared with shochu, wine, brewer's, alcohol or sake strains of *Saccharomyces cerevisiae*. *Journal of the Brewing Society Japan* 93(10): 833-840.

Hussein AH, Lisowska BK, Leak DJ. 2015. The genus *Geobacillus* and their biotechnological potential. *Advances in Applied Microbiology* 92: 1-48.

IFP Energies Nouvelles. http://www.ifpenergiesnouvelles.com/Research-themes/New-energies/Producing-fuels-from-biomass/Biocatalysts-one-of-IFPEN-s-expertise-field-Questions-to-Frederic-Monot-Head-of-the-Biotechnology-Department-at-IFPEN, Accessed 20. July 2016.

Ingram LO & Conway T. 1988. Expression of different levels of ethanologenic enzymes from *Zymomonas mobilis* in recombinant strains of *Escherichia coli*. *Applied and Environmental Microbiology* 54(2): 397-404.

Ingram LO, Conway T, Clark DP, Sewell GW, Preston JF. 1987. Genetic engineering of ethanol production in *Escherichia coli. Applied and Environmental Microbiology* 53(10): 2420-2425.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12): 1647-1649.

Kerfeld CA & Erbilgin O. 2015. Bacterial microcompartments and the modular construction of microbial metabolism. *Trends in Microbiology* 23(1): 22-34.

Kern D, Kern G, Neef H, Tittman K, Killenberg-Jabs M, Wikner C, Schneider G, Hübner G. 1997. How thiamine diphosphate is activated in enzymes. *Science* 275: 67-70.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgings DG. 2007. ClustalW and ClustalX version 2. *Bioinformatics* 23: 2947-2948.

Lawrence AD, Frank S, Newnham S, Lee MJ, Brown IR, Xue WF, Rowe ML, Mulvihill DP, Prentice MB, Howard MJ, Warren MJ. 2014. Solution structure of bacterial microcompartment targeting peptide and its application in the construction of an ethanol bioreactor. *ACS Synthetic Biology* 3: 454-465.

Liao HT & Kanikula AM. 1990. Increased efficiency of transformation of *Bacillus stearothermophilus* by a plasmid carrying a thermostable kanamycin resistance marker. *Current Microbiology* 21: 301-306.

Liao H, McKenze T, Hageman R. 1986. Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *PNAS* 83: 576-580.

Liao JC, Mi L, Pontrelli S, Luo S. 2016. Fuelling the future: microbial engineering for the production of sustainable biofuels. *Nature Reviews Microbiology* 14: 288-304.

Liu S, Dien BS, Cotta MA. 2005. Functional expression of bacterial *Zymobacter palmae* pyruvate decarboxylase gene in *Lactococcus lactis*. *Current Microbiology* 50: 324-28.

Long F, Vagin A, Young P, Murshudov GN. 2008. *BALBES*: a molecular replacement pipeline. *Acta Crystallographica Section D* 64: 125-132.

Loughran NB, O'Connell MJ, O'Connor B, Ó'Fágáin C. 2014. Stability properties of an ancient plant peroxidase. *Biochimie* 104: 156-159.

Lowe TM & Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25(5): 955-964.

Lowe SE & Zeikus JG. 1992. Purification and characterization of pyruvate decarboxylase from *Sarcina ventriculi. Journal of General Microbiology* 138: 803-807.

Lynd LR, Laser MS, Bransby D, Dale BE, Davison B, Hamilton R, Himmel M, Keller M, McMillar JD, Sheehan J, Wyman CE. 2008. How biotech can transform biofuels. *Nature Biotechnology* 26(2): 169-172.

Ma K & Adams MWW. 1999. An unusual oxygen-sensitive, iron- and zinc-containing alcohol dehydrogenase from the hyperthermophilic archaeon *Pyrococcus furiosus. Journal of Bacteriology* 181(4): 1163-1170.

Ma K, Hutchins A, Sung SJS, Adams MWW. 1997. Pyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon, *Pyrococcus furiosus,* functions as a CoA-dependent pyruvate decarboxylase. *PNAS* 94: 9608-9613.

Mauch TJ, Donohue TM, Zetterman KR, Sorrel MF, Tuma DJ. 1986. Covalent binding of acetaldehyde selectively inhibits the catalytic activity of lysine-dependent enzymes. *Hepatology* 6(2):263-269.

Meyer D, Neumann P, Parthier C, Friedemann R, Nemeria N, Jordan F, Tittmann K. 2010. Double duty for a conserved glutamate in pyruvate decarboxylase: evidence of the participation in stereoelectronically controlled decarboxylation and in protonation of the nascent carbanion/enamine intermediate. *Biochemistry* 49: 8198-8212.

Morant N. 2014. Novel thermostable polymerases for isothermal DNA amplification. PhD Thesis, University of Bath, Bath, UK.

Mosier N, Wyman C, Dale B, Elander R, Lee YY, Holtzapple M, Ladisch M. 2005. Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresource Technology* 96(6): 673-686.

Nagarajan N, Cook C, Di Bonaventura MP, Ge H, Richards A, Bishop-Lilly KA, DeSalle R, Read TD, Pop M. 2010. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics* 11: 242.

Nazina TN, Tourova TP, Poltaraus AB, Novikova EV, Grigoryan AA, Ivanova AE, Lysenko AM, Petrunyaka VV, Osipov GA, Belyaev SS, Ivanov MV. 2001. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. *International Journal of Systematic and Evolutionary Microbiology* 51: 433-446.

Okamoto T, Taguchi H, Nakamura K, Ikenaga H, Kuraishi H, Yamasato K. 1993. *Zymobacter palmae* gen. nov., sp. nov., a new ethanol-fermenting peritrichous bacterium isolated from palm sap. *Archives of Microbiology* 160**:** 333-337.

Olson DG, Sparling R, Lynd LR. 2015. Ethanol production by engineered thermophiles. *Current Opinion in Biotechnology* 33: 130-141.

Pace NR. 1991. Origin of life - facing up to the physical setting. *Cell* 65: 531-533.

Pauling L & Zuckerkandl E. 1963. Chemical paleogenetics, molecular restoration studies of extinct forms of life. *Acta Chemica Scandinavica* 17: 9-16.

Pei J, Erixon KM, Luisi BF, Leeper, FJ. 2010. Structural insights into the prereaction state of pyruvate decarboxylase from *Zymomonas mobilis*. *Biochemistry* 49: 1727-1736.

Pei J, Kim BH, Grishin NV. 2008. PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Research* 36(7): 2295-2300.

Piuelle L, Margo V, Hatchikian EC. 1997. Isolation and analysis of the gene encoding the pyruvate ferredoxin oxidoreductase of *Desulfovibrio africanus*, production of the recombinant enzyme in *Escherichia coli* and effect of carboxy-terminal deletions on its stability. *Journal of Bacteriology* 179(18): 5684-5692.

Pohl M, Grötzinger J, Wollmer A, Kula MR. 1994. Reversible dissociation and unfolding of pyruvate decarboxylase from *Zymomonas mobilis*. *European Journal of Biochemistry* 224: 651-661.

Pohl M, Mesch K, Rodenbrock A, Kula MR. 1995. Stability investigations on the pyruvate decarboxylase from *Zymomonas mobilis*. *Biotechnology and Applied Biochemistry* 22: 95-105.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* 25: 1253-1256.

Potterton E, Briggs P, Turkenburg M, Dodson E. 2003. A graphical user interface to the CCP4 program suite. *Acta Crystallographica Section D* 59: 1131-1137.

Raj KC, Talarico LA, Ingram LO, Maupin-furlow JA. 2002. Cloning and characterization of the *Zymobacter palmae* pyruvate decarboxylase gene (*pdc*) and comparison to bacterial homologues. *Applied and Environmental Microbiology* 68(6): 2869-2876.

RFA (Renewable Fuel Association). 2016. 2016 Ethanol industry outlook [online]. Retrieved 20. July 2016, from http://www.ethanolrfa.org/resources/publications/outlook/

RFA (Renewable Fuel Association). 2016. 2016 Industry statistics [online]. Retrieved 21. July 2016, from http://www.ethanolrfa.org/resources/industry/statistics/

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: 2899-2902.

Ritter SK. 2008. Lignocellulose: a complex biomaterial. *Chemical and Engineering News* 86(49): 15.

Ryan BJ, O'Connell MJ, Ó'Fágáin C. 2008. Consensus mutagenesis reveals that non-helical regions influence thermal stability of horseradish peroxidase. *Biochimie* 90: 1389-1396.

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular Cloning: A Laboratory Manual. 2nd ed. eds. J Sambrook and D W Russell. New York, USA: Cold Spring Habor Laboratory Press.

Sanchez OJ & Cardona CA. 2008. Trends in biotechnological production of fuel ethanol from different feedstocks. *Bioresource Technology* 99(13): 5270-5295.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19(2): 301-302.

Shaw AJ, Hogsett DA, Lynd LR. 2009. Identification of the [FeFe]-hydrogenase responsible for hydrogen generation in *Thermoanaerobacterium saccharolyticum* and demonstration of increased ethanol yield via hydrogenase knockout. *Journal of Bacteriology* 191: 6457-6464.

Shaw AJ, Podkaminer KK, Desai SG, Bardsley JS, Rogers SR, Thorne PG, Hogsett DA, Lynd LR. 2008. Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. *PNAS* 105(37): 13769-13774.

Sherwood D & Cooper J. 2011. Crystals, X-rays and Proteins. Comprehensive Protein Crystallography. Oxford, UK: Oxford University Press.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Weizhong L. 2011. Fast scalable generation of high-quality multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539.

Solem C, Dehli T, Jensen PR. 2013. Rewiring *Lactococcus lactis* for ethanol production. *Applied and Environmental Microbiology* 79(8): 2512-2518.

Sommer P, Georgieva T, Ahring BK. 2004. Potential for using thermophilic anaerobic bacteria for bioethanol production from hemicellulose. *Biochemical Society Transactions* 32: 283-289.

Spencer PS & Barral JS. 2012. Genetic code redundancy and its influence on the encoded polypeptides. *Computational and Structural Biotechnology Journal* 1(1): 1-8.

Spencer PS, Siller E, Anderson JF, Barral JS. 2012. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology* 422(3): 328-335.

Sterner R & Liebl W. 2001. Thermophilic adaptation of proteins. *Critical Reviews in Biochemistry and Molecular Biology* 36: 39-106.

Stetter KO. 1996. Hyperthermophiles in the history of life. *Ciba Foundation Symposium* 202: 1-10.

Stetter KO. 2006. History of discovery of the first hyperthermophiles. *Extremophiles* 10(3): 357-362.

Suzuki H, Kobayashi J, Wada K, Furukawa M, Doi K. 2015. Thermoadaptation-directed enzyme evolution in an error-prone thermophile derived from *Geobacillus kaustophilus* HTA426. *Applied and Environmental Microbiology* 81: 149-158.

Swings J & De Ley J. 1977. The biology of *Zymomonas*. *Bacteriological Reviews* 41(1): 1-46.

Talarico LA, Ingram LO, Maupin-Furlow JA. 2001. Production of the Gram-positive *Sarcina ventriculi* pyruvate decarboxylase in *Escherichia coli*. *Microbiology* 147: 2425-2435.

Taylor MP. 2008. Metabolic engineering of *Geobacillus* species for enhanced ethanol production. PhD Thesis, Imperial College, London, UK.

Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA. 2009. Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. *Trends in Biotechnology* 27(7): 398-405.

Taylor MP, Esteban CD, Leak DJ. 2008. Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. *Plasmid* 60(1): 45-52.

Thompson AH, Studholme DJ, Green EM, Leak DJ. 2008. Heterologous expression of pyruvate decarboxylase in *Geobacillus thermoglucosidasius*. *Biotechnology Letters* 30: 1359-1365.

Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews Genetics* 5: 366-375.

Tittman K, Golbik R, Uhleman K, Khailova L, Schneider G, Patel M, Jordan F, Chipman DM, Duggleby RG, Hübner G. 2003. NMR analysis of covalent intermediates in thiamin diphosphate enzymes. *Biochemistry* 42: 7885-7891.

