NATURAL EVOLUTION AND ENGINEERING OF TRANSKETOLASE

A thesis submitted to the

University of London

for the degree of

Doctor of Philosophy

Seán Joseph Costelloe

Department of Biochemical Engineering

University College London

2006

Dedicated to the memory of Sabina Costelloe

(1914 – 2004)

I, Seán J. Costelloe confirm that all work contained within the thesis entitled "Natural evolution and engineering of transketolase" is my own.

Signed:
Date:

Acknowledgement

I will begin by thanking my supervisor, Dr. Paul Dalby and my advisor Dr. John Ward. Their insight and experience has been of enormous benefit to my studies.

My fellow lab rats helped created a more pleasant working environment in Foster Court. Particular thanks to: Tarik, Ed, Waqar, Chrisine, Ollie, Janahan, Jaspreet, John, Richard, Jean, Julio, Evi, Jennifer, Jobin, Martina, Michael and Al. Dr. Mark Smith of Chemistry was a great help in during my final two years, as was Dr. Ziheng Yang of Biology, who helped with the user-rude phylogenetic software.

To my parents Seán and Carmel, my brothers Neal and Ciaran, my sister Céire and my Nana Lil, I am grateful for the unending support, particularly in the last few years and months. I wish them, and my extended family all the very best in the future.

To the Londoners: Eoin, Niamh, Eve, Caroline, Janice, Ray, Ben1, Ben2, Joanne, Seán, Mark, Jan; the Edinburghers: Dave, Nora, Colm and Mary; the Dubliners: Fowler, Elaine, Louise, Heinzy, Cathy & Sam, Grace, Gar, Fiona, Sonia, Andrew, Rosaline, Conor, Gillian and Gavin, Daoire, Gillian, Letitia, Karen, the Dunnes, Shane and Mel and those further afield: Kev, Fiadhna, Karen a massive thank you – I am blessed with my mates and I know it!

Finally, I'd like to thank Catherine Dunne in particular. Throughout the last four years she has shown superhuman patience and always been there for me.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC).

Abstract

Transketolase (TK) is an important metabolic enzyme in all organisms. The enantioselective carbon-carbon bond forming action of TK makes it significantly interesting for biocatalysis. TKs from different organisms exhibit varied substrate specificities mediated by a small number of differing residues, giving insight into a potential route for engineering new enzymes.

TK uses Thiamine Pyrophosphate (TPP) as cofactor, as do many other evolutionarily related enzymes. In Chapter 3, a phylogenetic analysis of the catalytic domains of TPP-dependent enzymes enabled the assembly of the evolutionary history of this enzyme family.

In Chapter 4, the evolution of the differing substrate specificities of *Eschericia coli* TK and *Saccharomyces cerevisiae* TK from their most recent common ancestor enzyme was analysed. A detailed phylogenetic analysis of TK was performed, yielding the amino acid sequences of the ancient TKs. TKs linking the common ancestor of *E. coli* and *S.cerevisiae* with extant *E. coli* TK were "resurrected" and assayed for the β-Hydroxypyruvate (β-HPA) + glycoaldehyde (GA) reaction. The common ancestor TK and *E. coli* TK were assayed for many reactions to define their substrate repertoires and to elucidate any evolutionary trends in substrate specificity.

β-HPA is the ideal donor substrate for TK, since it yields CO_2, making reactions irreversible. β-HPA is not readily available commercially and is very expensive. To be industrially viable, a cheaper donor is needed. Pyruvate is much cheaper than β-HPA yet TK has never been shown to use pyruvate as a donor. Chapter 5 describes a comparison of TK with the pyruvate utilising TPP-dependent enzymes DXPS and PDC, suggesting residues which may confer pyruvate usage. Mutants were generated and tested for activity with pyruvate.

In Chapter 6, the non-catalytic C-terminal domain of TK (TKC domain) was examined. The function of the TKC domain is currently undefined. In this chapter the TKC domain is removed and it is shown that the TK enzyme activity is retained.

Contents

Index of Figures

Index of Tables

Index of Alignments

Abbreviations

(*R*)-PAC	(*R*)-1-hydroxy 1-phenyl 2-propanone
.pdb	file extension for PDB file
[S]	sum of branch lengths
2-HPP	(*S*)-2-hydroxypropiophenone
2OXO	2-oxoisovalerate dehydrogenase
2OXOα	2-oxoisovalerate dehydrogenase alpha subunit
2OXOβ	2-oxoisovalerate dehydrogenase beta subunit
A_n	absorbance measured at *n* nm
A5P	arabinose 5-phosphate
ALS	acetolactate synthase
AMP$^+$	impregnated with 150 mg / L ampicilin (unless stated otherwise)
amp^R	ampicilin resistance gene
APS	ammonium persulfate
ATP	adenosine triphosphate
BAL	benzaldehyde lyase
β-HPA	β-Hydroxypyruvate
BFDC	benzoylformate decarboxylase
CAII	carbonic anhydrase II
CoA	coenzyme A
CoM	coenzyme M
CPU	central processing unit
ddH$_2$O	double distilled water
DHA	dihydroxyacetone
DHAP	dihydrohyacetone phosphate
DHETPP	α,β- dihydroxyethyl-TPP
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
dsDNA	double stranded DNA
dsRNA	double stranded RNA
DXP	D-xylulose 5-phosphate
DXPS	D-xylulose 5-phosphate synthase
E4P	erythrose 4-phosphate
EC	Enzyme Commission
ECD	electrochemical detector
ECP	eosinophil cationic protein
EDN	eosinophil-derived neurotoxin
EDTA	ethylene diaminetetraacetic acid
EF-Tu	elongation factor Tu
epPCR	error-prone PCR
F6P	fructose 6-phosphate
FAD	flavin adenine dinucleotide
FDA	Federal Drugs Agency
G3P	glyceraldehyde 3-phosphate
G6P	glucose 6-phosphate
GA	glycolaldehyde
GDP	guanosine diphosphate
Gly-Gly	glycyl-glycine
GXC	glyoxylate carboligase

HETPP	1-hydroxyethyl TPP
HGT	horizontal gene transfer
HNP	hydrophobic non-polar
HP	hydrophobic polar
HPLC	high performance liquid chromatography
IAA	indoleacetic acid
IPDC	indolepyruvate decarboxylase
IPP	isopentyl diphosphate
K_i	inhibition constant
K_m	Michaelis constant
L1	long interspersed (repetitive) element 1
LB	Luria Bertani
LS	least-square
m.y.a.	million years ago
MAP	methylaminopyridine
ME	minimum evolution
ML	maximum likelihood
M_r	relative molecular mass
mRNA	messenger RNA
MT	methythialonium
mtDNA	mitochondrial DNA
NADH	nicotinamide adenine dinucleotide, reduced form
NJ	neighbour joining
NMR	nuclear magnetic resonance
OCADC	oxalyl-CoA decarboxylase
ORF	open reading frame
PAM	percent accepted mutations
PB	bootstrap confidence value
PCR	polymerase chain reaction
PDB	RCSB Protein Data Bank
PDC	pyruvate decarboxylase
PDH	pyruvate dehydrogenase
PDH-E1	E1 subunit of the pyruvate dehydrogenase complex
PFRD	pyruvate ferredoxin reductase
PhPDC	phenylpyruvate decarboxylase
pK_a	-$\log_{10}[K_a]$
PKD	propionaldehyde ketodiol
PKL	phosphoketolase
PMA	phorbyl myristate acetate
PO	pyruvate oxidase
PPDC	phosphopyruvate decarboxylase
PPP	pentose phosphate pathway
P_t	probability of obtaining the true topology
R5P	ribose 5-phosphate
RAM	random access memory
RAMA	fructose 1,6-bisphosphate aldolase from rabbit muscle
RNA	ribonucleic acid
RNase	ribonuclease
rRNA	ribosomal RNA
*Sce*TK	transketolase from *Saccharomyces cerevisiae*
S7P	sedoheptulose 7-phosphate
SDM	site-directed mutagenesis

SDS	sodium dodecylsulphate
SDS-PAGE	sodium dodecylsulphate polyacrylamide gel electrophoresis
SPDC	sulphopyruvate decarboxylase
ssRNA	single stranded RNA
SSU rRNA	small subunit ribosomal RNA
$t_{½}$	half-life
TAE (buffer)	buffer containing Tris-acetate and EDTA
TDP	thiamine diphosphate
TEMED	N, N, N', N' - tetramethylethly-endiamine
TFA	trifluoroacetic acid
TK	transketolase
tkt E. coli	transketolase gene from *E. coli*
TLC	thin layer chromatography
TPP	thiamine pyrophosphate
Tris (buffer)	tris(hydroxymethyl)aminomethane (buffer)
UCL	University College London
UPGMA	unweighted pair-group method using arithmetic average
UV	ultraviolet
X5P	xylulose 5-phosphate

Nomenclature for TK enzymes and mutants

Since engineering is performed on *E. coli* TK, residues are referred to according to this species' numbering throughout, unless otherwise stated. Throughout the analyses in this thesis species will be referred to by a three letter abbreviation, defined in Table A in this section. When referencing a species of interest in the text, it s generally by the three letter code (e.g. *Sce* for *Saccharomyces cerevisiae*) thereafter. When discussing enzymes of interest from a given species, the three letter species abbreviation is immediately followed by the enzyme abbreviation. For example, transketolase (TK) from *S.cerevisiae* (*Sce*) is referred to as "*Sce*TK". Mutants of *Eco*TK are generated in Chapters 4 to 6. In chapter 4, ancient forms of TK are generated by mutation of *Eco*TK and named according to the node number at which they occur in a reconstructed TK phylogeny. For example, when the ancient TK corresponding with node 58 is generated, it is referred to as "N58TK".

Abb.	Species name	Abb.	Species name
Aae	Aquifex aeolicus	Cef	Cornyebacterium efficiens
Aan	Artemisia annua	Cjk	Corynebacterium jeikeium
Abr	Azospirillum brasilense	Chu	Cytophaga hutchinsonii
Aci	Acinetobacter sp. ADP1	Cje	Campylobacter jejuni
Afu	Archaeoglobus fulgidus	Cpn	Chlamydophila pneumonia
Ali	Azospirillum lipoferum	Cro	Catharanthus roseus
Ani	Aspergillus nidulans	Cte	Chlorobaculum tepidum
Aor	Aspergillus oryzae	Cth	Clostridium thermocellum
Apa	Acetobacter pasteurianus	Ctr	Chlamydia trachomatis
Ape	Aeropyrum pernix	Cvi	Chromobacterium violaceum
Api	Buchnera aphidicola	Cwa	Crocosphaera watsonii
Apl	Actinobacillus pleuropneumoniae	Dar	Dechloromonas aromatica
Ath	Arabidopsis thaliana	Dde	Desulfovibrio desulfuricans
Atu	Agrobacterium tumefaciens	Dps	Desulfotalea psychrophila
Ava	Anabaena variabilis	Dra	Deinococcus radiodurans R1
Avi	Azotobacter vinelandii	Dvu	Desulfovibrio vulgaris
Azo	Azoarcus sp	Eca	Erwinia carotovora
Bad	Bifidobacterium adolescentis	Ecl	Enterobacter cloacae
Bam	Bifidobacterium animalis	Eco	Escherichia coli
Ban	Bacillus anthracis	Efa	Enterococcus faecalis
Bce	Bacillus cereus	Efc	Enterococcus faecium
Bcp	Burkholderia cepacia	Ego	Eremothecium gossypii
Bfr	Bacteroides fragilis	Eni	Emericella nidulans
Bfu	Burkholderia fungorum	Gka	Geobacillus kaustophilus
Bga	Bifidobacterium gallinarum	Gme	Geobacter metallireducens
Bha	Bacillus halodurans	Gst	Geobacillus stearothermophilus
Bhe	Bartonella henselae	Gsu	Geobacter sulfurreducens
Brh	Bradyrhizobium sp	Gvi	Gleobacter violaceus
Bja	Bradyrhizobium japonicum	Hdu	Haemophilus ducreyi
Bli	Bacillus licheniformis	Hhe	Helicobacter hepaticus
Blo	Bifidobacterium longum	Hin	Haemophilus influenzae
Bma	Burkholderia mallei	Hma	Haloarcula marismortui
Bme	Brucella melletensis biovar Abortis	Hpy	Helicobacter Pylori
Bov	Bovine	Hso	Haemophilus somnus
Bpa	Bordetella parapertussis	Hum	Homo Sapiens
Bpb	Bordetella pertussis Tohama I	Huv	Hanseniaspora uvarum
Bpd	Bifidobacterium pseudolongum subsp. globosum]	Kla	Kluyveromyces lactis
Bps	Burkholderia pseudomallei	Kma	Kluyveromyces marxianus
Bpu	Bifidobacterium pullorum	Kpn	Klebsiella pneumoniae
Bqu	Bartonella quintana	Kte	Klebsiella terrigena
Bsi	Brucella suis	Lco	Lotus corniculatus
Bsp	Buchnera aphidicola str. APS	Les	Lycopersicon esculentum
Bsu	Bacillus subtilis	Lga	Lactobacillus gasseri
Bte	Bacteroides thetaiotaomicron	Lin	Listeria innocua
Bth	Bacillus thuringiensis	Lac	Lactobacillus acidophilus
Cab	Candida boidinii	Ljo	Lactobacillus johnsonii
Cac	Clostridium acetobutylicum	Lla	Lactococcus lactis
Cal	Candida albicans	Lme	Leishmania mexicana mexicana
Can	Capsicum annuum	Lmo	Listeria monocytogenes
Cbl	Candidatus blochmannia	Lms	Leuconostoc mesenteroides

Table A: Three letter abbreviations for species names used throughout thesis

Abb.	Species name	Abb.	Species name
Lpe	Lactobacillus pentosus	Pst	Pichia stipitis
Lpl	Lactobacillus plantarum	Psy	Pseudomonas syringae
Lsa	Lactobacillus sakei	Rat	Rat
Mac	Methanosarcina acetivorans	Rba	Rhodopirellula baltica
Mag	Magnetococcus sp. MC-1	Rca	Rhodobacter capsulatus
Mar	Methanococcus maripaludis	Reu	Ralstonia eutropha
Mba	Methanosarcina barkeri str. Fusaro	Rge	Rubrivivax gelatinosus
Mbe	Microbulbifer degradans	Ric	Rice
Mbu	Methanococcoides burtonii	Rme	Ralstonia metallidurans
Mca	Methylococcus capsulatus	Rnu	Roseovarius nubinhibins
Mes	Mesorhizobium sp. BNC1	Ror	Rhizopus oryzae
Mhe	Methanosaera thermophila	Rpa	Rhodopseudomonas palustris
Mhu	Methanospirillum hungatei	Rru	Rhodospirillum rubrum
Mfl	Methylobacillus flagellatus	Rso	Ralstonia solanacearum
Mge	Mycoplasma genitalium	Rsp	Rhodobacter sphaeroides
Mja	Methanocaldococcus jannaschii	Rxy	Rubrobacter xylanophilus
Mka	Methanopyrus kandleri	Sag	Streptococcus agalactiae
Mle	Mycobacterium leprae	Sau	Staphylococcus aureus
Mlo	Mesorhizobium loti	Sce	Saccharomyces cerevisiae
Mma	Methanosarcina mazei	Sel	Synechococcus elongatus
Mou	Mouse	Sen	Salmonella enterica
Mpe	Mycoplasma penetrans	Sep	Staphylococcus epidermidis
Mpu	Mycoplasma pulmonis	Sor	Streptococcus oralis
Mst	Methanospaera stadmanae	Sbo	Shigella boydii
Mth	Methanothermobacter thermautotrophicus	Sdy	Shigella dysenteriae
Mtm	Moorella thermoacetica	Sfl	Shigella flexneri
Mtr	Medicago truncatula	Shy	Streptomyces viridochromogenes
Mys	Mycobacterium sp.	Skl	Saccharomyces kluyveri
Mtu	Mycobacterium tuberculosis	Sip	Silicibacter pomeroyi
Mxp	Mentha x piperita	Sme	Sinorhizobium meliloti
Nar	Novosphingobium aromaticivorans	Smu	Streptococcus mutans
Ncr	Neurospora Crassa	Son	Shewanella oneidensis
Nme	Neisseria meningitidis	Spn	Streptococcus pneumoniae R6
Noc	Nitrosococcus oceani	Spo	Schizosaccharomyces pombe
Nos	Nostoc sp. PCC 7120	Sso	Sulfolobus solfataricus
Nps	Narcissus pseudonarcissus	Sth	Symbiobacterium thermophilum
Npu	Nostoc punctiforme	Sto	Sulfolobus tokodaii
Nsp	Nostoc sp. PCC 7120	Sac	Sulfolobus acidocaldarius
Nta	Nicotiana tabacum	Str	Streptococcus thermophilus
Oih	Oceanobacillus iheyensis	Stu	Solanum tuberosum
Ooe	Oenococcus oeni	Sty	Salmonella typhimurium
Oxo	Oxalobacter formigenes	Sve	Sarcina ventriculi
Pab	Pyrococcus abyssi	Svi	Streptomyces viridochromogenes
Pae	Pseudomonas aeruginosa	Swe	Atreptomyces wedmorensis
Pan	Pichia angusta	Syc	Synechocystis sp. PCC 6803
Par	Pyrobaculum aerophilum	Syn	Synechococcus sp. PCC 6301
Pfl	Pseudomonas fluorescens	Tac	Thermoplasma acidophilum
Pfu	Pyrococcus furiosus	Tde	Thiobacillus denitrificans
Pho	Pyrococcus horikoshii	Tel	Thermosynechococcus elongatus
Plu	Photorhabdus luminescens subsp. laumondii TTO1	Ter	Trichodesmium erythraeum
Pmo	Pueraria montana	Tet	Thermoanaerobacter ethanolicus
Pmu	Pasteurella multocida	Tma	Thermotoga maritima
Pol	Polaromonas sp. JS666	Tpa	Treponema pallidum
Ppe	Pediococcus pentosaceus	Tte	Thermoproteus tenax
Ppr	Photobacterium profundum	Tth	Thermus thermophilus
Ppu	Pseudomonas putida	Ttn	Thermoanaerobacter tengcongensis

Table A continued.

Abb.	Species name	Abb.	Species name
Tvo	*Thermoplasma volcanium*	***Xca***	*Xanthomonas campestris*
Uur	*Ureaplasma urealyticum*	***Xfa***	*Xylella fastidiosa*
Vch	*Vibrio cholerae*	***Xfl***	*Xanthobacter flavus*
Vpa	*Vibrio parahaemolyticus*	***Ype***	*Yersinia pestis biovar Medievalis*
Vvi	*Vitis vinifera*	***Yps***	*Yersinia pseudotuberculosis*
Vvu	*Vibrio vulnificus*	***Zma***	*Zea mays*
Wgl	*Wigglesworthia glossinidia*	***Zpa***	*Zymobacter palmae*
Xax	*Xanthomonas axonopodis*		

Table A continued.

In Chapter 5, mutants are named according to the mutation generated. For example, the *Eco*TK mutant where S385 is substituted by G385 is named "S385GTK". Finally in Chapter 6, truncated forms of TK are generated, by insertion of stop codons. Such mutants are referred to by the site of insertion of the stop codon. For example, where a stop codon is inserted at position G540, the resulting enzyme is referred to as "G540StopTK".

	T	**C**	**A**	**G**
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

Table B: Table of the Standard Genetic Code.

Units

Å	angstroms
AU	absorbance units
bp	base pair
fmoles	femtomoles
g	grams
GB	gigabyte
GHz	gigahertz
kbp	kilobase pair
kDa	kilodaltons
L	litre
M	molar
mg	milligram
mL	millilitre
mm	millimetre
mM	millimolar
mol	mole
ng	nanogram
nm	nanometre
°C	degrees centigrade
ODU	optical density units
pH	$-\log_{10}[H^+]$
rpm	revolutions per minute
U	units
V	volt
v / v	volume / volume
w / v	weight / volume
µL	microlitre
µm	micrometre

Chapter 1: Introduction

1.1 Transketolase

1.1.1 Function of transketolase

Transketolase (TK) (EC 2.2.1.1) is an intracellular thiamine diphosphate (TPP) dependent enzyme occupying a pivotal role in metabolic regulation, providing a link between glycolysis and the non-oxidative branch of the pentose phosphate pathway (PPP), where it performs two "carbon shuffling" reactions (Figure 1.1). TK supplies ribose units for nucleotide biosynthesis (Reaction 1, Figure 1.1) and supplies erythrose 4-phosphate (E4P) to the shikimate pathway for aromatic amino acid biosynthesis (Reaction 2, Figure 1.1) in microorganisms. Since TK was first identified in *S. cerevisiae* [1] it has been found in all organisms studied.

1.1.2 Structure of transketolase

TK was the first of the TPP-dependent enzymes to undergo successful crystallographic analysis. The TK crystal structure at both the 2.5Å [2] and 2.0Å [3] levels has yielded information on the overall topology of TPP-dependent enzymes.

TK is an obligatory homodimeric protein with *Sce*TK composed of two identical 74 kDa subunits [4,5]. Each subunit consists of three α/β type domains (Image A, Figure 1.2). The 320 residues of the amino-terminal form the PP domain which consists of a five strand β-sheet (parallel) flanked with helices (Image B, Figure 1.2 and Figure 1.4). The Pyr domain (residues 323-538) contains a six strand β-sheet (parallel), with α-helices (Image B, Figure 1.2).

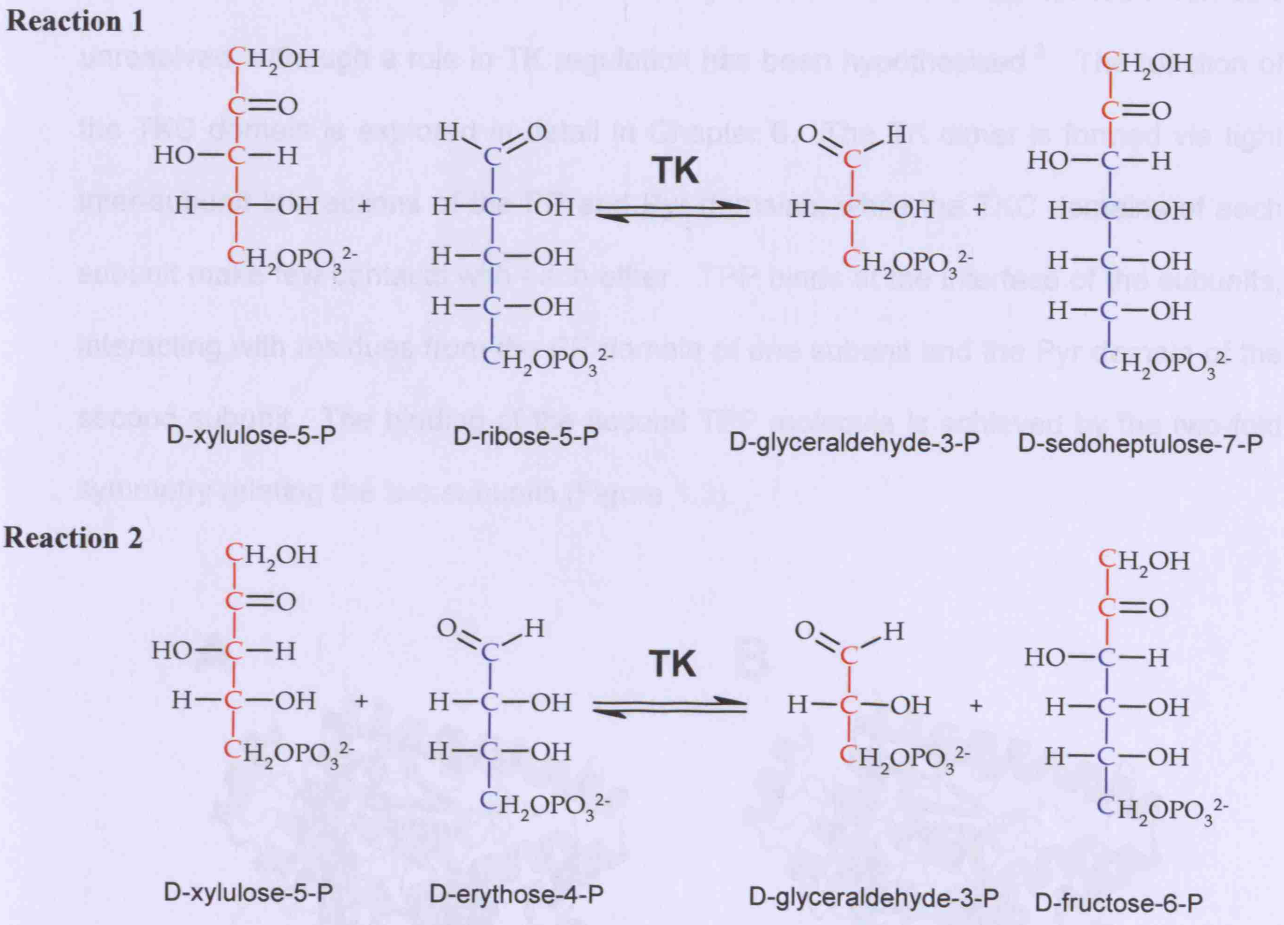


Figure 1.1: The physiological "carbon-shuffling" reactions of transketolase in the pentose phosphate pathway. In both cases, the carbon backbone is coloured so that the transfer of a two carbon unit (red) from the ketol donor, (xylulose-5-phosphate in both reactions) to the acceptor substrate (with the blue backbone) can be followed.

The PP and Pyr domains are common to all enzymes which use TPP as a cofactor. The two domains are structurally similar with 120 superimposable Cα atoms, suggesting an evolutionary relationship that extends to other TPP-dependent enzymes [2].

Both the PP and Pyr domains interact with TPP and thus their topology has been denoted the TPP-binding fold [6], (Section 1.1.3). The third carboxy-terminal domain (TKC domain) (residues 539-680) is composed of a mixed β-sheet with four parallel and one antiparallel strand (Image B, Figure 1.2).

The TKC domain contains no catalytic residues and its function remains unresolved, although a role in TK regulation has been hypothesised [3]. The function of the TKC domain is explored in detail in Chapter 6. The TK dimer is formed via tight inter-subunit interactions of the PP and Pyr domains, while the TKC domains of each subunit make few contacts with each other. TPP binds at the interface of the subunits, interacting with residues from the PP domain of one subunit and the Pyr domain of the second subunit. The binding of the second TPP molecule is achieved by the two-fold symmetry relating the two subunits (Figure 1.3).

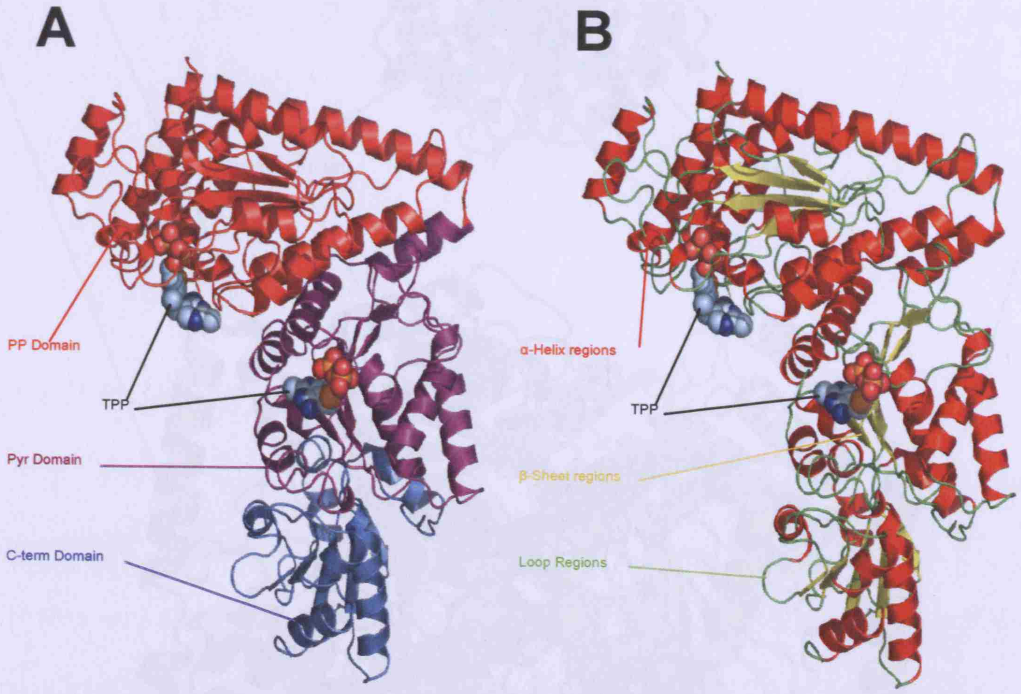


Figure 1.2: Structure of the transketolase monomer for *E. coli.* Image "A" shows the boundaries of the three domains of the TK monomer, the PP (red), Pyr (purple) and TKC (blue) domains. The TPP molecule is coloured according to a CNOS colouring scheme. In image "B" the TK monomer is coloured according to secondary structure. α-helices are coloured red, β-sheets are yellow, while loop regions are green. Images were generated using the 1QGD.pdb file in *Pymol*. Adapted from Schneider and Lindquist [7].

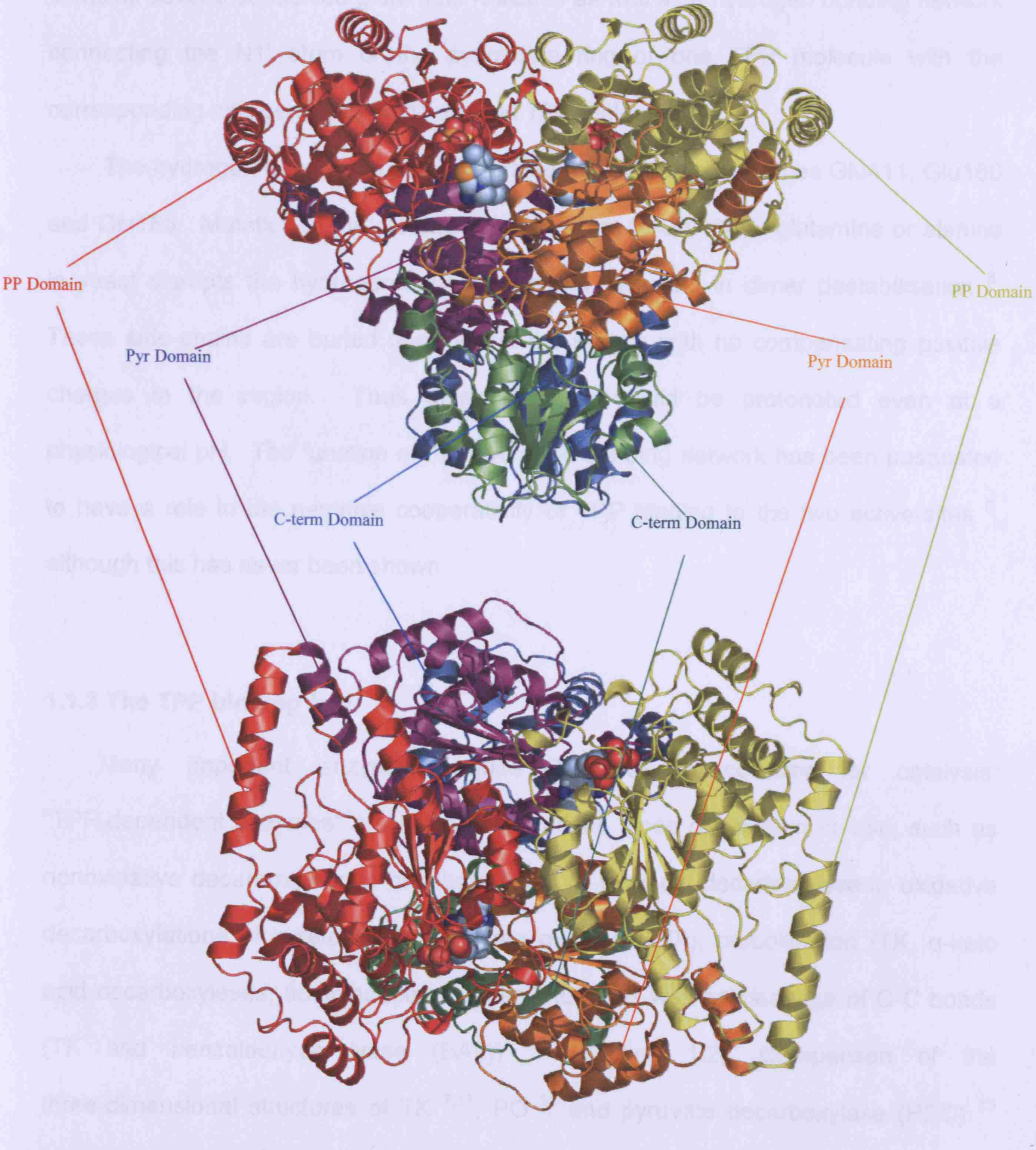


Figure 1.3: Views of the TK dimer. One subunit is coloured according to the scheme used in Figure 1.2, while the second is coloured so that the PP domain is yellow, the Pyr domain orange and the TKC domain is green. The TPP molecule is coloured according to a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

The interface between TK subunits contains several cavities, one of which forms a large solvent-filled channel running between the two TPP molecules. This channel contains several conserved glutamate residues as well as a hydrogen bonding network connecting the N1' atom of the pyrimidine ring of one TPP molecule with the corresponding nitrogen atom of the second TPP molecule [3].

The hydrogen-bond network incorporates the conserved residues Glu411, Glu160 and Glu165. Mutation of the residue corresponding to Glu160 to glutamine or alanine in yeast disrupts the hydrogen-bonding network, resulting in dimer destabilisation [8]. These side-chains are buried deep within the protein with no compensating positive charges in the region. Thus these residues would be protonated even at a physiological pH. The function of this hydrogen-bonding network has been postulated to have a role in the negative cooperativity of TPP binding to the two active-sites [9], although this has never been shown.

1.1.3 The TPP binding fold

Many important enzymes require TPP as a cofactor for catalysis. "TPP-dependent enzymes" are involved in various types of reaction *in vivo*, such as nonoxidative decarboxylations of α-keto acids (α-ketoacid decarboxylases), oxidative decarboxylations of α-keto acids (pyruvate oxidase (PO)), carboligation (TK, α-keto acid decarboxylases, acetolactate synthase (ALS)) as well as cleavage of C-C bonds (TK and benzaldehyde lyase (BAL)) [10] (Section 1.2). Comparison of the three-dimensional structures of TK [7,11], PO [12] and pyruvate decarboxylase (PDC) [13] have shown that their conformation and modes of TPP binding are very similar, despite considerable differences in primary structure [12].

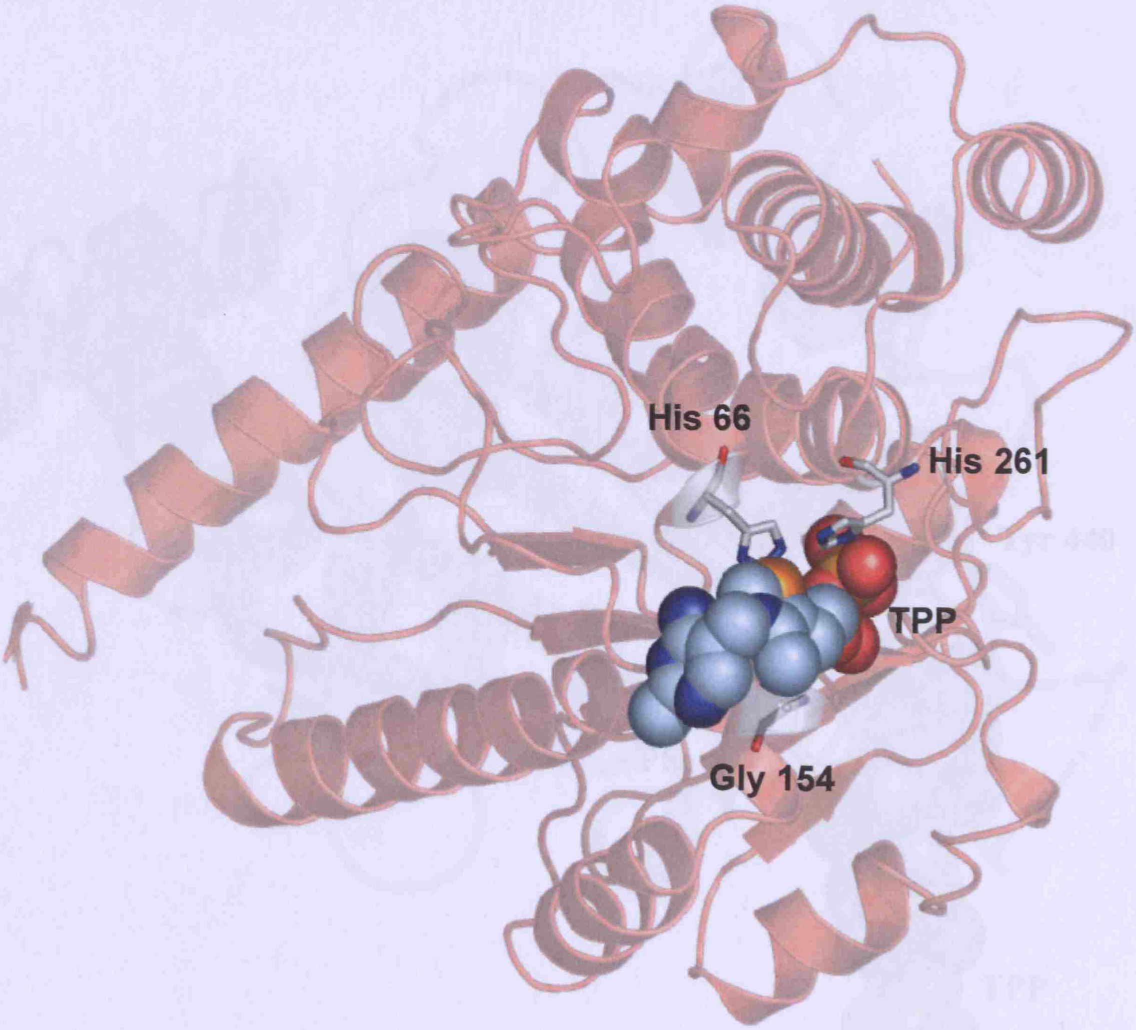


Figure 1.4: TPP binding fold of the PP domain. The proposed roles of residues His66, Gly156 and His261 are discussed in Section 1.1.3. The TPP cofactor is coloured according to a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

In TK, the diphosphate group of TPP (Figure 1.4) is bound at the switch point between strands 1 and 3 at the amino terminus of a small helix in the PP domain. Once bound, TPP is contorted into a V shaped tautomer, most likely required for catalysis [14]. Residues His 66, His261 and Gly154 form hydrogen bonds with the diphosphate group of TPP (Figure 1.4). Indirect interactions with the divalent metal cation also stabilise the diphosphate group, as discussed later in this section.

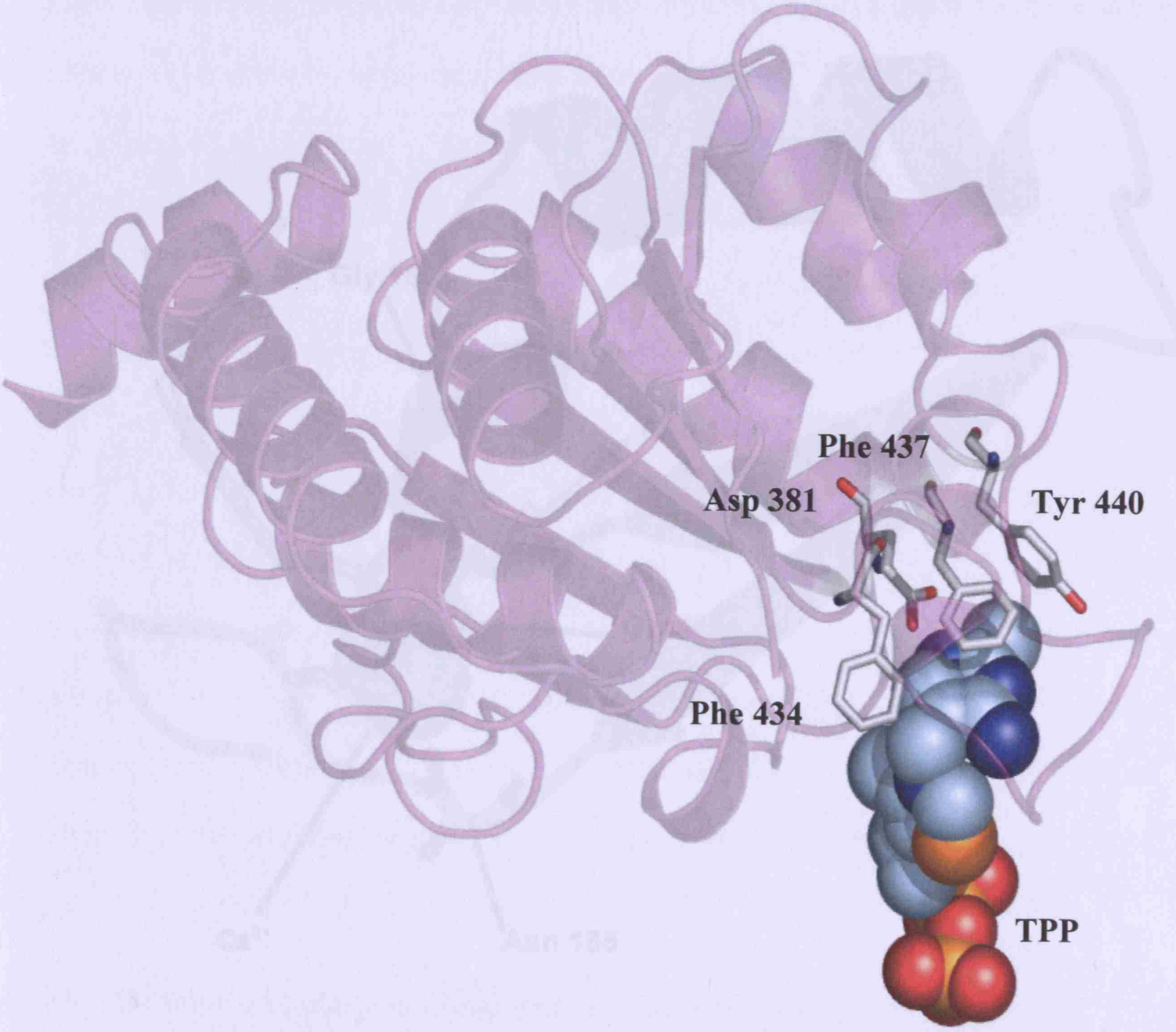


Figure 1.5: TPP binding fold of the Pyr domain. The proposed roles of residues Asp381, Phe434, Phe437 and Tyr440 are discussed in Section 1.1.3. The TPP cofactor is coloured according to a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

The methylthiazolium (MT) ring of TPP is found in the cleft between the PP domain of one subunit and the Pyr domain of the second, interacting with residues from both. It has been proposed that the highly conserved Asp381 stabilises the positive charge of the MT ring in TPP [8]. The position of Asp381 in the *Eco*TK crystal structure is shown in Figure 1.5.

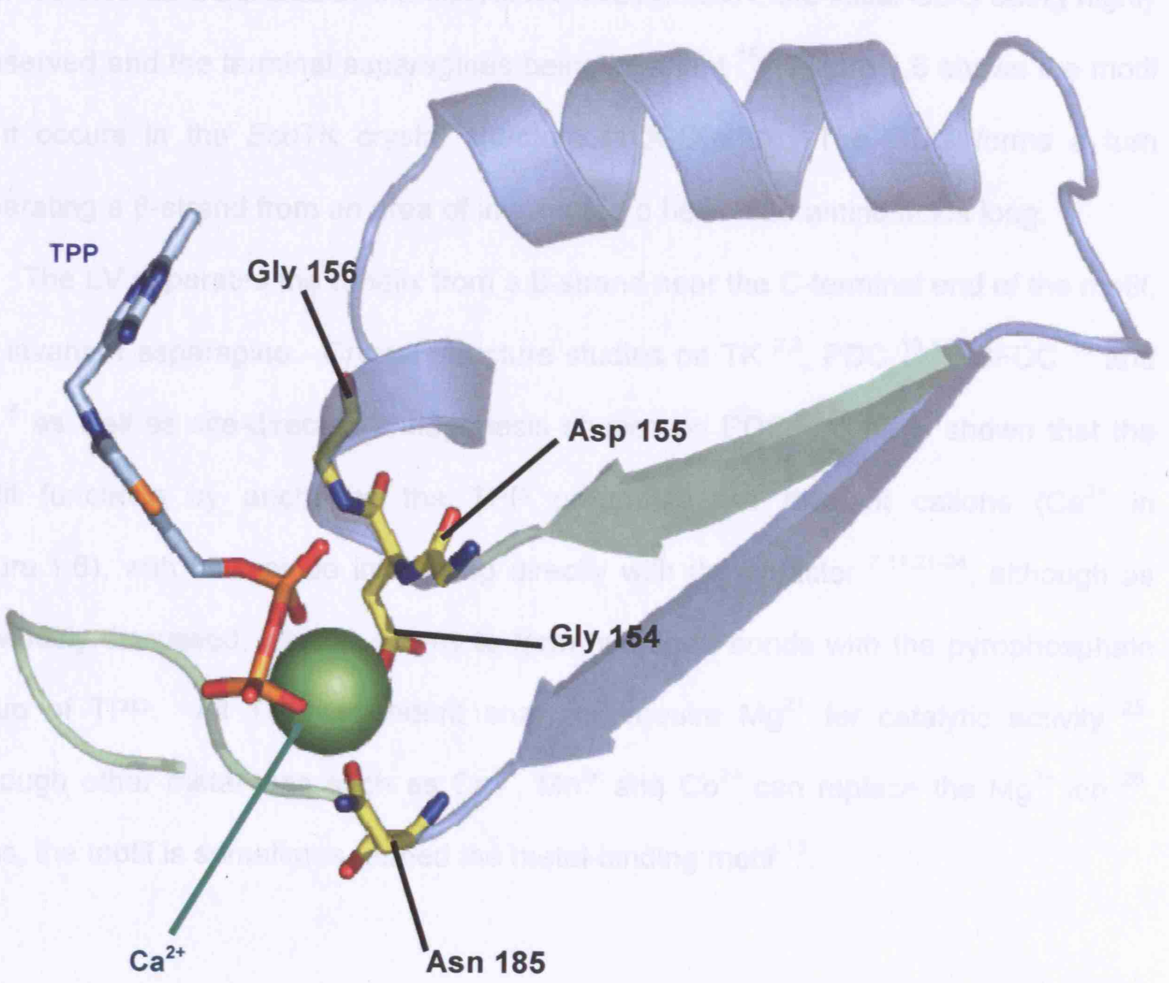


Figure 1.6: The βαβ metal-binding motif of transketolase. Highlighted residues are those found to be highly conserved among all TPP-dependent enzymes and are implicated in metal binding (Section 1.1.3). Regions of secondary structure forming the motif are coloured in blue. The TPP molecule is shown with a CNOS colouring scheme, while residues highlighted as sticks are shown with a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

The methylaminopyrimidine (MAP) ring of TPP binds in a hydrophobic pocket composed of aromatic residues from the Pyr domain in (Figure 1.5), namely Phe434, Phe437, and Tyr440.