Vagin AA, Steiner RS, Lebedev AA, Potterton L, McNicholas S, Long F, Murshudov GN. 2004. *REFMAC*5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallographica Section D* 60: 2284-2295.

van Zyl LJ, Schubert WD, Tuffin MI, Cowan DA. 2014a. Structure and functional characterization of pyruvate decarboxylase from *Gluconacetobacter diazotrophicus*. *BMC Structural Biology* 14: 21.

van Zyl LJ, Taylor MP, Eley K, Tuffin M, Cowan DA. 2014b. Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host *Geobacillus thermoglucosidasius*. *Applied Microbiology and Biotechnology* 98(3): 1247-1259.

Vasan PT, Piriya PS, Prabhu DI, Vennison SJ. 2011. Cellulosic ethanol production by *Zymomonas mobilis* harboring an endoglucanase gene from *Enterobacter cloacae*. *Bioresource Technology* 102: 2585-2589.

Waite CJ. 2010. Engineering a novel pyruvate decarboxylase for bio-ethanol production in *Geobacillus* spp., MRes Report, Imperial College, London, UK.

Wechsler C, Meyer D, Loschonsky S, Funk LM, Neumann P, Ficner R, Brodhun F, Muller M, Tittman K. 2015. Tuning and switching enantioselectivity of asymmetric carboligation in an enzyme through mutational analysis of a single hot spot. *Chembiochem* 16: 2580-2584.

Wheeler LC, Lim SA, Marqusee S, Harms MJ. 2016. The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology* 38: 37-43.

Williams LP. 2015. Alcohol dehydrogenases from the thermophile *Geobacillus thermoglucosidasius*. PhD Thesis, University of Bath, Bath, UK.

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D* 67: 235-242.

Xia X & Xie Z. 2001 DAMBE: software package for data analysis in molecular biology and evolution. *Journal of Heredity* 92(4): 371-373.

Yamaoka C, Kurita O, Kubo T. 2014. Improved ethanol tolerance of *Saccharomyces cerevisiae* in mixed cultures with *Klyveromyces lactis* on high-sugar fermentation. *Microbial Research* 169: 907-914.

Yanase H, Miyawaki H, Sakurai M, Kawakami A, Matsumoto M, Haga K, Kojima M, Okamoto K. 2012. Ethanol production from wood hydrolysate using genetically engineered *Zymomonas mobilis*. *Applied Microbiology and Biotechnology* 94: 1667-1678.

Yanase H, Sato D, Yamamotos K, Matsuda S, Yamamoto S, Okamoto K. 2007. Genetic engineering of *Zymobacter palmae* for production of ethanol from xylose. *Applied and Environmental Microbiology* 73: 2592-2599.

Yanase H, Yamamoto K, Sato D, Okamoto K, Yanase H. 2005. Ethanol production from cellobiose by *Zymobacter palmae* carrying the *Ruminococcus albus* β-glucosidase gene. *Journal of Biotechnology* 118: 35-43.

Yang Z. 2007. PAML: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586-1591.

Yano JK & Poulos TL. 2003. New understandings of thermostable and peizostable enzymes. *Current Opinion in Biotechnology* 14: 360-365.

Yao S & Mikkelsen MJ. 2010. Metabolic engineering to improve ethanol production in *Thermoanaerobacter mathranii*. *Applied Microbiology and Biotechnology* 88: 199-208.

Zaldivar J, Nielsen J, Olsson L. 2001. Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. *Applied Microbiology and Biotechnology* 56(1-2): 17-34.

Zhang H, Gao S, Lercher MJ, Hu S, Chen WH. 2012. EvolView: an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acid Research* 40: W569-W572.

Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural and Molecular Biology* 16(3): 274-280.

Zhou S & Ingram LO. 1999. Engineering endoglucanase-secreting strains of ethanologenic *Klebsiella oxytoca* P2. *Journal of Industrial Microbioly and Biotechnology* 22(6): 600-607.

Zwickl DJ. 2010. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, The University of Texas, Austin, TX, USA.

APPENDIX I

SEQUENCES

Wild type *Zymobacter palmae* PDC amino acid sequence

MYLAERLAQIGLKHHFAVAGDYNLVLLDQLLLNKDMEQVYCCNELNCGFSAEGYARARGA
AAAIVTFSVGAISAMNAIGGAYAENLPVILISGSPNTNDYGTGHILHHTIGTTDYNYQLE
MVKHVTCAAESIVSAEEAPAKIDHVIRTALRERKPAYLEIACNVAGAECVRPGPINSLLR
ELEVDQTSVTAAVDAAVEWLQDRQNVVMLVGSKLRAAAAEKQAVALADRLGCAVTIMAAA
KGFFPEDHPNFRGLYWGEVSSEGAQELVENADAILCLAPVFNDYATVGWNSWPKGDNVMV
MDTDRVTFAGQSFEGLSLSTFAAALAEKAPSRPATTQGTQAPVLGIEAAEPNAPLTNDEM
TRQIQSLITSDTTLTAETGDSWFNASRMPIPGGARVELEMQWGHIGWSVPSAFGNAVGSP
ERRHIMMVGDGSFQLTAQEVAQMIRYEIPVIIFLINNRGYVIEIAIHDGPYNYIKNWNYA
GLIDVFNDEDGHGLGLKASTGAELEGAIKKALDNRRGPTLIECNIAQDDCTETLIAWGKR
VAATNSRKPQA

59.4 kDa

Tagged *Zymobacter palmae* PDC amino acid sequence

MYTVGMYLAERLAQIGLKHHFAVAGDYNLVLLDQLLLNKDMEQVYCCNELNCGFSAEGYA
RARGAAAAIVTFSVGAISAMNAIGGAYAENLPVILISGSPNTNDYGTGHILHHTIGTTDY
NYQLEMVKHVTCAAESIVSAEEAPAKIDHVIRTALRERKPAYLEIACNVAGAECVRPGPI
NSLLRELEVDQTSVTAAVDAAVEWLQDRQNVVMLVGSKLRAAAAEKQAVALADRLGCAVT
IMAAAKGFFPEDHPNFRGLYWGEVSSEGAQELVENADAILCLAPVFNDYATVGWNSWPKG
DNVMVMDTDRVTFAGQSFEGLSLSTFAAALAEKAPSRPATTQGTQAPVLGIEAAEPNAPL
TNDEMTRQIQSLITSDTTLTAETGDSWFNASRMPIPGGARVELEMQWGHIGWSVPSAFGN
AVGSPERRHIMMVGDGSFQLTAQEVAQMIRYEIPVIIFLINNRGYVIEIAIHDGPYNYIK
NWNYAGLIDVFNDEDGHGLGLKASTGAELEGAIKKALDNRRGPTLIECNIAQDDCTETLI
AWGKRVAATNSRKPQALVPRGSGGGLEHHHHHH

Underlined: C-terminal thrombin cleavage site, 3x Glycine linker and hexa-histidine-tag

61.8 kDa

Wild type *Geobacillus thermoglucosidasius* ADH6 amino acid sequence

MNTFFLKPKIYFGNHSLNHLSDFNAGKVFIVTDQTMLKLGMAEKIIEKIKGAAFKIFPDV
EPNPSIETVKKAFECFLQEQPELVIALGGGSAIDAAKAMLLFYHYMKDISDIEMDLKKPL
LIAIPTTSGTGSEMTSYSVITDTTNHLKIPLRDERMLPDVAILDEQLTITVPPSVTADTG
MDVLTHAIEAYVSLNSSEFTDIFAERSIKMVFNYLLRAYRFGEDLDARGKLHIASCMAGI
AFTNSSLGINHSLAHAVGAKFHLPHGRTNAILLPYVIQYNSGLCDDTMDASPVAKRYTEI
SKMLGLPSSTLKEGVISLVTAIQFLNKKLDIPSSFKECDINETEFAKYIPSLAKDAMQDI
CTAGNPRKVTEKDFVYLLKWAYNG

42.5 kDa

Tagged *Geobacillus thermoglucosidasius* ADH6 amino acid sequence

MGSSHHHHHHSSGLVPRGSHMASNTFFLKPKIYFGNHSLNHLSDFNAGKVFIVTDQTMLK
LGMAEKIIEKIKGAAFKIFPDVEPNPSIETVKKAFECFLQEQPELVIALGGGSAIDAAKA
MLLFYHYMKDISDIEMDLKKPLLIAIPTTSGTGSEMTSYSVITDTTNHLKIPLRDERMLP
DVAILDEQLTITVPPSVTADTGMDVLTHAIEAYVSLNSSEFTDIFAERSIKMVFNYLLRA
YRFGEDLDARGKLHIASCMAGIAFTNSSLGINHSLAHAVGAKFHLPHGRTNAILLPYVIQ
YNSGLCDDTMDASPVAKRYTEISKMLGLPSSTLKEGVISLVTAIQFLNKKLDIPSSFKEC
DINETEFAKYIPSLAKDAMQDICTAGNPRKVTEKDFVYLLKWAYNG

Underlined: N-terminal hexa-histidine-tag and thrombin cleavage site

44.8 kDa

p778/p600 wt ZpPDC promoter region

Purple, lower case – M13 F
Restriction sites indicated in *Italics* and name given in brackets
Green, upper case – wild type *Z. palmae* PDC start codon

tgtaaaacgacggccagt(M13F)gccaagcttgcatgcctgcaggtcgac*TCTAGA*(*Xba*I)ggatccccgggtaccgcgggacg
gggagctgagtgctcccgttgtttgccgcggcgtctgtcatgaaatggacaaacaatagtcaaacaatcgccacaatcgcgcatgcatt
gcggtgcgcctttcgcgtaaaatatttatatgaaagtgttcgcattatattgagggaggatgaatcatggatccATG- wtZpPDC

pGR002 wtZpPDC

Restriction sites indicated in *Italics* and name given in brackets
Purple, lower case - M13 F
Black, lower case – *G. stearothermophilus* NCA1503 *ldh*A promoter
Black, upper case – *G. stearothermophilus* DSMZ6285 *phe*B RBS
Green, upper case – wild type *Z. palmae* PDC (wtZpPDC F1 primer underlined, F2 *Italics* and underlined)
Black, upper case, underlined – *G. stearothermophilus* DSMZ6285 *phe*B downstream region
Blue, upper case – M13 R

tgtaaaacgacggccagtgccaagcttgcatgcctgcaggcgggacgggagctgagtgctcccgttgtttgccgcggcgtctgtcatga
aatggacaaacaatagtcaaacaatcgccacaatcgcgcatgcattgcggtgcgcctttcgcgtaaaatatttatatgaaagtgttcgc
attatattgagggaggat*TCTAGA*(*Xba*I)TAAGGAGTGATTCGAATGTATACCGTTGGTATGTACTTGGCAGAA
CGCCTAGCCCAGATCGGCCTGAAACACCACTTTGCCGTGGCCGGTGACTACAACCTGGTGTTGCTTGAT
CAGCTCCTGCTGAACAAAGACATGGAGCAGGTCTACTGCTGTAACGAACTTAACTGCGGCTTTAGCGCC
GAAGGTTACGCTCGTGCACGTGGTGCCGCCGCTGCCATCGTCACGTTCAGCGTAGGTGCTATCTCTGCA
ATGAACGCCATCGGTGGCGCCTATGCAGAAAACCTGCCGGTCATCCTGATCTCTGGCTCACCGAACACC
AATGACTACGGCACAGGCCACATCCTGCACCACACCATTGGTACTACTGACTATAACTATCAGCTGGAA
ATGGTAAAACACGTTACCTGCGCAGCTGAAAGCATCGTTTCTGCCGAAGAAGCACCGGCAAAAATCGA
CCACGTCATCCGTACGGCTCTACGTGAACGCAAACCGGCTTATCTGGAAATCGCATGCAACGTCGCTGG
CGCTGAATGTGTTCGTCCGGGCCCGATCAATAGCCTGCTGCGTGAACTCGAAGTTGACCAGACCAGTGT

CACTGCCGCTGTAGATGCCGCCGTAGAATGGCTGCAGGACCGCCAGAACGTCGTCATGCTGGTCGGTA
GCAAACTGCGTGCCGCTGCCGCTGAAAAACAGGCTGTTGCCCTAGCGGACCGCCTGGGCTGCGCTGTC
ACGATCATGGCTGCCGCAAAAGGCTTCTTCCCGGAAGATCATCCGAACTTCCGCGGCCTGTACTGGGGT
GAAGTCAGCTCCGAAGGTGCACAGGAACTGGTTGAAAACGCCGATGCCATCCTGTGTCTGGCACCGGT
ATTCAACGACTATGCTACCGTTGGCTGGAACTCCTGGCCGAAAGGCGACAATGTCATGGTCATGGACA
CCGACCGCGTCACTTTCGCAGGACAGTCCTTCGAAGGTCTGTCATTGAGCACCTTCGCCGCAGCACTGG
CTGAGAAAGCACCTTCTCGCCCGGCAACGACTCAAGGCACTCAAGCACCGGTACTGGGTATTGAGGCC
GCAGAGCCCAATGCACCGCTGACCAATGACGAAATGACGCGTCAGATCCAGTCGCTGATCACTTCCGA
CACTACTCTGACAGCAGAAACAGGTGACTCTTGGTTCAACGCTTCTCGCATGCCGATTCCTGGCGGTGC
TCGTGTCGAACTGGAAATGCAATGGGGTCATATCGGTTGGTCCGTACCTTCTGCATTCGGTAACGCCGT
TGGTTCTCCGGAGCGTCGCCACATCATGATGGTCGGTGATGGCTCTTTCCAGCTGACTGCTCAAGAAGT
TGCTCAGATGATCCGCTATGAAATCCCGGTCATCATCTTCCTGATCAACAACCGCGGTTACGTCATCGAA
ATCGCTATCCATGACGGCCCTTACAACTACATCAAAAACTGGAACTACGCTGGCCTGATCGACGTCTTC
AATGACGAAGATGGTCATGGCCTGGGTCTGAAAGCTTCTACTGGTGCAGAACTAGAAGGCGCTATCAA
GAAAGCACTCGACAATCGTCGCGGT*CCGACGCTGATCGAATGT*AACATCGCTCAGGACGACTGCACTG
AAACCCTGATTGCTTGGGGTAAACGTGTAGCAGCTACCAACTCTCGCAAACCACAAGCGTAATCGCGCC
CCGAAAGGGGGCGTTTTTTTGC*GAGCTC*(*Sac*I)gaattcgtaatcatggtcatagctgtttCCTGTGTGAAATTGTT
ATCCGCTC

pGR002 2.0 (codon harmonized wt *Z. palmae pdc*)

Restriction sites indicated in *Italics* and name given in brackets
Purple, lower case – M13 F
Black, lower case – *G. stearothermophilus* NCA1503 *ldh*A promoter
Black, upper case – *G. stearothermophilus* DSMZ6285 *phe*B RBS
Orange, upper case – *Z. palmae* PDC 2.0 (2.0 F1 primer underlined, 2.0 F2 *Italics* and underlined)
Black, upper case, underlined – *G. stearothermophilus* DSMZ6285 *phe*B downstream region
Blue, upper case – M13 R

tgtaaaacgacggccagtgccaagcttgcatgcctgcaggcgggacgggagctgagtgctcccgttgtttgccgcggcgtctgtcatga
aatggacaaacaatagtcaaacaatcgccacaatcgcgcatgcattgcggtgcgcctttcgcgtaaaatatttatatgaaagtgttcgc
attatattgagggaggat*TCTAGA*(*Xba*I)TAAGGAGTGATTCGAATGTATACCGTTGGTATGTATCTAGCAGAA
CGCCTCGCGCAAATTGGCTTAAAACACCACTTCGCGGTGGCGGGTGATTATAATTTAGTGCTCCTAGAC
CAACTCTTATTAAATAAAGATATGGAGCAAGTCTATTGCTGTAATGAACTAAATTGCGGCTTCAGCGCG
GAAGGTTATGCTAGGGCAAGGGGTGCGGCGGCTGCGATTGTCACGTTTAGCGTAGGTGCTATTAGTGC
AATGAATGCGATTGGTGGCGCGTACGCAGAAAATTTACCGGTCATTTTAATTAGTGGCTCACCGAATAC
GAACGATTATGGCACTGGCCACATTTTACACCACACGATAGGTACTACTGATTACAATTACCAATTAGA
AATGGTAAAACACGTTACGTGCGCAGCTGAAAGCATTGTTAGTGCGGAAGAAGCACCGGCAAAAATTG
ATCACGTCATTAGGACGGCTCTAAGGGAACGCAAACCGGCTTACTTAGAAATTGCATGCAATGTCGCTG
GCGCTGAATGTGTTAGGCCGGGCCCGATTAACAGCTTATTAAGGGAACTAGAAGTTGATCAAACGAGT
GTCACTGCGGCTGTAGACGCGGCGGTAGAATGGTTACAAGATCGCCAAAATGTCGTCATGTTAGTCGG
TAGCAAATTAAGGGCGGCTGCGGCTGAAAAACAAGCTGTTGCGCTAGCCGATCGCTTAGGCTGCGCTG
TCACGATTATGGCTGCGGCAAAAGGCTTTTTTCCGGAAGACCATCCGAATTTTCGCGGCTTATATTGGG
GTGAAGTCAGCTCCGAAGGTGCACAAGAATTAGTTGAAAATGCGGACGCGATTTTATGTTTAGCACCG
GTATTTAATGATTACGCTACGGTTGGCTGGAATTCCTGGCCGAAAGGCGATAACGTCATGGTCATGGAT
ACGGATCGCGTCACTTTTGCAGGGCAATCCTTTGAAGGTTTATCACTAAGCACGTTTGCGGCAGCATTA
GCTGAGAAAGCACCTAGTCGCCCGGCAACGACTCAGGGCACTCAGGCACCGGTATTAGGTATAGAGG
CGGCAGAGCCAAACGCACCGTTAACGAACGATGAAATGACGAGGCAAATTCAATCGTTAATTACTTCC

GATACTACTTTAACTGCAGAAACTGGTGATAGTTGGTTTAATGCTAGTCGCATGCCGATACCTGG*CGGT GCTAGGGTCGAATTAG*AAATGCAGTGGGGTCATATTGGTTGGTCCGTACCTAGTGCATTTGGTAATGC GGTTGGTAGTCCGGAGAGGCGCCACATTATGATGGTCGGTGACGGCAGTTTTCAATTAACTGCTCAGG AAGTTGCTCAAATGATTCGCTACGAAATTCCGGTCATTATTTTTTTAATTAATAATCGCGGTTATGTCATT GAAATTGCTATTCATGATGGCCCTTATAATTATATTAAAAATTGGAATTATGCTGGCTTAATTGATGTCT TTAACGATGAAGACGGTCATGGCTTAGGTTTAAAAGCTAGTACTGGTGCAGAACTCGAAGGCGCTATT AAGAAAGCACTAGATAACAGGCGCGGTCCGACGTTAATTGAATGTAATATTGCTCAAGATGATTGCAC TGAAACGTTAATAGCTTGGGGTAAAAGGGTAGCAGCTACGAATAGTCGCAAACCACAGGCCTAATCGC GCCCCGAAAGGGGGCGTTTTTTTGC*GAGCTC*(*Sac*I)gaattcgtaatcatggtcatagctgtttCCTGTGTGAAATT GTTATCCGCTC