A sequence motif common to all TPP-dependent enzymes [15] has been proposed as having a role in TPP binding. The consensus sequence of this βαβ motif in TK has been defined as GDGxxxEGxxxExxxxAxxxxLxxLVxxxDxN, the initial GDG being highly conserved and the terminal asparagines being invariant [16]. Figure 1.6 shows the motif as it occurs in the *Eco*TK crystal structure (1QGD.pdb). The GDG forms a turn separating a β-strand from an area of interrupted α helix ~ 20 amino acids long.

The LV separates the α-helix from a β-strand near the C-terminal end of the motif, the invariant asparagine. Crystal structure studies on TK [2,3], PDC [13,17], BFDC [18] and PO [6] as well as site-directed mutagenesis studies on PDC [19,20] have shown that the motif functions by anchoring the TPP group via the divalent cations (Ca^{2+} in Figure 1.6), with no residue interacting directly with the cofactor [7,11,21-24], although as previously discussed, Gly154 seems to form hydrogen bonds with the pyrophosphate group of TPP. All TPP-dependent enzymes require Mg^{2+} for catalytic activity [25], although other metal ions such as Ca^{2+}, Mn^{2+} and Co^{2+} can replace the Mg^{2+} ion [26]. Thus, the motif is sometimes termed the metal-binding motif [15].

1.1.4 Substrate binding and catalysis in transketolase

It is suggested that the invariant residues His66 and His100 bind the C-1 hydroxyl group of the donor substrate in TK and stabilise the reaction intermediate [27,28]. This may explain why β-HPA is a substrate of TK and pyruvate (which differs from β-HPA by one hydroxyl group) is not.

The highly conserved Asp469 is thought to form a hydrogen bond with the C-2 hydroxyl group of the substrates [29], with donor substrates of the D-threo configuration and acceptor substrates of the C2 D-configuration preferred [23], indicating that Asp469 may be a determinant of enantioselectivity [30].

The proximity of Arg358, Arg520 and His461 to the phosphate group of E4P in the crystal structure of *Sce*TK and the observation that mutation of any of these residues to alanine results in TK mutants with elevated K_ms [29] suggests they are involved in binding of the phosphate groups of TK substrates, resulting in the observed preference for phosphorylated over non-phosphorylated substrates [31] (Figure 1.7).

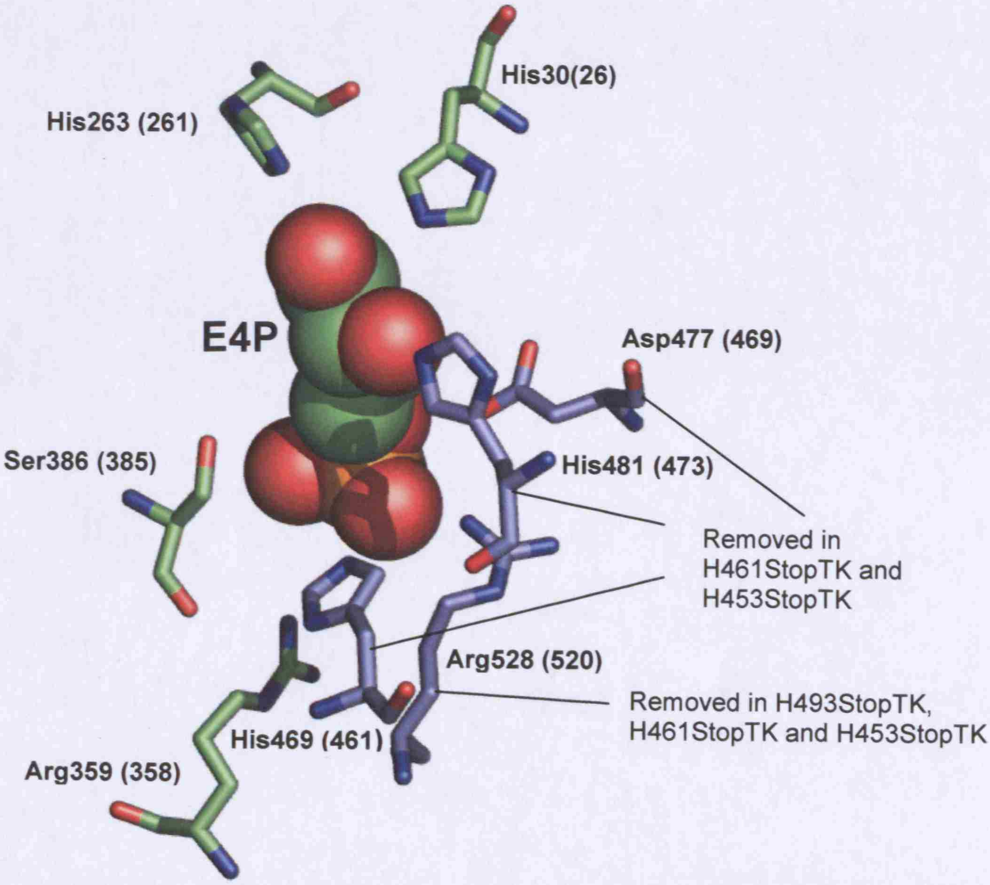


Figure 1.7: Residues believed to be involved in E4P binding in *Sce*TK. *Sce*TK numbering is used, with *Eco*TK numbering in brackets [8]. Residues in blue are removed by mutations indicated, and discussed in Chapter 6. Images were generated using the 1NGS.pdb file in *Pymol*.

The general scheme for the catalytic mechanism of TK is shown in Figure 1.8, as deduced for *Sce*TK [11], and *Eco*TK [32], and exhibits a bi-bi 'ping-pong' two-substrate mechanism, where the donor ketol substrate binds first. The mechanism of pyruvate decarboxylase is also discussed and compared with the TK scheme. The residues discussed as being catalytically important for TK and PDC are shown in Images A and B respectively in Figure 1.10. *Sce*PDC residue numbers are used throughout this section.

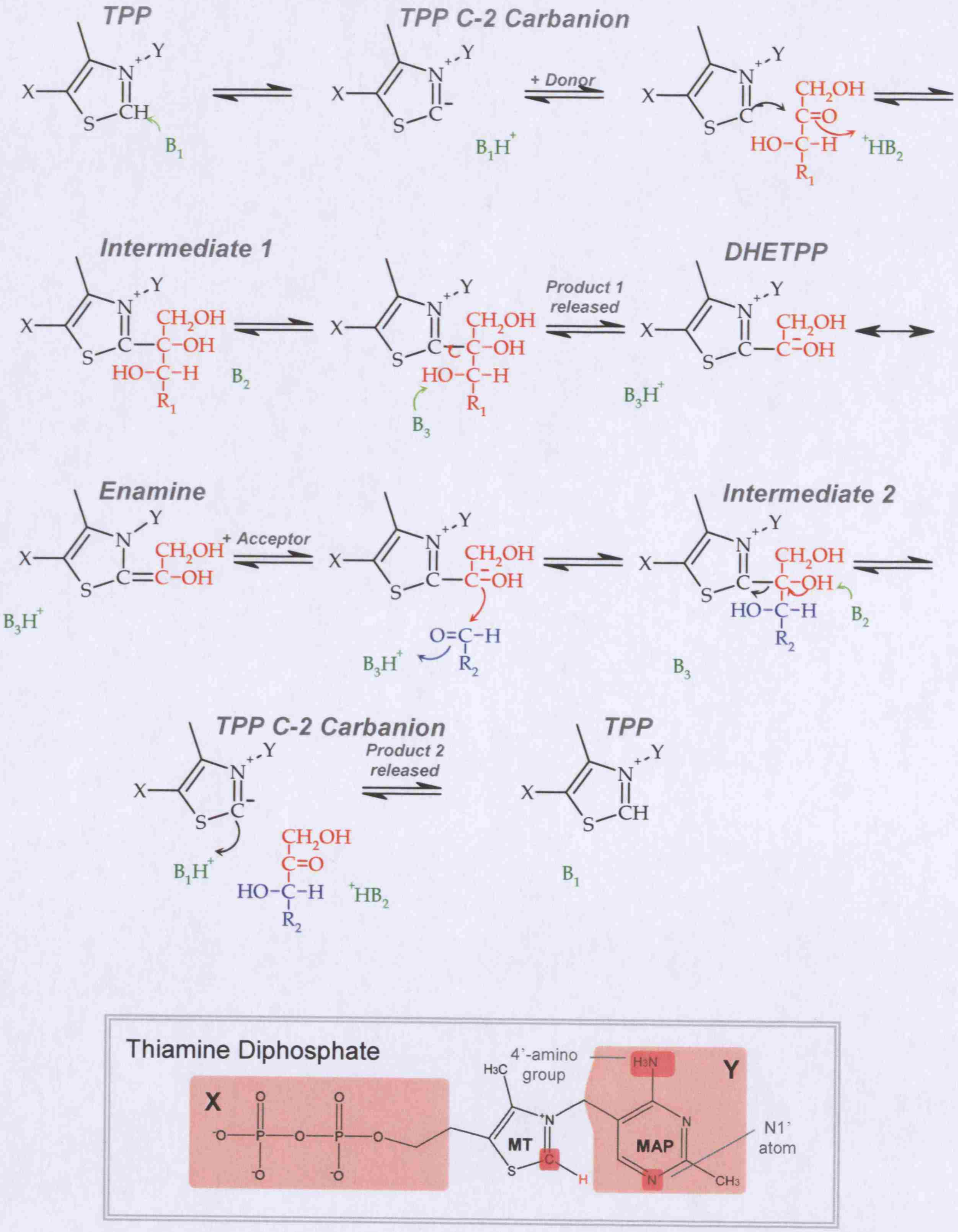


Figure 1.8: Reaction mechanism of transketolase. The donor substrate (a ketose) and the acceptor substrate (an aldose) are coloured red and blue respectively. B1 is most likely to be the 4'-imino group of the TPP pyrimidine ring. This group is also a candidate for B2; His473 is another possibility. His26 and His261 are the two candidates for B3. In the mechanism diagram, the TPP molecule is simplified. regions labelled X and Y are shown in the full molecule diagram for TPP in the grey box. Also highlighted are the N1' atom, the 4'- amino group and the C-2 atom of the MT ring. Adapted from Schneider and Lindqvist [7].

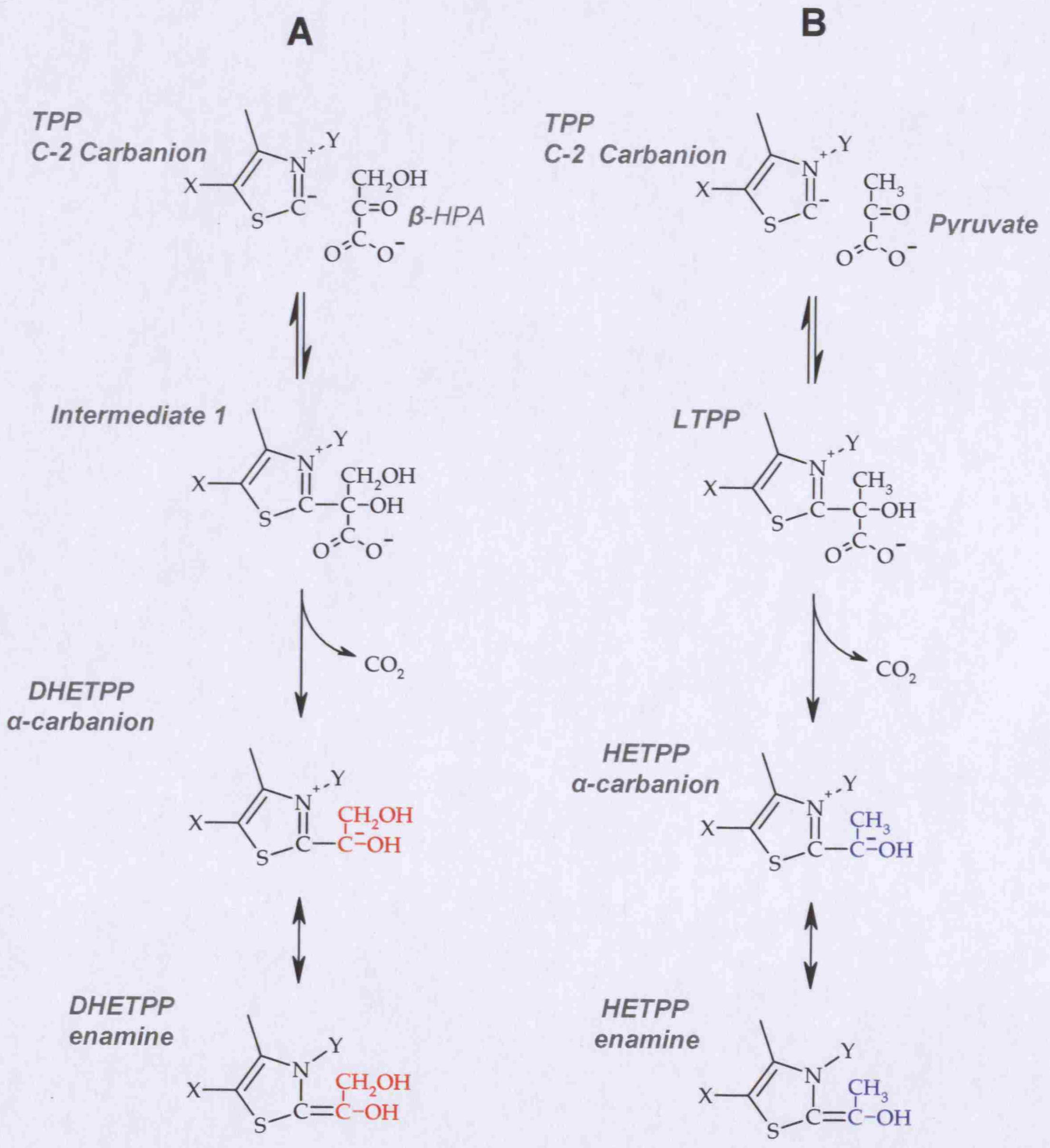


Figure 1.9: Formation of α-carbanion / enamine compounds from TPP and (A) β-HPA or (B) pyruvate. The α-carbanion/enamine formed in (A) is 1,2-dihydroxyethyl TPP (the 1,2-dihydroxyethyl group is coloured red) and in (B) is 1-hydroxyethyl TPP (the 1-hydroxyethyl group is coloured blue). The X and Y regions of the TPP molecule are as described in Figure 1.8. The presence of either DHETPP or HETPP as the α-carbanion intermediate determines whether a transferase or a decarboxylase reaction occurs in TK and PDC.

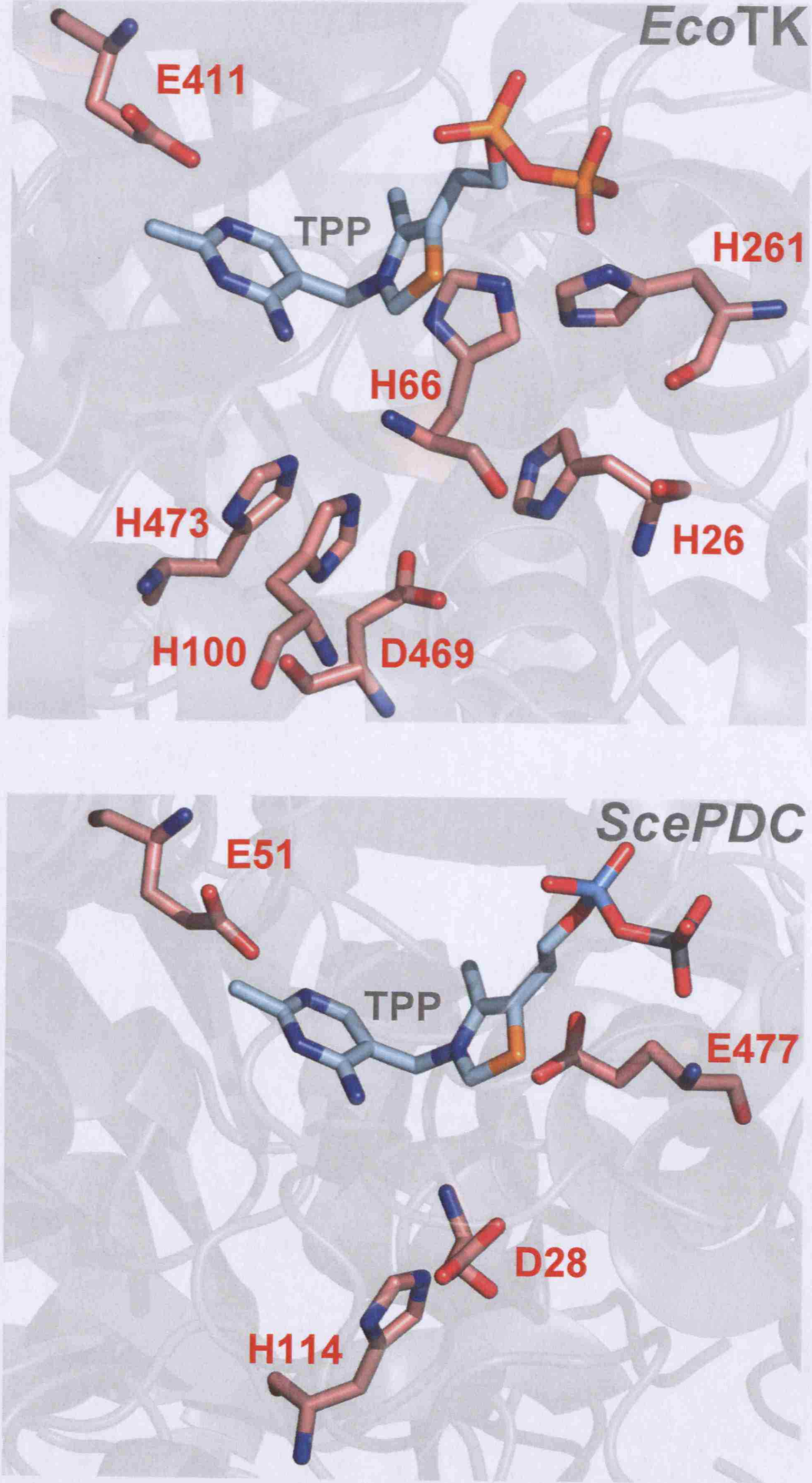


Figure 1.10: Residues implicated in TK and PDC catalysis. Image A is the structure of *Eco*TK while Image B is *Sce*PDC. Residues highlighted as sticks are those discussed as being catalytically important in Section 1.1.4. Images were generated using the 1QGD.pdb file in *Pymol*.

The first step of catalysis by TPP-dependent enzymes involves the activation of TPP by deprotonation of the C-2 atom (Figure 1.8). A proton relay system involving the 4'-amino group and the N1' atom of TPP as well as a universally conserved glutamate residue (Glu411 in *Eco*TK, Glu51 in *Sce*PDC) mediate this fast deprotonation step [33], forming a resonance form of TPP, the TPP C-2-carbanion, with a positively charged imino group at the 4' position. It had previously been suggested that in TK this imino-group may in turn be deprotonated by the neighbouring His473, resulting in an elevated pKa for the imino-group and the subsequent abstraction of the proton from C-2 of the thiazolium ring. This model for the deprotonation of C-2 cannot be applied to all TKs however since mammalian TKs contain a glutamine at this position. A human TK mutant with the glutamine replaced by histidine was generated by Singleton *et al.* [34], yielding a 70% decrease in activity relative to human wild-type TK. Thus the same amino acid position is likely to serve different roles in mammalian and non-mammalian TKs. This is however speculative, since as of yet no mammalian TK has been crystallised. In PDC from *Zymomonas mobilis*, it has been shown that the residues corresponding to *Sce*PDC His114 is essential for catalysis, but is not involved in the deprotonation step [35].

Studies in recent years have used ^{1}H-NMR methods to analyse the distribution of reaction intermediates in the TPP-dependent enzymes during catalysis [35], suggesting specific mechanistic roles for amino acids previously shown to affect catalytic rate.

As mentioned earlier in this section, the invariant His100, His66 and Asp469 residues may be necessary for the optimal positioning of the substrate for nucleophilic attack by the TPP C2-carbanion, by interaction with the hydroxyl groups of substrates in TK [27,28,29]. These residues are not conserved in PDC, where pyruvate lacks the C-1 hydroxyl group. In PDC another highly conserved glutamate residue Glu477 is implicated (along with Glu51) in the binding of substrate to the C2-carbanion, where it forms H-bonds with the oxygen of pyruvate as well as being involved in the subsequent decarboxylation of LTPP.

In TK, this nucleophilic attack, and protonation of the carbonyl oxygen of substrate 1, possibly by His473, leads to the formation of intermediate 1 (2-[2-(1,2,3,4,5-pentahydroxy)-pentyl]-TPP) [35] (Figure 1.8). Site-directed mutagenesis (SDM) of His481 caused a drastic decrease in the catalytic activity of TK [28]. The corresponding His114 residue in *Sce*PDC does not seem to be involved at this stage, where the previously mentioned Glu477 and Glu51 are implicated. In PDC, the proton relay system consisting of Glu51, the 4'-amino group and N1' atom of TPP, is implicated not only in the deprotonation and substrate binding steps of catalysis but also the cleavage of HETPP to free aldehyde.

In TK, deprotonation of the C-3-hydroxyl group of the substrate, possibly by His26 in concert with His261 [28], leads to the cleavage of the intermediate and release of product 1, an aldose sugar. Concomitantly, the α-carbanion of the intermediate is formed, α,β-dihydroxyethyl-TPP (DHETPP in Figures 1.8 and 1.9). In PDC, the Glu477 is implicated in the decarboxylation of LTPP to produce HETPP.

In PDC, acetaldehyde is released at this stage requiring the protonation of the α-carbanion and the deprotonation of the hydroxyl group of the HETPP. A catalytic dyad of His114 and Asp28 is implicated in the protonation of the α-carbanion, while the Glu477 is also involved in this step.

In the final catalytic step for TK, the DHETPP reacts with an aldose acceptor substrate to form intermediate 2 (2-[2-(1,2,3,4,5,6-hexahydroxy)-hexyl]-TPP) [35], before deprotonation allows release of the ketose sugar (product 2) and regeneration of the TPP.

Formation of a α-carbanion intermediate, where an aldehyde residue is bound to the C-2 of TPP is common to the catalytic mechanisms of all TPP-dependent enzymes.

While in TK, DHETPP is the intermediate, in PDC the intermediate is HETPP, which differs from DHETPP by a single OH group at the C-2 of the aldehyde residue. Due to the bi-bi ping-pong mechanism of TK catalysis (Figure 1.13), the absence of an acceptor substrate for the enzyme leads to the reaction halting at the DHETPP stage. When β-HPA is a substrate of PDC, the reaction can yield erythrulose, with DHETPP as intermediate. It has been demonstrated that TK can bind the HETPP as well, and in a manner similar to PDC produce free aldehyde [36].

Thus both TK and PDC can perform either a transferase or a decarboxylation reaction dependent on whether DHETPP or HETPP is bound in the intermediate stage. This explains why β-HPA, but not pyruvate can be used as an aldol donor in the TK transferase reaction, further discussed in Chapter 5.

Source	**Human**	***S. cerevisiae***	**Spinach**	***E. coli***
Xylulose 5-phosphate	0.49 mM	0.21 mM	not available	0.16 mM
Ribose 5-phosphate	0.53 mM	0.4 mM	0.4 mM	1.4 mM
Fructose 6-phosphate	7 mM	1.8 mM	3.2 mM	1.1 mM
Glyceraldehyde-3-phosphate	not available	4.9 mM	not available	2.1 mM
Erythrose 4-phosphate	0.36 mM	not available	not available	0.09 mM
β -HPA	no activity	33 mM	not available	18 mM

Table 1.1: K_m values for selected substrates of TK. Values represent mammalian, fungal, plant and bacterial enzymes. Data for the human enzyme were taken from Waltham [37]. All non human data are quoted from the collection of Sprenger and Pohl [10]. Spinach data is from Villefranca and Axelrod. [38]

1.1.6 Substrate specificity of transketolase

Substrate specificity is defined as the preference an enzyme expresses for one substrate over other competitor substrates. Since the binding of substrate is not sufficient for catalysis to occur, substrate specificity depends on both binding and catalytic turnover. TKs from yeasts, plants and bacteria have shown a broad range of substrate specificities. *Sce*TK can use the following sugars as donors: X5P, S7P, F6P and E4P as well as dihydroxyacetone phosphate (DHAP), dihydroxyacetone (DHA) and β-HPA. Acceptor substrates for SceTK include R5P, G3P, and GA [25,39].

TK from spinach leaves has a similar substrate repertoire to *Sce*TK [38] and in addition can catalyse the transferase reaction with non-phosphorylated acceptor sugars, when β-HPA is the ketol donor. Purified *Eco*TK exhibits similar substrate specificity to those reported for yeast and bacteria [31], while mammalian TKs are more specific, utilising only X5P, F6P and S7P as donors, and R5P, E4P, G3P and GA as acceptor substrates [40,41].

The *in vivo* concentrations of most TK substrates is 1-100 μM. However, the *K*m values for these substrates in most TK species are generally one order of magnitude larger (Table 1.1). Thus under physiological conditions, TK enzyme activity will be approximately proportional to substrate concentration.

The differences in substrate specificity between species is surprising given TKs important metabolic role and the high level of homology observed at the amino acid level. This observation makes TK a suitable candidate for protein engineering, since a small number of amino acid variations are likely to be responsible for altering the substrate specificity.

1.1.7 The use of transketolase in biocatalysis

Enzymes exhibit a high degree of regio- and stereospecificity, leading to their acceptance as catalysts in the fields of pure and applied chemistry, particularly in the pharmaceutical sector [10]. Biocatalysts minimise the problems of isomerisation, racemisation, epimerization and rearrangements common in chemical processes [42], while they also have the advantage that they can be recycled, reducing their environmental impact. The greatest driving force for the increased use of biotransformations in industry, particularly in the pharmaceutical sector, is the desire for optically-pure products.

The Food and Drug Administration (FDA), and equivalent agencies around the world, now demand pharmacological and toxicity data for each enantiomer of a chiral drug [43], providing pharmaceutical companies with a strong incentive to develop single enantiomer drugs Companies may also redevelop previously licensed racemic drugs as single enantiomers to extend their patent protection [44].

Chemical processes such as aldol condensation have the potential to produce similar chiral compounds to TK [45,46] but only when chiral auxiliaries are used. The expense and limited variety of these reagents make large-scale chemical processes more costly and less flexible than equivalent biotransformations.

During the past 15 years, the use of TK for generating asymmetric C-C bond synthesis has progressed greatly. The enzyme has been overexpressed in both *E. coli* and *S. cerevisiae*, providing sufficient quantities of the enzyme for industrial applications. The range of potential applications for TK has been expanded by widening the pool of successful substrates and exploring the synthetic opportunities of the products [7,11,21-24].

In order to achieve large-scale production using TK, the conditions which can affect the reaction need to be known, such as K_m for the substrates and the K_i for each of the products. Since TK has a bi-bi ping-pong mechanism, substrates can bind to the "incorrect" form of the enzyme, potentially causing substrate inhibition (Figure 1.11).

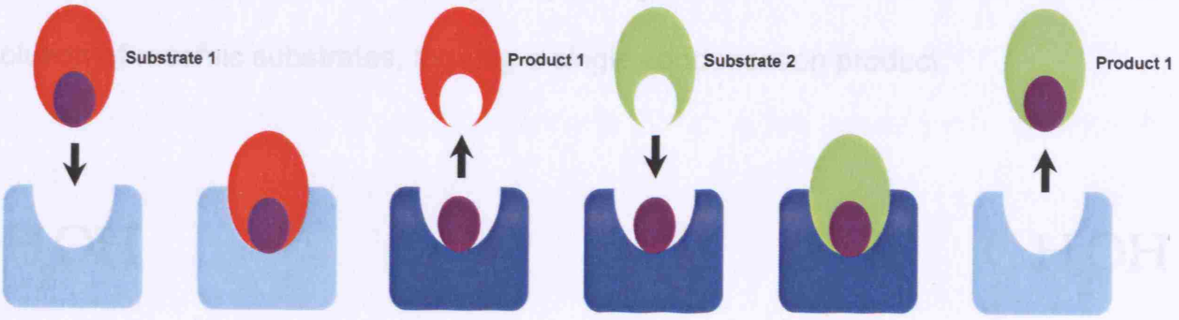


Figure 1.11: The mechanism of bi-bi ping-pong kinetics. The enzyme binds substrate 1 in the active-site. Product 1 is released, but in the process, a component of substrate 1 (purple dot) remains covalently bound to the enzyme, changing properties. This modified form of the enzyme is suggested by the darker shading in the figure. Substrate 2 binds to the modified enzyme (only this form) and reacts in the active-site with the fragment of substrate 1. Product 2 is finally released, and the enzyme returns to its normal form. Thus, substrate 1 and product 2 are in competition for the active-site of the enzyme in its normal form, while product 1 and substrate 2 are in competition for the altered form of the enzyme.

The transfer of a two-carbon ketol unit from X5P to R5P, generating S7P and G3P, catalysed by TK *in vivo* (Figure 1.1) has been used for industrial applications [47]. However, a complete reaction can be achieved more readily by the use of β-HPA as the ketol donor (Figure 1.12), resulting in the release of CO_2, essentially rendering the reaction irreversible. The reaction of β-HPA and GA to produce L-erythrose [21,48] has been adopted as the model TK reaction in this laboratory. The use of β-HPA at industrial scales is impractical due to its high price, as discussed in Chapter 5.

TK catalysed C-C bond formation is stereospecific, producing products with chiral centres in the (S)-configuration (the same configuration as C-3 of the natural donor substrate). The enzyme is also stereoselective, with a preference for α-hydroxyaldehydes with (R)-configuration at C-2. TK mediated condensations of α-substituted aldehydes with β-HPA produce enantiomerically pure chiral triols with D-threo stereochemistry. Stereospecificity is near total with *Sce*TK and *Eco*TK, and less so with other TKs such as spinach [47].

The selectivity of TK for (R)-α-hydroxyaldehydes means it is useful for the kinetic resolution of racemic substrates, forming a single condensation product.

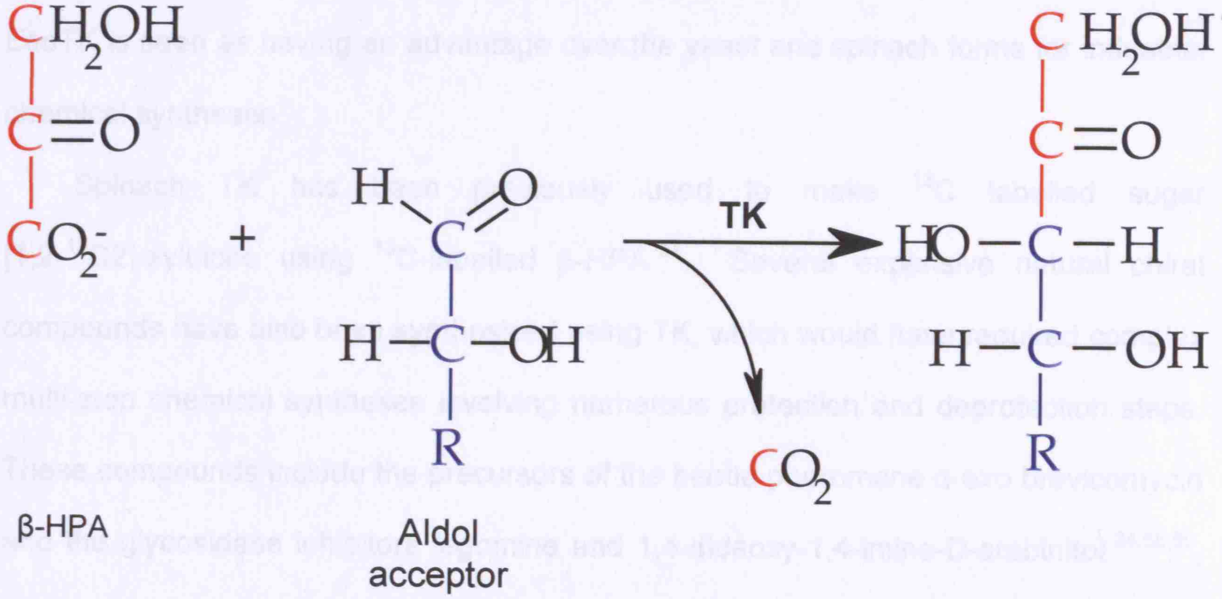


Figure 1.12: The TK reaction where β-HPA is the ketol donor. Liberation of CO_2 renders the reaction irreversible.

1.1.8 Applications of transketolase in organic synthesis

The ability of TK to utilise a wide variety of non-phosphorylated 2-hydroxyaldehydes *in vitro* as acceptor substrates, yielding ketoses with a 3*S*,4*R*-configuration [49] can be exploited to resolve racemates yielding L-hydroxyaldehydes with good yields [23,49,50]. TK also catalyses, reactions where the C2 of the acceptor substrate lacks an OH group, (but not pyruvate, as discussed in Chapter 5) although the rate is much lower than that for hydroxyaldehydes [51-54]. TKs from yeast and spinach have also been reported as using aromatic aldehydes, such as benzaldehyde and hydroxybenzaldehyde [51]. However, *Eco*TK does not favour aromatic aldehydes.

The ketol donor β-HPA is converted by *Eco*TK at a rate of 60 U/mg (1 U = the amount of enzyme activity that will catalyse the transformation of one micromole of the substrate per minute under standard conditions [55]), while the spinach and yeast forms of TK convert it at rates of 2 U/mg and 9 U/mg respectively [31,51,56]. For this reason, *Eco*TK is seen as having an advantage over the yeast and spinach forms for industrial chemical syntheses.

Spinach TK has been previously used to make ^{13}C labelled sugar [1,2-^{13}C2]-xylulose using ^{13}C-labelled β-HPA [57]. Several expensive natural chiral compounds have also been synthesised using TK, which would have required complex multi-step chemical syntheses involving numerous protection and deprotection steps. These compounds include the precursors of the beetle pheromone α-exo brevicomycin and the glycosidase inhibitors fagomine and 1,4-dideoxy-1,4-imino-D-arabinitol [24,58,59], all of which are chiral products of reactions involving racemic acceptor substrates and β-HPA.

TK has been used in multienzyme applications [47]. Spinach TK has been used in the production of 6-deoxy-L-sorbose, a precursor of the caramel flavoured furanol to 45 % yield. In this synthesis, β-HPA was produced from L-serine by the spinach serine glyoxylate aminotransferase. In another example, 4-deoxy-L-threose was produced by whole cells of *Comebacterium equi* or *Serratia liqufaciens* using 4-deoxy-L-erythrulose and β-HPA as TK substrates [60].

Other biotransformations using TK include: synthesis of the non-natural sugar 4-deoxy-D-fructose-6-phosphate to 52% yield using *Sce*TK [61]; X5P with 82% yield at the gram scale [62]; and 3-*O*-benzyl-D-xylulose, used in turn for the synthesis of the potential glycosidase inhibitor *N*-hydroxypyrrolidine [63].

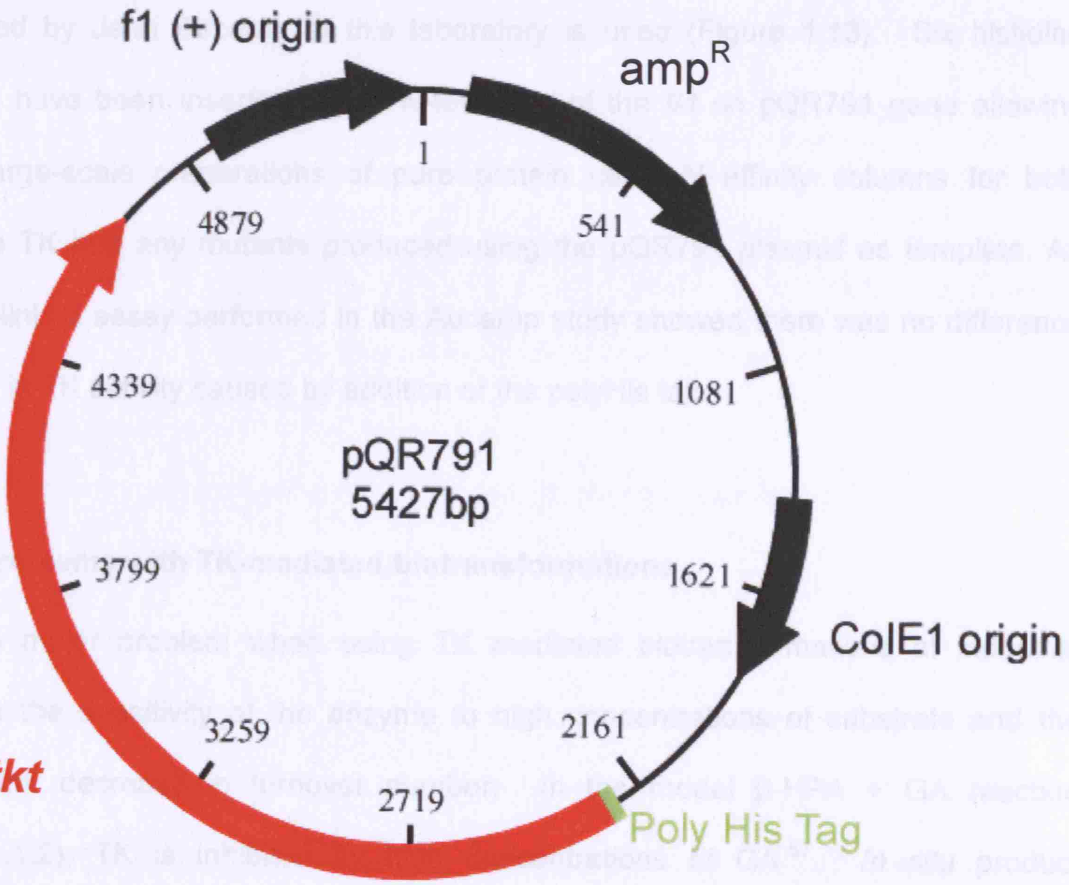


Figure 1.13: Map of the *E. coli* transketolase over expression vector pQR791. The orientation of each gene is represented by the direction of its arrow. amp^R is the ampicilin resistance gene and *tkt* is the transketolase gene, with the Poly histidine tag in green. Adapted from French and Ward [65].

1.1.9 Over expression of transketolase

The first uses of TK in biotransformations used low yields of enzymes from *Sce* and spinach. A high-yield recombinant expression system for TK was developed in the 1990s [64], where the *Eco*TK gene (*tkt*, 1991bp) was cloned on a 5kb fragment into the low copy number vector pBR325. Expression levels were improved by subcloning the *tkt* gene from this construct into various high copy-number expression vectors [65]. One such strain, *E. coli* JM107 pQR711, was grown to 20 g dry cell weight per litre, producing 4 kg of enzyme from a 1000 L glycerol-fed fermentation [66], thus permitting large scale TK biotransformations.

In this thesis a modified version of the pQR711 plasmid, the pQR791 plasmid developed by Jean Aucamp at this laboratory is used (Figure 1.13). Six histidine residues have been inserted at the N-terminus of the *tkt* on pQR791 gene allowing future large-scale preparations of pure protein using Ni-affinity columns for both *wild-type* TK and any mutants produced using the pQR791 plasmid as template. An enzyme-linked assay performed in the Aucamp study showed there was no difference between in TK activity caused by addition of the polyHis tail.

1.1.10 Problems with TK-mediated biotransformations

The major problem when using TK mediated biotransformations at industrial scales is the sensitivity of the enzyme to high concentrations of substrate and the consequent decrease in turnover number. In the model β-HPA + GA reaction (Figure 1.1.2), TK is inhibited by high concentrations of GA [67]. *In situ* product removal [68] and enzyme-membrane reactors [69] are engineering approaches which have been developed to keep the substrate and product concentrations low in TK biotransformations, ensuring that the $t_{½}$ of the enzyme and productivity of the biotransformation greatly increased. The possibility also exists of engineering a TK enzyme which can cope with higher substrate and product concentrations without the problematic decreases in turnover number.

1.1.11 Transketolase and Disease

TK has been implicated in a number of diseases. For example, reductions in TK activity have been implicated in cases of Chronic Uremia [70], similar to those caused by thiamine deficiency (beri beri). Modifications to TK appear in cultured dermal fibroblasts from Alzheimer's patients suggesting that this could be used as a marker for the disease [71].

TK has also been proposed as a potential target for anti-cancer treatments. Tumour cells heavily utilize the non-oxidative branch of the PPP for ribose production, and so inhibition of TK activity, or restriction of thiamine in the diet, could be used to modulate tumour cell growth. This hypothesis is supported by metabolic control analyses [72].

1.1.12 Summary of the important residues in transketolase and other TPP-dependent enzymes

Table 3.4 (Chapter 3) summarises the important residues in TK which are discussed during the course of this thesis. Each residue has a reference number to allow labelling of sequence alignments in Chapters 3, 4 and 5. The conservation of each residue among the different enzymes is summarised and refers to data collated in Chapter 3 (Section 3.4).

1.2 The TPP-dependent enzymes

TPP-dependent enzymes are involved in many metabolic pathways *in vivo*. Their common mechanisms of binding TPP and catalysis as well as high homology at the structural level suggest that these enzymes share a common evolutionary lineage. Many TPP-dependent enzymes have great potential in the realm of synthetic chemistry. An understanding of the evolution of structure, function and substrate specificity in the TPP-dependent enzyme family would be of interest for guiding the engineering of TPP-dependent enzyme activities.

Various TPP-dependent enzymes, that will be studied in subsequent chapters, are discussed below (TK is discussed in Section 1.1). Several enzymes are not discussed as they are excluded from our studies, for reasons given in Chapter 3 (Section 3.2).

1.2.1 Overview of the TPP-dependent enzymes

The TPP-dependent enzymes perform reactions involving both the synthesis and breaking of C-C bonds. Here, an introduction to each of the enzymes studied in this thesis is given. The potential and previous uses of various TPP-dependent enzymes are discussed where applicable (The industrial applications of TK were discussed in Sections 1.1.7 and 1.1.8). Using the tools of molecular biology as well as crystal structure information has the potential to modify the substrate repertoire of these enzymes and design new and more efficient chemical syntheses. Specifically, the application of knowledge of the TPP-dependent enzymes to the engineering of TK is of interest here.

1.2.1.1 D-xylulose 5 phosphate synthase

D-xylulose 5 phosphate synthase (DXPS) is widespread in eubacteria, plant chloroplasts as well as in green algae, where it is involved in the synthesis of DXP, an intermediate in the mevalonate independent synthetic pathway of isopentyl diphosphate (IPP), which in turn is used in isoprenoid synthesis. DXP is formed by a TK-like acyloin condensation of the (hydroxyethyl)thiamine derived from pyruvate with the C1 aldehyde group of G3P [73-76].

Synthetic applications of DXPS include the single pot multienzyme synthesis of the lithium salt of DXP [77], where *Eco*DXPS, along with fructose-1,6-bisphosphate aldolase from rabbit muscle (RAMA) and Triose phosphate isomerase (TPI) from *E. coli* were used.