Sequence alignment wtZp*pdc* vs 2.0:

```
wtZpPDC         ATGTACTTGGCAGAACGCCTAGCCCAGATCGGCCTGAAACACCACTTTGCCGTGGCCGGT
ZpPDC2.0        ATGTATCTAGCAGAACGCCTCGCGCAAATTGGCTTAAAACACCACTTCGCGGTGGCGGGT
                *****  * ********** ** ** ** *** * ********** ** ***** ***

wtZpPDC         GACTACAACCTGGTGTTGCTTGATCAGCTCCTGCTGAACAAAGACATGGAGCAGGTCTAC
ZpPDC2.0        GATTATAATTTAGTGCTCCTAGACCAACTCTTATTAAATAAAGATATGGAGCAAGTCTAT
                ** ** **  * *** * ** ** ** *** *  * ** ***** ******** *****

wtZpPDC         TGCTGTAACGAACTTAACTGCGGCTTTAGCGCCGAAGGTTACGCTCGTGCACGTGGTGCC
ZpPDC2.0        TGCTGTAATGAACTAAATTGCGGCTTCAGCGCGGAAGGTTATGCTAGGGCAAGGGGTGCG
                ******** ***** ** ******** ***** ******** *** * *** * *****

wtZpPDC         GCCGCTGCCATCGTCACGTTCAGCGTAGGTGCTATCTCTGCAATGAACGCCATCGGTGGC
ZpPDC2.0        GCGGCTGCGATTGTCACGTTTAGCGTAGGTGCTATTAGTGCAATGAATGCGATTGGTGGC
                ** ***** ** ******** **************   ********* ** ** ******

wtZpPDC         GCCTATGCAGAAAACCTGCCGGTCATCCTGATCTCTGGCTCACCGAACACCAATGACTAC
ZpPDC2.0        GCGTACGCAGAAAATTTACCGGTCATTTTAATTAGTGGCTCACCGAATACGAACGATTAT
                ** ** ********  * ********  * **   ************ ** ** ** **

wtZpPDC         GGCACAGGCCACATCCTGCACCACACCATTGGTACTACTGACTATAACTATCAGCTGGAA
ZpPDC2.0        GGCACTGGCCACATTTTACACCACACGATAGGTACTACTGATTACAATTACCAATTAGAA
                ***** ********  * ******** ** ********** ** ** ** **  * ***

wtZpPDC         ATGGTAAAACACGTTACCTGCGCAGCTGAAAGCATCGTTTCTGCCGAAGAAGCACCGGCA
ZpPDC2.0        ATGGTAAAACACGTTACGTGCGCAGCTGAAAGCATTGTTAGTGCGGAAGAAGCACCGGCA
                ***************** ***************** ***  *** ***************

wtZpPDC         AAAATCGACCACGTCATCCGTACGGCTCTACGTGAACGCAAACCGGCTTATCTGGAAATC
ZpPDC2.0        AAAATTGATCACGTCATTAGGACGGCTCTAAGGGAACGCAAACCGGCTTACTTAGAAATT
                ***** ** ********  * ********* * *****************  * *****

wtZpPDC         GCATGCAACGTCGCTGGCGCTGAATGTGTTCGTCCGGGCCCGATCAATAGCCTGCTGCGT
ZpPDC2.0        GCATGCAATGTCGCTGGCGCTGAATGTGTTAGGCCGGGCCCGATTAACAGCTTATTAAGG
                ******** ********************** * *********** ** *** *  *  *

wtZpPDC         GAACTCGAAGTTGACCAGACCAGTGTCACTGCCGCTGTAGATGCCGCCGTAGAATGGCTG
ZpPDC2.0        GAACTAGAAGTTGATCAAACGAGTGTCACTGCGGCTGTAGACGCGGCGGTAGAATGGTTA
                ***** ******** ** ** *********** ******** ** ** ********* *

wtZpPDC         CAGGACCGCCAGAACGTCGTCATGCTGGTCGGTAGCAAACTGCGTGCCGCTGCCGCTGAA
ZpPDC2.0        CAAGATCGCCAAAATGTCGTCATGTTAGTCGGTAGCAAATTAAGGGCGGCTGCGGCTGAA
                ** ** ***** ** ********* * ************ *  * ** ***** ******

wtZpPDC         AAACAGGCTGTTGCCCTAGCGGACCGCCTGGGCTGCGCTGTCACGATCATGGCTGCCGCA
ZpPDC2.0        AAACAAGCTGTTGCGCTAGCCGATCGCTTAGGCTGCGCTGTCACGATTATGGCTGCGGCA
                ***** ******** ***** ** *** * ***************** ******** ***
```

```
wtZpPDC         AAAGGCTTCTTCCCGGAAGATCATCCGAACTTCCGCGGCCTGTACTGGGGTGAAGTCAGC
ZpPDC2.0        AAAGGCTTTTTTCCGGAAGACCATCCGAATTTTCGCGGCTTATATTGGGGTGAAGTCAGC
                ******** ** ******** ******** ** ****** * ** ***************

wtZpPDC         TCCGAAGGTGCACAGGAACTGGTTGAAAACGCCGATGCCATCCTGTGTCTGGCACCGGTA
ZpPDC2.0        TCCGAAGGTGCACAAGAATTAGTTGAAAATGCGGACGCGATTTTATGTTTAGCACCGGTA
                ************** *** * ******** ** ** ** **  * *** * *********

wtZpPDC         TTCAACGACTATGCTACCGTTGGCTGGAACTCCTGGCCGAAAGGCGACAATGTCATGGTC
ZpPDC2.0        TTTAATGATTACGCTACGGTTGGCTGGAATTCCTGGCCGAAAGGCGATAACGTCATGGTC
                ** ** ** ** ***** ********** ***************** ** *********

wtZpPDC         ATGGACACCGACCGCGTCACTTTCGCAGGACAGTCCTTCGAAGGTCTGTCATTGAGCACC
ZpPDC2.0        ATGGATACGGATCGCGTCACTTTTGCAGGGCAATCCTTTGAAGGTTTATCACTAAGCACG
                ***** ** ** ********** ***** ** ***** ****** * *** * *****

wtZpPDC         TTCGCCGCAGCACTGGCTGAGAAAGCACCTTCTCGCCCGGCAACGACTCAAGGCACTCAA
ZpPDC2.0        TTTGCGGCAGCATTAGCTGAGAAAGCACCTAGTCGCCCGGCAACGACTCAGGGCACTCAG
                ** ** ****** * ***************  ****************** ********

wtZpPDC         GCACCGGTACTGGGTATTGAGGCCGCAGAGCCCAATGCACCGCTGACCAATGACGAAATG
ZpPDC2.0        GCACCGGTATTAGGTATAGAGGCGGCAGAGCCAAACGCACCGTTAACGAACGATGAAATG
                ********* * ***** ***** ******** ** ****** * ** ** ** ******

wtZpPDC         ACGCGTCAGATCCAGTCGCTGATCACTTCCGACACTACTCTGACAGCAGAAACAGGTGAC
ZpPDC2.0        ACGAGGCAAATTCAATCGTTAATTACTTCCGATACTACTTTAACTGCAGAAACTGGTGAT
                *** * ** ** ** *** * ** ******** ****** * ** ******** *****

wtZpPDC         TCTTGGTTCAACGCTTCTCGCATGCCGATTCCTGGCGGTGCTCGTGTCGAACTGGAAATG
ZpPDC2.0        AGTTGGTTTAATGCTAGTCGCATGCCGATACCTGGCGGTGCTAGGGTCGAATTAGAAATG
                  ****** ** ***  ************ ************ * ****** * ******

wtZpPDC         CAATGGGGTCATATCGGTTGGTCCGTACCTTCTGCATTCGGTAACGCCGTTGGTTCTCCG
ZpPDC2.0        CAGTGGGGTCATATTGGTTGGTCCGTACCTAGTGCATTTGGTAATGCGGTTGGTAGTCCG
                ** ********** ***************  ****** ***** ** ******  ****

wtZpPDC         GAGCGTCGCCACATCATGATGGTCGGTGATGGCTCTTTCCAGCTGACTGCTCAAGAAGTT
ZpPDC2.0        GAGAGGCGCCACATTATGATGGTCGGTGACGGCAGTTTTCAATTAACTGCTCAGGAAGTT
                *** * ******** ************** ***  *** **  * ******** ******

wtZpPDC         GCTCAGATGATCCGCTATGAAATCCCGGTCATCATCTTCCTGATCAACAACCGCGGTTAC
ZpPDC2.0        GCTCAAATGATTCGCTACGAAATTCCGGTCATTATTTTTTTAATTAATAATCGCGGTTAT
                ***** ***** ***** ***** ******** ** **  * ** ** ** ********

wtZpPDC         GTCATCGAAATCGCTATCCATGACGGCCCTTACAACTACATCAAAAACTGGAACTACGCT
ZpPDC2.0        GTCATTGAAATTGCTATTCATGATGGCCCTTATAATTATATTAAAAATTGGAATTATGCT
                ***** ***** ***** ***** ******** ** ** ** ***** ***** ** ***

wtZpPDC         GGCCTGATCGACGTCTTCAATGACGAAGATGGTCATGGCCTGGGTCTGAAAGCTTCTACT
ZpPDC2.0        GGCTTAATTGATGTCTTTAACGATGAAGACGGTCATGGCTTAGGTTTAAAAGCTAGTACT
                *** * ** ** ***** ** ** ***** ********* * *** * ******  ****

wtZpPDC         GGTGCAGAACTAGAAGGCGCTATCAAGAAAGCACTCGACAATCGTCGCGGTCCGACGCTG
ZpPDC2.0        GGTGCAGAACTCGAAGGCGCTATTAAGAAAGCACTAGATAACAGGCGCGGTCCGACGTTA
                *********** *********** *********** ** **  * ************ *

wtZpPDC         ATCGAATGTAACATCGCTCAGGACGACTGCACTGAAACCCTGATTGCTTGGGGTAAACGT
ZpPDC2.0        ATTGAATGTAATATTGCTCAAGATGATTGCACTGAAACGTTAATAGCTTGGGGTAAAAGG
                ** ******** ** ***** ** ** ***********  * ** ************ *

wtZpPDC         GTAGCAGCTACCAACTCTCGCAAACCACAAGCGTAA
ZpPDC2.0        GTAGCAGCTACGAATAGTCGCAAACCACAGGCCTAA
                *********** **   ************ ** ***
```

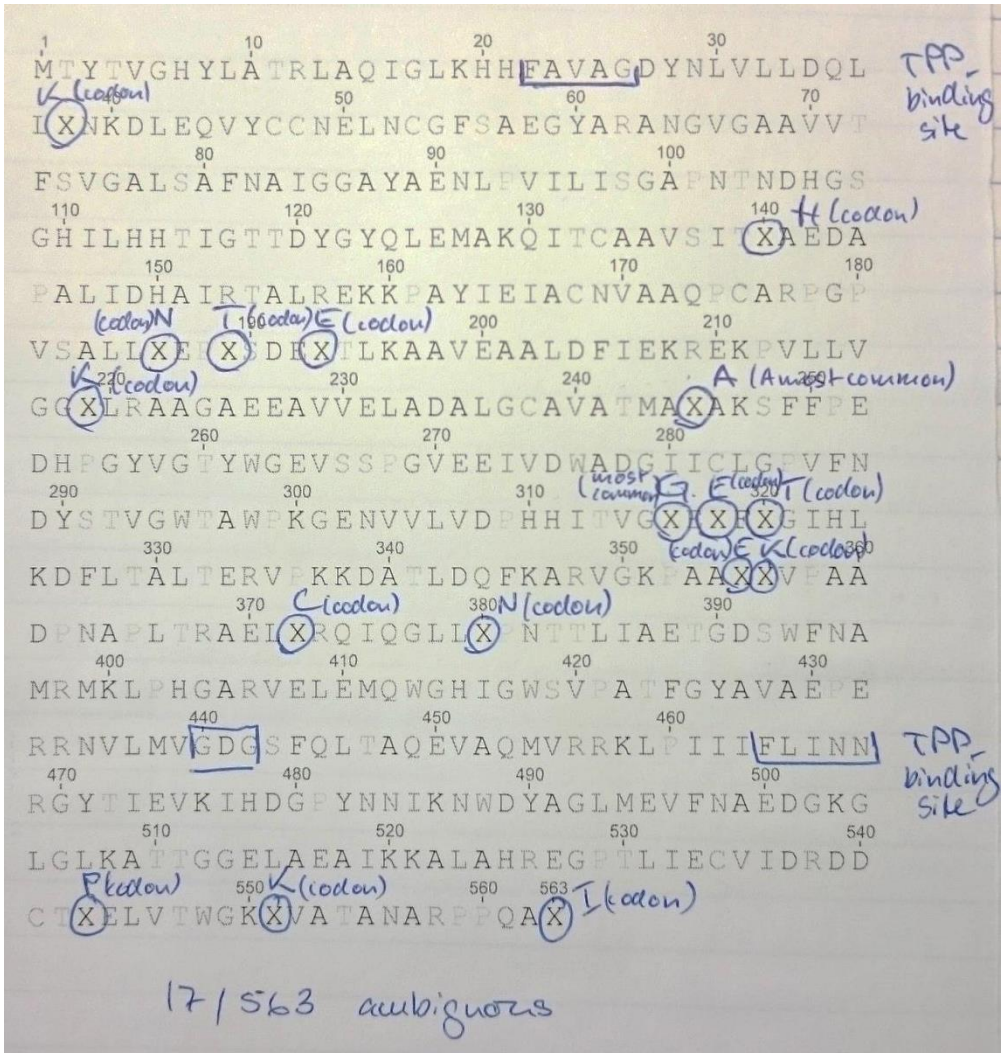
pUCGT PET

Restriction sites indicated in *Italics* and name given in brackets
Purple, lower case – M13 F
Black, lower case – *G. stearothermophilus* NCA1503 *ldh*A promoter
Black, upper case – *G. stearothermophilus* DSMZ6285 *phe*B RBS

Orange, upper case – ZpPDC 2.0 (2.0 F1 primer underlined, 2.0 F2 *Italics* and underlined)
Black, upper case – *G. stearothermophilus* DSMZ6285 *phe*B RBS
Light blue, lower case – *G. thermoglucosidasius* ADH6
Black, upper case, underlined – *G. stearothermophilus* DSMZ6285 *phe*B downstream region
Dark blue, upper case – M13 R

TGTAAAACGACGGCCAGTgccaagcttgcatgcctgcaggcgggacgggagctgagtgctcccgttgtttgccgcggcgtctgt
catgaaatggacaaacaatagtcaaacaatcgccacaatcgcgcatgcattgcggtgcgcctttcgcgtaaaatatttatatgaaagtg
ttcgcattatattgagggaggatt*TCTAGA*(*Xba*I)TAAGGAGTGATTCGAATGTACACGGTTGGTATGTATCTAG
CAGAACGCCTCGCGCAAATTGGCTTAAAACACCACTTCGCGGTGGCGGGTGATTATAATTTAGTGCTCC
TAGACCAACTCTTATTAAATAAAGATATGGAGCAAGTCTATTGCTGTAATGAACTAAATTGCGGCTTCA
GCGCGGAAGGTTATGCTAGGGCAAGGGGTGCGGCGGCTGCGATTGTCACGTTTAGCGTAGGTGCTAT
TAGTGCAATGAATGCGATTGGTGGCGCGTACGCAGAAAATTTACCGGTCATTTTAATTAGTGGCTCACC
GAATACGAACGATTATGGCACTGGCCACATTTTACACCACACGATAGGTACTACTGATTACAATTACCA
ATTAGAAATGGTAAAACACGTTACGTGCGCAGCTGAAAGCATTGTTAGTGCGGAAGAAGCACCGGCAA
AAATTGATCACGTCATTAGGACGGCTCTAAGGGAACGCAAACCGGCTTACTTAGAAATTGCATGCAAT
GTCGCTGGCGCTGAATGTGTTAGGCCGGGCCCGATTAACAGCTTATTAAGGGAACTAGAAGTTGATCA
AACGAGTGTCACTGCGGCTGTAGACGCGGCGGTAGAATGGTTACAAGATCGCCAAAATGTCGTCATGT
TAGTCGGTAGCAAATTAAGGGCGGCTGCGGCTGAAAAACAAGCTGTTGCGCTAGCCGATCGCTTAGGC
TGCGCTGTCACGATTATGGCTGCGGCAAAAGGCTTTTTTCCGGAAGACCATCCGAATTTTCGCGGCTTA
TATTGGGGTGAAGTCAGCTCCGAAGGTGCACAAGAATTAGTTGAAAATGCGGACGCGATTTTATGTTT
AGCACCGGTATTTAATGATTACGCTACGGTTGGCTGGAATTCCTGGCCGAAAGGCGATAACGTCATGG
TCATGGATACGGATCGCGTCACTTTTGCAGGGCAATCCTTTGAAGGTTTATCACTAAGCACGTTTGCGG
CAGCATTAGCTGAGAAAGCACCTAGTCGCCCGGCAACGACTCAGGGCACTCAGGCACCGGTATTAGGT
ATAGAGGCGGCAGAGCCAAACGCACCGTTAACGAACGATGAAATGACGAGGCAAATTCAATCGTTAAT
TACTTCCGATACTACTTTAACTGCAGAAACTGGTGATAGTTGGTTTAATGCTAGTCGCATGCCGATACCT
GG*CGGTGCTAGGGTCGAATTAG*AAATGCAGTGGGGTCATATTGGTTGGTCCGTACCTAGTGCATTTGG
TAATGCGGTTGGTAGTCCGGAGAGGCGCCACATTATGATGGTCGGTGACGGCAGTTTTCAATTAACTG
CTCAGGAAGTTGCTCAAATGATTCGCTACGAAATTCCGGTCATTATTTTTTTAATTAATAATCGCGGTTA
TGTCATTGAAATTGCTATTCATGATGGCCCTTATAATTATATTAAAAATTGGAATTATGCTGGCTTAATT
GATGTCTTTAACGATGAAGACGGTCATGGCTTAGGTTTAAAAGCTAGTACTGGTGCAGAACTCGAAGG
CGCTATTAAGAAAGCACTAGATAACAGGCGCGGTCCGACGTTAATTGAATGTAATATTGCTCAAGATG
ATTGCACTGAAACGTTAATAGCTTGGGGTAAAAGGGTAGCAGCTACGAATAGTCGCAAACCACAGGCC
TAA*CTCGAG*(*Xho*I)TAAGGAGTGATTCGAATGTACACGGTTGGTatgaatacattcttcttgaaaccaaaaatctact
tcggaaaccattcattaaatcatttgtctgattttaatgcaggaaaagtctttattgtaacggatcagacgatgctgaaactgggcatggc
agagaagattatcgaaaaaataaaaggtgctgcgtttaaaatttttccggatgtagagcctaatccgtccatagaaaccgtcaaaaag
gcttttgaatgttttttgcaagaacagccagagctggtgatagcgcttggcggtggttcagccattgatgctgctaaagcgatgttgctttt
ttatcactacatgaaagacatatctgatatagaaatggatttaaaaaaaccattattgattgcaatccccacaactagcggaacaggttc
agaaatgacatcttattcagtcattacggatacaacgaatcatttaaaaattcctttgcgtgatgaaaggatgctccctgatgttgccatt
ttagatgagcaattaacgataactgtgccaccttctgtcacagcggatacaggcatggatgtgctcactcatgccattgaagcatatgttt
ctttaaactcttcagaatttaccgatatatttgctgagcggtccattaaaatggtatttaattatctattaagggcatatcgttttggggaag
accttgatgccagagggaaattgcacatagcgtcctgtatggctggtattgcgtttaccaattcgtctttagggattaatcatagcctcgc
acatgcggttggcgcaaaatttcatttgccgcatggcagaactaatgctatttttattgccttatgttatccaatataatagcggtctttgcg
atgatacgatggatgcttctcctgtggcgaagaggtatacagaaatttcgaaaatgttaggcttgccaagctcaaccttaaaagaaggg
gtcataagtttggttactgccattcagtttcttaataaaaagctggatataccgtcaagtttcaaagaatgcgatattaacgaaaccgaat
ttgcaaaatatatcccttctttggccaaagacgcaatgcaagatatttgcacagctggtaatcctagaaaagtaacagaaaaagattttg
tctatttattaaaatgggcatataacggataaTCGCGCCCCGAAAGGGGGCGTTTTTTTGC*GAGCTC*(*Sac*I)gaattcg
taatcatgtcatagctgtttCCTGTGTGAAATTGTTATCCGCTC

Node 27 PDC consensus amio acid sequence



Node 27 PDC solved amino acid sequence

MTYTVGHYLATRLAQIGLKHHFAVAGDYNLVLLDQLLKNKDLEQVYCCNELNCGFSAEGY
ARANGVGAAVVTFSVGALSAFNAIGGAYAENLPVILISGAPNTNDHGSGHILHHTIGTTD
YGYQLEMAKQITCAAVSITHAEDAPALIDHAIRTALREKKPAYIEIACNVAAQPCARPGP
VSALLNEPTSDEETLKAAVEAALDFIEKREKPVLLVGGKLRAAGAEEAVVELADALGCAV
ATMAAAKSFFPEDHPGYVGTYWGEVSSPGVEEIVDWADGIICLGPVFNDYSTVGWTAWPK
GENVVLVDPHHITVGGEEFTGIHLKDFLTALTERVPKKDATLDQFKARVGKPAAEKVPAA
DPNAPLTRAELCRQIQGLLNPNTTLIAETGDSWFNAMRMKLPHGARVELEMQWGHIGWSV
PATFGYAVAEPERRNVLMVGDGSFQLTAQEVAQMVRRKLPIIIFLINNRGYTIEVKIHDG
PYNNIKNWDYAGLMEVFNAEDGKGLGLKATTGGELAEAIKKALAHREGPTLIECVIDRDD
CTPELVTWGKKVATANARPPQAILVPRGSGGGLEHHHHHH

Underlined: C-terminal thrombin cleavage site, 3x Glycine linker and hexa-histidine-tag

62.8 kDa

Node 27 PDC nucleotide sequence

ATGACGTATACGGTCGGCCATTATTTGGCGACGCGCTTGGCGCAAATTGGCTTGAAACATCATTTTGCG
GTCGCGGGCGATTATAACTTGGTCTTGTTGGATCAATTGTTGAAAAACAAAGATTTGGAACAAGTCTAT
TGCTGCAACGAATTGAACTGCGGCTTTAGCGCGGAAGGCTATGCGCGCGCGAACGGCGTCGGCGCGG
CGGTCGTCACGTTTAGCGTCGGCGCGTTGAGCGCGTTTAACGCGATTGGCGGCGCGTATGCGGAAAAC
TTGCCGGTCATTTTGATTAGCGGCGCGCCGAACACGAACGATCATGGCAGCGGCCATATTTTGCATCAT
ACGATTGGCACGACGGATTATGGCTATCAATTGGAAATGGCGAAACAAATTACGTGCGCGGCGGTCAG
CATTACGCATGCGGAAGATGCGCCGGCGTTGATTGATCATGCGATTCGCACGGCGTTGCGCGAAAAAA
AACCGGCGTATATTGAAATTGCGTGCAACGTCGCGGCGCAACCGTGCGCGCGCCCGGGCCCGGTCAGC
GCGTTGTTGAACGAACCGACGAGCGATGAAGAAACGTTGAAAGCGGCGGTCGAAGCGGCGTTGGATT
TTATTGAAAAACGCGAAAAACCGGTCTTGTTGGTCGGCGGCAAATTGCGCGCGGCGGGCGCGGAAGA
AGCGGTCGTCGAATTGGCGGATGCGTTGGGCTGCGCGGTCGCGACGATGGCGGCGGCGAAGAGCTTT
TTTCCGGAAGATCATCCGGGCTATGTCGGCACGTATTGGGGCGAAGTCAGCAGCCCGGGCGTCGAAGA
AATTGTCGATTGGGCGGATGGCATTATTTGCTTGGGCCCGGTCTTTAACGATTATAGCACGGTCGGCTG
GACGGCGTGGCCGAAAGGCGAAAACGTCGTCTTGGTCGATCCGCATCATATTACGGTCGGCGGCGAA
GAATTTACGGGCATTCATTTGAAAGATTTTTTGACGGCGTTGACGGAACGCGTCCCGAAAAAAGATGC
GACGTTGGATCAATTTAAAGCGCGCGTCGGCAAACCGGCGGCGGAAAAAGTCCCGGCGGCGGACCCG
AACGCGCCGTTGACGCGCGCGGAATTGTGCCGCCAAATTCAAGGCTTGTTGAACCCGAACACGACGTT
GATTGCGGAAACGGGCGATAGCTGGTTTAACGCGATGCGCATGAAATTGCCGCATGGCGCGCGCGTC
GAATTGGAAATGCAATGGGGCCATATTGGCTGGAGCGTCCCGGCGACGTTTGGCTATGCGGTCGCGG
AACCGGAACGCCGCAACGTCTTGATGGTCGGCGATGGCAGCTTTCAATTGACGGCGCAAGAAGTCGCG
CAAATGGTCCGCCGCAAATTGCCGATTATTATTTTTTTGATTAACAACCGCGGCTATACGATTGAAGTCA
AAATTCATGATGGCCCGTATAACAACATTAAAAACTGGGATTATGCGGGCTTGATGGAAGTCTTTAACG
CGGAAGATGGCAAAGGCTTGGGCTTGAAAGCGACGACGGGCGGCGAATTGGCGGAAGCGATTAAAA
AAGCGTTGGCGCATCGCGAAGGCCCGACGTTGATTGAATGCGTCATTGATCGCGATGATTGCACGCCG
GAATTGGTCACGTGGGGCAAAAAAGTCGCGACGGCGAACGCGCGCCCGCCGCAAGCGATTTAA

APPENDIX II

TRNA DATA

Fields marked yellow: no tRNA gene available, i.e., translated by "wobble" pairing

# of codons	Amino Acid		Codon	Anticodon	*G. thermoglucosidasius* genome (C56-YS93, NC_015660.1) # of tRNA genes	*E. coli* BL21 (DE3) genome (NC_012892.2) # of tRNA genes	*Z. palmae* genome (this study) # of tRNA genes
1	Methionine	M	ATG	CAT	6	7	3
1	Tryptophan	W	TGG	CCA	1	1	1
2	Asparagine	N	AAT	ATT	0	0	0
	Asparagine	N	AAC	GTT	4	4	2
2	Aspartic acid	D	GAT	ATC	0	0	0
	Aspartic acid	D	GAC	GTC	4	3	0
2	Cysteine	C	TGT	ACA	0	0	0
	Cysteine	C	TGC	GCA	2	1	0
2	Glutamine	Q	CAG	CTG	0	2	1
	Glutamine	Q	CAA	TTG	4	2	1
2	Glutamic acid	E	GAG	CTC	0	0	0
	Glutamic acid	E	GAA	TTC	5	4	4
2	Histidine	H	CAT	ATG	0	0	0
	Histidine	H	CAC	GTG	2	1	2
2	Lysine	K	AAG	CTT	0	0	0
	Lysine	K	AAA	TTT	4	6	4
2	Phenylalanine	F	TTT	AAA	0	0	0
	Phenylalanine	F	TTC	GAA	2	2	1
2	Tyrosine	Y	TAT	ATA	0	0	0
	Tyrosine	Y	TAC	GTA	2	3	1
3	Isoleucine	I	ATA	TAT	0	0	0
	Isoleucine	I	ATT	AAT	0	0	0
	Isoleucine	I	ATC	GAT	4	3	1
4	Alanine	A	GCG	CGC	1	0	0
	Alanine	A	GCA	TGC	5	3	1
	Alanine	A	GCT	AGC	0	0	0
	Alanine	A	GCC	GGC	1	2	3
4	Glycine	G	GGG	CCC	1	1	1
	Glycine	G	GGA	TCC	3	1	1
	Glycine	G	GGT	ACC	0	0	0
	Glycine	G	GGC	GCC	5	4	4
4	Threonine	T	ACG	CGT	1	1	1
	Threonine	T	ACA	TGT	3	1	1
	Threonine	T	ACT	AGT	0	0	0

					G. thermoglucosidasius genome (C56-YS93, NC_015660.1)	*E. coli* BL21 (DE3) genome (NC_012892.2)	*Z. palmae* genome (this study)
	Threonine	T	ACC	GGT	1	2	1
4	Proline	P	CCG	CGG	1	1	1
	Proline	P	CCA	TGG	3	1	1
	Proline	P	CCT	AGG	0	0	0
	Proline	P	CCC	GGG	0	1	1
4	Valine	V	GTG	CAC	0	0	0
	Valine	V	GTA	TAC	4	5	1
	Valine	V	GTT	AAC	0	0	0
	Valine	V	GTC	GAC	1	2	2
6	Arginine	R	AGG	CCT	0	1	1
	Arginine	R	AGA	TCT	2	1	1
	Arginine	R	CGG	CCG	1	1	1
	Arginine	R	CGA	TCG	0	0	0
	Arginine	R	CGT	ACG	4	4	1
	Arginine	R	CGC	GCG	0	0	0
6	Leucine	L	TTG	CAA	1	1	1
	Leucine	L	TTA	TAA	2	1	0
	Leucine	L	CTG	CAG	1	4	0
	Leucine	L	CTA	TAG	2	1	1
	Leucine	L	CTT	AAG	0	0	0
	Leucine	L	CTC	GAG	1	1	0
6	Serine	S	AGT	ACT	0	0	0
	Serine	S	AGC	GCT	2	1	1
	Serine	S	TCG	CGA	1	1	2
	Serine	S	TCA	TGA	2	1	1
	Serine	S	TCT	AGA	0	0	0
	Serine	S	TCC	GGA	1	2	1
				TCA	0	1	0
				Total:	90	85	51

APPENDIX III

CHARACTERIZATION OF *ZYMOBACTER PALMAE* ALCOHOL DEHYDROGENASES

In this chapter, characterization of *Z. palmae* ADHI and ADHII, previously undocumented, is presented. Kinetic characterization was carried out with the help of supervised in2scienceUK student Kishwar Khanum.