1.2.1.2 Dihydroxyacetone synthase

Dihydroxyacetone synthase (DHAS), or formaldehyde TK is involved in the xylulose monophosphate shunt, where it uses formaldehyde. The enzyme shares high homology with TK and is found in methylotrophic yeasts [78], and some bacteria. Given the high degree of sequence homology with TK, it is not surprising that DHAS can also use β-HPA as a donor substrate. DHAS has been used in industry to generate [1,3-^{13}C]-DHA. [1,3-^{13}C]-DHAP can be used as a precursor in the production of labelled sugars with a DHAP-dependent enzyme such as RAMA [77].

1.2.1.3 Phosphoketolase

Phosphoketolase (PKL) is involved in the pathways of glucose fermentation and glucose heterofermentation which yield lactate and ATP. In bacteria, the PKLs [79] catalyse the irreversible splitting of F6P and inorganic phosphate to form E4P and acetylphosphate or similarly the splitting of X5P and inorganic phosphate to yield G3P and acetylphosphate. Little is known about the side reactions of the phosphoketolases and their synthetic potential has not been investigated [80]. Certain forms of phosphoketolase have been reported as using β-HPA and GA, in a study where arsenate was used instead of phosphate, yielding acetate [81].

1.2.1.4 2-oxoisovalerate dehydrogenase

2-oxoisovalerate dehydrogenase (2OXO) acts on 2-keto-3-methyl-valerate in the isoleucine degradation pathway, about which not much is known. 2OXO also acts on 3-methyl-2-oxobutanoate and 4-methyl-2-oxopentanoate [82]. The crystal structure of 2OXO has been solved at the 2.8Å level (2BP7.pdb).

1.2.1.5 Pyruvate Decarboxylase

Pyruvate Decarboxylase (PDC) is a key enzyme in the fermentation of alcohol and has been isolated from plant, bacteria and yeast sources. PDC catalyses the non-oxidative conversion of pyruvate (or other 2-oxo acids) to acetaldehyde and CO_2 [83]. All PDCs thus far described for yeast and bacteria are tetrameric, while it has been proposed that plants can adopt higher oligomeric states.

Most studies of the substrate repertoire and the reaction mechanism of PDC have been performed on the *Sce*PDC and PDC from *Z. mobilis* PDC forms [84,85]. The crystal structure of PDC has also been solved for both *Sce*PDC (1PVD.pdb and 1PYD.pdb) and *Zmo*PDC (1ZPD.pdb) As in TK, there are differences in substrate specificity between species. When compared with PDC from yeast, the *Z. mobilis* PDC uses only unbranched aliphatic and *Ppu*PDC only unbranched aromatic substrates. The *Z. mobilis* PDC only uses C-4 (α-keto butanoic acid) and C-5 (α-keto pentanoic acid) keto acids [84,86]. *Ppu*PDC requires substrates with an aromatic ring connected directly to the α-carbonyl group.

The most intensively studied of these reactions is the production of (R)-1-hydroxy-1-phenyl-2-propanone ((R)-PAC), a precursor of ephedrine. R-PAC is produced in gram quantities by PDC using pyruvate and benzaldehyde. *Zmo*PDC is more stable than other forms of PDC at room temperature ($t_{½}$ >> 100 h), so it is perhaps the most suitable for production of R-PAC and activity has been enhanced by site-directed mutagenesis [85,87,88].

A point of interest is that the reaction of acetaldehyde with either pyruvate or glyoxylate resulted in different stereoisomers depending on which species' PDC was used [89-91].

1.2.1.6 Indolepyruvate Decarboxylase

Indolepyruvate decarboxylase (IPDC) is the key enzyme in the pathway leading from tryptophan to indole acetic acid (IAA) The enzyme is a homotetramer in the presence of TPP and Mg^{2+}, as confirmed by the crystal structure of *Ecl*PDC (1OVM.pdb). IPDC has high specificity and affinity only for its *in vivo* substrate indolepyruvate [92].

IPDC has been cloned from the plant associated bacterium *Enterobacter cloacae* [93,94] and expressed in *E. coli* cells. It can also decarboxylate pyruvate, but the activity is only 19 % of that when indolepyruvate is the substrate. IPDC seems to be highly specific, being inactive with indole-3-lactic acid and oxaloacetic acid [80] and has to date not been used for synthetic purposes.

1.2.1.7 Phenylpyruvate Decarboxylase

Phenylpyruvate decarboxylase (PhPDC) catalyses the decarboxylation of phenylpyruvate to phenylacetaldehyde in the anaerobic metabolism of L-phenylalanine to benzoyl-CoA [95]. PhPDC has also been shown to be active on indolepyruvate and is also involved in tryptophan metabolism [95].

PhPDC is the least studied among the TPP-dependent non-oxidative α-keto acid decarboxylases. *In vivo* the enzyme is inducible, being called into action when organisms were grown on phenylalanine, tryptophan or mandelate. The enzyme can use straight chain α-keto acids with lengths of six or more carbons.

PhPDC can catalyse the asymmetric acyloin condensation of phenylpyruvate and acetaldehyde to produce 3-hydroxy-1-phenyl-2-butanon [96].

1.2.1.8 Sulfopyruvate decarboxylase

Sulfopyruvate decarboxylase (SPDC) was the first enzyme involved in the coenzyme M biosynthetic pathway to be described. Coenzyme M is essential as a cofactor in methanogenesis and aliphatic alkene metabolism [97].

SPDC appears to be a TPP-dependent enzyme based on sequence homology with other members of the TPP-dependent family, having as it does the TPP motif [16] as well as the nature of the reaction it catalyses. As well as requiring TPP, SPDC also requires FAD, along with other TPP family members such as PO (Section 1.2.1.11).

1.2.1.9 Phosphonopyruvate decarboxylase

Phosphonopyruvate decarboxylase (PPDC) is involved in synthesising the 2-aminoethylphosphonate functional group of polysaccharide B, a component of the capsular polysaccharide complex in *Bacteroides fragilis.* The enzyme is promiscuous however, using pyruvate and sulfopyruvate at low levels [98].

1.2.1.10 Benzoylformate Decarboxylase

Benzoylformate decarboxylase (BFDC) catalyses the conversion of benzoylformate to benzaldehyde and carbon dioxide in the mandelate degradation pathway. This pathway is particularly important in several pseudomonads, which can use mandelic acid as a sole carbon source.

BFDC has been used previously to form α-hydroxyketones such as (*S*)-2-hydroxypropiophenone (2-HPP) from benzoylformate and acetaldehyde [99]. BFDC can use aldehydes that have not been previously decarboxylated in place of the more expensive α-keto acids [100]. *Ppu*BFDC is preferred since it works optimally within a broad pH range (5 - 8) and broad temperature range (20-40 °C). The crystal structure of *Ppu*BFDC has also been solved (1MCZ.pdb). The same enzyme can produce (*R*)-benzoin with high enantiomeric excess [10]. PDC and BFDC may complement each other as catalysts for organic syntheses [10], due to their having opposing stereoselectivity.

1.2.1.11 Pyruvate Oxidase

Pyruvate Oxidase (PO) is a membrane bound homotetrameric enzyme that catalyses the oxidative decarboxylation of pyruvate to form acetate and CO_2, requiring FAD, as well as TPP for catalysis [101]. The crystal structure of *Lpl*PO is known (1POW.pdb and 1POX.pdb).

The metabolic function of this reaction remains poorly understood, although it has been suggested that PO function in bacteria is growth phase dependent and involved in the switch from aerobic to anaerobic growth [102-104].

PO is activated by neutral lipids, specifically palmitic acid [105]. The enzyme appears to be coupled with the electron transport chain. The binding of C12 – C20 lipids has been shown to be important *in vivo* [106] and appears to affect the rate of flavin mediated electron transfer [107-109].

1.2.1.12 Acetolactate synthase

ALS is a TPP-dependent enzyme involved in the biosynthesis of isoleucine. ALS activity is required for the biosynthesis of alpha-aceto-alpha-hydroxybutyrate for the isoleucine pathway and for the biosynthesis of alpha-acetolactate for the valine pathway [110]. The crystal structure of ALS in complex with various sulfonylureal herbicides has been solved for *Ath*ALS (1YBH.pdb, 1YHY.pdb, 1YHZ.pdb, 1YIO.pdb, 1YI1.pdb and 1Z8N.pdb) and *Sce*ALS (1T9A.pdb, 1T9B.pdb, 1T9C.pdb and 1T9D.pdb).

The two major reactions of ALS are the condensation of pyruvate, either with itself to form α-acetolactate, or with α-ketobutyrate to form α-aceto-α-hydroxybutyrate. ALS II from *Salmonella typhimurium* also catalyses the self-condensation of α-ketobutyrate to yield α-propio-α-hydroxybutyrate and CO_2 at a rate 20 % that of the pyruvate self-condensation. The α-ketobutyrate self-condensation reaction was not observed with the ALS III from *E. coli* [93].

1.2.1.13 Glyoxylate Carboligase

Glyoxylate carboligase (GXC) is found in proteobacteria, where it is involved in both the glycolate and allantoin degradation pathways. *E. coli* can utilise either of the two carbon compounds glycolate or glyoxylate as sole carbon sources. Glycolate undergoes oxidation to yield glyoxylate, two molecules of which are condensed by glyoxylate carboligase to give tartronate semialdehyde. The enzyme requires both TPP and FAD for catalysis, though the need for FAD is unclear since the condensation reaction is non-oxidative and doesn't appear to require the oxidation-reduction process [111,112].

The GXC reaction is analogous to that of ALS. GXC, ALS and PO belong to a family of enzymes with the common feature that they require FAD as an additional cofactor. Both PO and GXC are able to catalyse the ALS reaction at a very low rate.

1.2.1.14 Benzaldehyde Lyase

Benzaldehyde Lyase (BAL) catalyses the lysis of benzoin into two benzaldehyde molecules. BAL has been used previously to form (*R*)-2-HPP at over 95 % yields as well as highly enantiomerically pure (*R*)-benzoin [113].

1.2.1.15 Oxalyl CoA Decarboxylase

Oxalyl CoA decarboxylase (OCADC) is a key enzyme in the catabolism of the highly toxic compound oxalate, catalysing the decarboxylation of oxalyl CoA to formyl CoA. OCADC has been isolated from *Bifidobacterium lactis*, a culture found in probiotic foodstuffs and is believed to have a role in oxalate degradation in the intestines. The crystal structure of *Oxo*OCADC is known (3C3I.pdb).

1.2.1.16 Pyruvate Ferredoxin Reductase

Pyruvate ferredoxin reductase (PFRD) catalyses the eleventh and final step in the fermentation of glucose to acetate. The arrangement of the various domains of the enzyme can vary (Section 1.2.2). At present, only the crystal structure for *Daf*PFRD has been solved (1BOP.pdb).

1.2.2 Domain arrangements in the TPP-dependent enzyme family

Where it is an enzyme cofactor, TPP is bound by the homologous PP and Pyr domains, both of which are common to all TPP-dependent enzymes described (Sections 1.1.2 and 1.1.3). Within the family of TPP-dependent enzymes the order in which these domains occur in the amino acid sequence varies. Here, the TPP-dependent enzymes are classified into 7 groups (Figure 1.14).

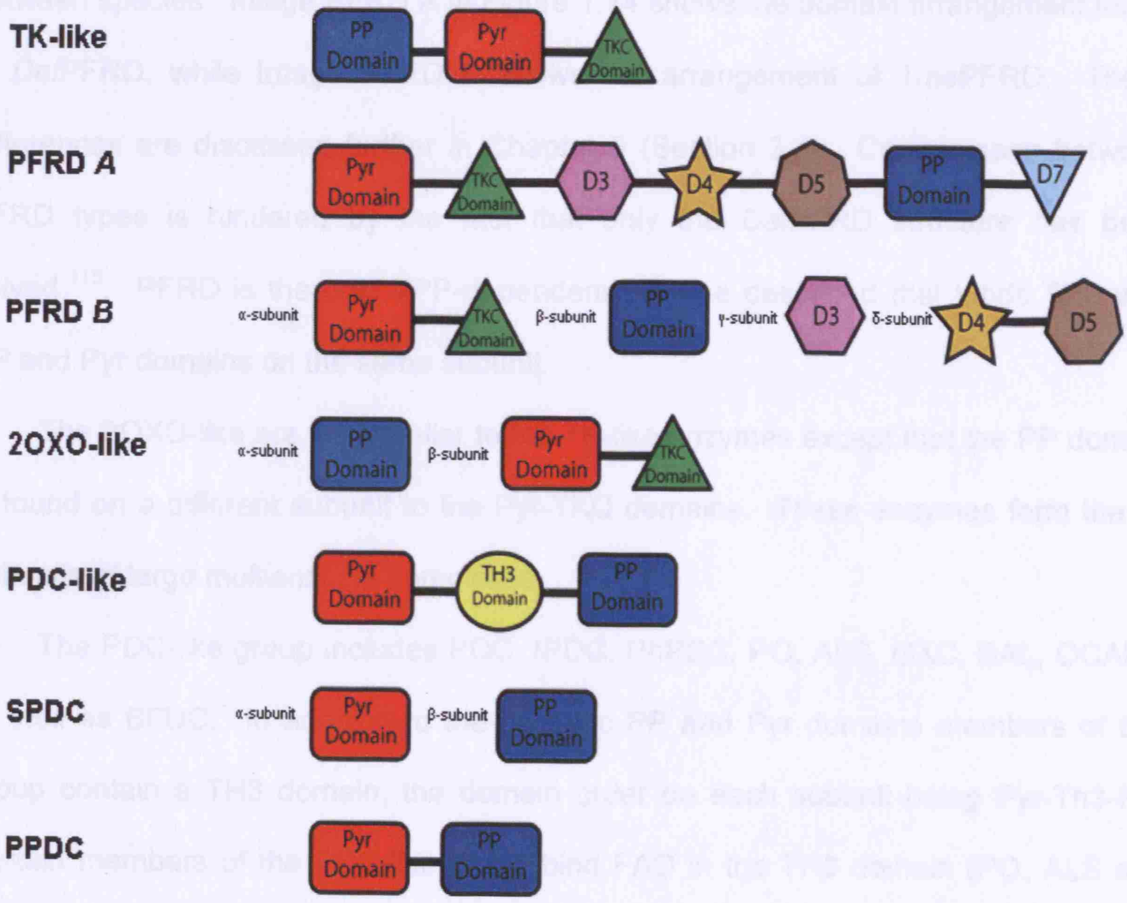


Figure 1.14: The domain arrangements found in the TPP-dependent enzyme family. PFRD Image *A* shows the arrangement of the PFRD type found in *Daf*-like PFRDs, while Image *B* Shows the arrangement found in the *Tma*-like PFRDs, as discussed in Section 1.2.2.

Enzymes of the TK-like group (TK, DXPS, DHAS, and PKL) contain three domains on the same subunit in the order PP, Pyr and TKC, in the same arrangement as found in TK (Section 1.2.2). The TK-like enzymes are homodimeric, although it has been reported that the PKLs adopt a homohexameric arrangement [114].

PFRD enzymes contain the PP and Pyr, the TKC as well as three additional domains domains, designated D3, D4 and D5. D4 is a ferredoxin-like domain, which binds 2 [4Fe-4S] clusters. The presence of these additional domains suggests a complex evolutionary past for PFRD. Indeed PFRD domain arrangements can differ between species. Image PFRD A in Figure 1.14 shows the domain arrangement found in *Daf*PFRD, while Image PFRD B shows the arrangement of *Tma*PFRD. These differences are discussed further in Chapter 3 (Section 3.4). Comparisons between PFRD types is hindered by the fact that only the *Daf*PFRD structure has been solved [115]. PFRD is the only TPP-dependent enzyme described that binds TPP with PP and Pyr domains on the same subunit.

The 2OXO-like are very similar to the TK-like enzymes except that the PP domain is found on a different subunit to the Pyr-TKC domains. These enzymes form the E_1 subunits of large multienzyme complexes.

The PDC-like group includes PDC, IPDC, PhPDC, PO, ALS, GXC, BAL, OCADC as well as BFDC. In addition to the catalytic PP and Pyr domains members of this group contain a TH3 domain, the domain order on each subunit being Pyr-Th3-PP. Certain members of the PDC-like group bind FAD in the TH3 domain (PO, ALS and GXC), possibly for structural reasons (Section 1.2.1.11). The PDC-like enzymes adopt a homotetrameric quaternary structure.

The two final domain arrangements discussed are those found in SPDC and the PPDC. SPDC and PPDC enzymes are very similar, sharing high sequence homology [97]. These represent the simplest domain arrangements found in the TPP-family, both consisting of PP and Pyr domains only. In SPDC, the PP and Pyr domains are found respectively on the α- and β-subunit, which adopt an $\alpha_2\beta_2$-heterotetrameric arrangement. In PPDC, the PP and Pyr domains are fused, with PPDC adopting a homotrimeric arrangement [98].

In all cases previously examined the actual TPP-binding unit is highly conserved [12] (Section 1.1.3). It has been suggested that the PP and Pyr domains may have independently fused onto the same chain after divergence from their ancestral domain [116]. The various possible mechanisms of protein function evolution are discussed in Section 1.3.

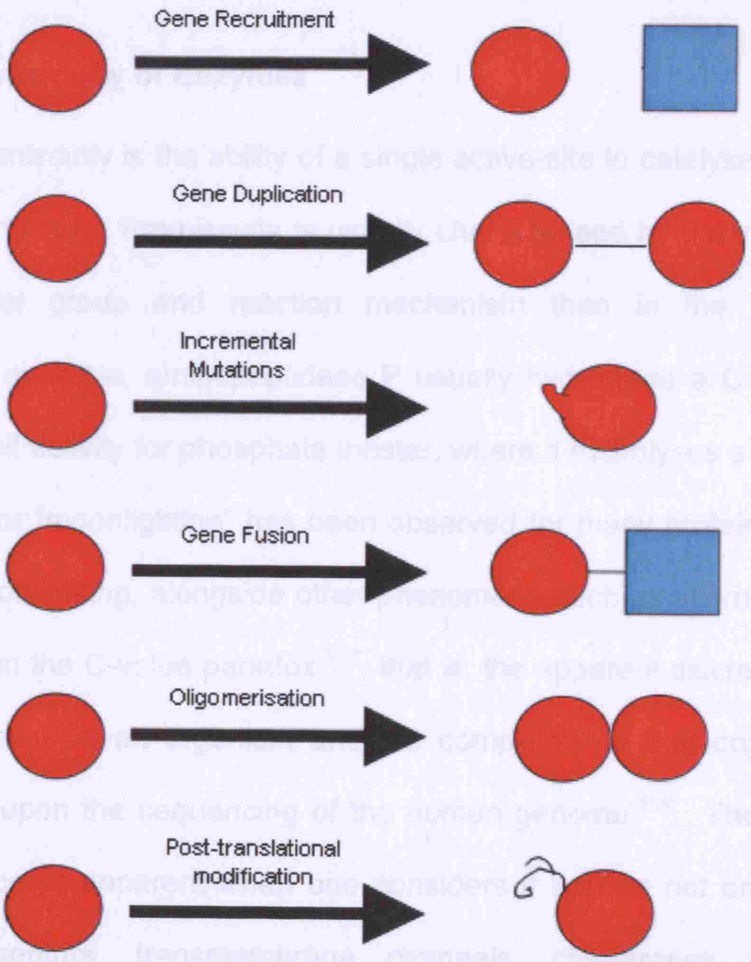


Figure 1.15: Different modes of functional divergence in proteins. The red circle represents a protein, which through gene duplication, incremental mutations, oligomerisation or post-translational modification can evolve new functionality. The blue box represents a second protein which interacts with the first by either gene recruitment or gene fusion, providing the context for new protein functions to evolve. Adapted from Todd et al. [116].

1.3 Mechanisms of protein evolution

Routes towards evolution of new functions in proteins include gene duplication, incremental mutations, gene fusions, oligomerisation and post-translational modification (Figure 1.15). Such routes will be considered when examining the evolution of the TPP-dependent enzymes in Chapter 3. Enzyme promiscuity is also of great importance for the evolution of new enzyme function (Section 1.4).

1.4 Catalytic Promiscuity of Enzymes

Catalytic promiscuity is the ability of a single active-site to catalyse more than one chemical transformation. Promiscuity is usually characterised by the involvement of a different functional group and reaction mechanism than in the native enzyme reaction [117]. For example, aminopeptidase P usually hydrolyses a C-N amide bond, but can also exhibit activity for phosphate triester, where it hydrolyses a P-O bond [117].

Promiscuity or "moonlighting" has been observed for many proteins. It has been proposed that moonlighting, alongside other phenomena such as alternative splicing of mRNA may explain the C-value paradox [118], that is, the apparent discrepancy between the number of genes in an organism and the complexity of that organism, a topic recently revisited upon the sequencing of the human genome [119]. The importance of moonlighting becomes apparent when one considers it applies not only to enzymes, but also to receptors, transmembrane channels, chaperones and ribosomal proteins [120-122].

The molecular mechanism of moonlighting may involve a change in cellular location, expression in a non-native cell type, the oligomeric state as well as the concentrations of product, ligand or cofactor in the cell. Promiscuity could also be explained if promiscuous reactions occurred on surfaces other than the enzyme primary active-site [123].

Enzymes evolve new functions dependent on apparently contradictory features, namely robustness (the reduced lethality of mutations) and plasticity (achieving a new function with relatively few amino acid changes) [124]. For example, carbonic anhydrase II (CAII) is considered one of the most efficient enzymes known and catalyses the reversible hydration of CO_2. The enzyme had previously been shown to be weakly active towards esters such as p-nitrophenylacetate. Directed evolution of CAII towards this promiscuous activity yielded an enzyme with a 10-40-fold increase in catalytic activity, accompanied by a relatively small 1.4–2-fold decrease in native activity [124].

Ahoroni *et al.* [124] performed six directed evolution experiments on the basis of which they propose that evolvability may rely on the ability of protein to retain robustness toward their original "native" function, while at the same time exhibiting high plasticity towards their promiscuous functions. This allows phenotypic robustness and evolvability to coincide and provides the context for gene duplication and divergence. Interestingly, those mutations found to increase promiscuity without affecting the native function were found primarily in the active-site. The existence of moonlighting proteins provides a mechanism by which a small number of mutations may have a large impact on the function of a protein. Kazlauskas [117] gives examples of previous work where existing catalytic promiscuity has been used in biocatalysis. For example, lipase B from *Candida antarctica* has a native function as a carboxyl ester hydrolase. It can also catalyse a carbon-carbon bond forming reaction in an aldol addition of hexanal in cyclohexane [125], although the reaction isn't enantioselective.

Where a disease results from a mutant in a given protein with moonlighting activity, this promiscuity may hinder development of treatments for the disease i.e. a treatment that targets only one function of a multifunctional enzyme may be insufficient. A striking example was observed where a mutated tumour suppressor gene resulted in inherited uterine and benign smooth muscle tumours. When identified, the gene coded for formate hydratase, an enzyme of the citric acid cycle [126].

Thus, the inherent evolvability of proteins makes them ideal subjects for engineering, where new functions may be obtained by mutation of relatively few residues.

1.5 Engineering new functions in enzymes

The engineering of proteins has been a reality since the first successful site-directed mutations in the early 1980s [127]. Protein engineering can be described as the use of site-directed or random mutagenesis to alter primary amino acid sequence and subsequently the tertiary structure of a protein, where the aim is to alter the properties of that protein or enzyme. The following sections contain brief summaries of the most commonly used methods in protein engineering.

1.5.1 Rational Design

A rational approach to protein engineering involves using information that is known about a protein, such as the crystal structure to target certain residues using site directed mutagenesis (SDM). Successful implementations in industrial biocatalysis have been rare [128,129]. The possible reasons for the problems encountered are summarized by Dalby [130].

1.5.2 Directed Evolution

Directed evolution overcomes the problem of understanding how an enzyme works [131] by mimicking natural evolution over several cycles of mutation and selection. This process is achieved using random mutagenesis or recombining gene fragments as discussed in Sections 1.5.2.1 and 1.5.2.2. The resulting “library” of DNA products is transformed into a host strain and the resulting mutant enzymes, "variants" are individually expressed.

After each cycle of mutagenesis variants are selected for a certain characteristic (e.g. altered substrate specificity). The optimal mutant from each cycle is used as the template for the next round of mutagenesis. The evolutionary pressure of each round of mutation and selection increases the chances of attaining the desired phenotype.

The main methods of directed evolution are described in Sections 1.5.2.1 to 1.5.2.2.

1.5.2.1 Non-recombinant Directed evolution methods

1.5.2.1a Random mutagenesis

Two main types of random mutagens are commonly used. Genes can be mutated through several rounds of division in a cell line that has a deficient DNA repair mechanism, such as the XL1 Red cell line from Stratagene. These systems are generally inefficient, since the burden on the mutator strains both from inadequate DNA repair and over expression of protein results in low growth and mutation rates.

A second commonly used method of random mutagenesis is Error-prone PCR (epPCR), where a DNA polymerase with no proofreading function, such as *Taq* polymerase is used to induce point mutations during DNA replication. To reduce the fidelity of the *Taq* DNA polymerase reaction, Mn^{2+} is added in small amounts to the reaction mixture [132,133]. The use of biased concentrations of dNTPs or non-natural dNTP analogues has also been used to increase the mutation rates of the epPCR reaction [134]. EpPCR is limited somewhat by the natural codon bias, whereby the number of amino acids accessible after a single point mutation is limited to between 2 and 6, depending on the codon.

1.5.2.1b Random oligonucleotide mutagenesis

To overcome the problem of limited amino acid access encountered when using single point mutation (discussed for epPCR in Section 1.5.2.1a), saturation mutagenesis can be used. The process is similar to the PCR reaction. Primers are used that are specific to a region of the target gene, with certain bases randomized to ensure access to a host of amino acid residues at target residue positions.

Saturation mutagenesis produces large libraries. For example, if 4 residues in a protein were randomized, a library of 1.6 X 10^5 would result.

A more rational implementation of saturation mutagenesis is possible. For example, if a specific amino acid site is known to affect specificity for an interesting substrate, saturation mutagenesis may reveal which amino acid residue gives the optimal activity at this position.

1.5.2.2 Recombinant Directed Evolution methods

The hypothesis of Fisher [135] states that sex and recombination are advantageous since they allow advantageous mutations from different individuals to be combined in the same individual.

In 1994, Stemmer [136] reproduced the phenomenon of recombination *in vitro*. The basis of Stemmer's "DNA shuffling" method involves the digestion of 2 closely related genes with DNase I to yield short dsDNA fragments (10 – 50bp). These fragments are purified and used as primers in a PCR-like reaction. During the reaction, full length genes are produced, where fragments of gene A prime to gene B and vice versa in a process termed "template-switching". These gene products are then screened as described in Section 1.5.2.

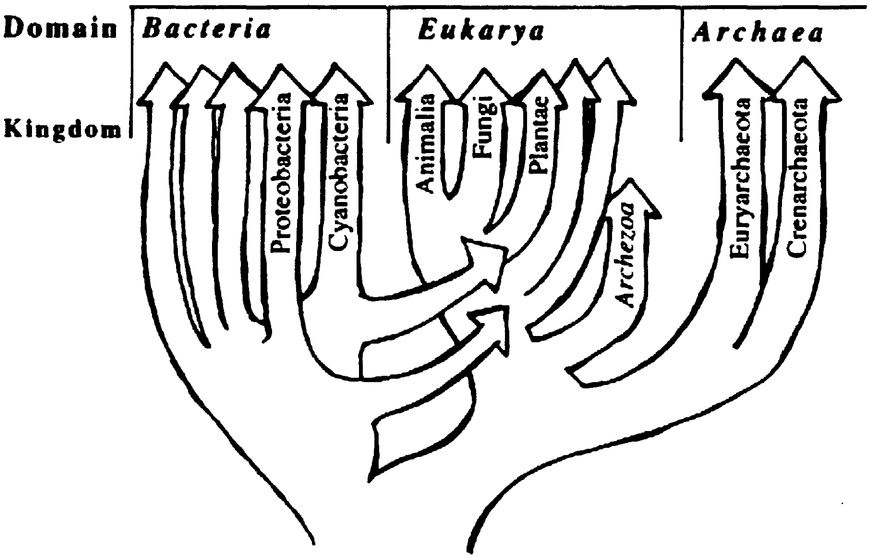


Figure 1.16: The Universal Tree of Life, as inferred from SSU rRNA. Figure taken from Doolittle [137].

In the intervening years, other non-recombinant methods have been developed for homologous and non-homologous recombination.

1.6 Focusing on the active-site for engineering Biocatalysis

Dalby [130] observed that mutants produced by directed evolution were often found in the active-site of enzymes. In a study by Morley and Kazlauskas [138], it was observed that in the cases where catalytic activity and thermal stability were the targets for mutagenesis, residues near and far from the active-site were found to be equally effective. However, for experiments where changes in enantioselectivity, substrate specificity or catalytic promiscuity are the desired result, random mutagenesis of the entire gene is unlikely to be the best strategy, since most residues in an enzyme are not close to the substrate binding site and thus most mutations won't occur where they would have the greatest effect. Mutations close to the substrate binding site were most effective in such cases.

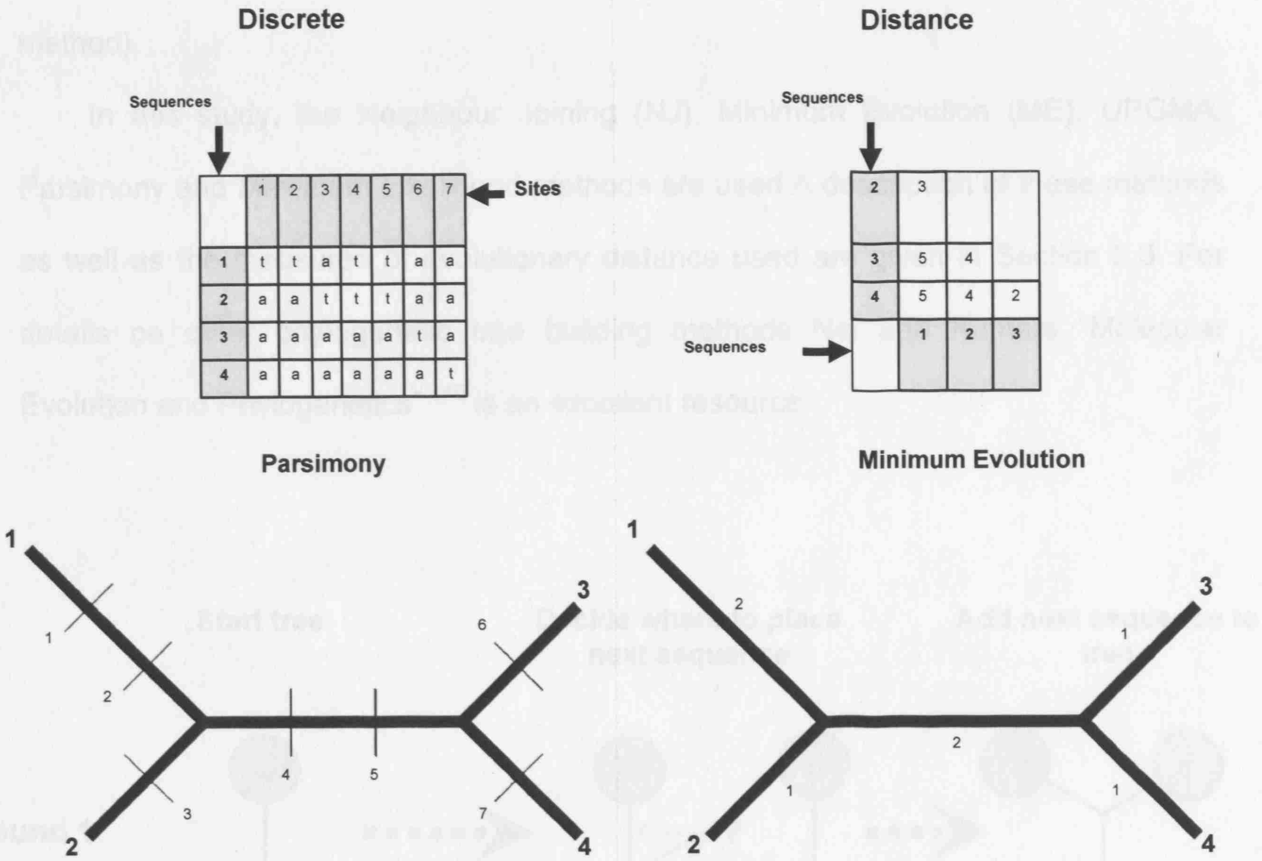
1.7 Phylogenetics

Understanding the evolution of enzymes can be useful as an aid to taxonomic classification and understanding the underlying genetic mechanisms of evolution. Phylogenetics, the study of relatedness among organisms, is potentially very useful in the realm of biocatalytic engineering. Since evolutionarily related enzymes have related modes of catalysis, conservation at the amino acid level mediates mechanistic conservation and allows the roles of certain amino acids to be proposed.

Phylogenetic trees can be constructed to illustrate the evolutionary relationships between groups of organisms or genes, with the most similar organisms clustered closest together in the tree. In molecular phylogenetics, trees are generally constructed using DNA or protein sequence information.

1.7.1 Commonly used tree-building methods

Phylogenetic tree-building methods commonly used include: Unweighted pair-group method with arithmetic mean (UPGMA); Neighbour-joining (NJ); Neighbourliness; Minimum-evolution (ME); Distance-Wagner (DW); Least-Squares (LS); Parsimony and the Maximum Likelihood (ML) method. Methods can be classified in terms of how they handle aligned sequence data and also on their approach when building phylogenetic trees [139]. In the case of the distance methods (UPGMA, NJ, Neighbourliness, ME, DW and LS), the sequences, once aligned are transformed into a pairwise distance matrix (Figure 1.17). Discrete methods (Parsimony, ML), on the other hand consider the character state of each site in a given sequence. Figure 1.17 shows an example where the same small dataset, results in the same topology when the parsimony and minimum evolution methods are used, despite the fact that parsimony is a discrete method and ME is a distance method.



	1	2	3	4	5	6	7
1	t	t	a	t	t	a	a
2	a	a	t	t	t	a	a
3	a	a	a	a	a	t	a
4	a	a	a	a	a	a	t

2	3		
3	5	4	
4	5	4	2
	1	2	3

Figure 1.17: The difference in data handling between discrete and distance tree-building methods. Parsimony is a discrete method, where each character state is examined individually. The simplest, most parsimonious tree is chosen (Section 2.3.1.3). The distance methods first convert sequence data to a numerical matrix. Here, the number of differences between the four species is calculated. On this basis, the ME algorithm groups species with the fewest differences together first, before addition of other species (Section 2.3.1.1b).

The two major approaches for tree building are clustering or optimality methods. Clustering methods follow an algorithm as illustrated in Figure 1.18. The example shows how, for example, beginning with five sequences, an unrooted tree is generated for three of them. A clustering method will then decide where to place the fourth sequence. In the next round the fifth sequence will be added to yield the final tree.

Methods using the optimality criterion consider all possible trees that could be generated from a given set of sequences. Considering five sequences, as illustrated in

Fig 1.19, there are fifteen possible trees to consider. The optimality method assigns a score or rank to each phylogeny and chooses on the basis of this score (e.g. ML method).

In this study, the Neighbour Joining (NJ), Minimum Evolution (ME), UPGMA, Parsimony and Maximum Likelihood methods are used A description of these methods as well as the measures of evolutionary distance used are given in Section 2.3. For details on other phylogenetic tree building methods Nei and Kumars "Molecular Evolution and Phylogenetics" [140] is an excellent resource.

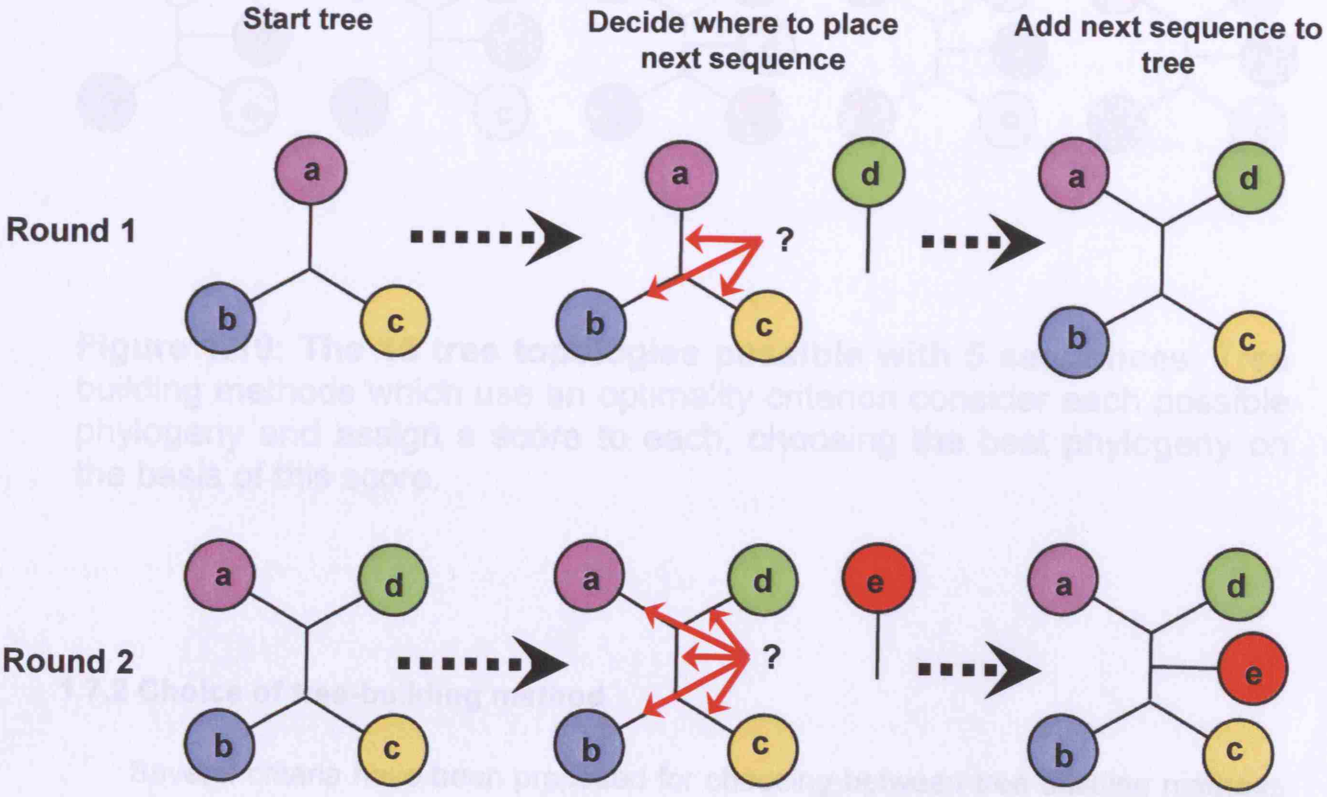

Start tree
Decide where to place next sequence
Add next sequence to tree
Round 1
Round 2
a
b
c
d
e
?

Figure 1.18: The method by which a clustering algorithm generates a phylogenetic tree.

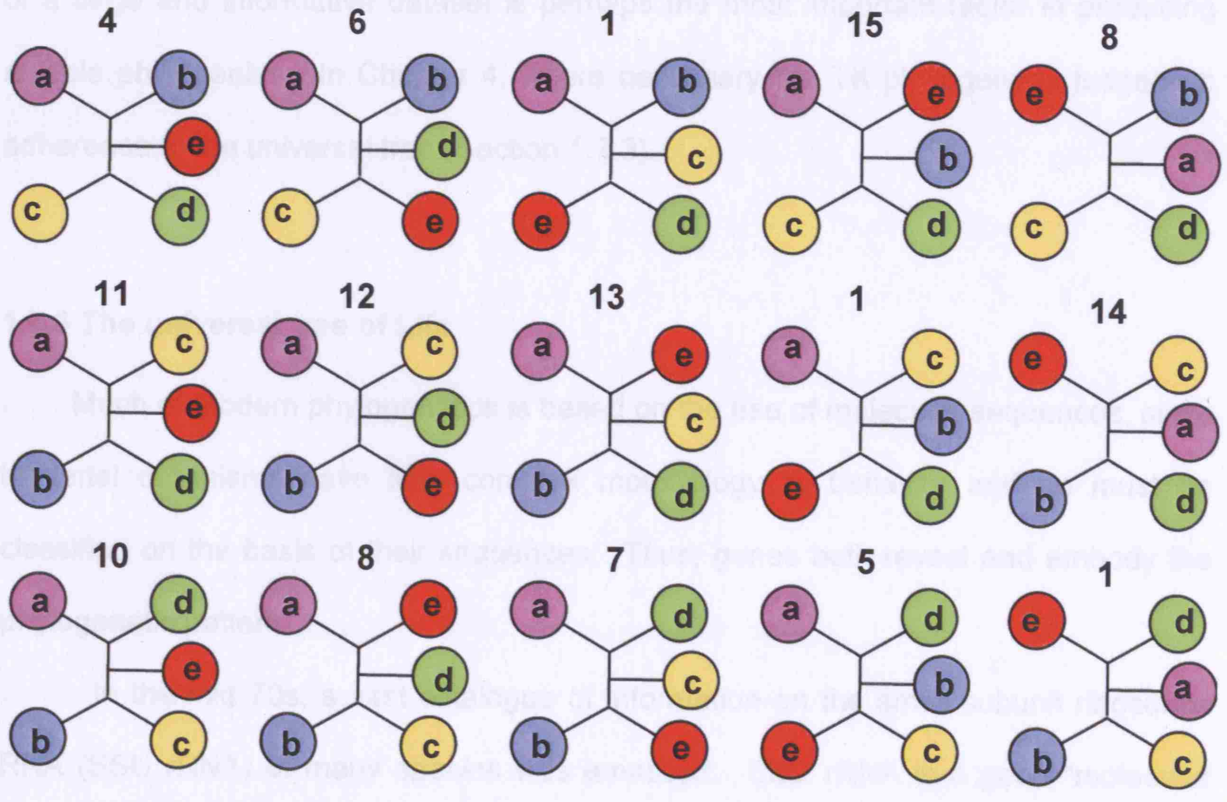


Figure 1.19: The 15 tree topologies possible with 5 sequences. Tree building methods which use an optimality criterion consider each possible phylogeny and assign a score to each, choosing the best phylogeny on the basis of this score.

1.7.2 Choice of tree-building method

Several criteria have been proposed for choosing between tree building methods Some of these include computational speed, consistency as an estimator of topology, statistical tests and the probability of obtaining the true phylogeny. However, most of these criteria are not useful in a practical sense and there may never be an agreement reached between phylogeneticists as to the best method of tree construction. The study of Russo *et al.* [141] showed that when amino acid sequences were used, the NJ, ME, MP (maximum parsimony), and ML method performed equally well. Thus the use of a large and informative dataset is perhaps the most important factor in producing reliable phylogenies. In Chapter 4, where necessary the TK phylogeny is judged on adherence to the universal tree (Section 1.7.3).

1.7.3 The universal tree of Life

Much of modern phylogenetics is based on the use of molecular sequences, since bacterial organisms have little complex morphology or behavior and so must be classified on the basis of their sequences. Thus, genes both reveal and embody the phylogenetic pattern.

In the mid 70s, a vast catalogue of information on the small subunit ribosomal RNA (SSU rRNA) of many species was amassed. SSU rRNA is a good "molecular chronometer" as it is widespread and abundant, is coded for by organellar, nuclear and prokaryotic genome, which contain slow and fast moving regions. The structure of SSU rRNA is universally conserved, reflecting an ancient ancestry.

SSU rRNA is also unlikely to undergo horizontal gene transfer (HGT) (Section 1.7.4) or other evolutionary anomalies since it has highly conserved interactions with many coevolved RNAs and proteins. Interspecial transfection of rRNA has been shown to deleterious [142,143].

The phylogeny using SSU rRNA is a good candidate for the universal organismal tree (or "true phylogeny"), which is backed up by the fact that whole-genome phylogenies correspond well with the SSU rRNA phylogeny (see Fig 1.16).

When examining the phylogenies of other enzymes, particularly those with ancient phylogenies, such as TK in Chapter 4, the SSU rRNA tree provides a good reference for assessing the quality of a phylogeny, where HGT is likely to have occurred.

There is however disagreement as to the significance of the universal tree of life. Two reasons for this are: (a) In protein and rRNA studies, artifacts relating to differences in evolutionary rate and mutational saturation can be misleading about the root of the phylogeny and (b) Many genes give plausible, but different phylogenies from the same organism. These observations piqued interest in phylogenetic anomalies, including HGT over a decade ago when the first genomes began to be sequenced.

1.7.4 Phylogenetic anomalies

Novel sequences can occur in a number of ways; by the evolution of adaptive alleles, by the divergence of gene duplications (paralogs) (Section 1.3) or by the acquisition of alien sequences (HGT).

These can cause phylogenetic anomalies, since most phylogenetic reconstruction methods assume a rate of mutation that is approximately constant and also that inheritance is strictly vertical. In addition, the rate of mutation is disrupted if intense selection is in operation or if evolution is occurring under biased mutation rates.

Anomalies can also be due to too little phylogenetic information (too few clades) or inadequate phylogenetic methods. HGT has been proposed as the "essence of phylogeny" [137]. This "rampant HGT paradigm" [144], which suggests that HGT is the dominant force in adaptive evolution, allowing ready-made responses to environmental change. HGT may have been very important before the emergence of the kingdoms of life, in the progenote population [145], where sequence evolution was intense and the chances of a mutation conferring selective advantage was high [145-147].