These ADHs were identified in the draft genome (containing 6 *adh* annotations) by sequence comparison with *Zymomonas mobilis* ADHA (Genbank: M32100.1, 75% nucleotide sequence identity to ADHI over 100% coverage) and ADHB (M15394.1, 73%) (*adh*I on contig 28, *adh*II on contig 7). The online tool ExPASyTranslate (http://web.expasy.org/translate/) was used to translate the identified nucleotide sequences into amino acid sequences and compare these to the ZmADHs. ZmADHA (GenBank: AAA27682.1) and ZmADHB (AAA27683.1) showed 81 and 77% amino acid sequence identity, respectively.

A.1 METHODS

A.1.1 CLONING FOR CHARACTERIZATION IN *E. COLI*

These putative alcohol dehydrogenases were amplified from *Z. palmae* genomic DNA using ZpADHI F/R (using 100 ng genomic DNA, annealing temperature 65°C) and ZpADHII F/R (using 10 ng genomic DNA, annealing temperature 62°C) (see Table A.1), and KAPA HiFi polymerase (see General Methods for details, extension time 1 min). The resulting PCR products were digested with *Nde*I and *Not*I. This double digest was also used on the destination vector pET28a. Digested vector and insert were purified from an agarose gel and ligated as described in General Methods. The ligation was transformed into *E. coli* BioBlue to yield N-terminal hexa-histidine-tagged constructs with a thrombin cleavage site (pET28 ZpADHI and pET28 ZpADHII). A correct clone, as confirmed by sequencing using T7F and T7R, was used to propagate the plasmid, which was reisolated and transformed into *E. coli* BL21 (DE3).

Table A.1 Primers used in cloning *Z. palmae* alcohol dehydrogenases.
Abbreviations are: F, forward; R, reverse; T_m , melting temperature. Restriction sites are underlined.

Name	T_m (°C)	Use	Sequence 5' → 3'
ZpADH I F	56	Cloning *Z. palmae adh*I, introduces *Nde*I site	GGG AAT TCC ATA TGA AAG CAG CAG
ZpADH I R	63	Cloning *Z. palmae adh*I, introduces *Not*I site	TTT TCC TTT TGC GGC CGC TTA CAG CGA G
ZpADH II F	56	Cloning *Z. palmae adh*II, introduces *Nde*I site	GGG AAT TCC ATA TGG CAT CTT CAA C
ZpADH II R	63	Cloning *Z. palmae adh*II, introduces *Not*I site	TTT TCC TTT TGC GGC CGC TTA GAA CGC TTC

A.1.2 Enzyme characterization – Kinetic and Thermal Properties

The protein was expressed in *E. coli* BL21 (DE3) and purified through small-scale metal-affinity chromatography detailed in General Methods.

Kinetic properties were determined at 30°C using a range of ethanol, acetaldehyde and NAD^+/NADH concentrations as described in General Methods measuring absorbance at 340 nm.

To assess temperature dependent ADH activity, standard assays using ethanol as the substrate were carried out in a 1 ml quartz cuvette at temperatures between 30 and 65°C.

The temperature at which the ADH was irreversibly denatured was assessed by incubating protein samples at fixed temperatures between 4 and 65°C for 30 min, then cooling the sample on ice and assaying using the standard assay at 30°C.

Denaturing temperatures were determined by a fluorescence based thermal shift assay using SYPRO®-Orange as described in General Methods.

A.2 Results

A.2.1 Cloning for Expression in *E.coli*

The genes were amplified adding appropriate up- and down-stream regions and sub-cloned into pET28a, thus creating pET28 ZpADHI and pET28 ZpADHII (see Figure A.1 for PCR results and Figure A.2 for a plasmid map). Sequencing confirmed the correct clone, which was then transformed into *E. coli* BL21 (DE3) for expression and purification.

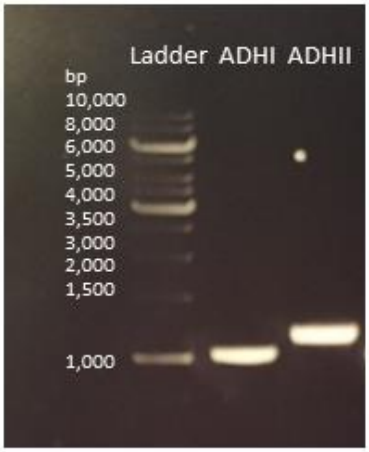


Figure A.1 Agarose gel electrophoresis of the *adhI* and *adhII* amplification products. The PCR-amplified products were visualized alongside the GeneRuler™ 1 kb ladder (Thermo Fisher Scientific).

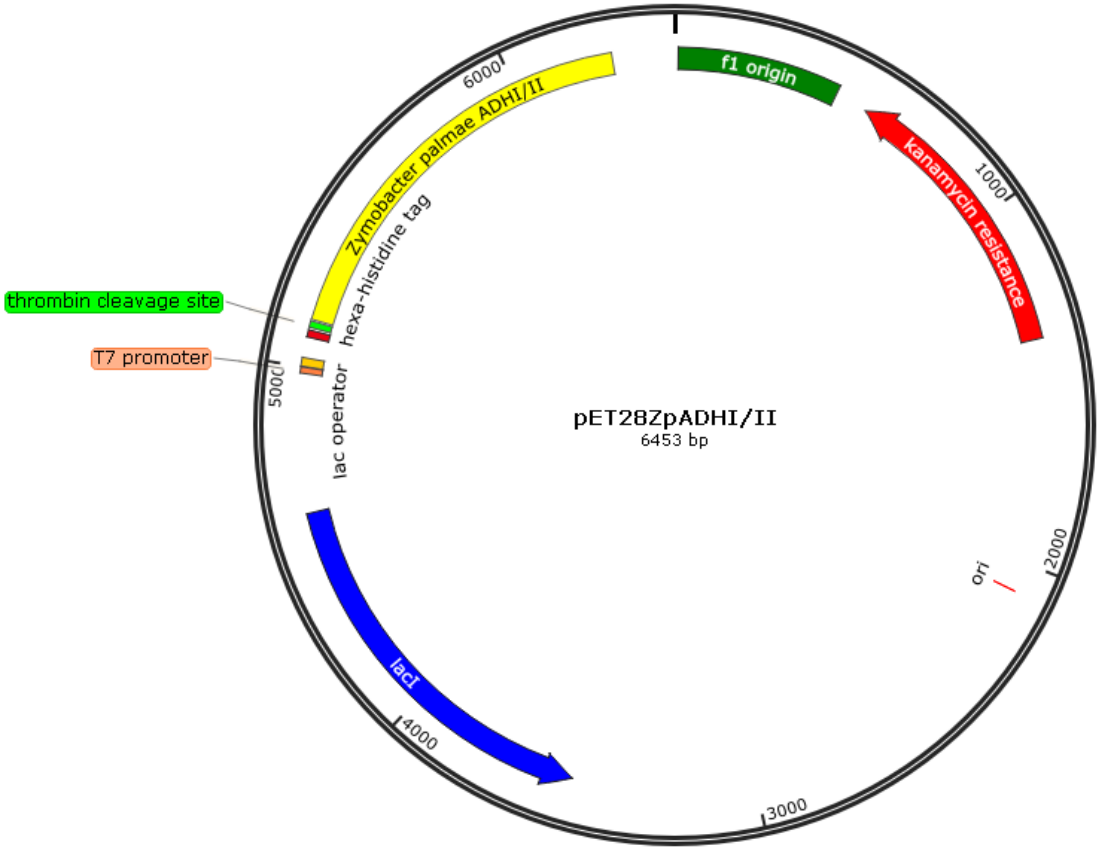


Figure A.2 Plasmid map of pET28 ZpADHI/II. The neomycin phosphotransferase gene (labelled kanamycin resistance) confers resistance to kanamycin. In the presence of a T7 RNA polymerase, the Zp*adh* is expressed from the T7 promoter under the control of the *lac* operator. *Lac*I encodes the *lac* operon repressor. f1 origin is the origin of replication from a f1 phage.

The ADH sequences are as follows:

Zp ADHI nucleotide sequence

ATGAAAGCAGCAGTTGTAGGCAAAGACCATAACGTTGAAATTCAGGACAAAAAACTGCGTCCGCTGGA
GCACGGCGAAGCCCTGCTGCAGATGGAGTGCTGTGGTGTGTGTCACACCGACCTGCACGTCAAAAATG
GTGATTTCGGTGACAAGACTGGCGTTATCCTGGGCCATGAAGGTGTAGGTGTCGTCAAAGAAGTTGGT
CCGGGCGTGACCTCTCTGAAAGTAGGCGACCGTGCAAGCGTTGCATGGTTCTTCCAAGGCTGTGGCCA
CTGCGAATACTGCATCAGCGGTAACGAAACGCTGTGCCGTAGCGTCAAAAACGCGGGTTACACCGTTG
ATGGTGGTATGGCGGAAGAATGCATCGTCACTGCTGACTACGCTGTCAAAGTACCGGACGGTCTGGAT
TCCGCAGCAGCCAGCAGCGTAACGTGCGCGGGTGTTACCACGTACAAAGCGATCAAAGTATCCAACAT
CAAAGCCGGCAAATGGATTGCCATCTACGGTCTGGGCGGTCTGGGTAACTTGGCGCTGCAGTATGCGA
AAAACGTCTTCAACGCTAAAGTCATCGCGATCGACGTCAACGATGAGCAGCTGAAACTGGCTCAGGAA
ATGGGCGCGGACATGGTCATTAACCCGGCCAAAGAAGACGCTGCAAAACTGATTCAGGAAAAAGTGG
GCGGTGCTCACGCGGCGGTCGTTACTGCCGTTGCCAAAGCCGCTTTCAACTCTGCGGTTGACGCTGTTC
GCGCTGGTGCGCGCATCGTAGCGGTTGGTCTGCCTCCGGAAGCCATGAGCCTCGACATTCCGCGCCTG
GTACTGGACGGTATCGAAGTCGTCGGCTCTCTGGTTGGTACGCGTGAAGACCTGGCTGAAGCCTTCCA
GTTCGCTGCTGAAGGCAAGGTCGTGCCGAAAGTGGCTAAACGCCCGATCGAAGACATCAACGATATCT
TCCACGAAATGGAACAAGGTAAGATCAAAGGCCGCATGGTTGTCGATTTCTCGCTGTAA

1014 nucleotides

Zp ADHI amino acid sequence

MKAAVVGKDHNVEIQDKKLRPLEHGEALLQMECCGVCHTDLHVKNGDFGDKTGVILGHEG
VGVVKEVGPGVTSLKVGDRASVAWFFQGCGHCEYCISGNETLCRSVKNAGYTVDGGMAEE
CIVTADYAVKVPDGLDSAAASSVTCAGVTTYKAIKVSNIKAGKWIAIYGLGGLGNLALQY
AKNVFNAKVIAIDVNDEQLKLAQEMGADMVINPAKEDAAKLIQEKVGGAHAAVVTAVAKA
AFNSAVDAVRAGARIVAVGLPPEAMSLDIPRLVLDGIEVVGSLVGTREDLAEAFQFAAEG
KVVPKVAKRPIEDINDIFHEMEQGKIKGRMVVDFSL