Woese [145-147] has described a system where a network that is heavily influenced by HGT "evolves" to a network that is not amenable to HGT. Thereafter, inheritance progresses through vertical lineages.

Despite controversy [148,149,144], the importance of HGT appears to have been overestimated in the past, with most imported genes not conferring selective advantage and being purged from the genome within a few million years. Other events such as the segregation of paralogs constitute more frequent challenges to the genome phylogeny. Thus, Darwinian lineages are the essence of modern organisms' evolution.

1.8 Palaeogenetics

"Paleobiochemistry" was the term coined in the early 1960s to describe the use of chemical sequences, such as those of DNA, RNA and proteins to infer the sequences of ancient, extinct biological molecules. Previous uses of ancestral sequences to "resurrect" ancient proteins are discussed later in this section, while in Chapter 4, the reconstruction of ancestral forms of TK is described in an attempt to apply Palaeogene tics to protein engineering and biocatalysis.

In their 1963 paper, "Chemical Paleogenetics: Molecular "Restoration Studies" of extinct forms of life", Zuckerkandl and Pauling put forward the following points:

(a) In haemoglobin, polypeptide chains from different vertebrate organisms commonly have the same amino acid residue at some positions. This, they argue is not likely to be due to multiple instances of convergent evolution from heterologous ancient proteins but by the existence of an ancient polypeptide ancestor, encoded by similarly ancient genes from which extant forms have evolved.

(b) On the basis of differences between extant forms of a given enzyme in different organisms, it is possible to get a very rough approximation of how long ago the enzyme forms diverged.

(c) The amino acid sequence of the common polypeptide ancestor can be determined using sequence information from extant forms.

(d) In instances where polypeptide chains differ at a given position it is possible to determine in which line of descent the difference has occurred.

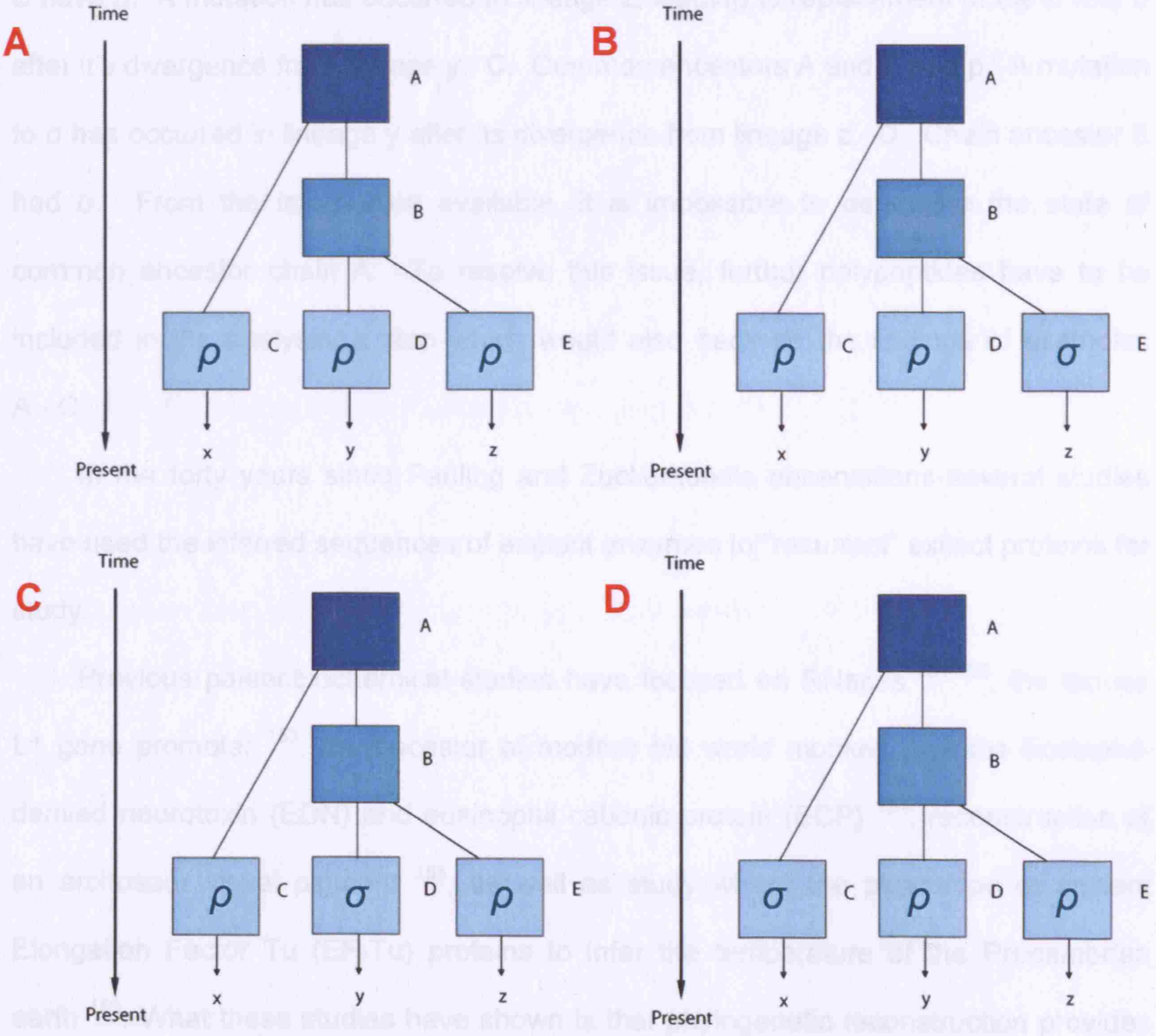


Figure 1.20: The different character states to illustrate the reconstruction of ancient polypeptides. Adapted from Zuckerkandl and Pauling [150].

In Figure 1.20, three lines of evolutionary descent are represented by arrows leading to x, y and z. Two branching points are shown by blue boxes A and B, which represent common ancestors of the extant polypeptides. In the paper, two amino acids, σ and ρ are considered at an equivalent sequence position in three extant polypeptides, shown in light blue boxes and labelled C, D. Several probable assertions are made about the four states A – D. These are: **A.** The common ancestors of A and B has residue ρ at the residue site under consideration. **B.** Common ancestors A and B have ρ. A mutation has occurred in lineage z, leading to replacement of the ρ with σ after it's divergence from lineage y. **C.** Common ancestors A and B had ρ. A mutation to σ has occurred in lineage y after its divergence from lineage z. **D.** Chain ancestor B had σ. From the information available, it is impossible to determine the state of common ancestor chain A. To resolve this issue, further polypeptides have to be included in the analysis, a step which would also back up the findings of examples A - C.

In the forty years since Pauling and Zuckerkandls observations several studies have used the inferred sequences of ancient enzymes to "resurrect" extinct proteins for study.

Previous palaeobiochemical studies have focused on RNases [151,152], the mouse L1 gene promoter [153], the ancestor of modern old world monkey proteins Eosinphil-derived neurotoxin (EDN) and eosinophil cationic protein (ECP) [154], reconstruction of an archosaur visual pigment [155], as well as study where the phenotype of ancient Elongation Factor Tu (EF-Tu) proteins to infer the temperature of the Precambrian earth [156]. What these studies have shown is that phylogenetic reconstruction provides ancestors that once generated exhibit plausible structural, catalytic and physical properties. Importantly, the characteristics of ancient proteins were not necessarily found to be an average of the characteristics of their descendents. Paleobiochemistry can thus yield information beyond that yielded by comparisons of descended proteins alone. The usefulness of structural information was also highlighted [157], underlining the usefulness of higher order analysis when understanding biochemical phenomena, rather than treating protein sequences as if they were a string of independent letters.

In order to make directed evolution and rational design more effective, knowledge of how enzymes have evolved over time would be of great use (Section 1.7). In the case of substrate specificity, a detailed understanding of how enzymes modulate substrate specificity during the course of evolution would provide an invaluable insight into structure-function relationships. This would allow for the rational mutation of enzymes for altered substrate specificity with a greater degree of predictability.

The probability of reconstructing the exact ancestral sequence is low, due to the accumulation of errors across the many sites. However, since a small number of replacements can lead to a change in substrate specificity, realising the ancestral phenotype is more likely than reconstructing the genotype.

Recent advances in molecular biology have allowed the determination of the DNA sequence of any given individual from any species. This has revolutionised the study of adaptation and allowed a much greater understanding of the interplay between various evolutionary forces and constraints. However, although selection can be inferred from the genetic variation between species, an understanding of adaptive change requires phenotypic information. Studying phenotypes within a natural population (e.g. a microbial population) has helped highlight underlying molecular mechanisms [158-161], but to address the problem of studying ancient adaptations, phylogenies are needed.

As described in Section 4.2.1, in Chapter 4, the TK phylogeny is reconstructed following a detailed phylogenetic study.

1.9 Aims

This thesis consists of four experimental chapters, which address the following aims:

Chapter 3: Phylogenetic study of the TPP-dependent enzymes. In Chapter 3, an investigation into the evolution of the TPP-dependent enzymes is performed. TPP-dependent enzymes have evolved diverse chemistries by mutation of enzyme residues, domain rearrangement, gene duplication, gene-splitting and recruitment of additional domains. The aim of this study is to see how divergent events have affected comparable regions of structure at the amino acid level. Seventeen TPP-dependent enzymes are examined, with the alignment of TK being examined in the greatest detail. Phylogenies are generated for each enzyme individually before the comparable regions from the PP and Pyr domains of all enzymes are compiled and subject to phylogenetic study. Using information from this study, future engineering opportunities for TK may present themselves.

Chapter 4: Reconstruction of ancestral transketolase enzymes. Chapter 4 describes how the TK phylogeny generated in Chapter 3 was used to resurrect ancient forms of the enzyme. The amino acid sequences of ancient forms of TK will be generated using the PAML program. Various lineages will be examined in an attempt to elucidate patterns of mutation in the active-sites of TKs from different organisms during their respective evolution. The key active-site residue mutations occurring during the evolution of *Eco*TK from its common ancestor with *Sce*TK are constructed and each mutant is characterised kinetically.

The change in activity of *Eco*TK during evolution will be examined, by characterising each mutant for the β-HPA + GA reaction. The substrate repertoires of *Eco*TK and the common ancestor of *Eco*TK and *Sce*TK will be examined for a host of donor, acceptor and non-natural substrates. In this way, it is hoped an understanding of the evolution of substrate specificity and modulation substrate specificity in general will be reached.

Chapter 5: Engineering a pyruvate utilising transketolase enzyme. Since β-HPA is a very useful, but industrially unviable substrate for TK, an attempt is made to engineer a TK that can use pyruvate, a cheaper, more readily available donor, differing from β-HPA by a single hydroxyl group. The potential pyruvate + GA TK reaction yields CO_2 as one of its products, in a similar manner to TK reactions with β-HPA as a donor, thus making biotransformations essentially irreversible. A comparison of TK and DXPS is performed to identify highly conserved residues in each enzyme. Firstly, TK and the pyruvate-utilising DXPS are compared to identify active-site residues potentially responsible for pyruvate utility in DXPS, or for preventing pyruvate use in TK. To focus the study on a smaller number of candidate positions, PDC is included in the analysis. The proximity of residues to erythrulose in the *Eco*TK crystal structure is also examined. Two active-site residues in TK are mutated from their TK character state to the character state found at the equivalent position in DXPS. Each mutant, is screened for activity for the both the model β-HPA + GA and desired pyruvate + GA reactions.

Chapter 6: Investigation of the role of the C-terminal domain of transketolase. The function of the TKC domain of TK is poorly understood. While other chapters of this thesis focus on the catalytic PP and Pyr domains of the TPP-dependent enzymes, Chapter 6 examines the affect on TK of excising the TKC domain by insertion of a stop codon into the TK gene. The PP-Pyr form of TK, composed only of the catalytic domains, can then be characterised for the β-HPA + GA reaction. Further successive truncations of TK from the C-terminal end will be performed, chopping away stretches of secondary structure so as to examine the shortest form of TK capable of catalysing the β-HPA + GA reaction. A form of TK capable of catalysis at a rate comparable with *Eco*TK, but which is shorter in length, has potential biocatalytic implementations, since such an enzyme would be cheaper to produce in fermentations and may be free of certain regulatory constraints imposed by the TKC domain.

Chapter 2: Materials and methods

The protocols described in this chapter are standard practices. Experimental techniques are described in Chapters 3–6.

2.1 Laboratory Materials

Unless otherwise stated, all materials were purchased from Sigma-Aldrich Ltd. Water was purified to 15 MΩ.cm^{-1} resistivity using an Elix 5 water purification system (Millipore Corp.).

2.2 Computer programs used throughout this thesis

Biological programs used throughout this thesis are available free of charge at the following locations:

Phylip: http://evolution.genetics.washington.edu/phylip.html

PAML: http://abacus.gene.ucl.ac.uk/software/paml.html

MEGA3: http://megasoftware.net

Bioedit: http://www.mbio.ncsu.edu/BioEdit/bioedit.html

Treeview: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

AnnHyb: http://bioinformatics.org/annhyb/

Pymol: http://pymol.sourceforge.net/

Deepview: http://www.expasy.org/spdbv/

Scorecons: http://www.ebi.ac.uk/thornton-ru/databases/valdarprograms/scorecons

2.3 Phylogenetic methods used throughout this thesis

2.3.1 Measures of evolutionary distance

As described in Section 1.7.1, the distance methods of phylogenetic reconstruction require first that sequence data be transformed into a numerical matrix before construction of phylogenies. The data in these matrices represent measures of evolutionary distance. Common measures of evolutionary distance are described here.

The proportion of different amino acids between two sequences can be calculated using:

$$\hat{p} = n_d / n \qquad \textbf{Equation 2.1}$$

where n_d is the number of amino acids that are different between the two sequences and n is the total number of amino acids. $\hat{P}$ is defined as the *p-distance.*

However, a plot of $\hat{P}$ against time is non-linear. As time progresses, multiple amino acid substitutions at the same site occur and the discrepancy between n_d and the actual number of amino acid substitutions increases. A more accurate measure is to use the Poisson distribution, such that the probability of k amino-acid substitutions occurring at a given site after a period of time t, is given by:

$$P\,(k;t) = e^{-rt}(rt)^k / k! \qquad \textbf{Equation 2.2}$$

where r is the rate of amino acid substitution, and hence the mean number of amino acid substitutions per site over t years is rt.

The number of amino acid substitutions is estimated by comparing two homologous sequences that diverged t years ago. Since the probability of no substitutions having occurred at a site in a sequence $P\,(0;t)$ is e^{-rt}, the probability *(q)*, that neither of the two homologous sites has undergone substitutions is given by:

$$q = (e^{-rt})^2 = e^{-2rt} \qquad \textbf{Equation 2.3}$$

Since $q = 1 - p$ and d (the total number of amino-acid substitutions per site for the two sequences) = $2rt$, the total number of amino-acid substitutions per site for the two sequences is as follows:

$$d = - ln\ (1 - p) \qquad \textbf{Equation 2.4}$$

Using $\hat{p}$ in the above equation gives $\hat{d}$, the *Poisson Correction (PC) distance*. The above methods assume that the rate of amino-acid substitution (r) is uniform across all amino acid sites. This doesn't usually hold true as the rate of successful substitution is usually higher at functionally less important sites [162].

Uzzell and Corbin showed that the distribution of the number of amino-acid substitutions per site is greater than the Poisson variance and that it roughly follows the negative binomial distribution [163]. It can be shown that when the rate of amino-acid substitutions (r) varies according to a gamma distribution, the observed number of substitutions per site (k) follows a negative binomial distribution [164]. Therefore, Uzzell and Corbins' observation suggests that substitution rates vary from site to site according to the gamma distribution [163]. The gamma function is defined as:

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1}.dt \qquad \textbf{Equation 2.5}$$

where the function shape is defined by a, the shape or gamma parameter. The distribution is determined by a. For example, if $a = \infty$, r is the same for all sites, whereas if *a* = 1, *r* follows an exponential distribution with *r* varying extremely from

Round 1

Table A	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Round 2

Table B	(A,B)	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

Round 3

Table C	(A,B)	C	(D,E)
C	4		
(D,E)	6	6	
F	8	8	8

Round 4

Table D	((A,B),C)	(D,E)
(D,E)	6	
F	8	8

Round 5

Table E	(((A,B),C),(D,E))
F	8

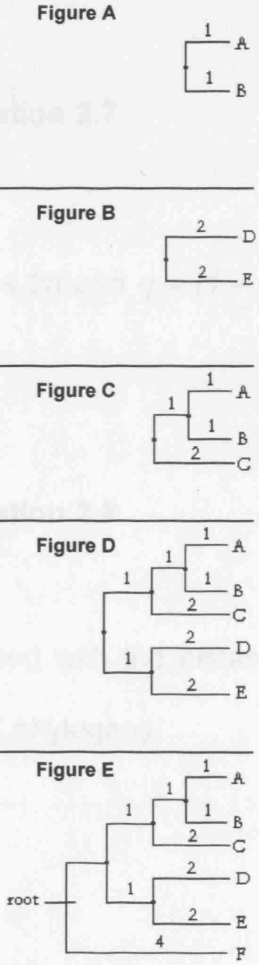


Figure 2.1: The UPGMA tree building algorithm. Successive rounds of clustering yield the final tree shown in Figure E, as described in the text.

amino acid site to amino acid site. When *a* < 1, distribution of r across all sites is even more skewed, with many sites showing *r* values close to 0.

When *r* varies following the gamma distribution, it is possible to estimate the number of amino-acid substitutions per site. We consider the probability of the identity of amino acids at a given site between two sequences at time t, given by the equation:

$$q = e^{-2rt} \qquad \textbf{Equation 2.6}$$

the average of *q* over all the sites is given by:

$$\dot{g} = \int_0^\infty qf(r)dr = \left[{}^{a}\!/\!_{(a+2rt)} \right]^a \qquad \textbf{Equation 2.7}$$

Since the total number of amino acid substitutions per site ($d\dot{g}$) is *2rt* and $q = (1 - p)$, $d\dot{g}$ defined as the *gamma distance,* becomes:

$$a\,[(1 - p)^{-1/a} - 1] \qquad \textbf{Equation 2.8}$$

In Chapter 4, such measures of evolutionary distance is used with the distance methods of tree construction (Section 2.3.2) when studying the TK phylogeny.

2.3.2 Distance methods of tree construction

2.3.2.1 Unweighted pair-group methods using arithmetic average

UPGMA uses an algorithm to analyse distance data and can be used to generate molecular phylogenies when the rate of gene substitution is approximately constant. Taxa are grouped based on increasing distance between each other. Firstly the two taxa with the least distance between them are grouped after which more distant taxa are progressively added to the group or to new groups.

Suppose there are six species: A, B, C, D, E and F, with distance data as shown in Table A of Figure 2.1. In round 1, the two species with the smallest distance between them, A and B in this case are joined. The branch point is half the distance between A and B.

Once A and B cluster, they are considered thereafter as a single taxa (A,B). The distances between (A,B) and all the other taxa are then calculated. For example, the distance between (A,B) and C is calculated as (distance AC + distance AB) / 2 = 4. In this way, the new distance Table B in Figure 2.1 is generated. D and E are grouped next on the basis that they have the least distance between them, generating cluster B.

The third cycle produces Table C. C groups with (A,B) as in Cluster C. In the fourth cycle, the ((A,B),C) group clusters with the (D,E) group as shown in cluster D.

After the fifth cycle, the final tree given by the UPGMA method for these 5 taxa is generated, as shown in Cluster E.

UPGMA assumes the data is ultrametric; i.e. a tree can be constructed so that the observed distance between 2 taxa is equal to the sum of the branch lengths joining them and that the tree is rooted so all taxa are equidistant from the root. However, topological errors often occur when using UPGMA if the rate of gene substitution isn't constant or if the number of genes or characters used is small.

2.3.2.2 Minimum Evolution

The ME model assumed that when unbiased estimates of evolutionary distance are used, the value of the sum of all branch length estimates, [S], becomes smallest for the true tree regardless of the number of sequences used [165]. Thus [S] is calculated for all tree topologies. The topology with the smallest value for S is chosen. Thus ME, although a distance method, has a similar philosophy to the parsimony method (Section 2.3.3).

[S] is estimated as:

$$[S] = \sum_{i}^{T} b_i$$ **Equation 2.10**

Where *b* = an estimate of the length of the ith branch and T is the total number of branches (2m – 3), for m taxa. For example, in Figure 2.2 below, [S] is given by $b_1 + b_2 + \cdots\cdots + b_7$.

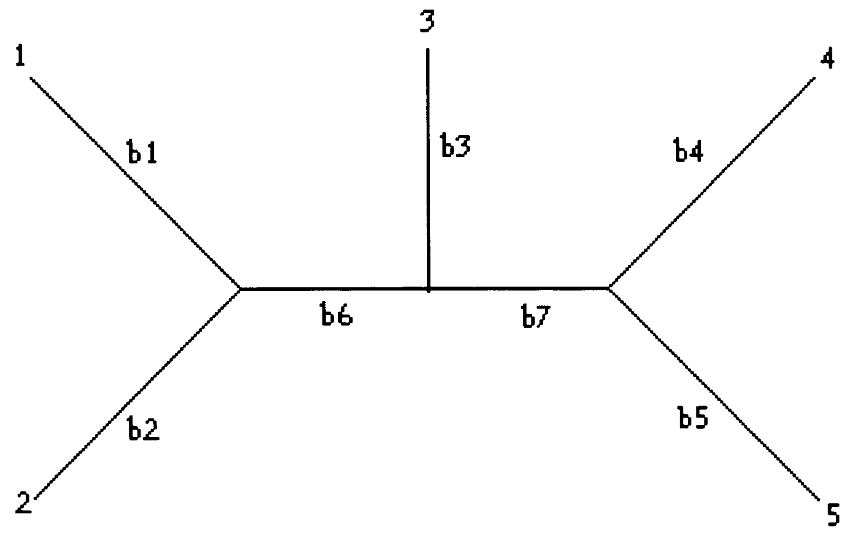


Figure 2.2: A Minimum Evolution Tree

To reduce computational time, when m is large, a NJ tree (Section 2.3.2.3) can first be calculated and topologies close to this tree examined, in an attempt to find a tree with a smaller [S]. If a tree is found, a new set of topologies close to this possible ME tree are examined. This process is continued until a tree with the smallest [S] is found.

2.3.2.3 Neighbour Joining

NJ, developed by Saitou and Nei [166] is an algorithm for analysing distance data and is based on the ME method. NJ doesn't examine all possible topologies, but at each stage of taxon clustering a ME principal is used. In a similar way to UPGMA, NJ groups the two taxa with the least distance between them and subsequently adds more distant taxa to the group or to new groups. The difference between NJ and UPGMA is that the distance matrix is modified at each step so that each group is adjusted on the basis of their average distance from all other groups. Therefore NJ assumes the data is merely additive rather than ultrametric as in the case of UPGMA, thus relaxing the assumption that all groups have been subject to the same rate of gene substitution.

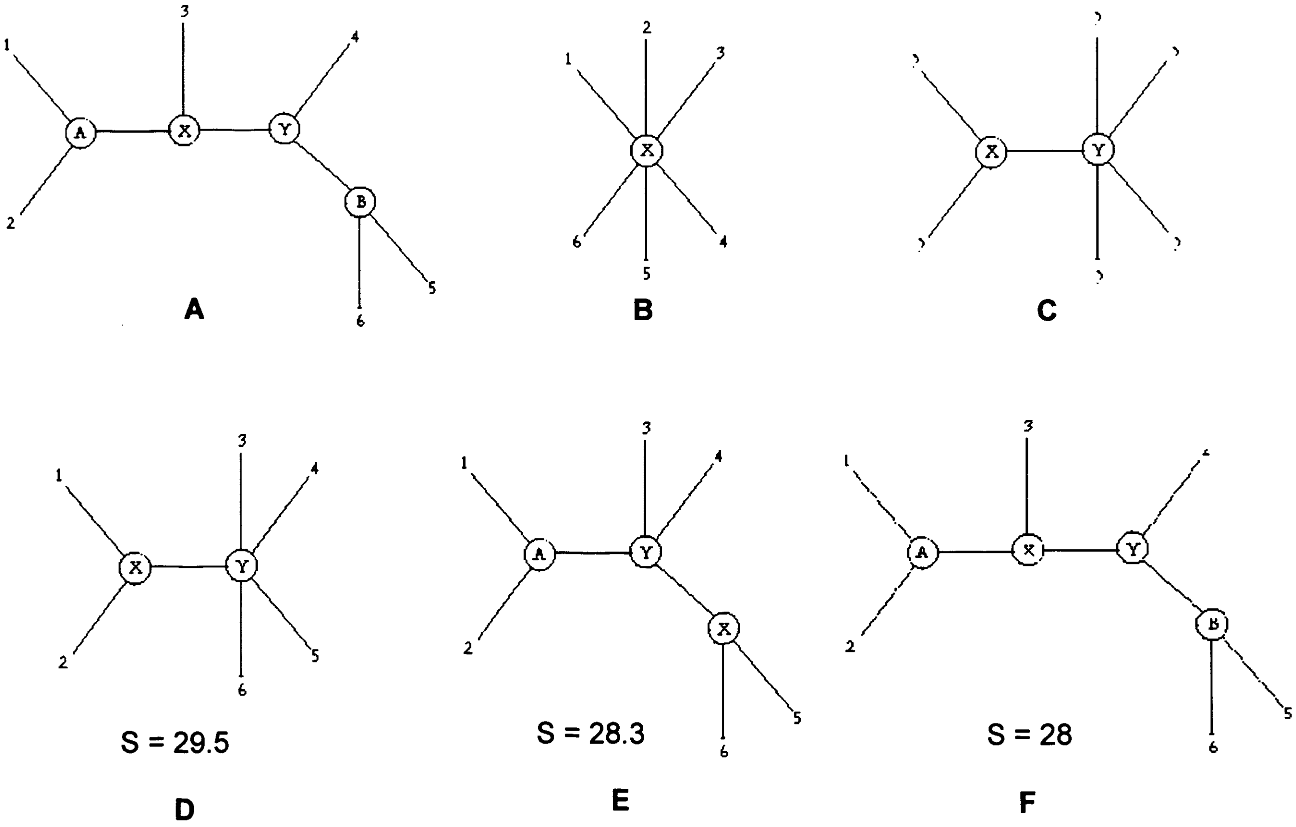


Figure 2.3: The star decomposition method of Neighbour Joining. Successive rounds of the NJ algorithm yield the final tree F.

Neighbours are defined as any two taxa connected by a single node in an unrooted tree. Thus in Image A of Figure 2.3, 1 and 2 are neighbours, as are 5 and 6. No other pairs of taxa are neighbours. However, if 1 and 2 are considered as one taxa (1,2), then (1,2) and 3 are now neighbours.

To generate a tree using NJ, firstly a star tree with no clustering is considered as shown in Image B of Figure 2.3. Methods that begin with such a star tree are often referred to as "star decomposition" methods. All taxa are considered potential neighbours.

[S_{in}], the sum of all branch lengths for the Ith and jth taxa are calculated, assuming that two taxa will group as shown in C of Figure 2.3. The taxa *j* and *i* which when clustered, show the smallest S_{in} value are chosen to cluster. In this case, it is 1 and 2. Thus the star becomes as is shown in D Figure 2.3.

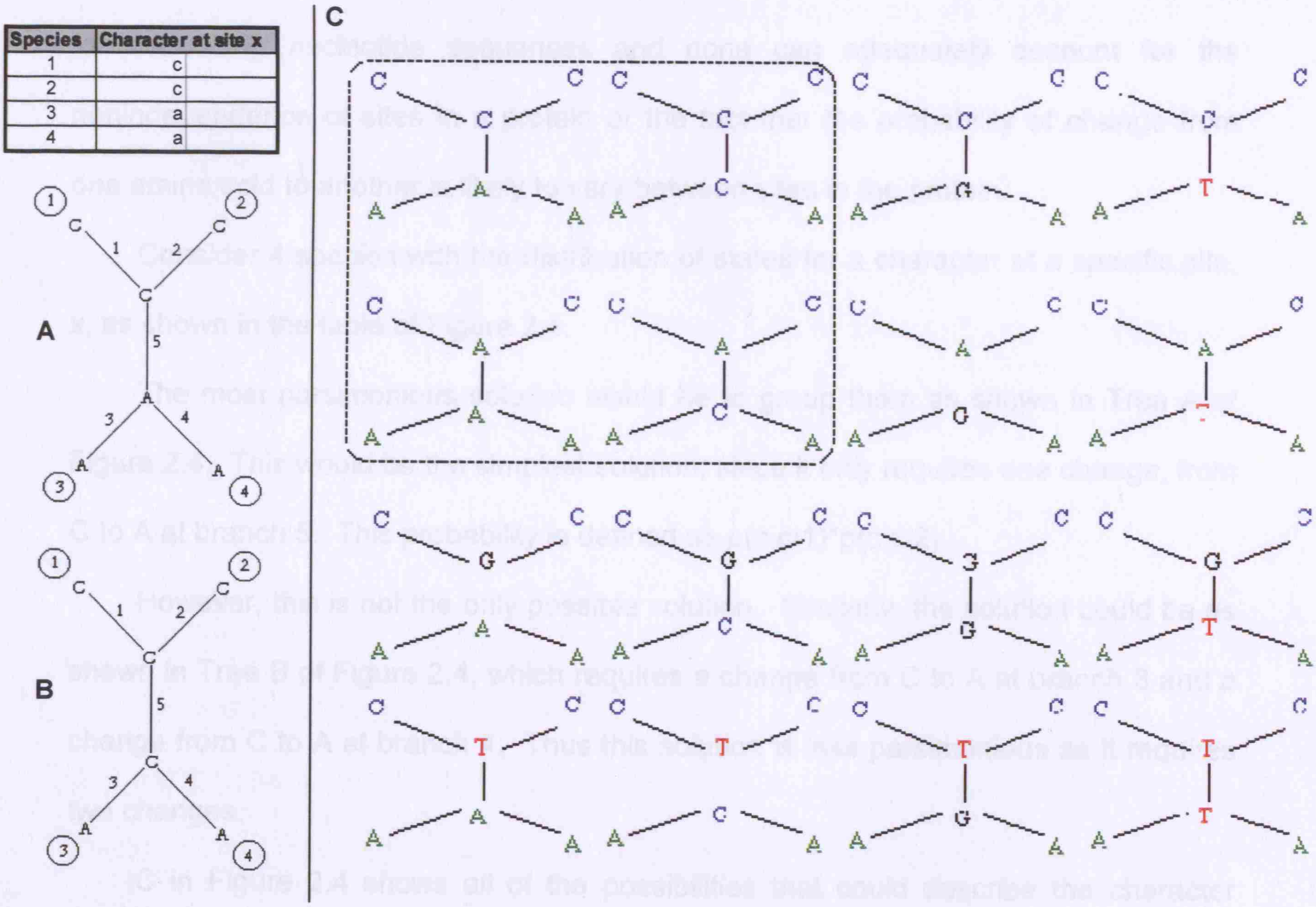
This process is repeated, generating new S_{ij} values and choosing the tree with the smallest S_{ij} at each cycle. Successive cycles generate the topologies shown for E and ultimately F in Figure 2.3.

2.3.3 The Parsimony method

The Parsimony method is commonly used, and yet it remains a controversial methodology. The two main arguments for the use of parsimony are as follows.

Firstly, parsimony tries to maximise the amount of similarity between sequences that is attributable to homologous similarity. In other words Parsimony explains a phylogeny so that maximum similarity is attributable to a common ancestry between species. If a given character state does not fit in with a tree, then any similarity that this character has with an equivalent character position in another sequence is attributed to homoplasy, such as parallel or convergent evolution. After examining all tree topologies, the tree with the fewest instances of homoplasy is chosen as being the most parsimonious.

The second main argument for parsimony is that evolutionary change is a rare occurrence. Thus the tree requiring the fewest character state changes during evolution is the best estimate of phylogeny.



Species	Character at site x
1	c
2	c
3	a
4	a

Figure 2.4: The different evolutionary possibilities that can explain the character states of 4 different species at a given position. Each possibility is considered when using the Maximum Likelihood method, as described in Section 2.3.4.

2.3.4 The Maximum Likelihood Method

ML is based on branch lengths obtained from all characters in an analysis, assuming a uniform rate of substitution. The ML method selects the tree that is most likely to have generated the observed data, given an explicit model of sequence evolution. In other words, ML works out the probability of any possible character being at any position in an ancestral sequence. Each possible tree is examined and the most probable tree is chosen. Models of amino acid evolution are not as well developed as for noncoding nucleotide sequences and none can adequately account for the nonindependence of sites in a protein or the fact that the probability of change from one amino acid to another is likely to vary between sites in the protein.

Consider 4 species with the distribution of states for a character at a specific site, x, as shown in the table of Figure 2.4.

The most parsimonious solution would be to group them as shown in Tree A of Figure 2.4. This would be the simplest solution, since it only requires one change, from C to A at branch 5. This probability is defined as p(c,cl1)*p(c,cl2).

However, this is not the only possible solution. Similarly, the solution could be as shown in Tree B of Figure 2.4, which requires a change from C to A at branch 3 and a change from C to A at branch 4. Thus this solution is less parsimonious as it requires two changes.

C in Figure 2.4 shows all of the possibilities that could describe the character distributions of the five species. Since each of these situations is possible, the probabilities for each state distribution at this character site need to be summed. This operation must be carried out for all character sites and the probability for each site multiplied together to give the likelihood of all of those characters being on that tree.

This entire procedure is carried out for each potential tree and the tree with the highest probability is chosen. ML is computationally intensive, but it has the advantage that it in calculating the phylogeny, the method considers the ancestral states of amino acid positions. The use of ancient protein sequences in biological studies is described in Section 1.8. An implementation of the ML method, the program PAML, is used in Chapter 4.

2.4 Preparation of buffers, media, and reagents

Unless otherwise stated, all buffers, media, and reagents were stored at 25 °c.

2.4.1 Luria Bertani medium

Luria Bertani (LB) medium was prepared by dissolving 10 g.L^{-1} tryptone, 10 g.L^{-1} NaCl, and 5 g.L^{-1} yeast extract in pure water. A concentrated solution of sodium hydroxide was used to adjust the pH to 7. Media were sterilised by autoclaving for 15 minutes (2 bar, 124 °C).

2.4.2 LB agar plates

LB agar was prepared by adding 20 g.L^{-1} select agar to LB medium. Agar preparations were sterilised by autoclaving. When required for making LB agar plates, solutions were melted in a microwave, transferred to a sterile Falcon tube, allowed to cool, but not solidify, before Ampicilin (AMP) was added as required. The solution was then transferred to a Petri dish and allowed to set.

2.4.3 Ampicillin

AMP was dissolved in pure water to a concentration of 150 g.L^{-1}. Stocks were sterilised by filtration and stored at -20 °C. AMP was used at a concentration of

150 mg.L^{-1} to select bacteria carrying the plasmid pQR791 and mutant variants thereof. Preparations containing this concentration of AMP were labelled 'AMP^{+}'.

2.4.4 0.5 M Tris buffer (pH 7.5)

0.5 M Tris buffer with a pH of 7.5 was prepared by dissolving 63.5 g.L^{-1} Tris hydrochloride and 11.8 g.L^{-1} Tris base in pure water.

2.4.5 Standard cofactor solution

Standard cofactor solution was prepared by dissolving 0.0915 g magnesium chloride hexahydrate (M_r = 203.3) and 0.0576 g TPP (M_r = 460.8) in 4.5 ml of pure water. A concentrated solution of NaOH was used to adjust the pH to 7.5. The cofactor solution was topped up to 5 ml with pure water and stored at 4 °C (for no more than 24 hours).

2.4.6 Substrate solutions

In the cases of pyruvate, propionaldehyde, benzaldehyde, hydroxybenzaldehyde and p-anisaldehyde solutions were prepared from commercial solutions. All substrate were prepared to a final concentration of 50 mM, except benzaldehde, hydroxybenzaldehyde and p-anisaldehyde, were used at 30mM. Such a concentration is greater than the K_m of *Eco*TK for the given substrate in most cases. Thus the rate observed is a fixed condition and not a direct measurement of K_m.

Final concentrations in g.L^{-1} for each substrate are as follows: arabinose (7.507 g.L^{-1} (M_r = 150.13)); A5P (13.705 g.L^{-1} (M_r = 274.1)); benzaldehyde (5.306 g.L^{-1} (M_r = 106.12)); erythrose (6.005 g.L^{-1} (M_r = 110.04)); E4P (10.005 g.L^{-1} (M_r = 200.1)); fructose (9.008 g.L^{-1} (M_r = 180.16)); F6P (17.005 g.L^{-1} (M_r = 340.1)); glucose

(9.008 g.L^{-1} (M$_r$ = 180.16)); G6P (19.773 g.L^{-1} (M$_r$ = 395.45)); glyceraldehyde (4.504 g.L^{-1} (M$_r$ = 90.08)); G3P (18.975 g.L^{-1} (M$_r$ = 379.5)); GA (6.005 g.L^{-1} (M$_r$ (dimeric) = 120.1)); hydroxybenzaldehyde (6.106 g.L^{-1} (M$_r$ = 122.12)); β-HPA (5.5 g.L^{-1} (M$_r$ = 109.99)); p-anisaldehyde (6.808 g.L^{-1} (M$_r$ = 136.15)); propionaldehyde (2.904 g.L^{-1} (M$_r$ = 58.08)); pyruvate (5.52 g.L^{-1} (M$_r$ = 110.04)); ribose (7.505 g.L^{-1} (M$_r$ = 150.1)); R5P (13.705 g.L^{-1} (M$_r$ = 274.1)); sedoheptulose (10.509 g.L^{-1} (M$_r$ = 210.18)); xylulose (7.507 g.L^{-1} (M$_r$ = 150.13)). G6P and G3P required sonication in order to dissolve fully.

2.5 Standard procedures

2.5.1 Streaked agar plates

Cultures were streaked out on a Petri dish of LB agar (AMP$^+$ if appropriate) using a wire loop. Plates were incubated overnight at 37 °C and stored at 4 °C.

2.5.2 Overnight cultures

Single colonies were picked from an agar plate into 5 ml of LB medium (AMP$^+$ if appropriate) in a 50 ml Falcon tube. These tubes were incubated for 16 hours at 37 °C with 220 rpm agitation.

2.5.3 Shake flask cultures

1 ml of an overnight culture was added to 49 ml of LB medium (Amp$^+$ if appropriate) in a sterile 500 ml shake flask. The shake flask was incubated for 16 hours at 37 °C with 220 rpm agitation.

2.5.4 Larger lysate preparations

Where a large amount of concentrated lysate was required, 1 L AMP$^+$ cultures were grown up in the same conditions as for the shake flask cultures. Prior to lysis, this culture was pelleted in batches in a Sorvall Super T21 centrifuge, using an SL-50T rotor at 13,000 rpm for 10 minutes. The concentrated pellet was resuspended in 50 mL of buffer (50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0) prior to sonication, as described in Section 2.5.6.

2.5.5 Glycerol stocks

A 20 % (v/v) glycerol stock was prepared by adding filter-sterilised 40 % (v/v) glycerol to an overnight culture in a one to one volume ratio and aliquots stored at - 80 °C.

2.5.6 Sonication

An open Falcon tube containing 20 ml of culture was packed into a beaker with ice. The probe of an MSE Soniprep 150 (Sanyo Europe Ltd.) was placed in the culture and the following program of sonication was used: eight cycles of 30 seconds on-time and 30 seconds off-time for 10 cycles. The amplitude of sonication was 8 angstroms.

2.5.7 Quantification of TK protein in filtered lysate by Bioanalysis

Filtered lysate was analysed on an Agilent 2100 Bioanalyser, using the Protein 200 Plus Kit (supplied by Agilent Technologies) [167,168] to detect the amount of TK enzyme present. Figure 2.5 shows the calibration curve, generated using commercially available TK from *E. coli* (Fluka). Bioanalysis compares favourably with other methods of protein quantification, such as densitometry as shown in Figure 2.6. Both methods suggest similar average concentrations for the same samples.

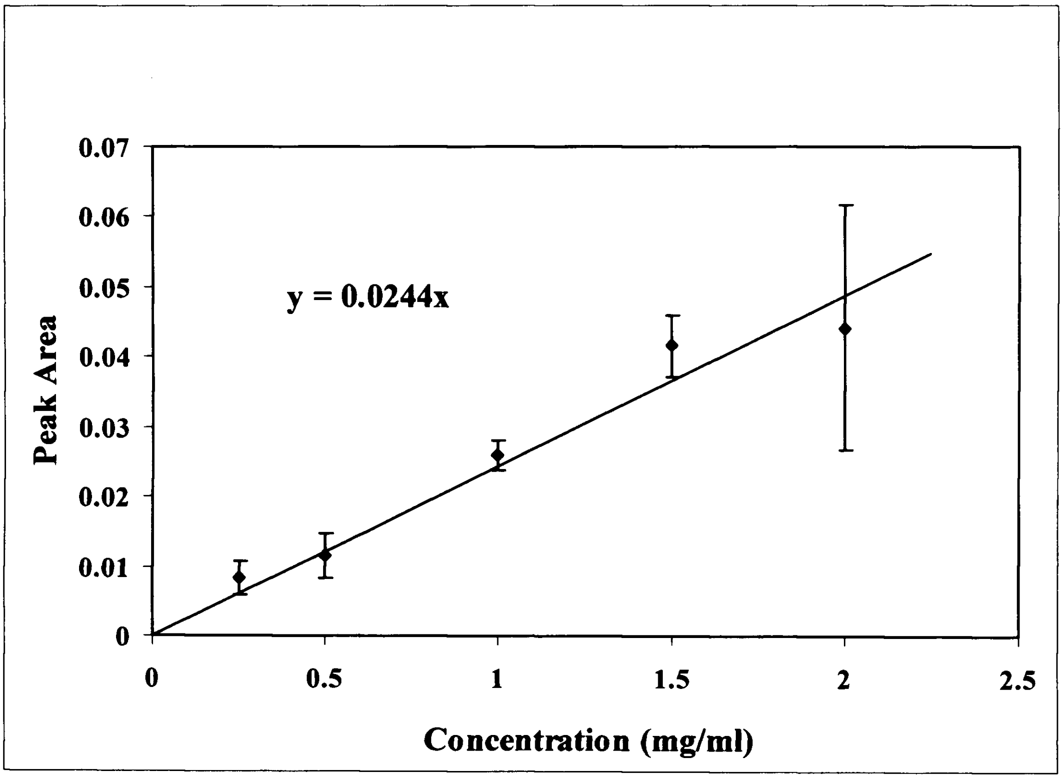


Figure 2.5: Calibration curve for commercial *EcoTK* (Fluka) on an Agilent 2100 Bioanalyser. Various concentrations of commercially available TK from *E. coli* (Fluka) were analysed using the Protein 200 Plus Kit (Agilent Technologies). Peak area refers to the average area of the characteristic TK peak at a given TK concentration.

2.5.8 Preparation of plasmid DNA

Plasmid DNA was prepared from overnight cultures using a QIAprep Spin Miniprep kit (QIAGEN Ltd.). The concentration of DNA in the product was determined by measuring its A_{260}. One A_{260} unit approximates to 50 μg.ml^{-1} dsDNA at neutral pH when a path length of 1 cm is used [169]. Plasmid DNA was stored at -20 °C.

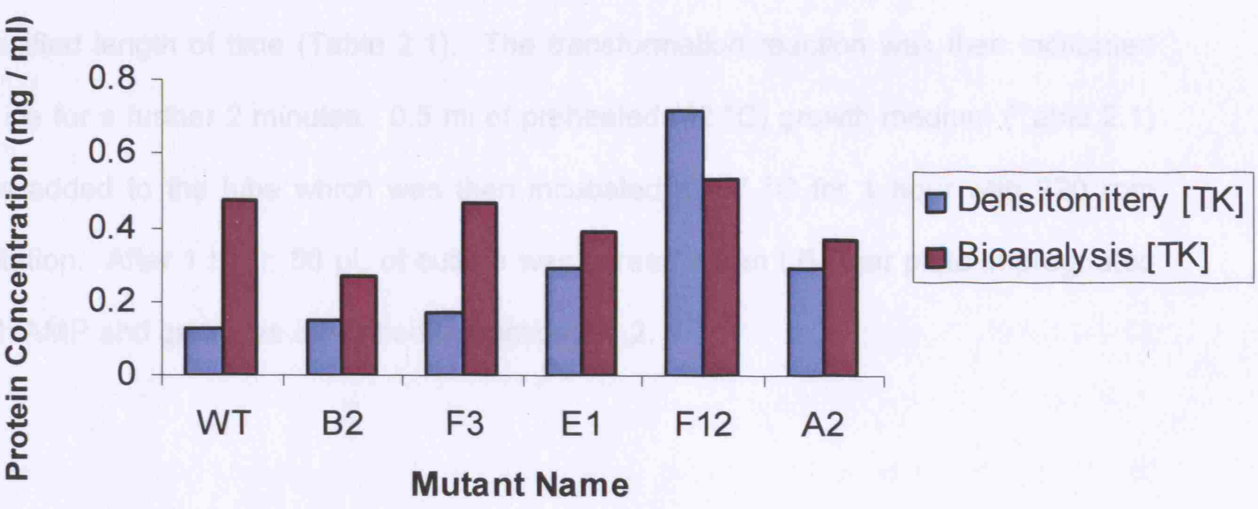


Figure 2.6: Comparison of bioanalysis and densitometry as methods of TK protein quantification. This graph was supplied by Tarik Senussi (Dept. Biochemical Engineering, UCL). Mutant names refer to another study.

E. coli strain	Aliquot (µl)	β-mercaptoethanol	Heat-shock (seconds)	Growth medium
XL1 Blue	50	-	45	LB
XL10-Gold	45	2 µl	30	LB

Table 2.1: Values for various parameters during the transformation of competent cells. The concentration of the β-mercaptoethanol solution provided with the competent XL10-Gold cells could not be discovered from Stratagene Ltd.

2.5.9 Transformation by heat-shock

Competent XL1 and XL10Gold cells are supplied by Stratagene Ltd. Cells were thawed on ice and an aliquot (Table 2.1) was transferred to a chilled 1.5 ml Eppendorf tube. If necessary, β-mercaptoethanol was added (Table 2.1) and the cells were incubated on ice for 10 minutes. 1 µl of the relevant plasmid DNA at a concentration of

50 ng/µL was added to the cells and mixed gently. The transformation reaction was incubated on ice for 30 minutes. Heat-shock was performed in a 42 °C water bath for a specified length of time (Table 2.1). The transformation reaction was then incubated on ice for a further 2 minutes. 0.5 ml of preheated (42 °C) growth medium (Table 2.1) was added to the tube which was then incubated at 37 °C for 1 hour with 220 rpm agitation. After 1 hour, 50 µL of culture was spread on an LB Agar plate impregnated with AMP and grown as described in Section 2.4.2.

2.5.10 Measurement of absorbance and optical density

Absorbance and optical density measurements were performed in a UV2 spectrophotometer (Unicam Ltd.). An ultra-micro quartz cuvette (Sigma-Aldrich Company Ltd.) was used when measuring absorbance in the UV region (200 - 400 nm wavelength). DNA samples were measured for absorbance at 260 nm, while protein absorbance was measured at 280 nm. Protein and DNA samples were sufficiently diluted to register an absorbency (or an optical density) of ≤1 AU (or ≤1 ODU) at their respective wavelengths.

2.5.11 Enzyme reactions

The standard reaction conditions for TK are described in Section 2.5.11.1. The model β-HPA + GA TK reaction is performed in Chapters 4, 5 and 6 and is described in Section 2.5.11.2 below, while the potential pyruvate + GA TK reaction attempted in Chapters 4 and 5 is described in Section 2.5.11.3. Enzyme reactions specific to Chapter 4 are described in Section 4.2.8. In all cases, reactions were quenched in the manner described in Section 2.5.12.1. Reactions were conducted at 25 °C, in quadruplicate, in 1.5 mL Eppendorf tubes. Prior to initiating each reaction, TK variants were incubated with 2.4 mM TPP, 9 mM $MgCl_2$ and 50 mM Tris-HCl (at pH 7), for 30 minutes at room temperature. TK was used at a concentration of 0.5 $mg.mL^{-1}$, based on TK quantification using the bioanalyser (Section 2.5.7).

2.5.11.1 Reaction conditions for transketolase

The principal TK-based biotransformation studied by was the synthesis of L-erythrulose from β-HPA and GA [67,52,66,170-173] adopted as the model TK reaction. Standard reaction conditions were used whenever possible to permit the comparison of data generated by different researchers (Table 2.2). The biocatalyst is derived from *E. coli* JM107 pQR791 (Section 1.1.9).

Variable	Optimum	Literature	UCL standard
pH	7.0–7.5		7.5
Preincubation period		3–30 min	30 min
Temperature	20–40 °C		25 °C
Cofactor concentrations			
Mg^{2+}		0.9–10 mM	9 mM
TPP		0.2–2.5 mM	2.5 mM
Substrate concentrations			
ß-HPA (donor)	<0.6 M		50 mM
Glycoaldehyde (acceptor)	<0.5 M		50 mM

Table 2.2: Reaction conditions for *E. coli* transketolase-based bioconversion of β-HPA and GA to L-erythrulose. "Optimum" ranges are for maximum transketolase activity [31,67]. "Literature" ranges are based on the conditions used by various laboratories around the world. "UCL standard" values are the conditions typically used by researchers at UCL.

The cofactors were used at saturating concentrations, and were incubated with the enzyme for 30 minutes prior to the addition of substrates to permit reconstitution of the holoenzyme. The GA concentration should not exceed 0.5 M because it has an inhibitory effect on TK [67]. The upper limit for β-HPA (0.6 M) is governed by its maximum solubility at room temperature and neutral pH [67].

2.5.11.2 The β-HPA + GA TK reaction

Reagents added to a 1.5 mL Eppendorf were as follows:

200 μL of a 3X (50 mM Tris-HCl with 50 mM β-HPA) stock solution

50 μL of a 12X (9 mM $MgCl_2$ with 2.4 mM TPP) stock solution

X μL of clarified lysate to give a TK concentration of 0.5 $mg.mL^{-1}$

(300 – X) μL of ddH_2O

These reagents were mixed by vortexing and left to incubate for 30 minutes at room temperature. The reaction was initiated by addition of 50 μL of a 12X GA stock solution. The mixture was then vortexed and kept at 25 °C. 100 μL samples of the reaction were taken at 5, 15, 30, 60 and 120 minute intervals and the reaction quenched as described in Section 2.5.12.1.

2.5.11.3 The potential pyruvate + GA TK reaction

Reagents added to a 1.5 mL Eppendorf were as follows:

200 μL of a 3X (50 mM Tris-HCl with 50 mM pyruvate) stock solution

50 μL of a 12X (9 mM $MgCl_2$ with 2.4 mM TPP) stock solution

X μL of clarified lysate to give a TK concentration of 0.5 $mg.mL^{-1}$

(300 – X) μL of ddH_2O

These reagents were mixed by vortexing and left to incubate for 30 minutes at room temperature. The reaction was initiated by addition of 50 µL of a 12X GA stock solution. The mixture was quickly vortexed and kept at 25 °C. 100 µL samples of the reaction were taken at 5, 30, 70, 120 and 1440 minute intervals and the reaction quenched as described in Section 2.5.12.1.

2.5.12 General HPLC methods

This protocol is based on a method developed by Christine Ingram (Department of Biochemical Engineering, UCL), but has been adapted for the 96 well plate format.

2.5.12.1 Sample preparation

The reaction sample was diluted 1:2 by the addition of 0.2 % (v/v) TFA. The addition of acid quenched the reaction by dropping the pH to well below the lower limit for TK activity (pH 6.5) [67]. The concentration of reagents in this quenched sample was determined by HPLC. Unless otherwise stated, the samples for assay on the HPLC were on a 96 well plate format. Samples for quenching were transferred to wells already containing 0.2 % TFA at the desired volume.

2.5.12.2 HPLC system configuration

The HPLC system consisted of an Endurance autosampler (Spark Holland BV), a GP50 gradient pump (Dionex Corp.), an LC30 chromatography oven (Dionex Corp.), a PC10 pneumatic controller (Dionex Corp.), and a Dionex ED40 Electrochemical Detector. A chromatography workstation running *PeakNet 5.1* (Dionex Corp.) was used to control the HPLC components and collect data from the detector.

2.5.12.3 Commonly used HPLC methods

The HPLC assay for TK activity on the model β-HPA + GA reaction used in Chapters 4, 5 and 6 is described in Section 2.4.12.3a below. The HPLC assay of TK activity for the potential pyruvate + GA reaction used in Chapter 4 and 5 is described in Section 2.5.12.3b. Assays specific to Chapter 4 are described in Sections 4.2.8 to 4.2.12.

2.5.12.3a HPLC assay for the β-HPA + GA TK reaction

The mobile phase was 0.1 % (v/v) TFA and the flow rate was 0.6 ml.min^{-1}. 10 μL of sample was injected onto a 300 mm Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories) and maintained at 60 °C. Concentrated NaOH was mixed with the output stream from the column prior to reaching the ECD.

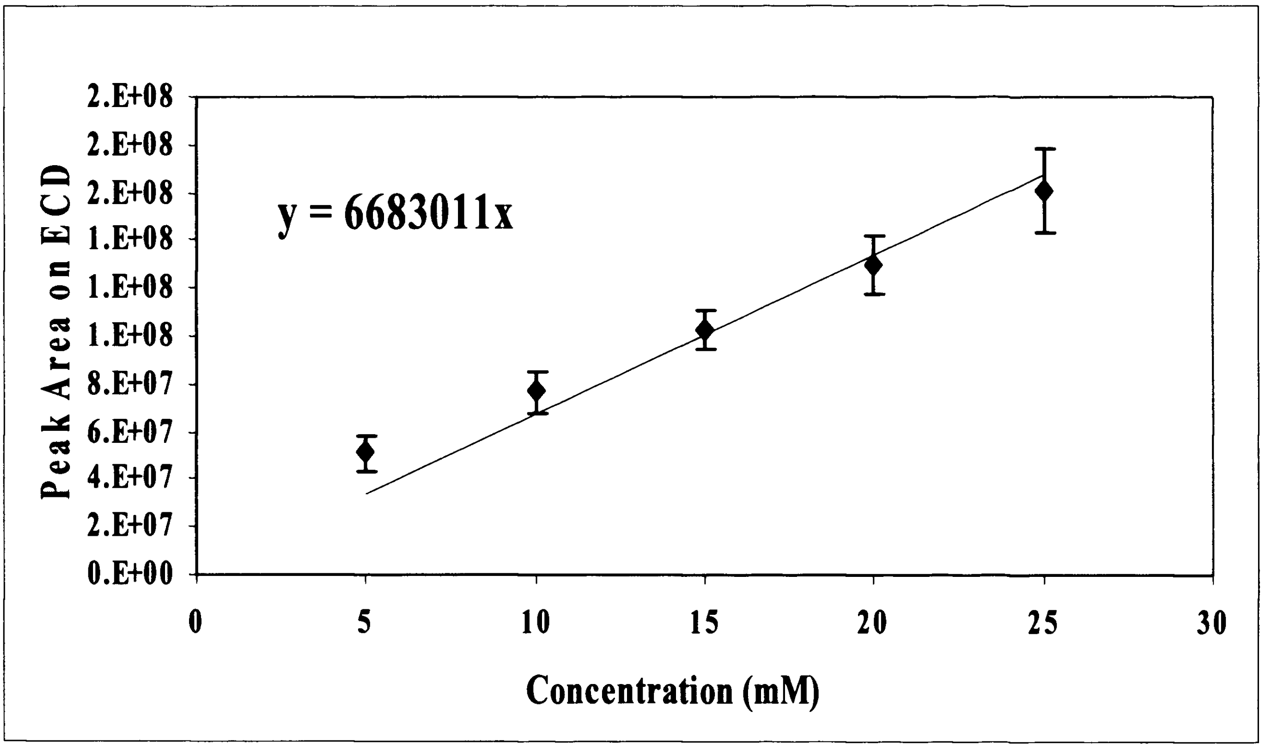


Figure 2.7: Calibration curve for L-erythrulose on 300 mm Aminex HPX-87H ion-exclusion column. Samples were prepared in the following conditions: 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 °C. Samples were then diluted 1:2 in 0.1% TFA and analysed on an Aminex HPX-87H ion-exclusion column.

The retention times of β-HPA and L-erythrulose and GA using this method are 5.3, 7.2 and 7.7 minutes, respectively. The progress of the reaction was followed by the appearance of L-erythrulose product, the peak area of which was used in subsequent analysis. Figure 2.7 illustrates the calibration curve for L-erythrulose on the 300 mm Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories). The relationship between injection concentration and peak area is linearly proportional up to 25 mM.

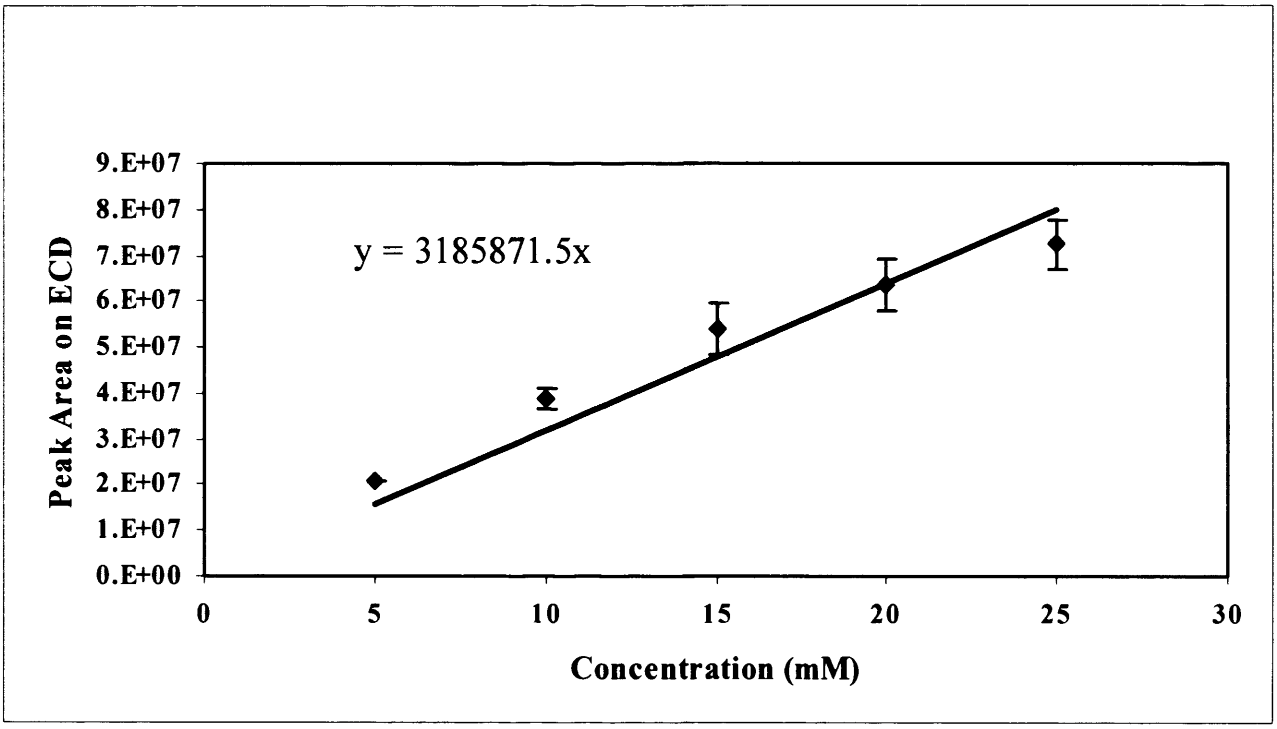


Figure 2.8: Calibration curve for pyruvate on a 300 mm Aminex HPX-87H ion-exclusion column. Samples were prepared in the following conditions: 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 °C. Samples were then diluted 1:2 in 0.1% TFA and analysed on an Aminex HPX-87H ion-exclusion column.

2.5.12.3b The pyruvate + GA reaction

The HPLC configuration was exactly as described in Section 2.5.12.3a except that the flow rate was 0.3ml.min^{-1}. The (*S*)-3,4-dihydroxybutan-2-one used to determine the retention time of the pyruvate + GA reaction product was supplied by Mark Smith of the Chemistry Department, UCL.

The retention times of pyruvate, GA and to (*S*)-3,4-dihydroxybutan-2-one using this method are 10.42, 12.80 and 14.25 minutes, respectively. The progress of the reaction was followed by the appearance of the (*S*)-3,4-dihydroxybutan-2-one product, although not enough of this substrate was available to generate a calibration curve. The disappearance of pyruvate was also monitored. Figure 2.8 illustrates the calibration curve for pyruvate on the 300 mm Aminex HPX-87H ion-exclusion column

(Bio-Rad Laboratories). The relationship between injection concentration and peak area is linear up to a 25 mM concentration of pyruvate.

2.5.13 Agarose gel electrophoresis

A GNA-100 system (Amersham Biosciences Ltd.) was used for agarose gel electrophoresis of DNA. The amount of agarose used in a particular gel depended upon the desired linear range of DNA fragment separation: 0.7 % (w/v) was used for 0.8–10.0 kbp fragments (plasmid DNA) and 2.0 % (w/v) was used for 800–2000 bp fragments (PCR products).

The appropriate amount of agarose was dissolved in 50 ml of 1× TAE buffer (40 mM Tris-acetate and 1 mM EDTA in pure water) by heating in a microwave. 0.5 mg.L^{-1} ethidium bromide was added to the gel to permit the visualisation of DNA by UV light. A comb was inserted at one end to form the sample wells. After the gel had set it was submerged in buffer containing Tris-acetate and EDTA (TAE buffer) in the gel tank and the comb was removed.

Samples were prepared for loading by adding 6× loading buffer (Bromophenol blue (0.05% w/v), sucrose (40% w/v), EDTA (0.1 M, pH 8.0), and SDS (0.5% w/v) (Sigma-Aldrich Company Ltd.). The wells of sample lanes were loaded with 3 µl of sample. The wells of marker lanes were loaded with 1.5 µl of Novagen 0.5–12.0 kbp Perfect DNA Markers (EMD Biosciences Inc.) or Novagen 50–2000 bp PCR Markers (EMD Biosciences Inc.).

2.5.14 SDS-PAGE

A Mini-Protean II system (Bio-Rad Laboratories) was used for SDS-PAGE of proteins. 12.5 % (w/v) acrylamide gels were used for all SDS-PAGE analyses.

2.5.14.1 Stock solutions

Gels were prepared using the methods of Laemmli (1970) [176]. The acrylamide solution was 29.2 % (w/v) acrylamide and 0.8 % (w/v) N,N'-methylene bisacrylamide in water (Bio-Rad Laboratories). The separating gel buffer was 1.5 M Tris buffer (pH 8.8). The stacking gel buffer was 0.5 M Tris buffer (pH 6.8). Ammonium persulphate, 10% (w/v) and TEMED were added at the last minute to initiate polymerisation. The running buffer was 0.05 M Tris-HCl, 0.38 M glycine, and 0.1 % (w/v) SDS in pure water, adjusted to pH 8.8. The staining solution was 0.05 % (w/v) Coomassie Brilliant Blue, 50 % (v/v) methanol, and 10 % (v/v) acetic acid in pure water.

2.5.14.2 Gel casting

The gel cassette was assembled according to the manufacturer's instructions. 12.5 % (w/v) separating and 6 % (w/v) stacking gels were prepared according to Table 2.3. The separating gel was poured first, overlaid with isopropanol to ensure a flat surface, and allowed to set. The solvent was carefully removed and the stacking gel was poured above the first gel. A comb was positioned in the stacking gel to form the sample wells. After polymerisation was complete, the comb was removed and the gel cassette was secured in the electrophoresis tank.

2.5.14.3 Sample preparation, electrophoresis and staining

Samples were mixed with 2× Laemmli Sample Buffer (Bio-Rad Laboratories), containing 62.5 mM Tris-HCl, pH 6.8, 25% glycerol, 2% SDS and 0.01% Bromophenol Blue. Samples were then heated to 100 °C for 2 minutes to denature the protein. The wells of sample lanes were loaded with 20 µl of sample. The wells of marker lanes were loaded with 5 µl of Precision Plus Protein Standards (Bio-Rad Laboratories).

Electrophoresis was performed at 100 V for 3 hours.

Component	Separating gel (ml)	Stacking gel (ml)
Acrylamide solution	4.2	2.0
Stacking gel buffer	0.0	2.5
Separating gel buffer	2.5	0.0
10 % (w/v) SDS solution	1.0	1.0
Pure water	2.3	4.5
10 % (w/v) APS	0.1	0.1
TEMED	0.01	0.01

Table 2.3. Components of 12.5 % (w/v) separating and 6 % (w/v) stacking gels. APS and TEMED were added to each gel at the last minute to initiate polymerisation.

Protein bands were visualised by staining in an aqueous staining solution containing 0.1% (w/v) Coomassie Blue R-250, 40% (v/v) methanol and 10% (v/v) acetic acid. The gel was placed in a plastic container, covered with 50 ml of staining solution, and microwaved on full power for 3 minutes. The stain was poured away and the gel was destained by boiling for 10 minutes in 1 litre of pure water. The gel was photographed using a Gel Doc 2000 system (Bio-Rad Laboratories).

2.3.15 DNA sequencing

Cycle sequencing [174] was performed at the Wolfson Institute for Biomedical Research (UCL). All DNA samples were given to the service at a concentration of 100 fmoles in 6 µl. Sequencing primers were generated for previous studies at this laboratory. When sequencing mutants where the amino acid targets were positions 23, 29, 64, 183 or 259, the TKN primer was used (5'-GATCCAGAGATTTCTGA-3'). When sequencing to check for mutations in the 381, 383, 384 or 385 positions, the LibCeA primer was used (5'-CAGCCGGTTGAGCAGGTCG-3'). When checking for mutations at the 453, 461, 492, 527 or 540 positions, the TKC primer was used for sequencing (5'-TATCTCCCTGCACGGTGGCTTCC-3').

Chapter 3: Phylogenetic study of the TPP-dependent enzymes

3.1 Introduction

TPP, the biologically active form of thiamine is the cofactor for many important enzymes (Section 1.2). The members of this TPP-dependent enzyme family perform diverse reactions, such as oxidative and non-oxidative decarboxylations, carboligations and carbon carbon bond cleavage. The variety of enzyme functions is reflected by the domain arrangements found in different enzymes of the family (Section 1.2.2), while the use of TPP as cofactor by these enzymes suggests a common ancestry. The diversity of substrates used by members of the family is of great interest industrially, as discussed in Section 1.2.1. An evolutionary analysis of the TPP-dependent enzymes may shed light on how domain arrangement and residue variation affects substrate specificity in different enzymes of the group, and in turn lead to insights into how different functions have arisen during evolution and how new functions may be engineered in TK.

A previous study detailing the cloning of the PDC from *S. ventriculi* into *E. coli* contained a dendrogram summarising the relationships between 66 TPP-dependent enzymes representing TK, PDH-E1, ALS, PDC and IPDC [175]. That phylogeny was constructed using the *MultAlin* program [177], which operates using hierarchical clustering. In the paper, no allowance seems to be made for the differing domain arrangements of TK and the other enzymes (Section 1.2.2). A recent study on the domain relationships between TPP-dependent enzymes suggested possible evolutionary routes for the enzymes [178]. An evolutionary study based on amino acid sequence could suggest the finer details of evolution in the enzyme family, particularly within enzyme groups and is thus timely.

In this study, seventeen TPP-dependent enzymes, TK, DXPS, DHAS, PKL, 2OXO, PFRD, PDC, IPDC, PhPDC, PO, ALS, GXC, BFDC, BAL and OCADC, SPDC and PPDC are examined. These enzymes represent the six TPP-dependent enzyme groups described in Section 1.2.2. A multiple sequence alignment and phylogenetic tree is constructed for each enzyme individually, before the phylogeny of all TPP-dependent enzymes together is constructed. Information from TK phylogeny and the evolutionary study of the TPP-dependent enzymes will be applied to the biocatalytic engineering of TK in subsequent chapters. Thus a more detailed analysis of the TK sequence alignment and phylogeny is performed, examining the distribution of highly conserved residues in TK. Finally the 17 TPP-dependent enzyme are examined together. Since domain arrangements differ between enzymes, the PP and Pyr domains for the enzymes under study are compiled and aligned separately. Crystal structures representing four of the six enzyme groups (N.B. both PDC and PO are studied) are examined structurally and equivalent regions are compared between enzymes. Concatenation of structurally equivalent regions of highest sequence homology from the PP and Pyr domains allows the evolutionary tree for the TPP-dependent enzyme family to be constructed.

3.2 Methods

In choosing the seventeen enzymes for study, each of the six domain arrangements for TPP-dependent enzymes was represented. Discussed in Section 1.2.2, these arrangements are referred to as TK-like, 2OXO-like, the PFRD arrangement PDC-like, the SPDC arrangement and the PPDC arrangement. 2OXO represents the only complex associated TPP-dependent enzyme in this study. Other complex associated TPP-dependent enzymes, all of which comprise of α and β subunits, like 2OXO were excluded, since they are poorly represented in the literature and in some cases, as with PDH-E1, homology with the other TPP-dependent enzymes was found to be very low [179]. Thus, 2-oxoglutarate dehydrogenase, PDH-E1 and acetoin dehydrogenase were excluded. Other enzymes excluded from the analysis were pyruvate-formate lyase, *N*^{2}-(2-carboxyethyl)arginine synthase, CDP-4-aceto-3,6-dideoxygalactose synthase, sulfoacetaldehyde acetyltransferase, 2-ketoisovalerate decarboxylase, 2-hydroxyphytanoyl-CoA lyase and (1R,6R)-2-succinyl-6-hydroxy-2,4-cyclohexadiene-carboxylate synthase.

Low homology is defined as <30 % sequence identity or similarity, a range of values sometimes referred to as the "twilight zone" [180]. Above 30 % the case for divergent evolution is strong, with homologous enzyme exhibiting strong similarities not only at the sequence level, but also in three-dimensional structure [181].

Below the 30 % homology level, the inference of evolutionary relationships between enzymes cannot rely on sequence data alone. While structural homology between the different TPP-dependent enzyme groups is observed, particularly for regions involved in TPP-binding (Section 1.1.3), sequence homology between TPP-dependent enzymes of different types is low. Within each group, sequence homology is high enough to infer evolutionary relationships. Thus, the approach presented here is to align enzymes using sequence homology within groups and then, using structural studies, compare enzymes from different groups. In this way, we can elucidate which regions of secondary structure are common to all TPP-dependent enzymes and infer phylogeny based on the most homologous stretches of sequence within these comparable regions.

3.2.1 Sequence Alignments for individual TPP-dependent enzymes

In the cases of DXPS, DHAS, PKL, 2OXO, PFRD, PDC, IPDC, PhPDC, PO, ALS, GXC, BFDC, BAL, OCADC, SPDC and PPDC all sequences were retrieved using a *BLAST* search (Section 2.2 for URL) with default parameters. Sequences for each enzyme were exported in the FASTA format and aligned using *ClustalW* [81,183]. Alignments were refined by eye where necessary, bearing in mind residues thought to be functionally important (Table 3.4), then converted to the Phylip [182] format for use in phylogenetic studies. For the purposes of the individual alignments, the SPDC alpha (Pyr) and beta (PP) subunits were concatenated. Similarly, for construction of the 2OXO tree, the alpha (PP) and beta (Pyr-TKC) were concatenated. In the case of the *Tma*-like PFRDs, domain I (Pyr) and domain VI (PP) were aligned separately with the PP and Pyr domains from the *Daf*-like PFRDs, before the PP and Pyr domains were concatenated.

3.2.1.1 Sequence alignment for TK

There is a discrepancy between the residue numbers in the *E. coli* TK sequence used in the TK alignment (and subsequent studies) and those found in the crystal structure for *E. coli* TK (1QGD.pdb). This is discussed fully in Appendix 1 (Section A1.3).

3.2.1.1a Choice of transketolase sequences to study

Initially, TK amino acid sequences were obtained from a *BLAST* search where *Eco*TK (X680125) was the query sequence. Fifty-four TK sequences representing Bacteria, Fungi and Plants, as well as one Trypanosome were chosen, exported in FASTA format and viewed using the *Bioedit* software [183]. An in depth description of how these fifty-four sequences were chosen is given in Appendix 1 (Section A1.1).

3.2.1.1b Sequence alignment of 54 TK sequences

Sequences were aligned in *Bioedit* using the *ClustalW* [79] algorithm with the default parameters. Sequences were further aligned by eye, avoiding gaps in the secondary structure of *Eco*TK. The N- and C-terminal regions of TK amino acid sequences varied somewhat in length between different species. Those amino acid residues that aligned with residues 10 – 627 of *Eco*TK were selected, while amino acids outside of this region were deleted from the alignment and not considered in subsequent analyses.

3.2.1.1c Alignment of PP and Pyr domains from the TPP-dependent enzymes.

As discussed in Section 3.2.1 the TPP-dependent enzymes examined fall into six topological categories. On the basis of these groupings, individual alignments for each TPP-dependent enzyme group were compiled with the ultimate aim of generating PP and Pyr domain alignments for all seventeen TPP-dependent enzymes together.

Firstly the TK-like enzymes were aligned. The next Alignment generated was for the PDC-like enzymes. The PPDC SPDC alpha and beta subunits were aligned, using the study of Graupner *et al.* [97] as a guideline. The SPDC / PPDC subalignment was then aligned with the PDC-like enzyme alignment. For four of the six TPP-dependent groups under examination at least one crystal structure has been solved. No structure exists for SPDC or PPDC, but their sequences share high homology with the PDC-like enzyme group. Examination of the crystal structures of *Eco*TK (1QGD.pdb), *Ppu*2OXO (2BP7.pdb), *Daf*PFRD (1BOP.pdb), *Sce*PDC (1PVD.pdb) and *Lpl*PO (POX.pdb) using the *Pymol* program allowed the identification of regions of secondary structure common to all TPP-dependent enzyme types in the PP and Pyr domains.

Due to the differing arrangements of the catalytic PP and Pyr domains in the TPP-dependent enzymes (Section 1.2.2), alignments for the PP and Pyr domains were generated separately. In the case of the alignment of PDC-like enzymes, residues found to align with *Sce*PDC residues 2 - 169 were deemed to represent the Pyr domain regions for these enzymes, while residues 360 - 540 were defined as the PP domain. For the TK-like enzymes, the PP domain was defined as those residues aligning with *Eco*TK residues 25 - 245, while the Pyr domain was defined as the regions aligning with residues 357 - 520 from the same enzyme.

For each of the PP and Pyr domains, the TK-like, PDC-like (prealigned with the SPDC and PPDC sequences) were manipulated along with the PFRD and 2OXO sequences, ensuring that structurally equivalent regions aligned. Where residues from different enzyme types were found to align in a region of common secondary structure, the three dimensional location of these residues was compared in the crystal structures, ensuring that sequence homology was backed up by structural equivalence.

While aligning, attention was also paid to residues known to be functionally important. Once the PP and Pyr alignments were constructed, regions were chosen for use in the phylogenetic analysis. Only residues found in regions of secondary structure common to al TPP-enzymes examined were used in the analysis. The informative regions were compiled from the PP and Pyr domain alignments, concatenated and used in the generation of the phylogenetic tree, as discussed in Section 3.2.3.

3.2.2.1 Phylogenetic analyses of individual TPP-dependent enzymes

Tree phylogenies for each enzyme were obtained using the Phylip program *ProML* [182], which uses the ML method. Resulting trees were viewed using the *Mega3* program [184]. Since the TK phylogeny is used to generate ancient TK enzymes in Chapter 4, an in depth description of the generation of the TK phylogeny is given in Section 3.2.2.2.

3.2.2.2 Phylogenetic analysis of transketolase

3.2.2.2a Programs used for Phylogenetic inference

NJ, ME and UPGMA distance trees were constructed in the *Mega2* [184] software package using the p-distance, the gamma distance, the number of differences and the Poisson correction distance.

Additional NJ and UPGMA trees were constructed in *Phylip* [182] using the Jones Taylor Thornton, Dayhoff PAM, Kimura and Categories models (all with and without assuming a gamma distribution). Distance data was calculated using the *Phylip* [182] program *Protdist* and resulting datasets were input into the Phylip *Neighbour* program for iteration of the UPGMA and NJ algorithms. In the categories model a Transition / Transverions ratio of 0.521 was used, corresponding to the average Transition / Transverions ratio for the 54 sequences as calculated by *Mega2*.

Parsimony and ML trees were constructed using the Phylip [182] programs *Propars* and *ProML* respectively. All trees were viewed using the *Treeview* program [185]. For illustrative purposes, some trees were edited using *Adobe Illustrator CS*.

3.2.2.2b Choice of most suitable phylogenetic tree for transketolase

Trees were assessed on the basis of how TKs from different species clustered and how they agreed with the universal tree (Section 1.7.3). In general, it is expected that TKs will group according to species, since TK it is likel to be a slow evolving enzyme. The choice of the best TK phylogeny is detailed in Appendix 1 (Section A1.2).

3.2.3 Construction of the phylogenies for the PP and Pyr domains of TPP-dependent enzymes

The concatenated alignment of the most homologous, structurally equivalent residues from the PP and Pyr domains (Section 3.2.1.1c) was input into the Phylip program *Protdist*, which generated data matrices for the alignments. These matrices were then input into the program *Neighbour*, which constructed the trees. The more computationally intensive ML method was also used, implemented by the Phylip program *ProML*. Resulting phylogenies were viewed in *Mega3*. To simplify the analysis of the trees, certain clades were collapsed in *Mega3*. For example if 2 firmicutes grouped together in a clade they will be replaced by one branch labelled "Firmicutes (2)". Where a more detailed discussion of individual sequences is required, the overall, uncollapsed tree can be referenced.

Species Name	Abbreviation	Accession no.
Pseudomonas aeruginosa	(*Pae*)	NC 002516
Mesorhizobium loti	(*Mlo*)	NC 002678
Listeria monocytogenes strain EGD	(*Lmo1*)	NC 003210
Listeria monocytogenes EGD-e	(*Lmo4*)	NC 003210
Agrobacterium tumefaciens str. C58	(*Atu*)	AE009144
Bacillus halodurans	(*Bha*)	NC 002570
Brucella melitensis biovar Abortus	(*Bme1*)	NC 003318
Buchnera aphidicola str. APS [Acyrthosiphon pisum]	(*Bsp*)	NC 002528
Bacillus subtilis subsp. subtilis str. 168	(*Bsu*)	NC 000964
Clostridium acetobutylicum	(*Cac1*)	NC 003030
Clostridium acetobutylicum	(*Cac2*)	NC 003030
Clostridium acetobutylicum	(*Cac3*)	NC 003030
Candida albicans	(*Cal*)	NC 003030
Capsicum annuum	(*Can*)	T09541
Campylobacter jejuni subsp. jejuni NCTC 11168	(*Cje*)	NC 002163
Chlamydophila pneumoniae CWL029	(*Cpn1*)	NC 000922
Chlamydia trachomatis	(*Ctr2*)	NC 000117
Deinococcus radiodurans R1	(*Dra*)	NC 001263
Escherichia coli	(*Eco1*)	X68025
Haemophilus influenzae Rd KW20	(*Hin*)	NC 000907
Helicobacter pylori J99	(*Hpy*)	NC 000921
Kluyveromyces lactis	(*Kla*)	Q12630
Listeria innocua	(*Lin2*)	NC 003212
Listeria innocua	(*Lin3*)	NC 003212
Lactococcus lactis subsp. Lactis	(*Lla*)	NC 002662
Leishmania mexicana mexicana	(*Lme*)	CAD20572
Mycoplasma genitalium	(*Mge*)	U39686
Mycobacterium leprae	(*Mle*)	NC 002677
Mycoplasma pulmonis	(*Mpu*)	NC 002771
Neurospora crassa	(*Ncr*)	CAC18218
Neisseria meningitidis MC58	(*Nme*)	NC 003112
Nostoc sp. PCC 7120	(*Nsp*)	NP 487384
Pichia stipitis	(*Pst*)	Z26486
Rhodobacter capsulatus	(*Rca*)	JC4637
Ralstonia eutropha	(*Reu*)	NC 005241
Ralstonia solanacearum	(*Rso*)	NC 003295
Rhodobacter sphaeroides	(*Rsp*)	P29277
Staphylococcus aureus subsp. aureus Mu50	(*Sau*)	NC 002745
Saccharomyces cerevisiae	(*Sce1*)	NP 015399
Saccharomyces cerevisiae	(*Sce2*)	NP 009675
Salmonella enterica subsp. enterica serovar Typhi	(*Sen1*)	NC 003198
Salmonella enterica subsp. enterica serovar Typhi	(*Sen2*)	NC 003198
Sinorhizobium meliloti	(*Sme*)	NC 003078
Streptococcus pneumoniae R6	(*Spn*)	NC 003098
Schizosaccharomyces pombe	(*Spo*)	C 0034231
Solanum tuberosum	(*Stu*)	S58083
Thermosynechococcus elongatus BP-1	(*Tel*)	P 682660
Treponema pallidum subsp. pallidum str. Nichols	(*Tpa*)	NC 000919
Ureaplasma urealyticum	(*Uur*)	NC 002162
Vibrio cholerae O1 biovar eltor str. N16961	(*Vch*)	NC 002505
Xylella fastidiosa 9a5c	(*Xfa*)	NC 002488
Xanthobacter flavus	(*Xfl*)	U29134
Yersinia pestis CO92	(*Ype*)	NC 003143

Table 3.1: Species used in the phylogenetic analysis of TK. Abbreviations given are used throughout the analysis. *BLAST* accession numbers are given in the far right column.

3.3 Results

3.3.1 Sequence Alignments for individual TPP-dependent enzymes

The individual sequence alignments (Alignments A3.1 to A3.17) are shown on the CD-ROM in FASTA format. For each alignment the mean sequence identity between two species in that alignment was calculated. These values were as follows: TK: 42.30%; DXPS: 47.35 %; DHAS: 41.10 %; PKL: 42.69 %; 2OXO: 35.82 %; PFRD: 39.05 %; PDC: 39.62 %; IPDC: 42.41 %; PhPDC: 32.30 %; PO: 29.87 %; ALS: 33.75 %; GXC: 64.98 %; BFDC: 29.56 %; BAL: 34.47 %; OCADC: 34.48 %; SPDC: 44.24 %; PPDC: 42.00 %.

3.3.2 Further analysis of 54 transketolase sequences

3.3.2.1 Choice of TK sequences for study

The procedure leading to the choice of the 54 TK sequences (shown in Table 3.1) for study is described in Appendix 1 (Section A1.2).

3.3.2.2 Sequence alignment for transketolase

As mentioned in Section 3.3.1, the alignment of the 54 TK sequences is found in Alignment 3.1 on the CD-ROM. The 100% conserved (by identity or similarity) residues in Alignment 3.1 are listed below and summarised in Table 3.2. Underlined residues are those known to be important for TK function from the literature while those residues italicised are for cases where the residue overwhelmingly conserved at this site is different from the character state found in *Eco*TK (the numbering for which is used throughout the analysis). The presence of *Bme*TK and *Cac2*TK caused very few amino

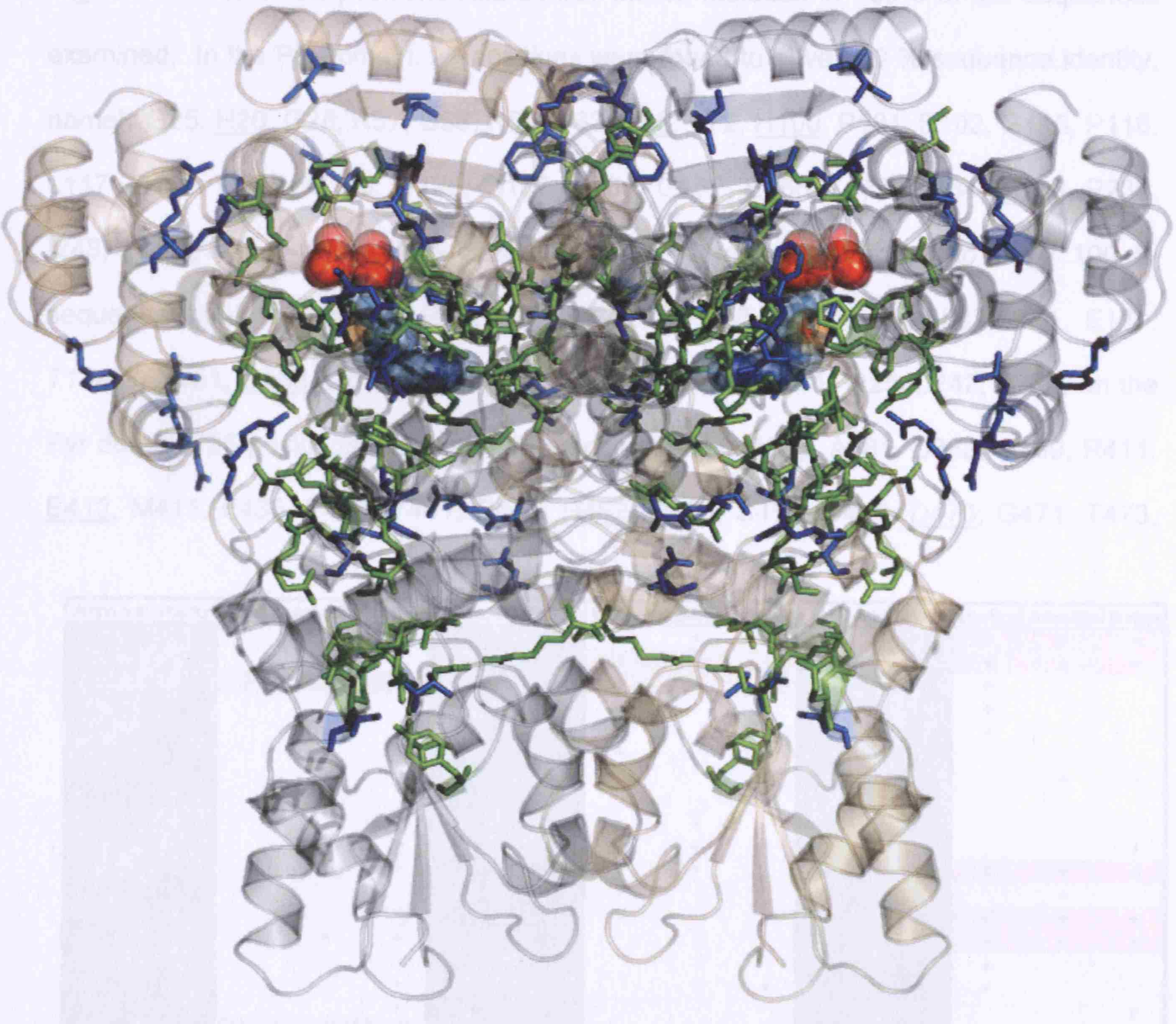


Figure 3.1: Distribution of 100 % conserved residues from TK alignment (Table 3.1). Those residues with 100 % sequence identity are shown in green, while those with 100 % sequence similarity appear in blue. Subunit A is shaded grey, while subunit B is shaded light orange. The TPP moiety is represented with a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

acid positions to be found as 100 % conserved. Removing *Cac2*TK and *Bme*TK from our analysis yielded an alignment with 93 highly conserved positions.