35.4 kDa (37.2 kDa with the N-terminal tag, HHHHHHSSGLVPRGSH), 336 amino acids

Zp ADHII nucleotide sequence

ATGGCATCTTCAACATTCTATGTGCCACCAGTAAACGAAATGGGCGAAGGCTCTTTGGAAAAAGCCATC
GGTGACCTGAAAGGCCGCGGTTTCAACCGTGCGCTGATCGTTACCGACGCGTTCATGAATTCCTGCGGT
ACGGCAGGTAAAGTGGCATCTCTGTTGGACGCGGCGGGTATCCCTTCCGTCATTTTTGATGGCGTTATG
CCTAACCCGACAGTCGGCAGCGTGTTACAGGGTCTCGAACTGCTGAAAGAAAACGACGCTGATCTGGT
TGTTTCCGTCGGTGGTGGTTCACCGCACGACTGCGCTAAAGCGGTTGCTCTGGTGGCTACCAACGGTG
GTGAAGTCAAAGATTACGAAGGTATTGATCGTTCCAGCAAAGCCGCGCTGCCGCTGATCTCGATCAAC
ACCACGGCCGGTACGGCGTCCGAAATGACGCGTTTCTGCATCATCACTGATGAAGAACGTCACGTGAA
AATGGCGATCGTCGACCGTCATGTCACTCCGATCGTTTCCGTCAACGACCCGATTCTGATGATGGGCAT
GCCGAAAGGCCTGTCAGCGGCAACCGGTATGGATGCGCTGACGCACGCTTTCGAAGCCTACGTTTCTA
CCGCTGCAACGCCGCTGACCGACGTCTGCGCGCTGAAAGCGGCTGAACTGATCGCTCGTTTCCTGCCGA
TCGCTTGCGAAGACGGCAGCAACATGGAAGCGCGTGAAGCCATGGCTTACGCACAGTTCATGGCCGGT
ATGGCGTTCAACAACGCCTCTCTGGGCTATGTACACGCCATGGCACACCAGCTGGGCGGTTACTACAAC
CTGCCTCACGGTGTTTGCAACGCCGTACTGCTGCCGCACGTACTGCGCTACAACGCTGAAGTCGCCGCT
GCTCGCATCAAGGATATGGGTGCTGCAATGGGCGTCGACGTTGCGGGTCTGAGTGACCGCGACGGTG
CTAACGCCACTATCGCTGCCGTTGAAGCATTGTCCAAACGCATCGACATTCCGGCAACGCTGACGGATC

TGGGTGCTAAACGTGAAGACGTCCAGATTCTGGCTGACCACGCGCTGAAAGACGCTTGTGCACTGACC
AACCCGCGTCAGGGTTCTCAGGAAGAAGTCGAAGCGCTGTTCCTGGAAGCGTTCTAA

1152 nucleotides

Zp ADHII amino acid sequence

MASSTFYVPPVNEMGEGSLEKAIGDLKGRGFNRALIVTDAFMNSCGTAGKVASLLDAAGI
PSVIFDGVMPNPTVGSVLQGLELLKENDADLVVSVGGGSPHDCAKAVALVATNGGEVKDY
EGIDRSSKAALPLISINTTAGTASEMTRFCIITDEERHVKMAIVDRHVTPIVSVNDPILM
MGMPKGLSAATGMDALTHAFEAYVSTAATPLTDVCALKAAELIARFLPIACEDGSNMEAR
EAMAYAQFMAGMAFNNASLGYVHAMAHQLGGYYNLPHGVCNAVLLPHVLRYNAEVAAARI
KDMGAAMGVDVAGLSDRDGANATIAAVEALSKRIDIPATLTDLGAKREDVQILADHALKD
ACALTNPRQGSQEEVEALFLEAF

40 kDa (41.8 kDa with the N-terminal tag, HHHHHHSSGLVPRGSH), 383 amino acids

A.2.2 Enzyme Characterization

The Zp*adh*s were expressed in *E. coli* BL21 (DE3) from pET28 ZpADHI/II using the T7 expression system. The recombinant protein was tagged with an N-terminal hexa-histidine-tag, with a predicted monomer size of 37 kDa and 42 kDa for ADHI and ADHII, respectively. His-tagged ZpADHs were purified by metal-affinity chromatography; see Figure A.3 and A.4 for SDS-PAGE analysis. ADH activity in the protein containing fractions was confirmed and the pure protein used in enzyme characterization studies.

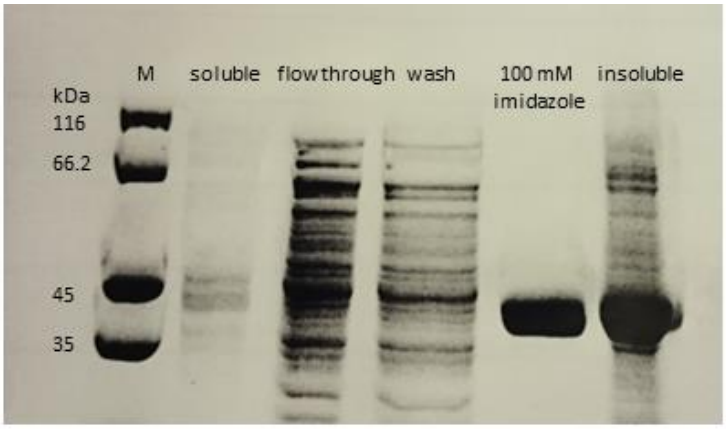


Figure A.3 SDS-PAGE analysis of ZpADHI metal-affinity chromatography fractions. His-tagged ZpADHI (monomer size of 37 kDa) was purified by metal-affinity chromatography. Lane M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific).

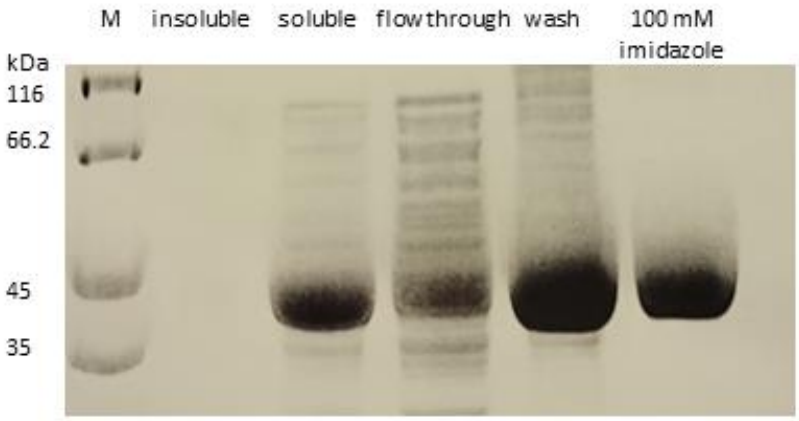


Figure A.4 SDS-PAGE analysis of ZpADHII metal-affinity chromatography fractions. His-tagged ZpADHII (monomer size of 42 kDa) was purified by metal-affinity chromatography. Lane M is the protein size marker, with sizes given in kDa (unstained protein molecular weight marker, Thermo Fisher Scientific).

Kinetic properties were analysed based on the standard assay at 30°C. The data were analysed using non-linear fit model from the enzyme kinetics module in SigmaPlot (Figure A.5 to A.12), resulting in a V_{max} and K_M as summarised in Table A.2.

Table A.2 ZpADHI and ADHII kinetics data. V_{max} is in mmol/min/mg, K_M is in mM unless otherwise indicated.

	Ethanol		NAD^+		Acetaldehyde		NADH
	V_{max}	K_M in M	V_{max}	K_M	V_{max}	K_M	
ADHI	1.4 ±0.1	0.043 ±0.01	1.9 ±0.1	0.22 ±0.07	1.3 ±0.2	0.1 ±0.04	NA
ADHII	0.34 ±0.06	0.08 ±0.07	0.44 ±0.04	0.20 ±0.08	0.04 ±0.002	0.06 ±0.01	NA

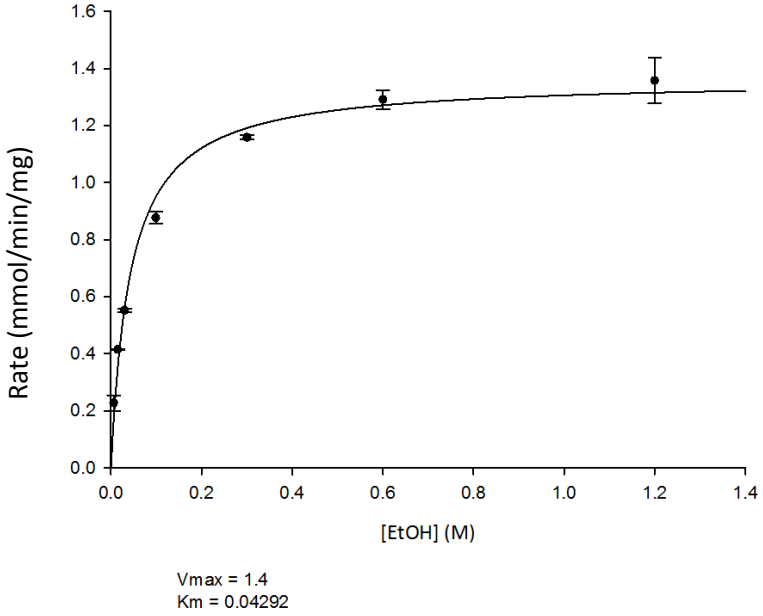


Figure A.5 Dependence of ZpADHI activity on the concentration of ethanol. The relationship between specific activity and ethanol concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 2 independent experiments with 3 measurements each.

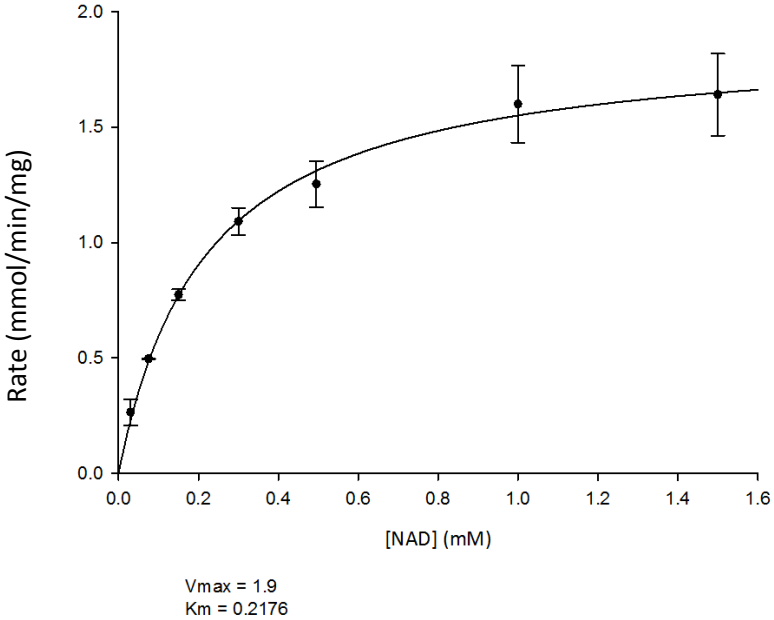


Figure A.6 Dependence of ZpADHI activity on the concentration of NAD^+. The relationship between specific activity and NAD^+ concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 2 independent experiments with 3 measurements each.