At 61 positions the same amino acid residue was found in all 52 TK species aligned. A further 32 positions had similar amino residues in 100% of the sequences examined. In the PP domain, 29 positions were found to have 100 % sequence identity, namely G25, H26, G28, R57, D58, R59, S63, H66, Y72, H100, P101, E102, G115, P116, L117, G118, Q119, G120, G126, G157, E161, G162, E166, L176, D184, N186, P239, I248, H262. A further 22 positions in the PP domain were found to have 100 % sequence similarity, namely R12, M31, A34, L40, W54, V61, E86, L87, R91, E111, *T144(S)* Y151, D156, I163, Y183, I190, D191, *A208(S)*, W211, I225, L242, F285. In the Pyr domain, 25 positions had 100% sequence identity: R359, A381, D382, G409, R411, E412, M415, F435, F438, Y441, R447, H462, D463, S464, G468, D470, G471, T473,

Amino Acid	Position	Interface	Active Site	Surface
R	12			
G	25		+	+
H	26	+	+	+
G	28			
M	31			
A	34			
L	40			
W	54			
R	57			+
D	58			+
R	59			+
V	61			
S	63			+
H	66		+	+
Y	72			
E	86			+
L	87			+
R	91	+	+	+
H	100	+	+	+
P	101			
E	102	+		+
E	111			
T(S)	113	+		
G	114	+	+	+
P	115	+	+	+
L	116	+	+	+
G	117	+		
Q	118			
G	119			
G	125			
Y	150			
D	155		+	
G	156	+	+	+
E	160	+	+	+
G	161	+		
I	162			
E	165	+		
L	175			+
Y	182			
D	183		+	
N	185		+	+
I	189	+	+	
D	190	+		
A(S)	207	+		
W	210			
I	224			
P	238			+
L	241			
I	247		+	+
H	261		+	+
F	284			+
R	358		+	+
A	380		+	+
D	381	+	+	+
L	382	+	+	+
G	408	+		
V	409	+	+	+
R	410	+		
E	411	+	+	+
M	414			
N	419			
F	434		+	+
F	437	+	+	+
Y	440	+	+	+
R	446	+		
A	449	+		
L	450	+		
T	460			
H	461		+	+
D	462			
S	463			
G	467			+
D	469	+	+	+
G	470		+	
T	472	+	+	+
H	473	+	+	+
Q	474	+		
E	477	+		
R	483			+
R	492			
D	495			
E	498			
S	519			
R	520			+
Q	521			+
Y	541			+
T	557			
G	558			+
S	559			+
E	560			+
V	561			
S	583			
E	613			

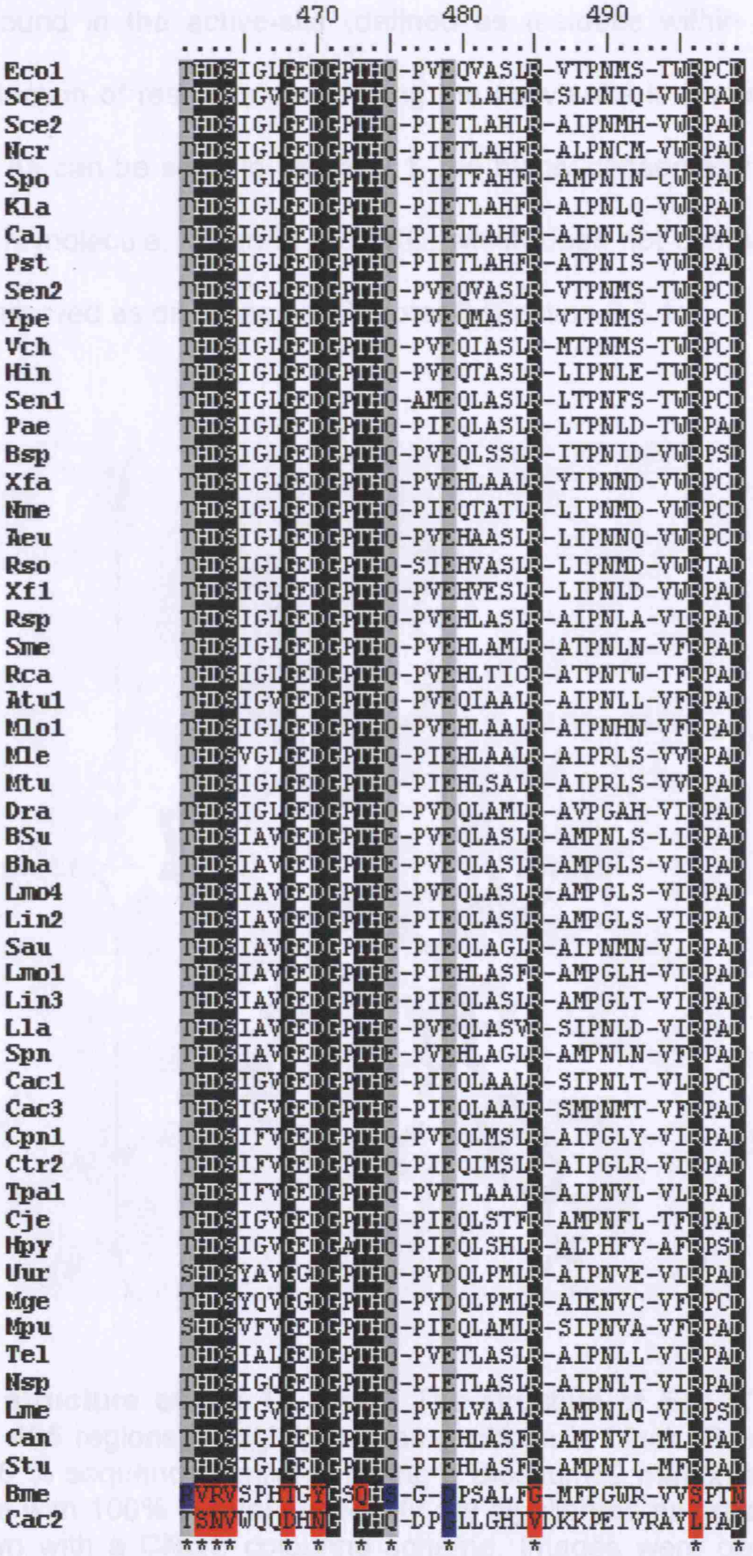
Table 3.2: Locations of highly conserved residues, either at the subunit interface, the active-site or the surface of the TK molecule. Those residues that can be categorised as being present at all three locations are

H474, R484, R493, D496, E499, R521, and Q522. The following 9 positions had sequence similarity of 100 %: L383, V410, N420, A450, L451, T461, Q475, E478, and S520. The TKC domain of the alignment contained 6 positions 100 % sequence identity: Y542, G559, S560, E561, S583 and E613, while 2 positions had 100% sequence similarity: T558 and V562.

Title	**Sequence**
TK motif	(S/T), H, (D/C), (S/G), X_3, **G**, X_2, **G**, P, (S/T), (Q/H), X_9, R, , X_8, (R/Y), P, X_1, **D**
NADH-binding like motif	**G**, X_1, **G**, X_2, **G**, ------------X_{24}---------, **D**
Consensus sequence for 52 TK sequence alignment (excluding *Cac2*TK and *Bme*TK)	(S/T), H, D, S, X_3, **G**, X_2, **G**, (P/A), T, H, X_9, R, , X_8, R, (P/T), X_1, **D**
Consensus sequence for 54 TK sequence alignment	(S/T/R), (H/V/S), (D/R/N), (S/V), X_3, (**G**/T/D), X_2, **G**, (P/A/S/Y), (T/Q), H, X_9, (R/G/V), (gap Cac2 D), X_8, (R/S/C), (P/T), X_1, (**D/N**)
Modified TK motif (excluding *Cac2*TK and *Bme*TK)	(S/T), H, (D/C), (S/G), X_3, **G**, X_2, **G**, (P/A), (S/T), (Q/H), X_9, R, , X_8, (R/Y), (P/T), X_1, **D**
Modified TK motif considering all 54 TK sequences	(S/T/R), (H/V/S), (D/C/R/N), (S/G/V), X_3, (**G**/T/D), X_2, **G**, (P/A/S/Y), (S/T/Q), (Q/H), X_9, (R/G/V), X_8, (R/Y/S/C), (P/T), X_1, (**D/N**)

Table 3.3: Comparison of the previously published TK motif and the consensus sequence for the 54 TK alignment (Alignment 3.1). The TK motif definition was modified accordingly. The affect of *Cac2*TK and *Bme*TK on the definition of the TK motif is also examined.

The distribution of conserved residues is shown in the crystal structure of TK (Figure 3.1). Examining the TK structure, of the conserved residues 32 were found at the interface of the 2 TK subunits, 43 were found at the surface of the TK molecule,



```
                    470       480       490
            ....|....|....|....|....|....|....|....
Eco1        THDSIGLGEDGPTHQ-PVEQVASLR-VTPNMS-TWRPCD
Sce1        THDSIGVGEDGPTHQ-PIETLAHFR-SLPNIQ-VWRPAD
Sce2        THDSIGLGEDGPTHQ-PIETLAHLR-AIPNMH-VWRPAD
Ncr         THDSIGLGEDGPTHQ-PIETLAHFR-ALPNCM-VWRPAD
Spo         THDSIGLGEDGPTHQ-PIETFAHFR-AMPNIN-CWRPAD
Kla         THDSIGLGEDGPTHQ-PIETLAHFR-AIPNLQ-VWRPAD
Cal         THDSIGLGEDGPTHQ-PIETLAHFR-AIPNLS-VWRPAD
Pst         THDSIGLGEDGPTHQ-PIETLAHFR-ATPNIS-VWRPAD
Sen2        THDSIGLGEDGPTHQ-PVEQVASLR-VTPNMS-TWRPCD
Ype         THDSIGLGEDGPTHQ-PVEQMASLR-VTPNMS-TWRPCD
Vch         THDSIGLGEDGPTHQ-PVEQIASLR-MTPNMS-TWRPCD
Hin         THDSIGLGEDGPTHQ-PVEQTASLR-LIPNLE-TWRPCD
Sen1        THDSIGLGEDGPTHQ-AMEQLASLR-LTPNFS-TWRPCD
Pae         THDSIGLGEDGPTHQ-PIEQLASLR-LTPNLD-TWRPAD
Bsp         THDSIGLGEDGPTHQ-PVEQLSSLR-ITPNID-VWRPSD
Xfa         THDSIGLGEDGPTHQ-PVEHLAALR-YIPNND-VWRPCD
Nme         THDSIGLGEDGPTHQ-PIEQTATLR-LIPNMD-VWRPCD
Aeu         THDSIGLGEDGPTHQ-PVEHAASLR-LIPNNQ-VWRPCD
Rso         THDSIGLGEDGPTHQ-SIEHVASLR-LIPNMD-VWRTAD
Xf1         THDSIGLGEDGPTHQ-PVEHVESLR-LIPNLD-VWRPAD
Rsp         THDSIGLGEDGPTHQ-PVEHLASLR-AIPNLA-VIRPAD
Sme         THDSIGLGEDGPTHQ-PVEHLAMLR-ATPNLN-VFRPAD
Rca         THDSIGLGEDGPTHQ-PVEHLTICR-ATPNTW-TFRPAD
Atu1        THDSIGVGEDGPTHQ-PVEQIAALR-AIPNLL-VFRPAD
Mlo1        THDSIGLGEDGPTHQ-PVEHLAALR-AIPNHN-VFRPAD
Mle         THDSVGLGEDGPTHQ-PIEHLAALR-AIPRLS-VVRPAD
Mtu         THDSIGLGEDGPTHQ-PIEHLSALR-AIPRLS-VVRPAD
Dra         THDSIGLGEDGPTHQ-PVDQLAMLR-AVPGAH-VIRPAD
BSu         THDSIAVGEDGPTHE-PVEQLASLR-AMPNLS-LIRPAD
Bha         THDSIAVGEDGPTHE-PVEQLASLR-AMPGLS-VIRPAD
Lmo4        THDSIAVGEDGPTHE-PVEQLASLR-AMPGLS-VIRPAD
Lin2        THDSIAVGEDGPTHE-PIEQLASLR-AMPGLS-VIRPAD
Sau         THDSIAVGEDGPTHE-PIEQLAGLR-AIPNMN-VIRPAD
Lmo1        THDSIAVGEDGPTHE-PIEHLASFR-AMPGLH-VIRPAD
Lin3        THDSIAVGEDGPTHE-PIEQLASLR-AMPGLT-VIRPAD
Lla         THDSIAVGEDGPTHE-PVEQLASVR-SIPNLD-VIRPAD
Spn         THDSIAVGEDGPTHE-PVEHLAGLR-AMPNLN-VFRPAD
Cac1        THDSIGVGEDGPTHE-PIEQLAALR-SIPNLT-VLRPCD
Cac3        THDSIGVGEDGPTHE-PIEQLAALR-SMPNMT-VFRPAD
Cpn1        THDSIFVGEDGPTHQ-PVEQLMSLR-AIPGLY-VIRPAD
Ctr2        THDSIFVGEDGPTHQ-PIEQIMSLR-AIPGLR-VIRPAD
Tpa1        THDSIFVGEDGPTHQ-PVETLAALR-AIPNVL-VLRPAD
Cje         THDSIGVGEDGPTHQ-PIEQLSTFR-AMPNFL-TFRPAD
Hpy         THDSIGVGEDGATHQ-PIEQLSHLR-ALPHFY-AFRPSD
Uur         SHDSYAVGGDGPTHQ-PVDQLPMLR-AIPNVE-VIRPAD
Mge         THDSYQVGGDGPTHQ-PYDQLPMLR-AIENVC-VFRPCD
Mpu         SHDSVFVGEDGPTHQ-PIEQLAMLR-SIPNVA-VFRPAD
Tel         THDSIALGEDGPTHQ-PVETLASLR-AIPNLL-VIRPAD
Nsp         THDSIGQGEDGPTHQ-PIETLASLR-AIPNLT-VIRPAD
Lme         THDSIGVGEDGPTHQ-PVELVAALR-AMPNLQ-VIRPSD
Can         THDSIGLGEDGPTHQ-PIEHLASFR-AMPNVL-MLRPAD
Stu         THDSIGLGEDGPTHQ-PIEHLASFR-AMPNIL-MFRPAD
Bme         RVRVSPHTGYGSQHS-G-DPSALFG-MFPGWR-VVSPTN
Cac2        TSNVWQQDHNG THQ-DPGLLGHIVDKKPEIVRAYLPAD
            ****   * *  * *   *     *        *  *
```

Alignment 3.18: The TK motif, as it appears in the 54 TK alignment. Those residues with 100% conserved sequence identity are highlighted in black, while those positions with 100 % sequence similarity are shown in grey. Where residues of TKs Bme and Cac2 differ from the remaining 52 TK sequences, they are highlighted in red (differing from a position conserved by identity) and blue (differing from a position conserved by similarity). Numbering refers to the residue number for *Eco*TK. Comparison of the while 30 were found in the active-site (defined as residues within 10 Å of the TPP cofactor).The selection of residues constituting the active-site is discussed in Chapter 4 (Section 4.3.2). As can be seen in Figure 3.1, the highly conserved residues are found throughout the TK molecule. Even in the TKC, which does not contribute to the active-site, is highly conserved as discussed in Chapter 6 (Section 6.3.1).

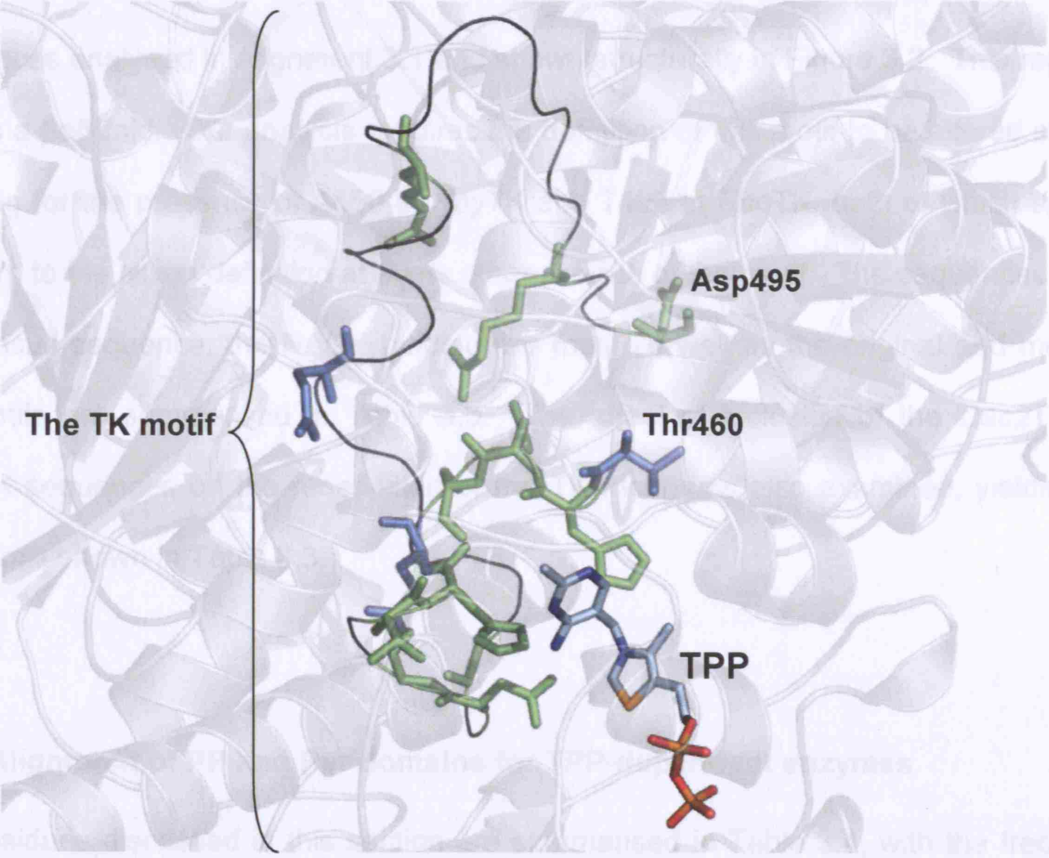


Figure 3.2: The structure of the TK motif. The structure of *EcoTK* (1QGD.pdb) was used. The 460 – 495 regions is highlighted as a ribbon in black. Positions of the motif found to have 100 % sequence similarity (using a Blossum62 matrix) are shown as blue sticks, while those with 100% sequence identity are highlighted as green sticks. The TPP molecule is shown with a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*. This information refers to Alignment 3.1 (simplified in Alignment 3.18) and excludes *Cac2*TK and *Bme*TK.

Table 3.2 summarises the location of the 93 conserved residues. At 17 positions, residues could be categorised as being in the active-site, at the interface and at the surface. 32 positions were found to be neither at the interface, in the active-site nor at the surface.

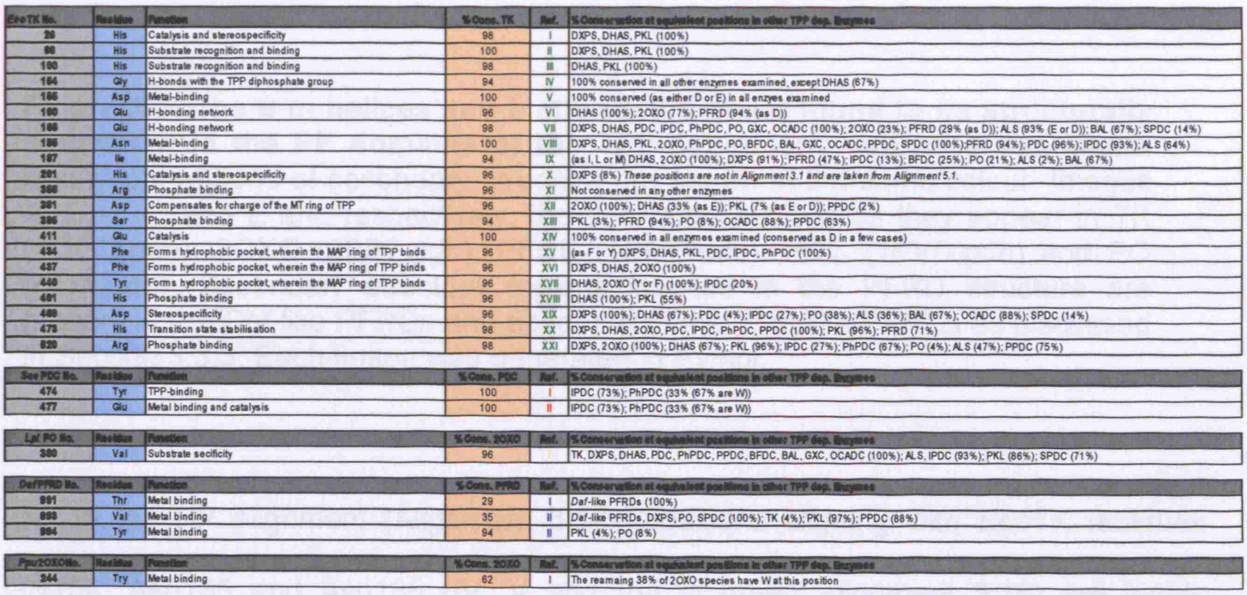
3.3.2.3 Presence of the TK motif

The presence of previously described "TK motif" [16] (residues 460 - 495) found in the Pyr domain is confirmed. The motif is shown in Alignment 3.18 as it appears in the 54 sequences analysed in Alignment 3.1 and shown structurally in Figure 3.2. This region is adopts a βαβ fold. Our analysis requires the definition of this motif to be altered slightly, allowing for the presence of A450 in *Hpy*TK and T493 in *Rso*TK, both of which species conform to the Nixon definition at every other residue of the motif. The sequences of our consensus sequence, the NADH binding like motif as well as the original and modified TK motifs are summarised in Table 3.3. The affect of inclusion of the *Cac2*TK and *Bme*TK sequences on the redefinition of the TK motif was also examined, yielding the definitions shown in Table 3.3.

3.3.3 Alignment of PP and Pyr domains for TPP-dependent enzymes

The residues discussed in this section are summarised in Table 3.4, with the frequency of their occurrence summarised. The overall alignment for the PP domains of 371 TPP-dependent enzyme is found in Alignment 3.19 on the CD-ROM in the FASTA format. From studying the PP domains of the 5 crystal structures described in Section 3.2.1.1c, six α-helices and a five stranded β-sheet were found to be common to all of the TPP-dependent enzyme types examined (Figure 3.3). Alignment 3.19 shows how the sequences corresponding with the crystal structure examined align in the PP alignment (Alignment 3.18). Regions of secondary structure are highlighted in the sequences of Alignment 3.19, using the same colour scheme used in the structural view of Figure 3.3.

Figure 3.4 illustrates how identical aligning residues from *Eco*TK and *Sce*PDC PP domains correspond in structural space. It was observed in the alignment of seventeen PFRD sequences (Section 3.2.1) and in the structural comparisons of Section 3.2.1.1c that domain VI (PP domain) of the *Tma*-like PFRDs lack two α-helices and one β-strand found in the *Daf*-like PFRDs and present in all other TPP-dependent enzyme structures examined (labelled in Figure 3.3 as Helices I and II and Sheet I). Since the phylogenetic study examines only regions found to be structurally equivalent in all enzymes, this presented a problem. Removal of the regions from the analysis could have resulted in a loss of ~10 % of the overall informative regions in the analysis. To avoid this, the *Tma*-like PFRDs were removed from the phylogenetic analysis of the TPP-dependent enzyme family.



EcoTK No.	Residue	Function	% Cons. TK	Ref.	% Conservation at equivalent positions in other TPP dep. Enzymes
26	His	Catalysis and stereospecificity	98	I	DXPS, DHAS, PKL (100%)
66	His	Substrate recognition and binding	100	II	DXPS, DHAS, PKL (100%)
100	His	Substrate recognition and binding	98	III	DHAS, PKL (100%)
154	Gly	H-bonds with the TPP diphosphate group	94	IV	100% conserved in all other enzymes examined, except DHAS (67%)
155	Asp	Metal-binding	100	V	100% conserved (as either D or E) in all enzyes examined
160	Glu	H-bonding network	96	VI	DHAS (100%); 2OXO (77%); PFRD (94% (as D))
165	Glu	H-bonding network	98	VII	DXPS, DHAS, PDC, IPDC, PhPDC, PO, GXC, OCADC (100%); 2OXO (23%); PFRD (29% (as D)); ALS (93% (E or D)); BAL (67%); SPDC (14%)
185	Asn	Metal-binding	100	VIII	DXPS, DHAS, PKL, 2OXO, PhPDC, PO, BFDC, BAL, GXC, OCADC, PPDC, SPDC (100%);PFRD (94%); PDC (96%); IPDC (93%); ALS (64%)
187	Ile	Metal-binding	94	IX	(as I, L or M) DHAS, 2OXO (100%); DXPS (91%); PFRD (47%); IPDC (13%); BFDC (25%); PO (21%); ALS (2%); BAL (67%)
261	His	Catalysis and stereospecificity	96	X	DXPS (8%) *These positions are not in Alignment 3.1 and are taken from Alignment 5.1.*
358	Arg	Phosphate binding	96	XI	Not conserved in any other enzymes
381	Asp	Compensates for charge of the MT ring of TPP	96	XII	2OXO (100%); DHAS (33% (as E)); PKL (7% (as E or D)); PPDC (2%)
385	Ser	Phosphate binding	94	XIII	PKL (3%); PFRD (94%); PO (8%); OCADC (88%); PPDC (63%)
411	Glu	Catalysis	100	XIV	100% conserved in all enzymes examined (conserved as D in a few cases)
434	Phe	Forms hydrophobic pocket, wherein the MAP ring of TPP binds	96	XV	(as F or Y) DXPS, DHAS, PKL, PDC, IPDC, PhPDC (100%)
437	Phe	Forms hydrophobic pocket, wherein the MAP ring of TPP binds	96	XVI	DXPS, DHAS, 2OXO (100%)
440	Tyr	Forms hydrophobic pocket, wherein the MAP ring of TPP binds	96	XVII	DHAS, 2OXO (Y or F) (100%); IPDC (20%)
461	His	Phosphate binding	96	XVIII	DHAS (100%); PKL (55%)
469	Asp	Stereospecificity	96	XIX	DXPS (100%); DHAS (67%); PDC (4%); IPDC (27%); PO (38%); ALS (36%); BAL (67%); OCADC (88%); SPDC (14%)
473	His	Transition state stabilisation	98	XX	DXPS, DHAS, 2OXO, PDC, IPDC, PhPDC, PPDC (100%); PKL (96%); PFRD (71%)
520	Arg	Phosphate binding	98	XXI	DXPS, 2OXO (100%); DHAS (67%); PKL (96%); IPDC (27%); PhPDC (67%); PO (4%); ALS (47%); PPDC (75%)

Sce PDC No.	Residue	Function	% Cons. PDC	Ref.	% Conservation at equivalent positions in other TPP dep. Enzymes
474	Tyr	TPP-binding	100	I	IPDC (73%); PhPDC (33% (67% are W))
477	Glu	Metal binding and catalysis	100	II	IPDC (73%); PhPDC (33% (67% are W))

Lpl PO No.	Residue	Function	% Cons. 2OXO	Ref.	% Conservation at equivalent positions in other TPP dep. Enzymes
380	Val	Substrate secificity	96	I	TK, DXPS, DHAS, PDC, PhPDC, PPDC, BFDC, BAL, GXC, OCADC (100%); ALS, IPDC (93%); PKL (86%); SPDC (71%)

DafPFRD No.	Residue	Function	% Cons. PFRD	Ref.	% Conservation at equivalent positions in other TPP dep. Enzymes
991	Thr	Metal binding	29	I	*Daf*-like PFRDs (100%)
993	Val	Metal binding	35	II	*Daf*-like PFRDs, DXPS, PO, SPDC (100%); TK (4%); PKL (97%); PPDC (88%)
994	Tyr	Metal binding	94	II	PKL (4%); PO (8%)

Ppu2OXONo.	Residue	Function	% Cons. 2OXO	Ref.	% Conservation at equivalent positions in other TPP dep. Enzymes
244	Try	Metal binding	62	I	The reamaing 38% of 2OXO species have W at this position

Table 3.4: A summary of the important residues in TK and other TPP-dependent enzymes discussed throughout thesis. The frequency of occurrence of residues at equivalent positions in other enzymes are tabulated from alignments described in Chapters 3 and 5. The reference column refers to how these positions are labelled in the summary alignments of Chapter 3.

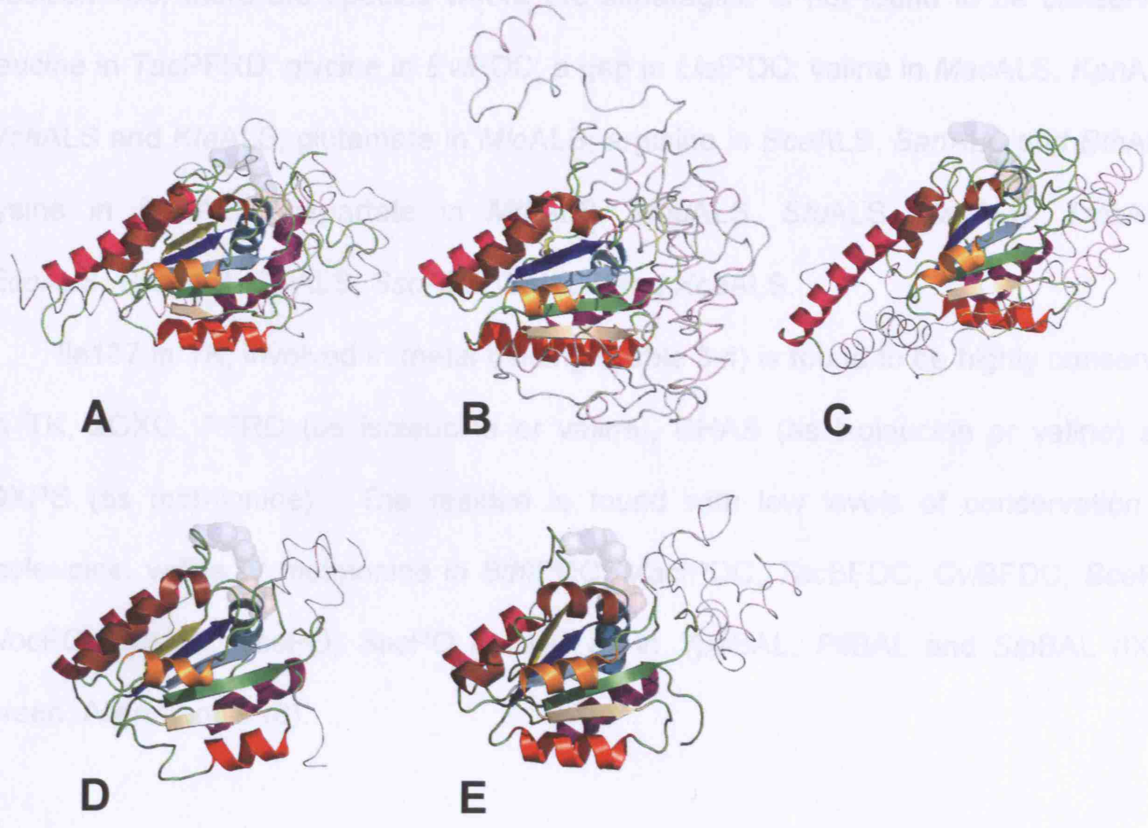
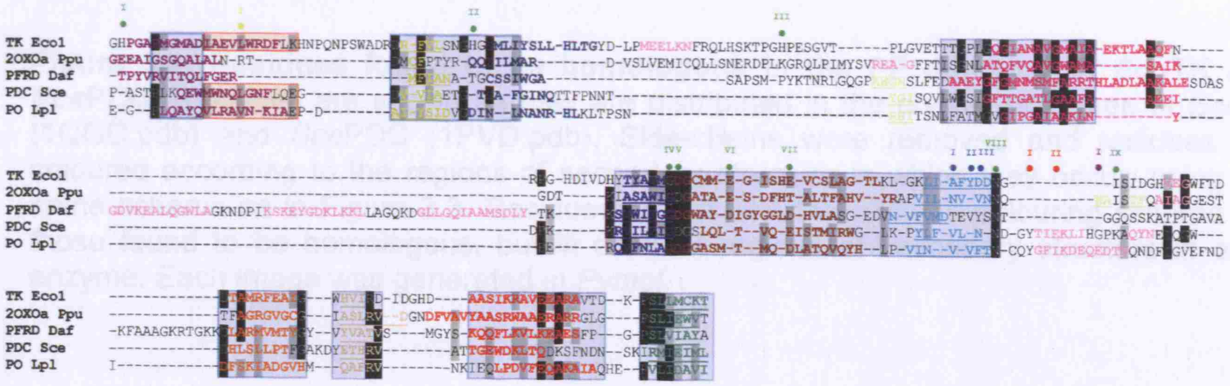


Figure 3.3: The 6 α- helices and 5 stranded β-sheet found to be structurally equivalent in the PP domains of the TPP-dependent enzymes examined. Equivalent regions of secondary structure are shown in the same colour. Images A, B, C, D and E are, respectively, the equivalent regions as they appear in TK (1QGD.pdb), 2OXO (2BP7.pdb), PFRD (1BOP.pdb), PDC (1PVD.pdb) and PO (POX.pdb). Secondary structure regions labelled in the PFRD structure are present in the *Daf*-like PFRDs and absent from the *Tma*-like PFRDs as discussed in Section 3.3.3. Each image was generated in *Pymol*.

In the PP alignment (Alignment 3.18), positions aligning with *Eco*TK residues Gly154, Asp155 and Asn185 (IV, V and VIII in green Alignment 3.18) are highly conserved in all seventeen enzymes, reflecting their critical roles in TPP-binding (Section 1.1.3). Some TK sequences such as *Mle*TK, *Mtu*TK, *Sau*TK as well as *Mys*DHAS have Ser at position 154, the reason for which is unclear. Less unusual is the observation of glutamate instead of the similar aspartate at position 155 in *Tpa1*TK, *Bja*2OXO, *Bce*2OXO, *Bsu*2OXO, *Sau*2OXO, *Bov*2OXO, *Mac*IPDC and *Psy*ALS. For position 185, there are species where the asparagine is not found to be conserved: leucine in *Tac*PFRD; glycine in *Evi*PDC; a gap in *Lla*IPDC; valine in *Mac*ALS, *Kpn*ALS, *Vch*ALS and *Kte*ALS; glutamate in *Mlo*ALS; arginine in *Bce*ALS, *Ban*ALS and *Bth*ALS; lysine in *Mca*ALS; aspartate in *Mth*ALS, *Mbu*ALS, *Stu*ALS, *Xax*ALS, *Tma*ALS, *Eco*ALS, *Sen*ALS, *Sty*ALS, *Sso*ALS, *Spl*ALS and *Xca*ALS.

Ile187 in TK, involved in metal binding (Table 3.4) is found to be highly conserved in TK, 2OXO, PFRD (as isoleucine or valine), DHAS (as isoleucine or valine) and DXPS (as methionine). The residue is found with low levels of conservation as isoleucine, valine or methionine in *Bth*IPDC, *Mac*IPDC, *Tac*BFDC, *Cvi*BFDC, *Bce*PO, *Noc*PO, *Bps*PO, *Bce*PO, *Sac*PO as well as in *Ype*BAL, *Pfl*BAL and *Sip*BAL (IX in green, Alignment 3.18).



Alignment 3.20: Alignment of the PP domain sequences from the crystal structures examined. Regions of α-helix are in bold, while regions of β-sheet are in bold and underlined. Regions of secondary structure are coloured as in Figure 3.3. Arrows indicate where regions of sequence found in the overall PP alignment (Alignment 3.19) have been removed to minimise this alignment. Grey boxed regions are those which are used in the phylogenetic analysis. The red box indicates a region excluded from the phylogenetic analysis which shows high homology between TK, PDC and PO. This region is included in the analysis described in Chapter 5 (Section 5.2.1). Functionally important residues, summarised in Table 3.4, are indicated with roman numerals.

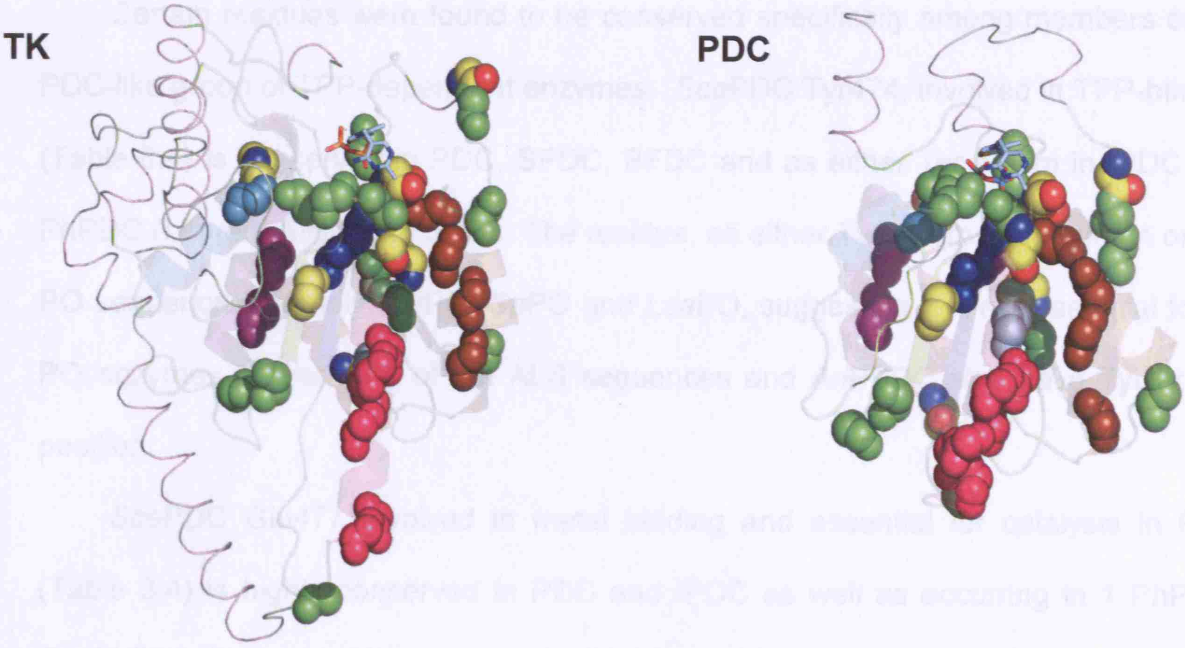


Figure 3.4: Residues found to be homologous in the PP domains of *Eco*TK and *Sce*PDC. Residues are shown as they are distributed in the crystal structures of *Eco*TK (1QGD.pdb) and *Sce*PDC (1PVD.pdb). Side-chains were removed and residues are coloured according to the regions of secondary structure in which they occur, using the same scheme as in Figure 3.3. Residues shown with the CHNOS colouring scheme are those found to be homologous, but in different regions of secondary structure in each enzyme. Each image was generated in *Pymol*.

His26 and His66, involved in catalysis and substrate specificity and recognition respectively in TK (Section 1.1.4) are found to be highly conserved in all TK-like enzymes (I and II in green, Alignment 3.18), while His100 is conserved in the TK-like enzymes TK, DHAS and PKL, but not in DXPS (III in green, Alignment 3.18). This position is discussed further in Chapter 5 (Section 5.3.2).

Glu160, which is part of a H-bonding network in TK is highly conserved in TK, DHAS, 2OXO and PFRD (as Asp) (VII in green, Alignment 3.18). Glu165, also part of the H-bonding network is highly conserved in TK, DXPS, DHAS, PDC, IPDC, PhPDC, PO, GXC, OCADC and BAL, while it is found with low frequency in 2OXO, PFRD and SPDC (VII in green, Alignment 3.18).

Certain residues were found to be conserved specifically among members of the PDC-like group of TPP-dependent enzymes. *Sce*PDC Tyr474, involved in TPP-binding (Table 3.4) is conserved in PDC, SPDC, BFDC and as either Tyr or Trp in IPDC and PhPDC (I in red, Alignment 3.18). The residue, as either Tyr or Trp is present in only 3 PO sequences, namely *Sor*PO, *Sp*PO and *Lsa*PO, suggesting it is not essential for all PO enzymes. Over 50% of the ALS sequences and *Ani*PPDC also have Tyr at this position.

*Sce*PDC Glu477 involved in metal binding and essential for catalysis in PDC (Table 3.4) is highly conserved in PDC and IPDC as well as occurring in 1 PhPDC, *Sce*PhPDC (II in red, Alignment 3.18). The PO residue Val380 has been shown to be involved in altering substrate specificity towards pyruvate and α-ketobutyrate [186]. Either Val, Ile or Met is found at the corresponding position in all enzymes examined, except PFRD and 2OXO (I in yellow, Alignment 3.18). *Daf*PFRD residues thr991, val993 and Tyr994 are known to be involved in metal-binding in PFRD (Table 1.3) (I, II and III in blue, Alignment 3.18). Thr991 is conserved in all *Daf*-like PFRDs as Thr, but as Met in the *Tma*-like PFRDs. Val993 is highly conserved in *Daf*-like PFRD, PKL (as Leu), DXPS (as Leu) as well as in SPDC, PPDC, PO and BFDC. Tyr994 is conserved in all PFRDs, as well as occurring in nearly 60% of PO sequences as either Phe or Tyr.

Positions corresponding to *Ppu*2OXO residue Trp244, involved in metal-binding (Table 3.4) are found in all 2OXO as either Trp or Tyr, but occurs in no other enzymes.

The overall alignment for the Pyr domains of 383 TPP-dependent enzymes is found in Alignment 3.20 on the CD-ROM. For the Pyr domain, five α-helices and four stranded β-sheet are found to be common to all of the TPP-dependent enzyme types. Figure 3.5 shows these regions of secondary structure colour-coded in the crystal structures examined. Alignment 3.21 shows how the sequences corresponding with the five enzyme crystal structures align, with regions of secondary structure highlighted and coloured under the same scheme as in Figure 3.5. To illustrate the correlation between sequence homology and structural equivalence the residues found to be identically conserved in the Pyr domains of *Eco*TK and *Sce*PDC are shown in the respective crystal structures in Figure 3.6.

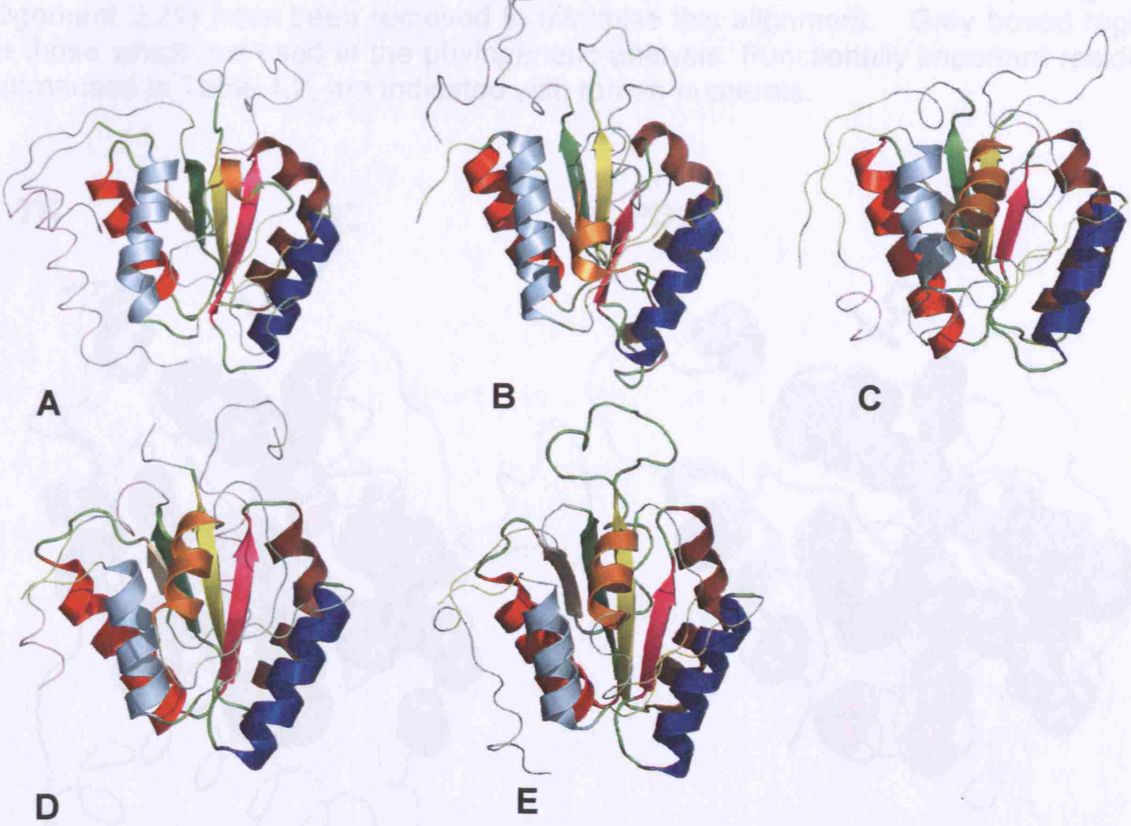
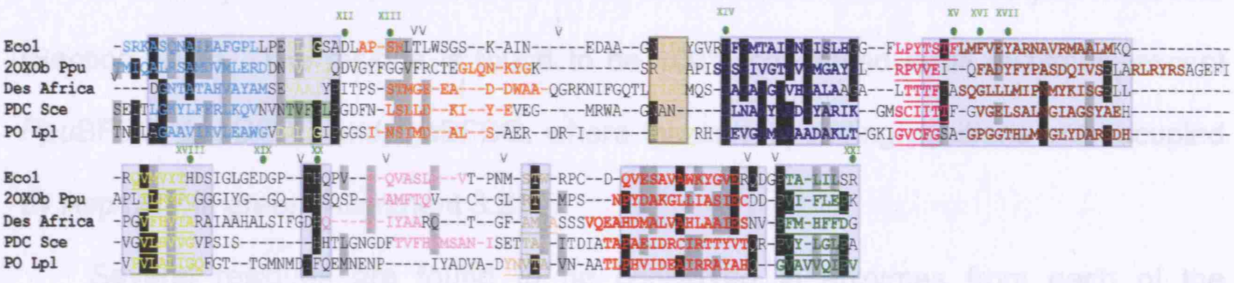


Figure 3.5: The 5 α- helices and 4 stranded β-sheet found to be structurally equivalent in the Pyr domains the TPP-dependent enzymes examined. Equivalent regions of secondary structure are shown in the same colour. Images A, B, C, D and E are, respectively, the equivalent regions as they appear in TK (1QGD.pdb), 2OXO (2BP7.pdb), PFRD (1BOP.pdb), PDC (1PVD.pdb) and PO (POX.pdb). Each image was generated in *Pymol*.



Alignment 3.22: Alignment of the Pyr domain sequences from the crystal structures examined. Regions of α-helix are in bold, while regions of β-sheet are in bold and underlined. Regions of secondary structure are coloured as in Figure 3.5. Arrowheads indicate where regions of sequence found in the overall Pyr alignment (Alignment 3.21) have been removed to minimise this alignment. Grey boxed regions are those which are used in the phylogenetic analysis. Functionally important residues, summarised in Table 1.2, are indicated with roman numerals.

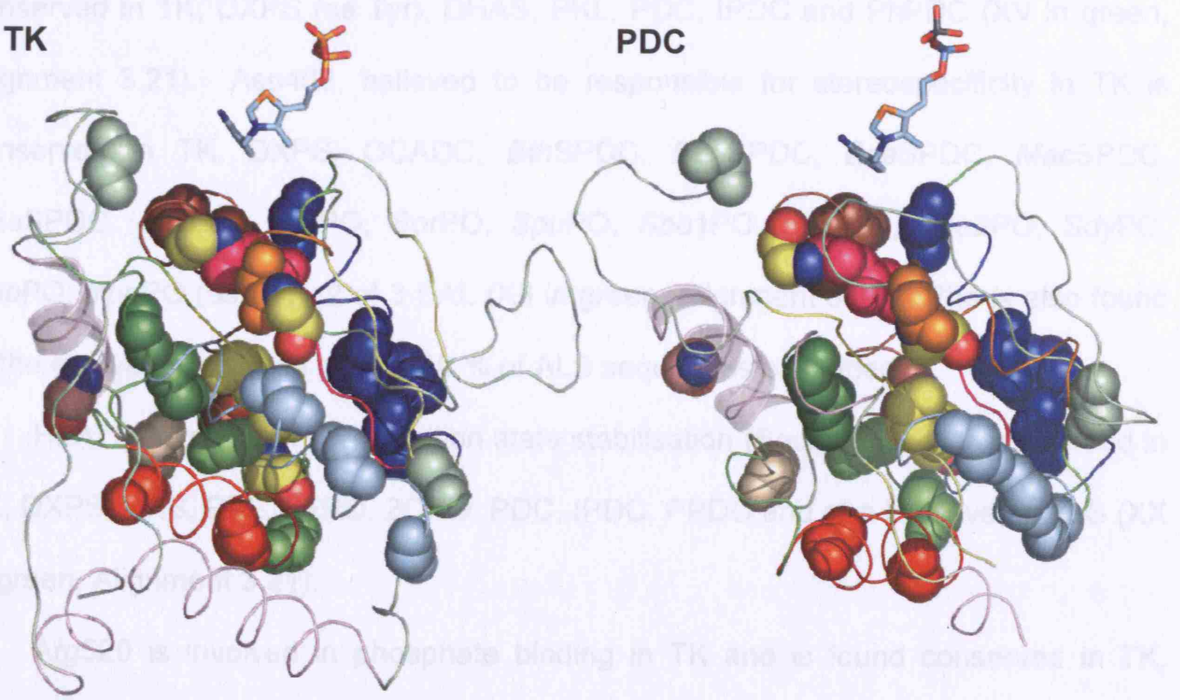


Figure 3.6: Residues found to be homologous in the Pyr domains of *Eco*TK and *Sce*PDC. Residues are shown as they are distributed in the crystal structures of *Eco*TK (1QGD.pdb) and *Sce*PDC (1PVD.pdb). Side-chains were removed and residues are coloured according to the regions of secondary structure in which they occur, using the same scheme as in Figure 3.5. Residues coloured according to the CHNOS scheme are those found to be homologous, but in different regions of secondary structure in each enzyme. The pink helix is present in both TK and PDC, but excluded form the phylogenetic analysis, due to low homology at the sequence level. The conserved arginine residue found in the helix in both enzymes is coloured according to the CHNOS scheme. Images were generated in *Pymol*.

In the Pyr domain, residue Glu411 is essential for TPP catalysis to occur (Section 1.1.4). The residue is found to be 100% conserved in all enzymes, except *Ppu*BFDC, *Lpn*BFDC and *Bja*BFDC, where the corresponding positions are occupied by Asp (XIV in green, Alignment 3.21).

Several residues are found to be conserved in enzymes from each of the TPP-dependent enzyme groups. Ser385, believed to be involved in phosphate binding in TK is conserved in TK, the *Daf*-like PFRDs (thr in the *Tma*-like PFRDs), PPDC (as Thr) and OCADC (as Thr), (XIII in green, Alignment 3.21). The residue is not conserved in DXPS, which is discussed further in Chapter 5 (Section 5.3.2). Phe434, which forms part of the hydrophobic pocket wherein the MT ring of TPP binds is conserved in TK, DXPS (as Tyr), DHAS, PKL, PDC, IPDC and PhPDC (XV in green, Alignment 3.21). Asp469, believed to be responsible for stereospecificity in TK is conserved in TK, DXPS, OCADC, *Bth*SPDC, *Bau*SPDC, *Bce*SPDC, *Mac*SPDC, *Mka*SPDC (as Glu), *Bja*PO, *Sor*PO, *Spu*PO, *Sbo1*PO, *Bma*PO, *Sbo2*PO, *Sdy*PO, *Rpa*PO, *Ype*PO (as Glu), 2 of 3 BAL (XII in green, Alignment 3.21). Phe is also found at the equivalent position in over 20 % of ALS sequences examined.

His473, involved in TK transition state stabilisation (Section 1.1.4) is conserved in TK, DXPS, DAS, PKL, PFRD, 2OXO, PDC, IPDC, PPDC and at a low level in ALS (XX in green, Alignment 3.21).

Arg520 is involved in phosphate binding in TK and is found conserved in TK, DXPS, DHAS, PKL, 2OXO, PPDC, PO, and also in IPDC and ALS at a low level of conservation (XXI in green, Alignment 3.21).

Phe437, which along with Phe434 and Tyr440 in TK forms the hydrophobic MT ring-binding pocket (Section 1.1.3) is conserved in TK, DXPS, DHAS and 2OXO (XVI in green, Alignment 3.21). Tyr440 is conserved in TK, DHAS, 2OXO (as Tyr or Phe) and in some of the *Tma*-like PFRDs (XVII in green, Alignment 3.21).

Asp381 is involved in TPP binding (Section 1.1.3) is conserved in TK and 2OXO (XII in green, Alignment 3.21). This residue is discussed further in Chapter 5 (Section 5.3.3). His461, which has a role in phosphate binding (Table 3.4) is also found to be conserved in TK and 2OXO only (XVIII in green, Alignment 3.21).

3.3.4 Phylogenetic analyses of individual TPP-dependent enzymes

The individual phylogenies are not shown for BAL, PhPDC and DHAS since they contained only three sequences each and their phylogenies would be uninformative. However, the positions of each of these enzymes in the overall phylogenetic trees for the PP and Pyr domains, are still of interest. Individual trees are discussed at the level of taxa. Where grouping by taxon is not clearly observed, then the subdivisions are discussed.

3.3.4.1 Phylogenetic analysis of transketolase

Phylogenetic analyses were performed for 54 TK sequences as described in Appendix 1 (Section A1.2). This resulted in 29 phylogenetic trees. 26 trees were rejected, as described in Section A1.2, as they did not show good clustering by species, leaving 3 candidate trees for consideration. The candidate trees are shown in Figures 3.7, 3.8 and 4.10 (Chapter 4).

The first and second candidates are NJ trees, one generated using the categories model in *Protdist* (Figure 3.7), the other using the Dayhoff PAM matrix with the gamma parameter (Figure 3.8). Resulting distance matrices were input into the *Neighbour* program. The third candidate tree was the ML tree generated with *ProML* (Section 3.2.2.2a).

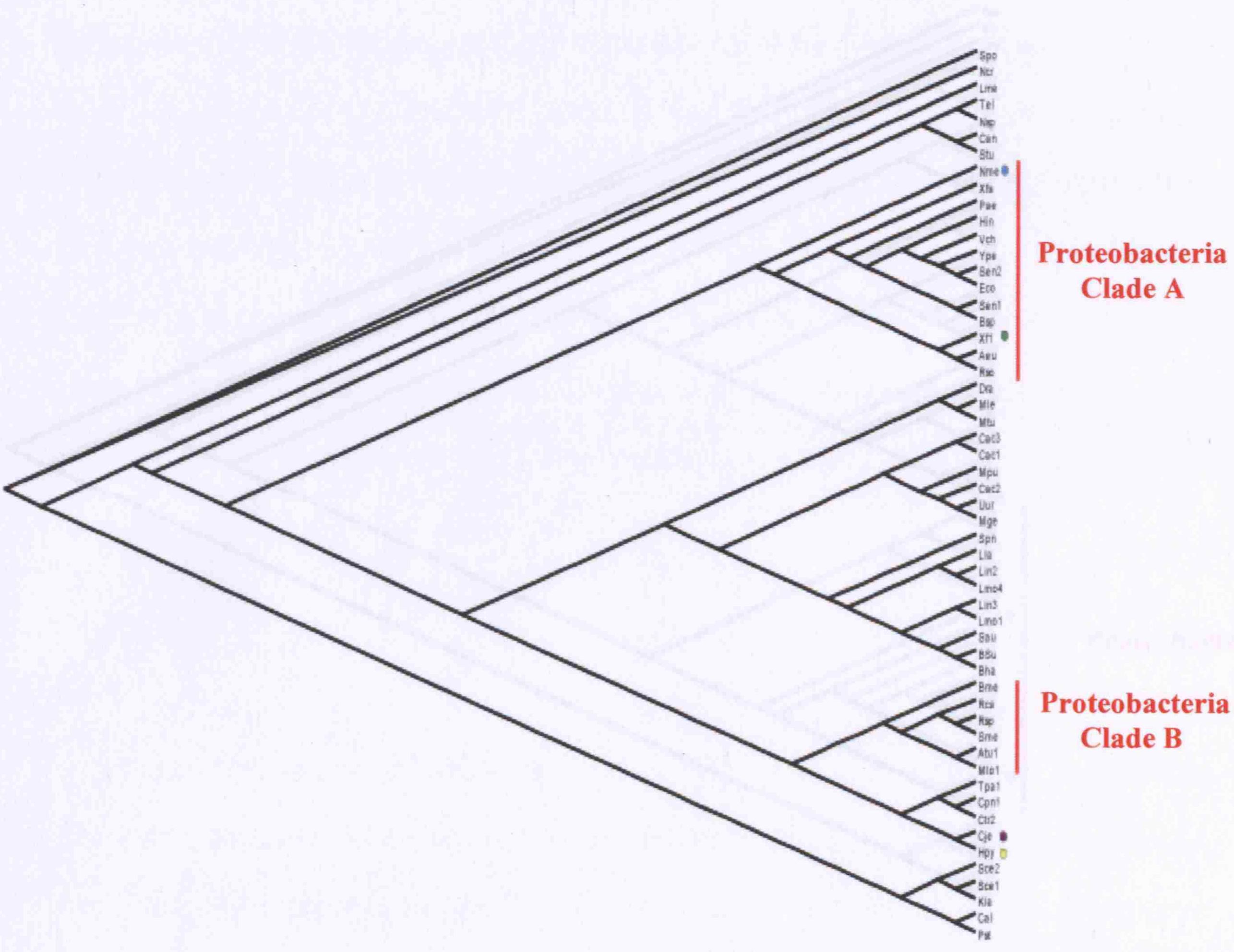


Figure 3.7: NJ tree constructed for transketolase using the Categories model. Constructed using the *Neighbour* program from the Phylip suite. Viewed in *Treeview.* Outliers, discussed in the text, *Nme*TK, *Xfl*TK, *Hpy*TK and *Cje*TK are marked with circles coloured blue, green, yellow and purple respectively. The grouping of proteobacteria sequences into two separate clades, A and B, is shown in red.

A choice was made to use the ML tree as it was thought that for the purpose of this study, where ancestral reconstruction was the ultimate aim, the examination of each character site individually (see Section 1.7.1), was most suitable. Thus the ML tree was subject to ancestral reconstruction. The ML tree is shown in Figure 4.10 in Chapter 4.

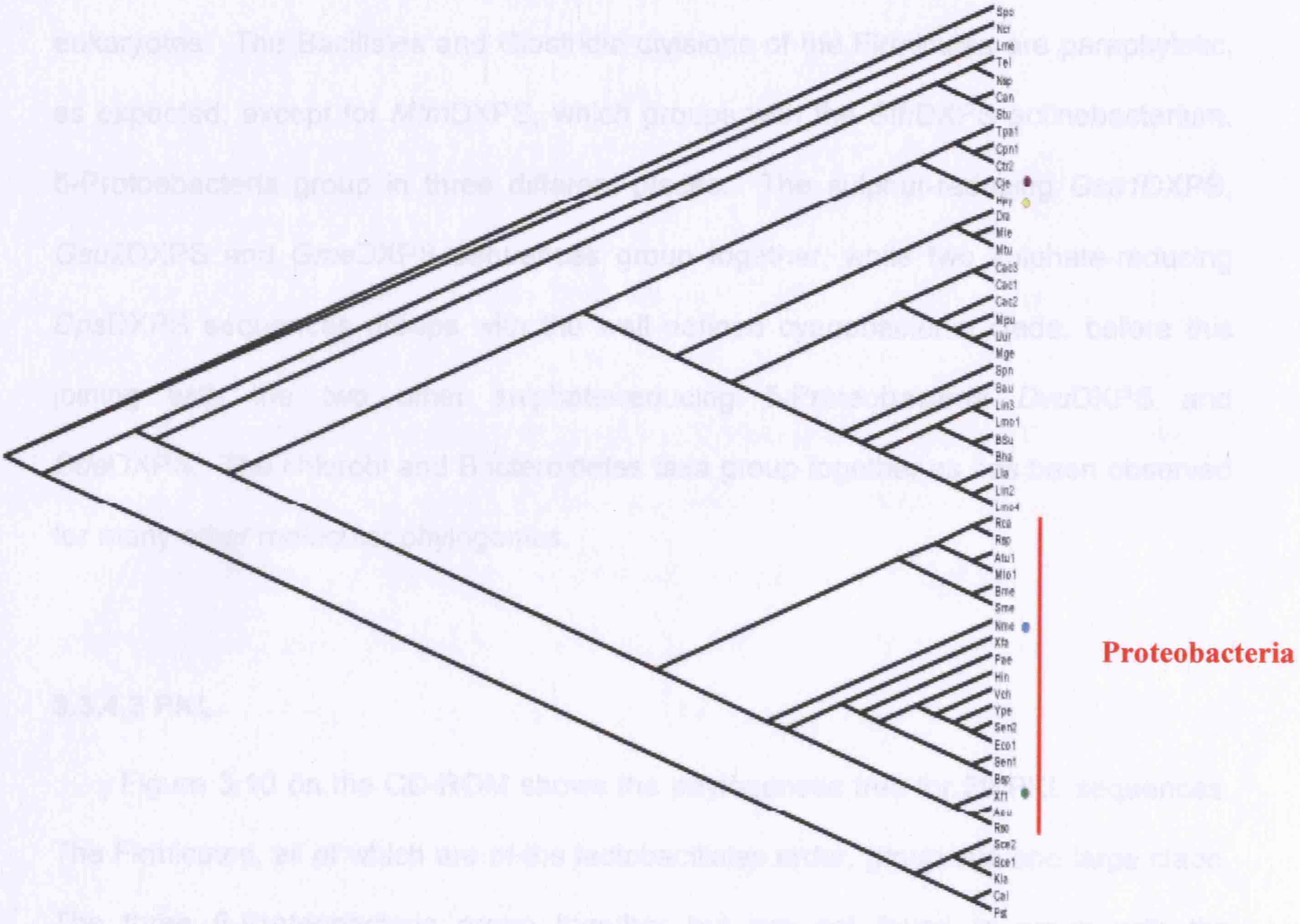


Figure 3.8: NJ tree constructed for transketolase using the Dayhoff PAM model with the gamma parameter. Constructed using the *Neighbour* program from the *Phylip* suite. Viewed in *Treeview*. Outliers, *Nme*TK, *Xfl*TK, *Hpy*TK and *Cje*TK are marked with circles coloured blue, green, yellow and purple respectively. The grouping of proteobacteria sequences into one clade, is denoted in red.

3.3.4.2 DXPS

The phylogenetic tree for DXPS is shown in Figure 3.9 on the CD-ROM. In general, species group well according to taxa. Most of the γ-proteobacteria group together, while several, of the order Xanthomonadales group with the β-proteobacteria. As expected the γ and β subdivisions are found to be paraphyletic. All of the α-Proteobacteria form a distinct clade, which in turn groups with the viridiplantae clade, reflecting the endosymbiotic relationship between the α-protoebacterial subdivision and eukaryotes. The Bacillales and Clostridia divisions of the Firmicutes are paraphyletic, as expected, except for *Mtm*DXPS, which groups with the *Sth*DXPS actinobacterium. δ-Protoebacteria group in three different places. The sulphur-reducing *Gsu1*DXPS, *Gsu2*DXPS and *Gme*DXPS sequences group together, while two sulphate-reducing *Dps*DXPS sequences groups with the well defined cyanobacterial clade, before this joining with the two other sulphate-reducing δ-Proteobacteria *Dvu*DXPS and *Dde*DXPS. The chlorobi and Bacteroidetes taxa group together as has been observed for many other molecular phylogenies.

3.3.4.3 PKL

Figure 3.10 on the CD-ROM shows the phylogenetic tree for 29 PKL sequences. The Firmicutes, all of which are of the lactobacillales order, group into one large clade. The three β-Proteobacteria group together but are not found to group with the γ-proteoacteria as expected. Neither do all of the α-protoebacteria all group together. The α- and γ-proteobacterial sequences group with the cyanobacteria (which themselves don't all group) in a manner that isn't clearly defined.

3.3.4.4 2OXO

The tree for the concatenated α and β subunits of the 2OXO enzymes is shown in Figure 3.11 on the CD-ROM. The three firmicute species group together well, while the γ- and β-proteobacteria are paraphyletic, as expected. All of the α-protoebacteria except *Bja*2OXO (which is found closest to the firmicute clade) group together. While the apparent α-proteobacterial origins of the mitochondrion may suggest that the group should join with the eukaryotic metazoan species, this is not observed. Instead the metazoan clade joins with the β- and γ- proteobacterial clade.

3.3.4.5 PFRD

The PFRD phylogeny (Figure 3.12 on the CD-ROM) contains the two types of PFRD discussed in Section 1.2.3, namely the *Daf*-like PFRDs (*Daf*PFRD, *Dvu*PFRD, *Dde*PFRD, *Mtm*PFRD and *Tet*PFRD) and the *Tma*-like PFRDs. The *Daf*-like PFRDs form their own clade, suggesting that the two different PFRD types diverged, following which the fusion of the separate domains lead to the modern *Daf*-like PFRDs. Soon after the divergence of the *Tma*-like PFRDs, the loss of the regions of secondary structure from domain VI discussed in Section 3.3.3 occurred. Within the *Daf*-like clade, the firmicutes and δ-proteobacteria form distinct clades. In the *Tma*-like clade, the firmicute *Mtm*PFRD groups within a larger clade of euryarchaeota sequences. Thermotogae *Tma*PFRD and ε-proteobacterium *Hpy*PFRD form a clade, which in turn groups with the chrenarchaeota clade.

3.3.4.6 PDC

The PDC phylogenetic tree is shown in Figure 3.13 on the CD-ROM. Fungi form a large clade, with the ascomycota and zygomycota species forming distinct clades. The viridiplantae group (all of which are streptophyta) group well. The fungal and plant clades do not group together. Rather, the plant clade groups with the proteobacteria, before the well resolved firmicute group joins and then the *Mtu*PDC actinobacterial species.

3.3.4.7 IPDC

Figure 3.14 on the CD-ROM shows the IPDC phylogeny. α- and γ-proteobacteria cluster well, but not together in the IPDC phylogeny. All firmicutes cluster together as well. The solitary archaeal sequence groups between the sole cyanobacteria and actinobacteria sequences.

3.3.4.8 PO

The phylogeny for the PO sequences is shown in Figure 3.15 on the CD-ROM. The Firmicutes, which are either Bacillales or Lactobacillales, group together into a single clade. The proteobacteria do not resolve well into subgroups, but apart from the actinobacterial *Cjk*PO sequence, they group together well. The sole archaeal sequence in the analysis joins at the point where firmicute clade joins the other bacterial clade.

3.3.4.9 ALS

The phylogeny of ALS, shown in Figure 3.16 on the CD-ROM. Of all the enzymes examined, the ALS phylogeny is the least well resolved. While some enzymes group according to species type, such as the larger firmicutes and γ-proteobacterial clades, clustering of most of the other species is poor. Most surprising is the manner in which archaeal species are interspersed with bacterial species throughout the tree, with no clear bacterial / archaeal resolution.

3.3.4.10 GXC

The GXC phylogeny is shown in Figure 3.17 on the CD-ROM. All but one of the sequences are proteobacteria. The β- and γ-Proteobacteria do not form distinct clades, but they primarily group with each other (except for the β-Protoebacteria *Pol*GXC). The α-Proteobacteral sequence *Mlo*GXC, the β-Proteobacterial sequence *Pol*GXC and the Actinobacterial sequence *Rxy*GXC sequences form a smaller clade. Overall the distribution of sequences is puzzling, although the extremely high conservation observed in the GXC alignment (Section 3.3.1) may suggest that there are insufficient informative sites in the GXC alignment to infer the phylogeny accurately.

3.3.4.11 BFDC

Thermoplasmata and γ-proteobacteria group well within the BFDC phylogeny illustrated in Figure 3.18 on the CD-ROM. Two β-proteobacteria sequences group with the Thermoplasmata sequences, while other β-proteobacteria sequence, *Bma*BFDC groups with the α-proteobacterium *Bja*BFDC.

3.3.4.12 OCADC

The OCADC phylogeny as seen in Figure 3.19 on the CD-ROM shows no obvious pattern of clustering. the α-proteobacterial *Bja*1OCADC and *Bja*2OCADC cluster. Otherwise, the firmicutes, actinobacteria and proteobacterial sequences do not cluster well. The only archaeal sequence, *Ath*OCADC is found to be most similar to the *Bja*3OCODC α-proteobacterial sequence.

3.3.4.13 SPDC

The SPDC phylogeny, found in Figure 3.20 on the CD-ROM consist of six euryarchaeotas and just one α-proteobacterium, *Rnu*SPDC, which clusters with *Mka*SPDC. The remaining five euryarchaeotal species group together well.

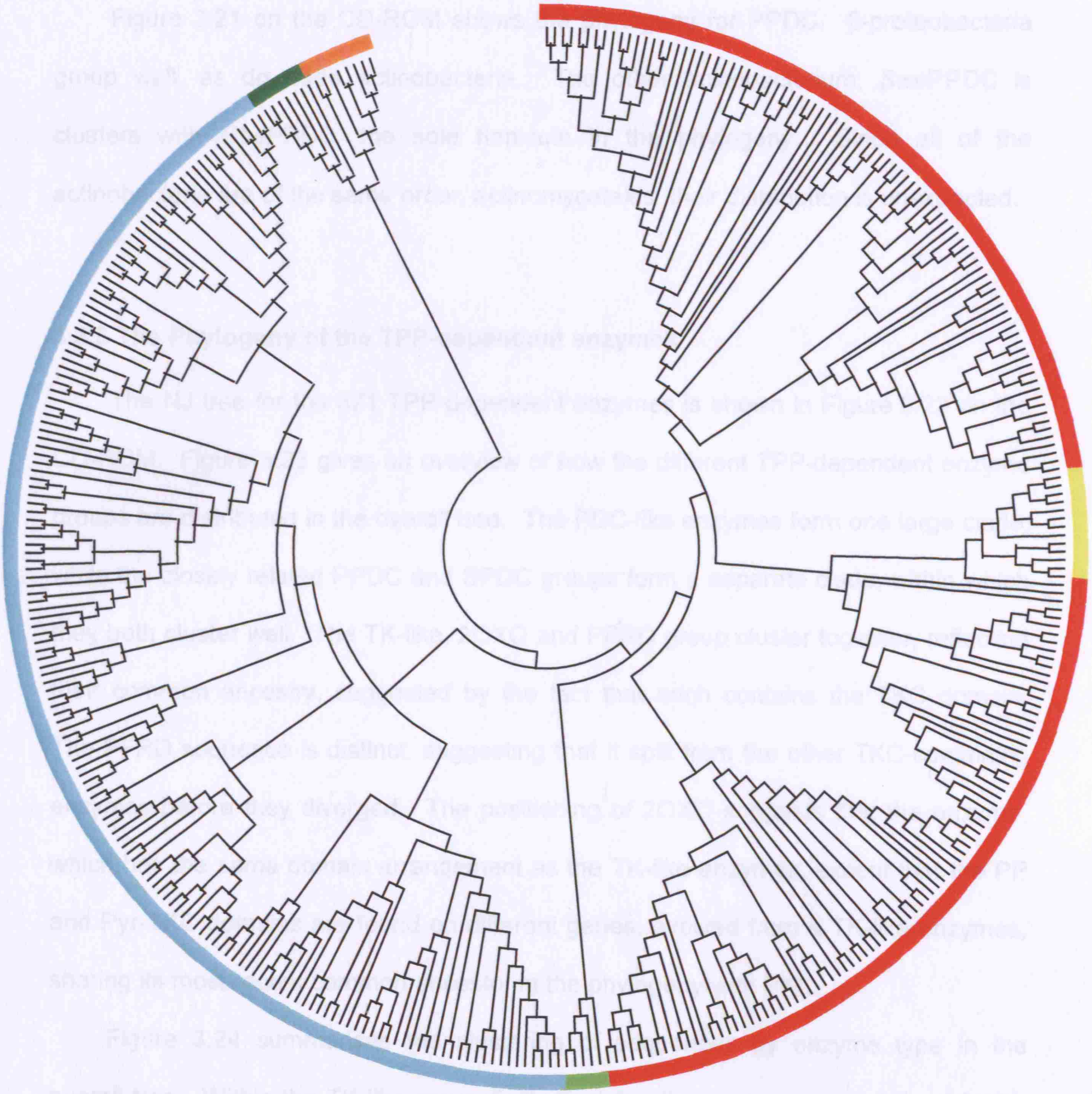


Figure 3.23: Overview of the TPP-dependent enzyme phylogeny coloured by enzyme group. Enzymes are coloured according to the group of TPP-dependent enzymes to which they belong, as defined in Section 1.2.2. TK-like enzymes are coloured red. PDC-like enzymes are in cyan, PFRD enzymes in the phylogeny are coloured light green, the 2OXO enzymes are in yellow, SPDC enzymes are dark green, while the PPDC enzymes are in orange.

3.3.4.14 PPDC

Figure 3.21 on the CD-ROM shows the phylogeny for PPDC. β-proteobacteria group well, as do three actinobacteria. The other acinobacterium, *Swe*PPDC is clusters with *Cbe*PPDC, the sole firmicute in the phylogeny. Since all of the actinobacteria are of the same order, actinomycetales, their distribution is unexpected.

3.3.5 The Phylogeny of the TPP-dependent enzymes

The NJ tree for the 371 TPP-dependent enzymes is shown in Figure 3.22 on the CD-ROM. Figure 3.23 gives an overview of how the different TPP-dependent enzyme groups are distributed in the overall tree. The PDC-like enzymes form one large clade, while the closely related PPDC and SPDC groups form a separate clade, within which they both cluster well. The TK-like, 2OXO and PFRD group cluster together, reflecting their common ancestry, suggested by the fact that each contains the TKC domain. The PFRD sequence is distinct, suggesting that it split from the other TKC-containing enzymes before they diverged. The positioning of 2OXO suggests that the enzyme, which has the same domain arrangement as the TK-like enzymes, except that the PP and Pyr-TKC domains are found on different genes, evolved from a TK-like enzymes, sharing its most recent common ancestor in the phylogeny with PKL.

Figure 3.24 summarises the clustering of sequences by enzyme type in the overall tree. Within the TK-like group (including, for discussion purposed the 2OXO), all of the DXPS sequences form a distinct clade. The DHAS sequences are found distributed amongst the main TK clade. Actinobacerial *Mys*DHAS groups with *Mle*TK and *Mtu*TK, both of which are actinobacteria. The 2 fungal *Pan*DHAS and *Cab*DHAS do not however group with the TK fungal sequences. The placement of the DHAS sequences remains difficult to explain. While most of the TK sequences group together (with the DHAS interspersed), there are two outlier, *Bme*TK and *Cac2*TK.

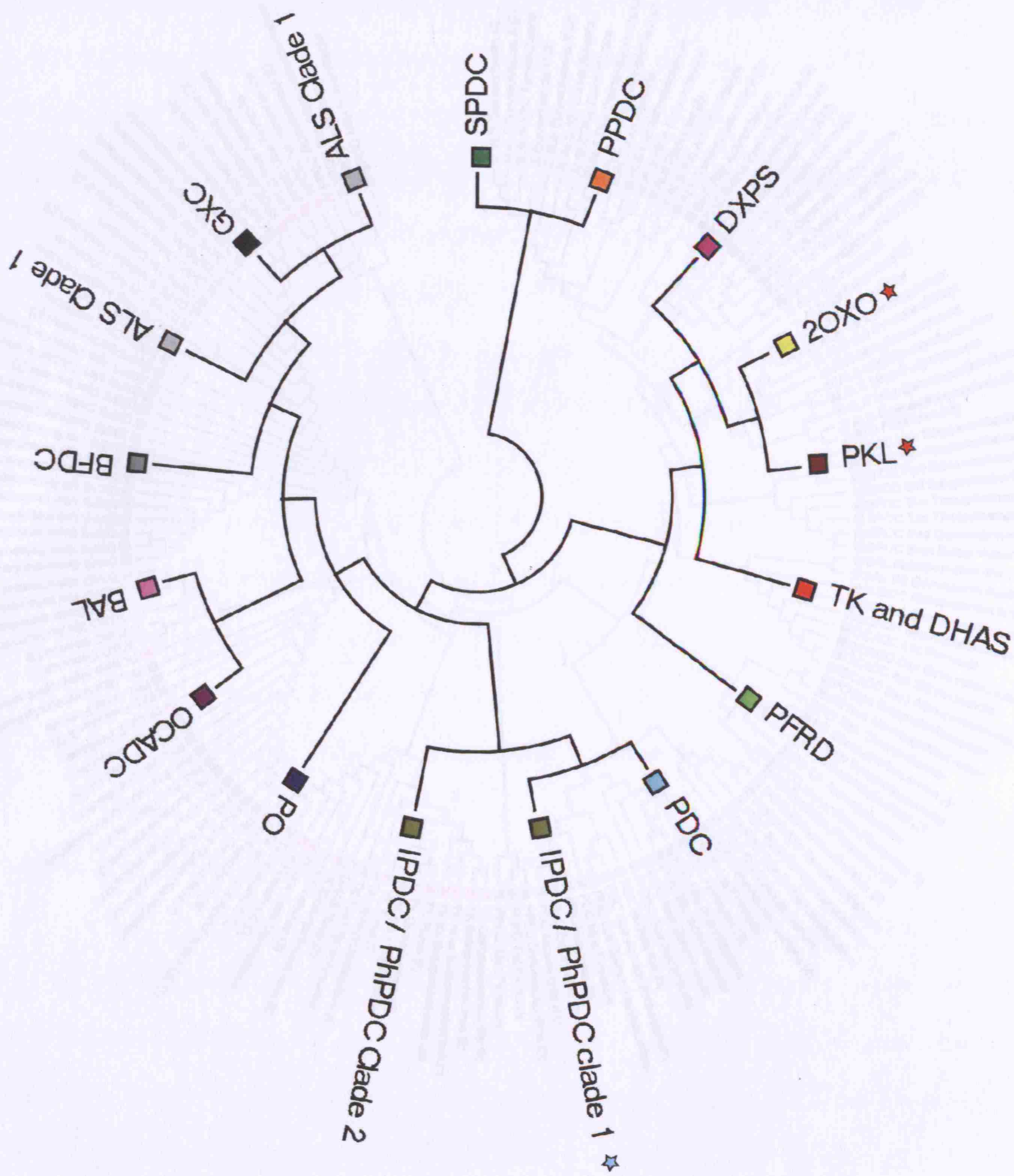


Figure 3.24: Simplified version of the NJ tree for the TPP-dependent enzymes. Individual enzymes are coloured as in the overall tree (Figure 3.22). Stars indicate where outliers are found and coloured according to the outliers' enzyme type.

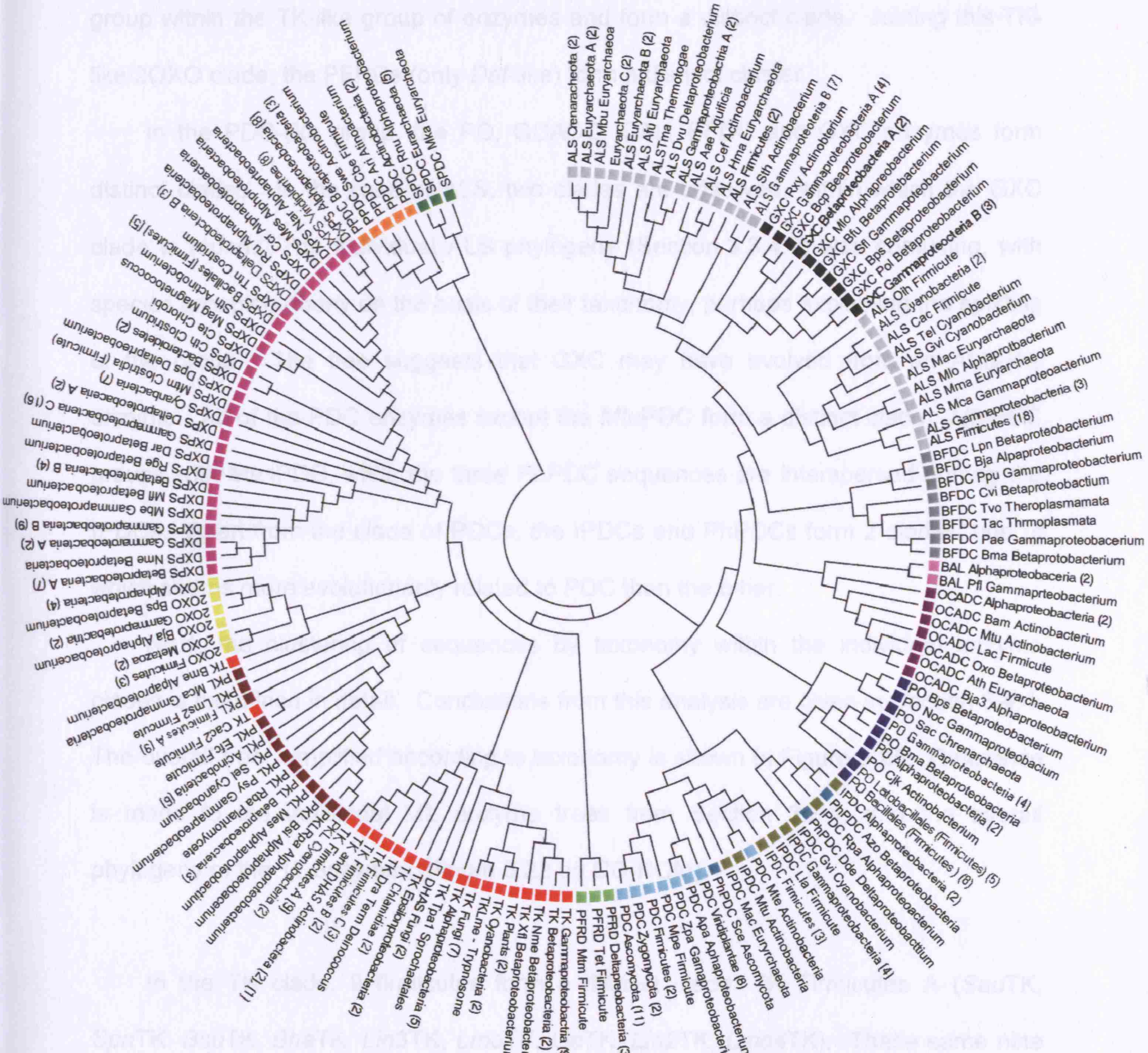

ALS Crenaracheota (2)
ALS Euryarchaeota A (2)
ALS Mbu Euryarchaeoa
Euryarchaeota C (2)
ALS Euryarchaeota B (2)
ALS Afu Euryarchaeota
ALSTma Thermotogae
ALS Dvu Deltaproteobacterium
ALS Gammaproteobacteria A (2)
ALS Aae Aquificia
ALS Cef Actinobacterium
ALS Hma Euryarchaeota
ALS Firmicutes (2)
ALS Sth Actinobacterium
ALS Gammaproteobacteria B (7)
GXC Rxy Actinobacterium
GXC Gammaproteobacteria A (4)
GXC Bcp Beaproteobacterium
GXC Betaproteobacteria A (2)
GXC Mlo Alphaproteobacterium
GXC Bfu Betaproteobacterium
GXC Sfi Gammaproteobacterium
GXC Bps Betaproteobacterium
GXC Pol Betaproteobacterium
GXC Gammaproteobacteria B (3)
ALS Olh Firmicute
ALS Cyanobacteria (2)
ALS Cac Firmicute
ALS Tel Cyanobacterium
ALS Gvi Cyanobacterium
ALS Mac Euryarchaeota
ALS Mlo Alphaprotbacterium
ALS Mma Euryarchaeota
ALS Mca Gammaproteoacterium
ALS Gammaproteobacteria (3)
ALS Firmicutes (18)
BFDC Lpn Betaproteobacterium
BFDC Bja Alpaproteobacterium
BFDC Ppu Gammaproteobacterium
BFDC Cvi Betaproteobactium
BFDC Tvo Theroplasmamata
BFDC Tac Thrmoplasmata
BFDC Pae Gammaproteobacerium
BFDC Bma Betaprotobacterium
BAL Alphaproteobaceria (2)
BAL Pfl Gammaprteobacterium
OCADC Alphaproteobacteria (2)
OCADC Bam Actinobacterium
OCADC Mtu Actinobacterium
OCADC Lac Firmicute
OCADC Oxo Betaproteobacterium
OCADC Ath Euryarchaeota
OCADC Bja3 Alphaproteobacterium
PO Bps Betaproteobacterium
PO Noc Gammaproteobacium
PO Sac Chrenarchaeota
PO Gammaproteobacteria (4)
PO Bma Betaproteobacteria
PO Alphaproteobacteria (2)
PO Cjk Actinobacterium
PO Lactobacillales (Firmicutes) (5)
PO Bacillales (Firmicutes) (6)
IPDC Alphaproteobacteria (2)
PhPDC Azo Betaproteobacteria
PhPDC Rpa Alphaproteobacterium
PhPDC Dde Deltaproteobactium
IPDC Gvi Cyanobacterium
IPDC Gammaproteobacteria (4)
IPDC Lla Firmicute
IPDC Firmicutes (3)
IPDC Mte Actinobacteria
IPDC Mtu Actinobacteria
IPDC Mac Euryrchaeota
PhPDC Sce Ascomycota
PDC Viridiplantae (6)
PDC Apa Aphaproteobacterium
PDC Zpa Gamaproteobacterium
PDC Mpe Firmicute
PDC Firmicutes (2)
PDC Zygomycota (2)
PDC Ascomycota (11)
PFRD Deltapreobacteria (3)
PFRD Tet Firmicute
PFRD Mtm Firmicute
TK Gammaproteobactria (9)
TK Betaproteobacteria (2)
TK Nme Betaproteobacterium
TK Xfl Betaproteobacterium
TK Plants (2)
TK Cyanobacteria (2)
TKLme - Trypnosome
TK Fungi (7)
TK Alphaproteobacteria (5)
TK Tpa1 Sprochaetales
DHAS Fungi (2)
TK Epsilonproteobacteria (2)
TK Chlamidiae (2)
TK Dra Term Deinococcus
TK Firmicutes C (3)
TK Firmicutes B (2)
TK and DHAS Actinobactera (2), (1)
TK Firmicutes A (9)
PKL Cyanobacteria (2)
PKLRpa Alphaproteobacterium
PKL Bsi Alphaproteobacterium
PKL Mes Alphaproteobactium
PKL Betaproteobacteria (3)
PKL Rba Plantomycete
PKL Psy Gammapreobacterium
PKL Sel Cyanobacrium
PKL Acinobacteria (6)
PKL Efc Firmicute
TK Cac2 Firmicute
PKL Firmicutes A (9)
PKL Lms2 Firmcute
PKL Mca Gmmaproteobacteria
TK Bme Alpaproteobacterium
2OXO Firmicutes (3)
2OXO Metazoa (2)
2OXO Bja Alphaproteobacerium
2OXO Gammaprotebactria (2)
2OXO Bps Betaproteobacterium
2OXO Alphaprotebacteria (4)
DXPS Betaproteobacteria A (7)
DXPS Nme Betaproteobacteria
DXPS Gammaroteobacteria A (2)
DXPS Gammaproteobacteria B (9)
DXPS Mbe Gammaproteobacterium
DXPS Mfl Betaproteobacterium
DXPS Betaproteobacteria B (4)
DXPS Rge Betaproteobacterium
DXPS Dar Betaproteobacterium
DXPS Gammaproteobacteria C (16)
DXPS Deltaproteobaceria A (2)
DXPS Cyanbacteria (7)
DXPS Mtm Clostridia (Firmicute)
DXPS Dps Deltaprteobacterium
DXPS Bacteroidetes (2)
DXPS Cth Clostridum
DXPS Cte Chlorobi
DXPS Mag Magnetococcus
DXPS Sth Actinobcterium
DXPS Bacillales (Firmicutes) (5)
DXPS Ttn Costridium
DXPS Deltaproteobcteria B (3)
DXPS Rru Alphaproteobacteria
DXPS Ccr Alphaproteobacteria
DXPS Mes Alphaprteobacteria
DXPS Nar Alphaproteobacteria
DXPS Viridiplantae (8)
DXPS Alphaproteobacteria (6)
PPDC Betaproteobacteria (3)
PPDC Swe Actinobacterium
PPDC Cbe Firmicute
PPDC Ari Atinoacterium
PPDC Actinobacteria (2)
SPDC Rnu Alphaproteobacterium
SPDC Eurarchaeota (5)
SPDC Mka Euryarcheota

Figure 3.25: Overall NJ tree for the TPP-dependent enzymes simplified according to taxonomy.

*Bme*TK groups with the 2OXO clade, while the firmicutes *Cac2*TK groups with the firmicutes *Efc*PKL within the PKL clade. Apart from the presence of the *Cac2*TK outlier, the PKL sequences form a distinct group. As mentioned the 2OXO sequences group within the TK-like group of enzymes and form a distinct clade. Joining this TK-like/2OXO clade, the PFRDs (only *Daf*-like) form a distinct cluster.

In the PDC-like group, the PO, OCADC, CAL, BFDC and GXC enzymes form distinct clades. In the case of ALS, two clades are observed, within which the GXC clade is found. The individual ALS phylogeny (Section 3.3.4.9) was confusing, with species grouping poorly on the basis of their taxonomy, perhaps explaining the splitting of the group. The tree suggests that GXC may have evolved from an ALS-like enzyme. All of the PDC enzymes except the *Mtu*PDC form a distinct clade. *Mt*uPDC groups with *Mtu*IPDC, while the three PhPDC sequences are interspersed among the IPDCs. Apart from the clade of PDCs, the IPDCs and PhPDCs form 2 clades, one of which seems more evolutionarily related to PDC than the other.

Next, the clustering of sequences by taxonomy within the individual enzyme clades is examined in detail. Conclusions from this analysis are given in Section 3.4.4. The overall tree, simplified according to taxonomy is shown in Figure 3.25. Reference is made to the individual ML enzyme trees from Section 3.3.4, and the overall phylogeny of the 17 enzymes (Figure 3.22 on CD-ROM).

In the TK clade, 9 firmicutes form a distinct clade, TK Firmicutes A (*Sau*TK, *Spn*TK, *Bsu*TK, *Bha*TK, *Lin*3TK, *Lmo*TK, *Lla*TK, *Lin2*TK, *Lmo4*TK). These same nine sequences form a clade in the individual TK tree. However, in Figure 3.25, this clade is joined first by the TK Firmicutes B (containing three sequences), then TK Firmicutes C (containing two sequences). The TK Firmicutes B and C clades are found in the overall tree, with sequences in the individual clades in the same order as in the individual TK tree, except that in the overall tree, a DHAS actinobacterium and two TK actinobacteria group first with TK Firmicutes A, before TK Firmicutes B and C join, whereas in the individual TK phylogeny, the actinobacterial clade groups elsewhere.

The seven fungal sequences group together in the overall phylogeny, as they do in the individual TK phylogeny, although the ordering is different. Next, the fungal clade is joined by the lone trypanosome sequence, *Lme*TK, before the clade formed by the two plant and two cyanbacteria join. Thus, the plant, fungal and cyanaoacterial clades group in the same order in both phylogenies.

The nine γ-proteobacteria form distinct clade in both phylogenies, with different ordering. The γ- and β-proteobacteria are paraphyletic in both trees, although in the individual TK tree, all four β-proteobacteria for a distinct clade, whereas in the overall tree, only *Rso*TK and *Aeu*TK cluster. In the TK tree, all six α-protoebacteria group in their own clade, while in the overall tree, five of these group, while *Bme*TK is an outlier which groups with the well resolved 2OXO clade.

The two ε-protoebacteria group in both trees close to the two chlamidia sequences, which form their own clade. The spirochatales sequence, *Tpa*TK is, however located differently in each phylogeny. The one Thermodeinococcus sequence is located close to the ε-protoebacteria and chlamidia clades in both trees.

In the PKL phylogeny, eleven of the firmicutes group together. Nine of these same firmicutes group in the overall tree, while *Lms2*PKL groups wit h the PKL Firmicutes A clade. *Lfc*PKL groups with the TK outlier *Cac2*TK, which is also a firmicute. The six actinobacterial species form a distinct clade in both the individual PKL and the overall trees. The three β-proteobacteria group together in both trees, while the cyanobacterial sequences *Ava*PKL and *Npu*PKL also group in both trees, while the third cyanobacterial sequence *Sel*PKL group close to the γ-proteobacterial *Psy*PKL in both trees, grouping with γ-proteobacteria *Mca*PKL in the PKL tree. In the overall tree, *Mca*PKL is found elsewhere on the tree.

Plantomycete *Rba*PKL groups differently in both trees. α-proteobacteria *Bsi*PKL and *Mes*PKL group in the PKL tree, with the other α-proteobacteria, *Rpa*PKL grouping with the γ-proteobacteria *Psy*PKL. In the overall tree, the three α-proteobacteria are found to be closely related, although they don't form their own clade.