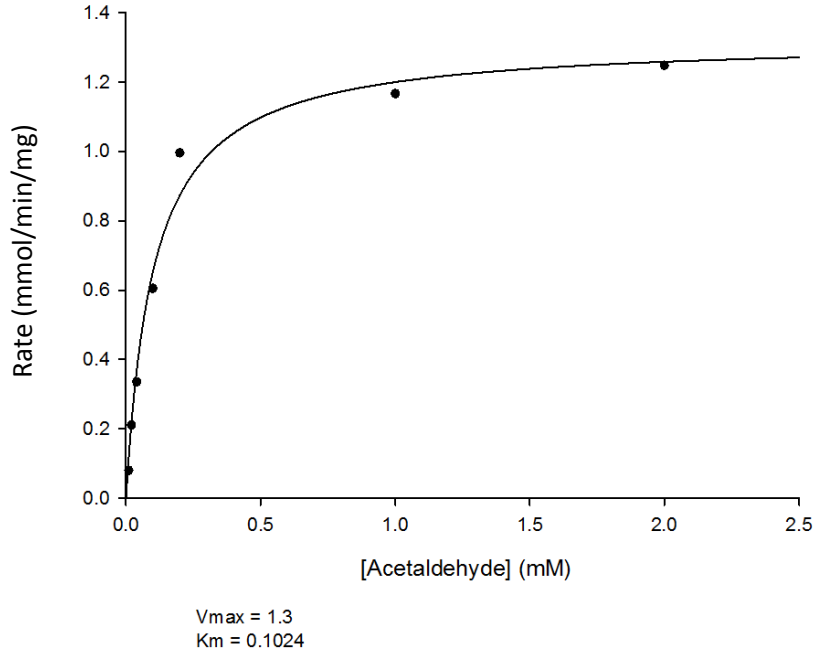


Figure A.7 Dependence of ZpADHI activity on the concentration of acetaldehyde. The relationship between specific activity and acetaldehyde concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 3 measurements.

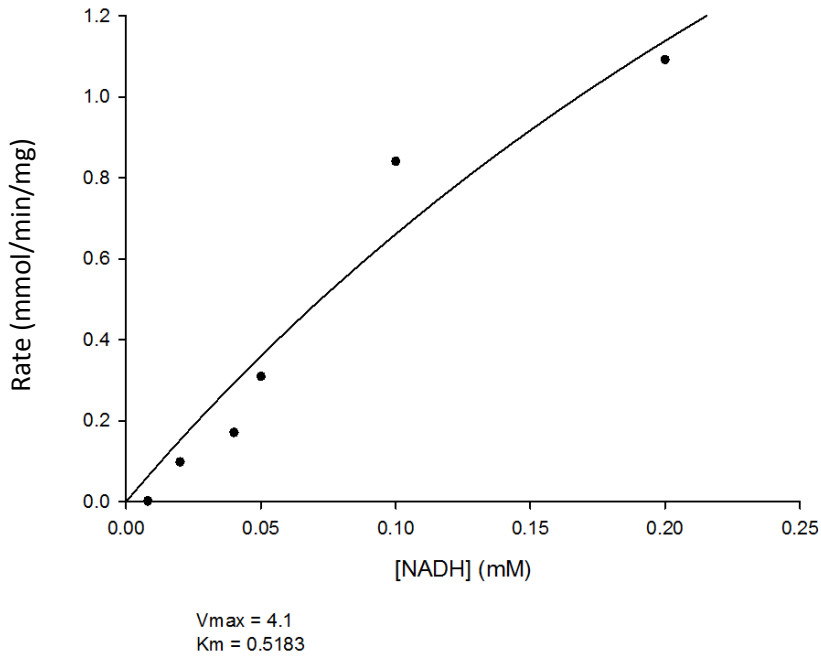


Figure A.8 Dependence of ZpADHI activity on the concentration of NADH. The relationship between specific activity and NADH concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 3 measurements.

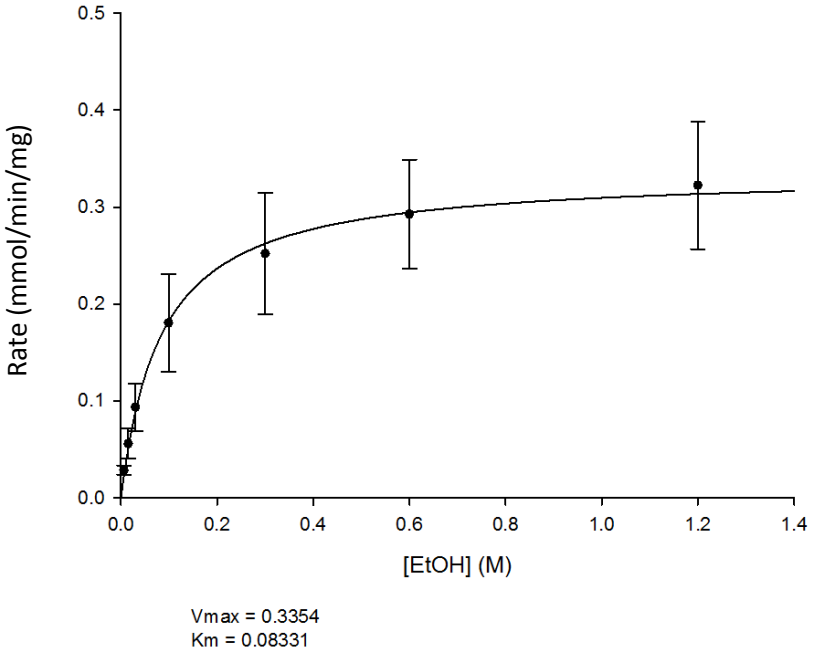


Figure A.9 Dependence of ZpADHII activity on the concentration of ethanol. The relationship between specific activity and ethanol concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 2 independent experiments with 3 measurements.

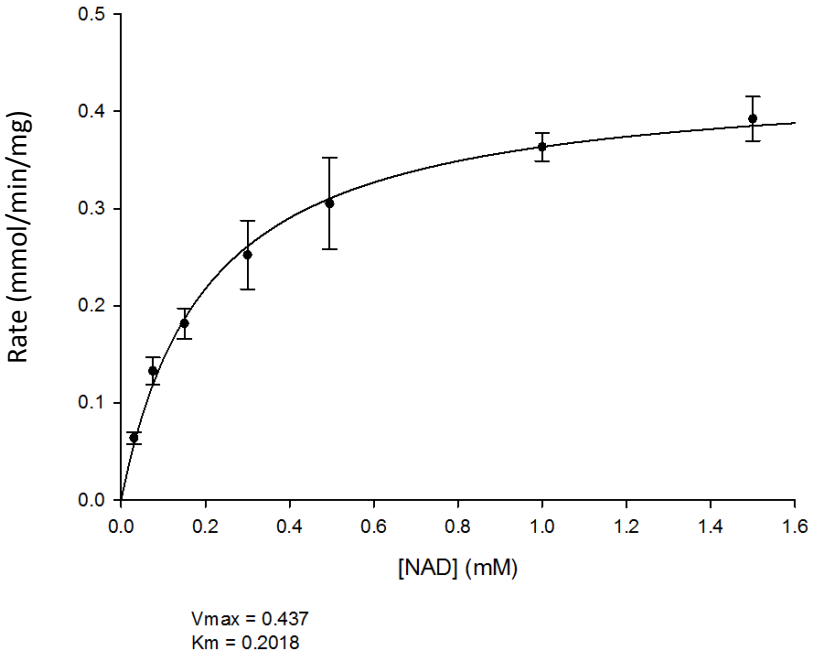


Figure A.10 Dependence of ZpADHII activity on the concentration of NAD^+. The relationship between specific activity and NAD^+ concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 2 independent experiments 3 measurements.

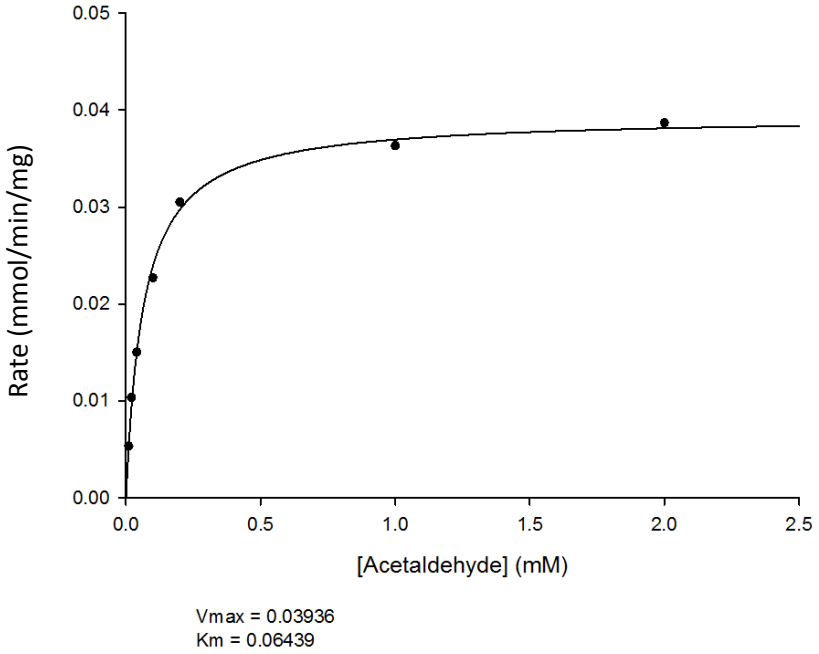


Figure A.11 Dependence of ZpADHII activity on the concentration of acetaldehyde. The relationship between specific activity and acetaldehyde concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 3 measurements.

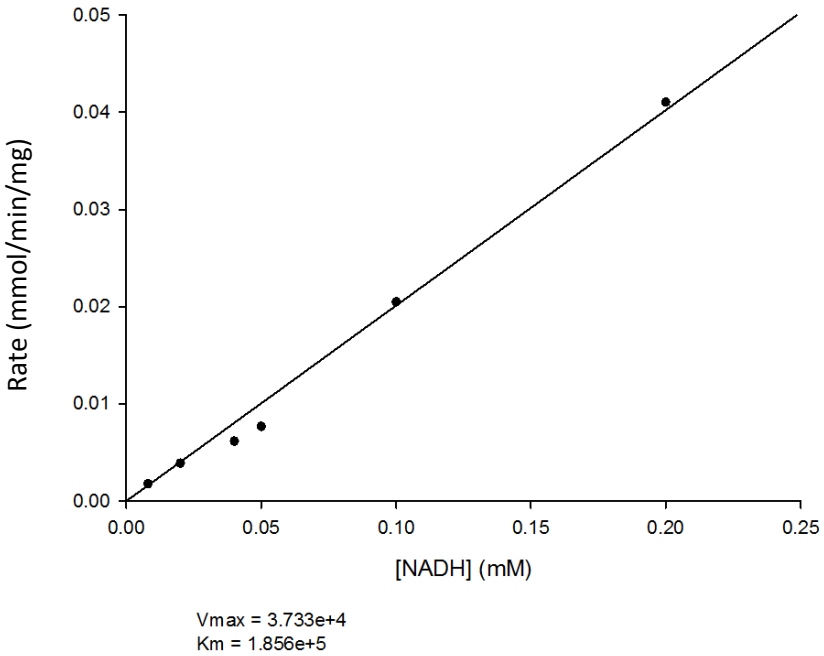


Figure A.12 Dependence of ZpADHII activity on the concentration of NADH. The relationship between specific activity and NADH concentrations is displayed as a Michaelis-Menten plot. Error bars are standard error based on 3 measurements.

Furthermore, thermal properties were investigated. The temperature optimum was found to be around 45°C for both enzymes. Irreversible denaturation analysis was carried out by incubation of the protein at 50 to 65°C for 30 min, and assaying for retained activity at 30°C. After exposure to 50°C, ZpADHI retained 20% activity, ADHII retained 15%. Using thermal shift assays, the denaturing temperature was determined to be 59°C and 58°C for ADHI and II, respectively.

A.3 Discussion

Metals were not investigated, as these ADHs are very similar to the well-studied ZmADHs. Based on sequence similarity it is likely that ADHI is a zinc-containing enzyme and ADHII is an iron-containing enzyme.

This is new data on previously not studied ADHs, but they were not appropriate for use in *Geobacillus* spp. without major improvements in thermoactivity.

APPENDIX IV

PUBLICATION: CRYSTAL STRUCTURE OF PYRUVATE DECARBOXYLASE FROM *ZYMOBACTER PALMAE*

Buddrus L, Andrews ES, Leak DJ, Danson MJ, Arcus VL, Crennell SJ. 2016. Crystal structure of pyruvate decarboxylase from *Zymobacter palmae*. *Acta Crystallographica F* 72(9): 700-706.