The eight viridiplantae form a distinct clade in both trees, as do the five bacillales firmicutes and seven cyanobacterial sequences. In the DXPS tree, the viridiplantae clade joins with a clade of twelve α-protoebacteria. However, in the overall tree, not all of the α-protobacteria cluster. Eight form their own clade, which joins the viridiplantae clade. Then the *Nar*DXPS, *Mes*DXPS, *Ccr*DXPS and *Rru*DXPS sequence join successively. In both trees, the β- and γ-proteobacterial sequences are found to cluster. In the DXPS tree, there is one major cluster of twenty-five γ-proteobacteria and a smaller group of three γ-protoebacteria, which group with the β-proteobacteria.

In the overall tree, the clustering for each the β- and γ-subdivisions of proteobacteria is less clear. There are three γ-proteobacterial clades: A (*Pbd*DXPS and *Aci*DXPS); B (*Mca*DXPS, *Xfa*DXPS, *Xax*DXPS, *Xca*DXPS, *Avi*DXPS, *Ppu*DXPS, *Pfl*DXPS, *Psy*DXPS and *Pae*DXPS); C (*Apl*DXPS, *Hdu*DXPS, *Vvu*DXPS, *Ppr*DXPS, *Vpa*DXPS, *Vch*DXPS, *Son*DXPS, *Cbl*DXPS, *Wgl*DXPS, *Bsp*DXPS, *Hin*DXPS, *Hso*DXPS, *Pmu*DXPS, *Sfl*DXPS, *Sen*DXPS, *Sty*DXPS). There are also two β-proteobacterial clades: A (*Bpb*DXPS, *Bpa*DXPS, *Pol*DXPS, *Tde*DXPS, *Cvi*DXPS, *Reu*DXPS, *Rme*DXPS); B (*Bfu*DXPS, *Bps*DXPS, *Bcp1*DXPS, *Bcp2*DXPS). Along with other γ- and β–proteobacterial sequences, which don't join clearly defined clades based on subdivision, these clades are interspersed with each other and not as well resolved as n the DXPS tree.

The δ-proteobacteria in both trees group similarly in three places. *Bfr*DXPS and *Chn*DXPS form a clade (A in Figure 3.25), while *Gsu1*DXPS, *Gsu2*DXPS and *Gme*DXPS form a clade together (B in Figure 3.25). *Dps*DXPS doesn't group with other δ-proteobacteria in either tree while clades A and B are found in different positions in the respective trees.

In neither tree does *Mtm*DXPS group with other firmicutes. The bacteroidetes group together in each tree, but in the individual DXPS tree, this clade groups with the clostridium sequence *Cte*DXPS, while in the overall tree *Cte*DXPS groups with the other clostridium *Cth*DXPS. In DXPS, the *Cth*DXPS groups with the other Clostridium *Tth*DXPS, which itself groups with the chlorobi *Cte*DXPS and Magnetococcus *Mag*DXPS in the overall tree.

In the individual 2OXO tree, the position of sequence is very similar to that found in the overall tree. In both, the two metazoan, four of the five α-proteobaceria, the three firmicutes each form distinct clades. In both trees the one β- and two γ-proteobacterial sequences are paraphyletic, but in the overall tree, the two γ-proteobacteria sequences form their own clade. *Bja2*OXO, an α-proteobacterium doesn't group with the other α-proteobacteria in either tree. The position of the metazoan clade varies somewhat between trees, grouping with the γ- and β-proteobacteria in the individual 2OXO tree and with the firmicutes in the overall tree.

In the individual PDC tree as well as in the overall tree, eight viridiplantae, eleven ascomycota and two zygomycota respectively form two distinct clades. In both trees, the α-proteobacterium *Apa*PDC and γ-proteobacterium *Zpa*PDC form a clade. In the individual PDC tree, the three firmicutes group, while in the overall tree, only *Sme*PDC and *Cac2*PDC form a clade, with *Mpe* grouping nearby. The topological relationship of these clades is remarkably similar in both trees.

In the individual PPDC tree, as well as in the overall tree, the three β-proteobacteria group, while the *Cbe*PPDC firmicutes and *Sme*PPDC actinobacteria sequences group together as well. In the overall tree, the *Svi*PPDC and *Shy*PPDC actinobacterial sequences group, with the actinobacterium *Ari*PPDC close by. In the individual PPDC tree, these actinobacteria do not form clades, but are found close to each other.

The distribution of sequences in the individual SPDC tree and the overall tree is identical. In both trees, a clade of five euryarchaeota sequences group, joined in turn by the α-proteobacterium *Rnu*SPDC and the euryarchaeota *Mka*SPDC.

In the individual GXC tree, γ-proteobacteria *Kpn*GXC and *Sen*GXC group, while *Eco*GXC and *Stu*GXC also group. The *Kpn*GXC:*Sen*GXC clade groups first with the β-proteobacterium *Bcp*GXC, before the *Eco*GXC:*Sty*GXC clade joins. In the overall tree, the four γ-proteobacteria (*Kpn*GXC, *Sen*GXC, *Eco*GXC and *Stu*GXC) group together, into one clade (γ-proteobacteria A in Figure 3.25), before the *Bcp*GXC joins.

β-proteobacteria *Reu*GXC and *Rme*GXC group in both trees (β-proteobacteria in Figure 3.25), closer to the *Bsp*GXC in the overall tree than in the individual GXC tree. In both trees, *Pae*GXC, *Ppu*GXC and *Pfu*GXC form a γ-proteobacterial clade (γ-proteobacteria B, in Figure 3.25). For the other sequences, the topologies in either tree are difficult to describe, with no clear grouping by subdivision amongst the proteobacteria.

In both the individual BFDC tree and the overall tree, clades are formed by the thermoplasmata, γ-proteobacteria, and the *Lpn*BFDC and *Cvi*BFDC β-proteobacteria. However, in the overall tree, no species are found to group into clades based on taxonomy.

There is no individual BAL tree (Section 3.34). In the overall tree, the two α-proteobacteria *Sip*BAL and *Rpa*BAL group before being joined by the γ-proteobacterium *Pfl*BAL.

In the individual PO tree, the bacillales and lactobacillales firmicutes sequences group well. In the overall tree, these groups form two tight clades respectively. The γ-proteobacteria *Ype*PO, *Sdy*PO, *Sbo1*PO and *Sbo2*PO form clade in both trees. In both trees this clade joins the β-proteobacterium *Bma*PO next, followed by the α-proteobacterial clade found in both, containing *Rpa*PO and *Bja*PO. *Cjk*PO then joins in both trees. In both trees the β-proteobacterium *Bps*PO and γ-proteobacterium *Npc*PO cluster, then being joined by the chrenarchaeota *Sac*PO.

In both the overall and the individual OCADC trees, the α-proteobacteria *Bja1*OCADC and *Bja2*OCADC group, while the other α-proteobacterium, *Bja3*OCADC groups with the euryarchaeota *Ath*OCADC. In the OCADC tree, actinobacterium *Bam*OCADC and firmicutes *Cac*OCADC group, while this is not observed in the overall tree, where the *Bam*OCADC and *Mtu*OCADC actinobacteria are found close to each other without forming a clade.

In the overall tree, the ALS sequences from two distinct clades. Clade 1 groups with the GXC clade, before being joined by ALS clade 2.

Clade A of Figure 3.25 contains the same sequences as Clade A of Figure 3.16, the individual ALS tree. In ALS Clade A of both trees, seven γ-proteobacteria (*Eca*ALS, *Sty*ALS, *Sen*ALS, *Sfl*ALS, *Eco*ALS, *Xca*ALS and *Xax*ALS) group (γ-proteobacteria B in Figure 3.25). Two chrenarchaeota *Sso*ALS and *Sno*ALS cluster and join with a euryarcaeota clade (euryarchaeota A in Figure 3.25) containing *Pab*ALS and *Pfu*ALS in both. Other euryarchaeota sequences *Mma2*ALS and *Mac*ALS form a clade in both (euryarchaeota B in Figure 3.25). In ALS tree clade A, *Mbu*ALS, *Mth*ALS and *Mja*ALS form another eryarchaeota clade. In the overall tree, *Mth*ALS and *Mja*ALS cluster (euryarchaeota C in Figure 3.25), but not with *Mbu*ALS, which is close by. Other clade common to both trees are the γ-proteobacteria *Ype*ALS and *Psy*ALS clustering (γ-proteobacteria A in Figure 3.25) and the *Bth*ALS and *Ban*ALS firmicutes clustering.

Sequence from ALS clade B in the overall tree correspond with the sequences from clade B of the individual ALS tree. Within these clade in both tree, eighteen of the firmicutes group (except *Cac*ALS). Four γ-proteobacteria group in Clade B of the ALS tree. Of these, three (*Vch*ALS, *Kpn*ALS and *Kte*ALS) form a clade in the overall tree, while the fourth, *Mca*ALS groups closeby. In clades of both trees, four cyanobacteria and the firmicutes *Cac*ALS group together. However, in the individual ALS tree, the four cyanobacteria group together before *Cac*ALS joins, while in the overall tree, the *Cac*ALS clusters amongst the cyanobacterial sequences. The α-proteobacteria *Mlo*ALS and euryarchaeota *Mma*ALS group in both trees.

As mentioned earlier in this section, the three PhPDC sequences are found to group amongst the IPDC sequences, while the outlier *Mtu*PDC is also found amongst the IPDCs. The IPDC / PhPDC sequences separate into two clades (Clades 1 and 2 in Figure 3.25). Clade 1 joins the well defined PDC clade, before IPDC/PhPDC Clade 2 joins. In the IPDC tree as in the IPDC / PhPDC Clade 1 of the overall tree, the four γ-proteobacteria group together. In both trees, the four firmicutes group well, although only three form a clade in the overall tree, with *Lla*IPDC grouping closest to the γ-proteobacterial clade.

The sole actinobacterial IPDC in the individual tree is closest to γ-proteobacteria. However, in the overall tree, *Mtu*IPDC groups with the PDC outlier *Mtu*PDC from the same bacterium.

The IPDC sequence found in PhPDC / IPDC Clade 2 correspond to the three α-proteobacteria and one cyanobacteria sequences which group well in the individual IPDC tree. The clustering of the three α-proteobacterial IPDC sequences (*Abr*IPDC, *Ali*IPDC and *Rpa*IPDC) in the overall tree is interrupted by the β-proteobacterium *Azo*PhPDC. The cyanobacterium *Gvi* is similarly related to the α-proteobacteria in both trees, except that the δ-proteobacterium *Dde*IPDC groups first with the other sequences in PhPDC / IPDC Clade 2.

3.4 Discussion

3.4.1 Individual sequence Alignments for the TPP-dependent enzymes

In the case of the PO and BAL alignments, mean sequence identity is found to be just under the 30 % threshold, although the values are in the upper range of the "twilight zone". In all other cases, homology is >30 %, suggesting their divergent evolution can be studied confidently.

3.4.1.1 Sequence Alignment for TK

3.4.1.1a Choice of TK sequences to study

The details of how the 54 TK sequences analysed were chosen is given in Appendix 1 (Section A1.1), but eventually they consisted of one protozoan, two plant, seven yeast and forty-four bacterial sequences. The bias to bacterial and yeast sequences was intended to improve the accuracy of reconstruction within these groups. The final fifty-four sequences in Table 3.1 represent the optimum set of data for this analysis.

3.4.1.1b Sequence alignment of 54 TK sequences

Alignment of 54 TK sequences showed an enzyme with a highly conserved sequence. Removal of *Cac2*TK and *Bme*TK from the alignment lead to an alignment with 61 fully conserved positions with an additional 29 positions of 100 % similarity. This supports the idea of TK being an essential "housekeeping enzyme" subject to a slow rate of evolution. The level of conservation observed suggests the Alignment 3.1 is robust and can confidently be used for subsequent analyses, such as those performed in Chapters 4 and 5.

In modelling the highly conserved residues of Table 3.2 into the crystal structure of *Eco*TK (1NGS.pdb), the location of each residue was explored. In general conserved residues are spatially spread out through the whole TK molecule, as seen in Figure 3.1, suggesting that the high levels of sequence identity and homology extend to all parts of the TK molecule, including the TKC domain.

In this analysis, three locations for conserved residues were examined; the subunit interface, the surface of the TK molecule and the active-site of TK.

Residues at the interface of the two TK subunits are essential for forming the TK active site (Section 1.1.2). Such residues may also have roles in dimer formation and stability. These residues may vary depending on the organisms environment (e.g. in extremophiles), to improve stability. However, 32 residues are 100 % conserved in the Alignment 3.1, which contains both extremophile and mesophile sequences. These sequences are likely conserved regardless of environmental factors. It is proposed that these residues are essential in all bacterial, yeast and plant TKs for dimer formation, structure formation and stability.

Surface residues are in contact with the solvent and as such can have roles in molecule stability as well as being in direct contact with cofactors and substrates. Of the 43 highly conserved residues found at the surface of the TK molecule, 26 are also found within the active-site, defined as those residues within 10 Å of the TPP molecule (Section 1.6).

These twenty-six residues are likely to be involved in catalysis, either by interaction with substrates or cofactor, or by involvement with the active-site conformation. There are thirty 100 % conserved residues in the active-site, the majority of the 52 positions within 10 Å of the TPP molecule.

This suggests that the catalytic mechanism is highly conserved among TKs and that any differences in substrate specificity between different species of TK will be due to a very small number of active-site differences.

As previously mentioned, 17 residues are found to be at the interface, in the active-site and at the surface of TK. These residues could be conserved for several reasons. Perhaps, for example some are conserved since they are essential for dimer stabilisation and are favoured over other equally stabilising residues for steric reasons.

3.4.1.1c Presence of the "TK motif"

As expected, the TK motif [16] was found in the residues aligning with *Eco*TK residues 461 – 496 (Alignment 3.1). In the Schenk paper, a smaller number (21 TKs and 1 putative TK) of TK sequences were examined, but there were several animal TK sequences, including rat and human TKs. Therefore the motif described in this study for region 461 – 496 in *Eco*TK agrees with the necessarily more loose definition provided by Schenk at most positions when *Cac2*TK and *Bme*TK are excluded form the analysis. Two positions, the A450 in *Hpy*TK and the T493 in *Pse*TK (*Eco*TK numbering) require the definition of the motif to be relaxed in two positions, as summarised in Table 3.3. In the Schenk paper, as in this analysis, the motif is compared with the NADH-binding like domain. The fact that 14 residues of the motif are 100 % conserved either by identity or similarity, suggests that among plants, bacteria and yeasts, many more positions are conserved than if animal species were to be included. Addition of the *Cac2*TK and *Bme*TK sequences to the analysis requires the motif to be redefined as shown in Table 3.3.

3.4.2 The overall Alignments of the PP and Pyr domains of the TPP-dependent enzymes

The alignment of the functionally important residues inspires confidence in the PP and Pyr alignments. when the structurally equivalent regions are compiled and concatenated from the PP and Pyr domain alignments, the *Eco*TK and *Sce*PDC sequences are found to share over 15% homology. While this is firmly in the "twilight zone" (Section 3.2), the structurally equivalent residues, particularly those of known functional importance suggest that such homology is definitely the result of evolution form a distant common ancestor enzyme. Thus homology between TPP-dependent enzyme groups allows the phylogenetic study to proceed.

3.4.3 Phylogenetic analysis of transketolase

Many different methods were employed to generate the 29 phylogenetic trees described in Section 3.2.2.2. Some of the trees were rejected outright but many gave plausible phylogenies on the whole with, in some cases, just a single species grouping within an incorrect kingdom to raise questions about its reliability.

In this study the choice came down to NJ versus ML. The NJ method requires that all sequence data be converted to numerical data and the resulting matrix used to generate a phylogeny by iterations of the NJ algorithm. ML on the other hand examines each character state individually. Since the ultimate aim was to generate mutants based on an accurate analysis of how individual character states evolved during evolution, the more computationally intensive ML method seemed like a suitable choice. With this in mind, the phylogenies of the three candidate trees were examined.

The three candidate trees, Figures 3.7, 3.8 and 4.10, were chosen, as described in Appendix 1 (Section A1.2), as they conform best with the species tree for the organisms examined (the perceived slow evolution of TK suggests its phylogeny should mirror that of the universal tree closely). In each of the three trees there were outliers. In the NJ tree using the categories model, there were *Nme*TK (a β-proteobacterium which doesn't group with the other β-proteobacteria in the tree), *Xff*TK (an α-proteobacterium which groups with β-proteobacteria) and the *Hpy*TK and *Cje*TK pair (γ- and δε-proteobacteria respectively) which group together with chlamydiae and other non-proteobacteria species. Of note in this tree is the fact that the proteobacteria in general separate into two clades (A and B in Figure 4.10), except for the outlying pair of *Hpy*TK and *Cje*TK.

The NJ tree using the Dayhoff PAM with the gamma parameter (Figure 3.7) generated a tree very similar to the NJ with categories (Figure 3.8), with the same outliers. However, in the NJ using the Dayhoff PAM and gamma parameter, all of the proteobacteria group into one clade, except for the outlying pair of *Hpy*TK and *Cje*TK.

The ML tree is very similar to the NJ tree using Dayhoff PAM with the gamma parameter. Outliers are *Hpy*TK and *Cje*TK as previously described, as well as *Xff*TK grouping with the β proteobacteria. The proteobacteria group into one clade, excepting the outliers *Hpy*TK and *Cje*TK. However, in the ML tree, the *Nme*TK species groups with the other β proteobacteria. Thus, the ML tree is chosen as our tree for study, showing how the universal tree is a useful tool for choosing between cladograms (Section 1.7.2). What is shown by this comparison is that NJ yields tree topologies very similar to ML. This observation supports our choice of the NJ method for generating the overall phylogeny for the TPP-dependent enzymes in Chapter 3, where use of the ML method was found to produce an erroneous topology.

3.4.4 The phylogenies of the PP and Pyr domains of the TPP-dependent enzymes

Overall, the phylogeny for the TPP-dependent enzymes is remarkably well resolved. In most cases, sequences cluster well on the basis of domain arrangement, enzyme type and taxonomy. Within enzyme clades, this resolution is, in many cases, as good as in the individual enzyme trees which were generated using the more computationally intensive ML method. Thus, the comparable regions of secondary structure produce a powerful phylogenetic signal. The implication of this is that these regions can be compared between enzymes in the phylogeny for the purposes of protein engineering. In Chapter 5, the comparable regions of TK and PDC are compared to prompt targets for mutagenesis experiments. For the overall tree, NJ was found to produce the best phylogeny, while ML produced an erroneous topology. This suggests that where a relatively small dataset (~320 positions are informative in Alignments 3.19 and 3.21), NJ is the more powerful method. Thus the phylogeny can be said to be robust.

Interesting observations include the apparent divergence of 2OXO from a TK-like enzyme and the scattering of PhPDC species in the IPDC clade. The observation that IPDC and PhPDC are mixed in their clade and the observation that PDC and IPDC forms of *Mtu* form a distinct cluster in the overall TPP-dependent enzyme tree highlights the care that is needed when classifying enzymes of similar type. Since PhPDC can decarboxylate indolepyruvate and is also implicated in tryptophan metabolism, like IPDC (Section 1.2.1.7), the placing of these sequences in the tree is perhaps unsurprising. It would seem however, that most PhPDC and IPDC enzymes have been classified on their ability to catalyse one certain reaction.

Using the data from this study, the evolutionary "story" of the TPP-dependent enzyme family can be assembled (Figure 3.26). This story is similar to that described recently by Duggleby, based on structural comparisons [178]. The earliest ancestor of the TPP-dependent enzymes may have contained regions for both the pyrophosphate and pyrimidine regions of the TPP molecule. In Chapter 6 it is suggested that this protein may be more similar to the modern PP than to modern Pyr domains

(Section 6.4). Dimerisation of this enzyme would have improved catalysis and TPP may have bound as shown in Box A of Figure 3.26. Gene duplication followed by divergence may have resulted in two domains, specialising in binding different parts of the TPP molecule, the ancient PP and Pyr domains. The $\alpha_2\beta_2$-heterotetramer assembly, as seen in Box B of Figure 3.26 survives in modern SPDC enzymes. Fusion of the PP and Pyr domain genes, could have resulted in a homodimeric architecture, as seen in Box C of Figure 3.26, with subunits similar to modern PPDC enzymes, although modern PPDC adopts a homotrimeric state.

From Box C, the other TPP-dependent enzymes evolve. The TH3 domain is recruited by the ancestor of the PDC-like enzymes. This may have coincided with their adoption of the homotetrameric assembly. The PDC clade evolves from here as described previously in Section 3.3.4. Addition of the TKC domain by an ancient enzyme such as shown in Box C formed the ancestor of the TK-like, 2OXO-like and the PFRD enzymes. It appears that PFRD diverged and recruited three additional domains. Evolution within the PFRD clade however, remains controversial [178]. The modern TK-like enzymes then diverge, from which the 2OXO enzymes emerged by splitting of the TK-like subunit into two genes. At the same time as the emergence of 2OXO, the PKL enzymes seem to have evolved, adopting the homohexameric structure.

3.5 Conclusions

For a given TPP-dependent enzyme, sequence homology is high enough to infer phylogeny. For TK, the mean level of sequence homology is 42 %. This suggests that in the active site, where conservation is likely to be even higher than average, a relatively few variable amino acid residues will be responsible for defining substrate specificity. In general in TK, the most highly conserved residues were found at the subunit interface, the surface or the active-site of the molecule, with the TK motif conserved in all TKs examined. In compiling the PP and Pyr alignments, regions of equivalent secondary structure, common to all of the TPP-dependent enzymes examined, were identified and shown to contain strong phylogenetic signals. Such regions can thus be compared and used in the rational design of proteins.

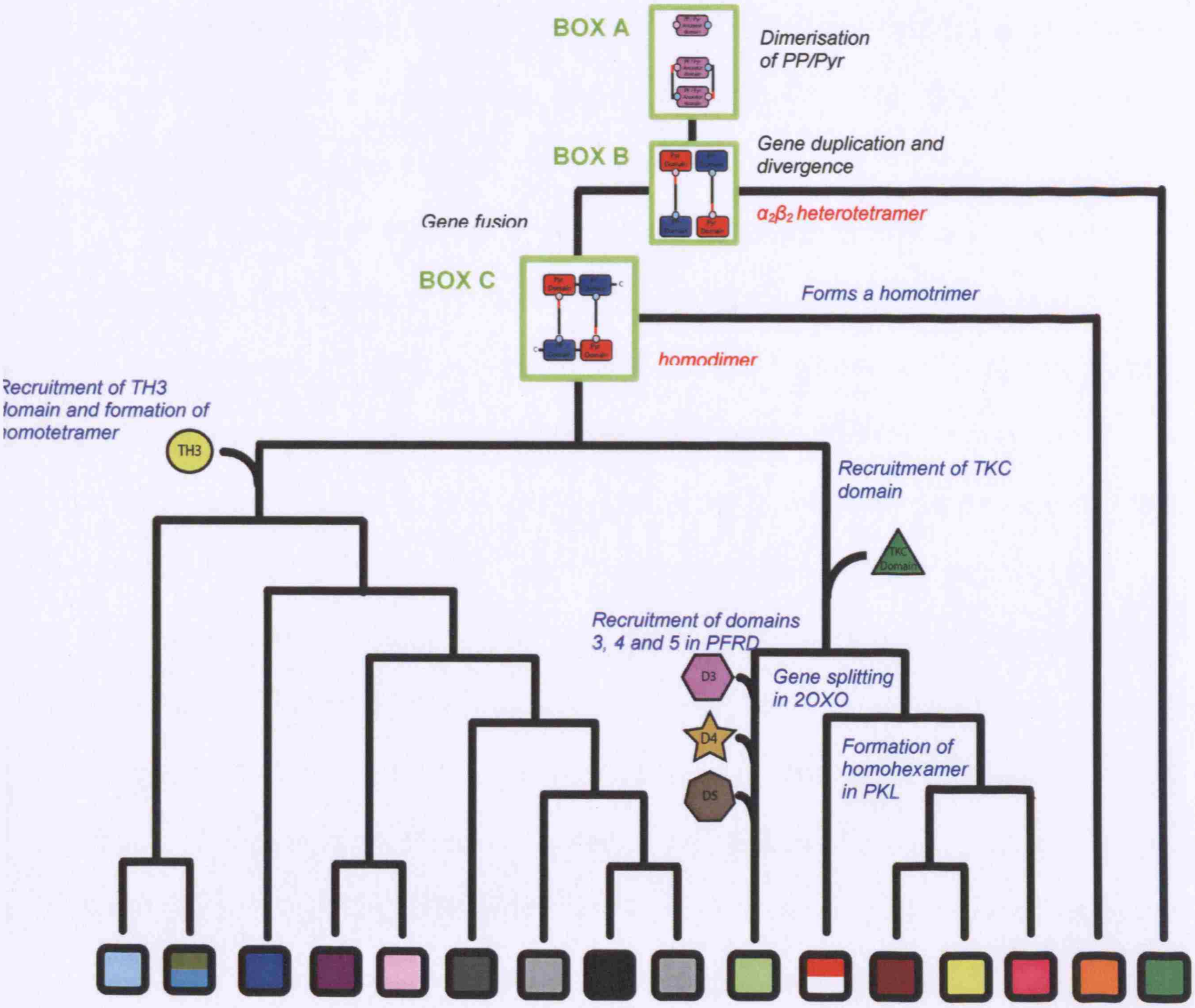

BOX A
Dimerisation of PP/Pyr
BOX B
Gene duplication and divergence
$\alpha_2\beta_2$ heterotetramer
Gene fusion
BOX C
Forms a homotrimer
homodimer
Recruitment of TH3 domain and formation of homotetramer
TH3
Recruitment of TKC domain
TKC Domain
Recruitment of domains 3, 4 and 5 in PFRD
D3
D4
D5
Gene splitting in 2OXO
Formation of homohexamer in PKL

Figure 3.26: Proposed evolutionary history of the TPP-dependent enzyme family.

The individual phylogenies of the TPP-dependent enzyme were, in general, robust. The TK phylogeny agrees well with the universal tree, as expected. The TK tree can thus be used for the reconstruction studies described in Chapter 4.

The overall TPP-dependent family is very ancient. In all but a few instances (GXC, DHAS, PhPDC) enzymes form into distinct clades. Thus their common ancestor is likely to have existed in the progenote population. Remarkably, the phylogenetic signal from the regions compared is still strong after such a long period of evolutionary time.

Chapter 4: Reconstruction of ancestral transketolase enzymes

4.1 Introduction

In the past 15 years, phylogenetic inference followed by reconstruction of ancient phenotypes has been performed. The rationale behind this "resurrection" of ancient proteins is that understanding a proteins evolutionary past can help to better understand its present biochemical function.

Previous examples of such studies in "Palaeobiochemistry" are summarised in Section 1.8. Here a similar study is performed on TK. The aim is to monitor how substrate specificity has evolved in this industrially important enzyme.

TK lends itself well to such a study, since substrate specificity varies considerably between species (Section 1.1.6). The most detailed analysis of the evolutionary history of TK conducted thus far was by Schenk *et al.* [16], where the DNA and translated protein sequences of 21 TKs and 1 putative TK were analysed, defining the "TK motif" in the process. Of the TK phylogenies generated in the study their tree found to best reflect the universal tree of life was a Neighbour Joining (NJ) tree constructed using DNA sequences. In general, TK phylogenies mirrored the "true phylogeny" of the universal tree underlying the housekeeping role of TK. The phylogenetic analysis of 44 bacterial, 7 yeast , 2 plant and 1 protozoan TK protein sequences described in Chapter 3 allows reconstruction of ancestral sequences. Focussing on two industrially useful forms of TK, *Eco*TK and *Sce*TK, their most recent common ancestor is generated. Mutants are then generated that link the common ancestor TK with modern *Eco*TK. Each mutant represents a step along the pathway of the evolution of TK in this lineage. *Eco*TK and *Sce*TK have both been characterised extensively by x-ray crystallography [7,29,187] and by kinetic studies with a wide variety of substrates [67,31,38,5,21-24,32,39,47,49,51-53,57,31,171,188-190]. The two enzymes share 47% sequence identity, suggesting that relatively few residues will be responsible for their differing substrate repertoires. The evolution of *Eco*TK from this common ancestor is studied for the model β-HPA + GA reaction. To assess whether the common ancestor has a broader or narrower substrate specificity than extant *Eco*TK, their respective activities are studied for a host of natural and non-natural substrates.

Here the use of phylogenetic reconstruction, followed by site-directed rational design is shown to generate TK mutants with increased activity for certain reactions as well as altered substrate specificities.

4.2 Methods

4.2.1 Ancestral Reconstruction of the TK phylogeny

4.2.1.1 File format for PAML

The *PAML* [191] program requires both a file describing the phylogenetic tree, previously generated in Chapter 3 (Section 3.2.2.2) and the original sequence Alignment.

4.2.1.2 Reconstruction of TK ancestry using *Codeml*

Using *Codeml* of the *PAML* [191] suite of programs with the default parameters, the ancestral nodes of the Phylogenetic tree for TK were reconstructed. Reconstructed sequences were viewed in *Bioedit*, while the reconstructed tree, with reconstructed node numbers corresponding to ancestral amino acid sequences, was viewed using *Treeview* [185].

4.2.1.3 Choosing active-site residues to study

Residues closest to the substrate binding site are most effective when attempting to engineer substrate specificity in enzymes [117,130,138] (Section 1.6). Residues within 10 Å of the TPP cofactor in the active-site of TK were considered for study. Using the crystal structure of *Sce*TK (1NGS.pdb) viewed with the *Pymol* software [192], 52 residues were chosen within 10 Å of the TPP cofactor.

4.2.2 Tracing Lineages

In this study, lineages will be named according to the convention where, for example, the lineage linking *Eco*TK and *Sce*TK with their most recent common ancestor will be referred to as the "*Eco*:*Sce*" lineage. Nodes encountered along branches of a lineage represent ancestral forms of TK. These ancestral TKs are named according to the node number at which they occur. For example, the ancestral TK at node 58 is named "N58TK". For each lineage, the sequences at each node were aligned in order as shown for the example of Alignment 4.1 on the CD-ROM, discussed in Section 4.3.3. The order of mutations occurring within the 52 active-site residues chosen were compiled for each lineage. The entire phylogenetic analysis was repeated with the TK sequences *Bme*TK and *Cac2*TK omitted to ensure they didn't adversely affect our analysis, as described in Appendix 2 (Section A2.1)

Mutating residue positions in each lineage were highlighted in the crystal structure of *Sce*TK (1NGS.pdb). The TK active-site was examined for each lineage, to identify any potential patterns to steric, hydrophobicity and polarity changes occurring during evolution.

4.2.3 Construction of ancestral TK mutants

Ancestral mutants were generated in the opposite temporal direction to evolution, beginning with extant *Eco*TK and generating successive mutations by SDM, each mutation taking the enzyme one node back in the phylogenetic tree. The initial template used for the K23N mutation was the pQR791 plasmid.

4.2.4 Mutagenic primer design for generating ancestral TKs

Primers were designed using the *Annhyb* program and obtained from Qiagen Ltd. Mutagenic codons are in bold format. Primers used were:

***Eco*1 K23Nv3(F)** CATGGACGCAGTACAGAAAGCC**AAT**TCCGGTCACCCGGGGGCCCCTAT

***Eco*1 K23Nv3(R)** ATAGGGGCCCCCGGGTGACCGGA**ATT**GGCTTTCTGTACTGCGTCCATGC

***Eco*1 P384G(F)** CCTCGGCGGTTCTGCTGACCTGGCG**GGG**TCTAACCTGACCCTGTGGTCTGG

***Eco*1 P384G(R)** CCAGACCACAGGGTCAGGTTAGA**CCC**CGCCAGGTCAGCAGAACCGCCGAGG

***Eco*1 A29Mv2(F)** CAATTCCGGTCACCCGGGT**ATG**CCTATGGGTATGGCTGACATTGC

***Eco*1 A29Mv2(R)** GCAATGTCAGCCATACCCATAGG**CAT**ACCCGGGTGACCGGAATTG

***Eco* 1 A383T(F)** CGGCGGTTCTGCTGACCTG**ACG**GGGTCTAACCTGACCCTGTGGTC

***Eco* 1 A383T(R)** GACCACAGGGTCAGGTTAGACCC**CGT**CAGGTCAGCAGAACCGCCG

***Eco*1 N64A(F)** CGTGACCGCTTCGTGCTGTCC**GCC**GGCCACGGCTCCATGCTGATCTAC

***Eco*1 N64A(R)** GTAGATCAGCATGGAGCCGTGGCC**GGC**GGACAGCACGAAGCGGTCACG

D259KF-BOT(F) CGGTACCCAC**AAA**TCCCACGGTG

D259KR-BOT(R) CACCGTGGG**TTT**AGTGGGTACCG

***Eco*1 G384S(F)** CCTCGGCGGTTCTGCTGACCTGACG**TCC**TCTAACCTGACCCTGTGGTCTGG

***Eco*1 G384S(R)** CAGACCACAGGGTCAGGTTAGA**GGA**CGTCAGGTCAGCAGAACCGCCGAGG

4.2.5 Site Directed mutagenesis (SDM)

SDM was carried out using a Quickchange kit (Stratagene), with protocols outlined below.

4.2.5.1 Site Directed Mutagenesis Protocol 1

SDM Protocol 1 was used to generate the K23N, P384G, A29M and A383T mutations. Each SDM experiment was performed at a range of DNA template (plasmid) concentrations including: 50, 37.5, 25, 12.5 and 5 ng per 50 µL reaction. Initially, separate reactions were carried out for the forward and reverse primer reactions.

Reaction mixtures were prepared as follows:

5 µL of 10X *Pfu* buffer

2 µL of dNTP mixture (4 mM)

1 µL of 50, 37.5, 25, 12.5 or 5 ng.µL^{-1} template plasmid.

1µL DMSO

40.8 µL ddH$_2$O

0.2 µL of 50 mM stock of the appropriate Primer (Forward or Reverse)

To each reaction mixture, 1 µL of *Pfu*Turbo DNAP was added. Reaction mixtures underwent 20 cycles of replication as shown in Figure 4.1.

The Forward primer and the reverse primer reactions were combined and an additional 2 µL of *Pfu*Turbo was added for an additional 20 cycles of PCR as previously described in Figure 4.1.

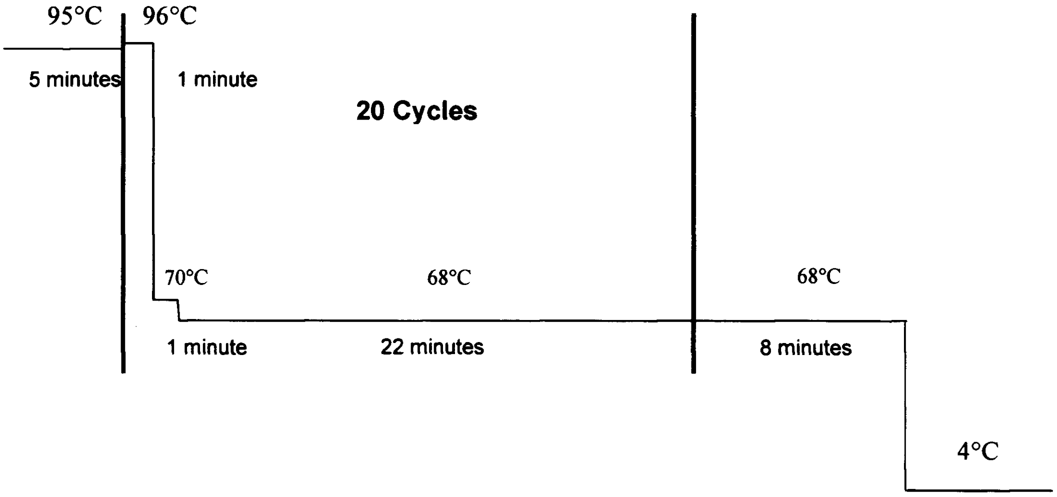


Figure 4.1: Thermocycling parameters for Protocol 1

4.2.5.2 Site Directed Mutagenesis Protocol 2

The D259K / N64A double mutant and G384S mutations were successively generated, using protocol 2. Please note that in generating the D259K and N64A mutations, the N64A mutation was generated first and then used as a template for the generation of the D259K mutation.

5 µL 10X *Pfu* buffer

1 µL of dNTP mixture (4 mM)

1 µL of 50, 37.5, 25, 12.5 or 5 ng.µL^{-1} template plasmid

1 µL DMSO

40.8 µL ddH$_2$O

0.2 µL of 50 mM Forward Primer

0.2 µL of 50 mM Reverse Primer

These reagents were mixed and 1 µL of *Pfu*Turbo DNAP was added to each reaction mixture. The mixture then underwent 20 cycles of replication as shown in Figure 4.2, yielding the required mutation. PCR reactions were confirmed by agarose gel electrophoresis as described in Section 2.5.13.

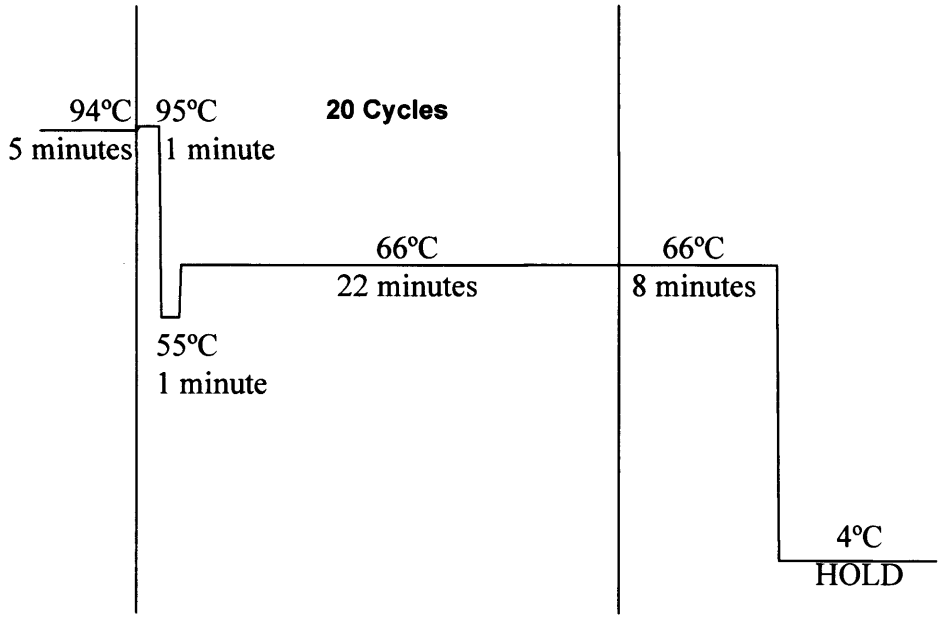


Figure 4.2 : Thermocycling parameters for protocol 2

4.2.6 Digestion of Parental DNA

Once each SDM reaction was complete, the parental, methylated DNA plasmid was digested by addition of 1 µL of *Dpn*1 restriction enzyme to each 50 µL PCR product and incubation at 37 °C for 60 minutes. Agarose gel electrophoresis of post *Dpn*1 digested DNA and transformation of plasmid mutant plasmids into Xl1 Blue competent cells was performed as described in Sections 2.5.13 and 2.5.9 respectively.

4.2.7 Characterisation and storage of TK ancestral mutants

Colonies from the XL1 Blue transformants were picked, grown overnight in LB media (150 mg.mL^{-1} AMP) and plasmid DNA obtained using a miniprep kit (Qiagen), as described in Section 2.5.8. DNA was quantified spectrophotometrically (Section 2.5.10) and confirmed by DNA sequencing. Plasmids containing the correct mutations were transformed into XL-10 Gold cells and plated on LB agar (AMP^{+}) as described in Section 2.5.9. Colonies were picked, grown overnight in LB and glycerol stocks were made for storage at -80 °C. Cultures for assay were grown in shake flasks, lysed, centrifuged and filtered using a 0.45 μm filter to remove debris. The TK concentration of this clarified lysate was quantified on an Agilent 2100 bioanalyser as described in Section 2.5.7. This clarified lysate was used in subsequent assays, described in Section 4.2.8.

4.2.8 Enzyme Assays

Assays were performed with *Eco*TK and ancestral TK mutants for the β-HPA + GA reaction as described in Section 2.5.11.2. Assays for *Eco*TK and N58TK activity on the potential pyruvate + GA reaction were conducted as detailed in Section 2.5.11.3.

4.2.8.1 Reactions of *Eco*TK and N58TK with β-HPA and a range of acceptor substrates

The following reactions were performed using *Eco*TK and N58TK:

β-HPA + erythrose, β-HPA + glyceraldehyde, β-HPA + G3P, β-HPA + E4P, β-HPA + ribose, β-HPA + R5P, β-HPA + arabinose, β-HPA + A5P, β-HPA + glucose and β-HPA + G6P.

Reagents added to a 1.5 mL Eppendorf were as follows:

200 μL of a 3X(50 mM Tris-HCl with 50 mM β-HPA) stock solution.

50 μL of a 12X(9 mM $MgCl_2$ with 2.4 mM TPP) stock solution.

X μL of clarified, bioanalysed lysate to give a TK concentration of 0.5 $mg.mL^{-1}$

(300 – X) μL of ddH_2O

These reagents were mixed by vortexing and incubated for 30 minutes at room temperature. The reaction was initiated by addition of 50 µL of a 12X(50 mM) stock solution of the aldol acceptor substrate. The mixture was quickly vortexed and kept at 25 °C. 100 µL samples of the reaction were taken at 5, 15, 30, 60 and 120 minute intervals. For the reactions of β-HPA with G3P and G6P, reactions were conducted with constant mixing at setting 400 on an IKA-VIBRAX-VXR machine. This was a precaution given that both G3P and G6P required sonication in order to dissolve. However, in neither reaction was any precipitation observed. Reactions were quenched as described in Section 2.5.12.1.

4.2.8.2 Reactions of *Eco*TK and N58TK with GA and a range of donor substrates

The following protocols are similar to those described in Section 2.5.11.2 for the β-HPA + GA reaction. In each case, a stock solution is made at 3X concentration containing 3X 50 mM Tris-HCl and 3X 50 mM concentration of the ketol donor, which is adjusted to pH 7.0 using concentrated NaOH. The following reactions were performed using *Eco*TK and N58TK:

GA + Xylulose, GA + Sedoheptulose, GA + Fructose, and GA + F6P.

Reagents added to a 1.5 mL Eppendorf were as follows:

200 µL of a 3X(50 mM Tris-HCl with 50 mM Ketol donor substrate) stock solution.

50 µL of a 12X(9 mM $MgCl_2$ with 2.4 mM TPP) stock solution.

X µL of clarified lysate to give a TK concentration of 0.5 mg.mL^{-1}

(300 – X) µL of ddH_2O

These reagents were mixed by vortexing and left to incubate for 30 minutes at room temperature. The reaction was initiated by addition of 50 µL of a 12X stock solution of GA. The mixture was quickly vortexed and kept at 25 °C. 100 µL samples of the reaction were taken at 5, 15, 30, 60 and 120 minute intervals, quenched as described in Section 2.5.12.1, by pipetting into a 96 well plate.

4.2.8.3 Reactions of *Eco*TK and N58TK with β-HPA and a range of non-natural substrates

This Section describes the general protocol for the following reactions: benzaldehyde + β-HPA , propionaldehyde + β-HPA , hydroxybenzaldehyde + β-HPA and p-anisaldehyde + β-HPA . Each susbtrate was used at 30mM, except popionaldehyde, which was used at 50mM.

For each different aldehyde reaction, a solution is made up containing 3X 2.4 mM TPP, 3X 9mM $MgCl_2$ and 3X aldehyde. This is termed the “aldehyde solution”. A second solution of 3X 50 mM β-HPA with 50 mM Gly-Gly is used in all reactions. This is termed the “β-HPA :Buffer” solution. The following reagents were added to a 1.5 mL Eppendorf tube:

100 µL of clarified lysate (neat).

100 µL of Aldehyde solution

The reaction mixture was vortexed and allowed to incubate for 30 minutes at room temperature. Reactions were initiated by addition of 100 µL of β-HPA:Buffer solution. Eppendorfs were shaken for the duration of the reaction at setting 400 on an IKA-VIBRAX-VXR machine. Reactions were quenched at 19 hours with 300 µL of 0.2 % TFA for those samples to be examined using TLC (Section 4.2.9).

The above reaction was repeated a second time for propionaldehyde so a reaction profile could be generated using the HPLC assay described in Section 3.2.6.8. The protocol was exactly the same as above, except that all of the volumes used were doubled and enzyme concentration was adjusted to 0.5 $mg.ml^{-1}$. Samples were taken at 5 minutes, 1 hour, 3, 6 and 25 hour time points, when 100 µL aliquots were quenched in 96 well plate wells containing 100 µL of 0.2 % TFA (Section 2.5.12.1).

4.2.9 Thin Layer Chromatography Method for analysing TKs for activity towards non-natural substrates

Samples of the reaction mixtures, where the reaction had reached completion as described in Section 4.2.8.3 were deposited as a spot on plastic plate (stationary phase) using a glass capillary tube. Samples were dried using a hot airgun. The bottom edge of the plate was placed in a solvent reservoir of 100% ethyl acetate, allowing the solvent to move up the plate by capillary action. When the solvent front reached the upper edge of the stationary phase, the plate was removed from the solvent reservoir. Any excess solvent was left to dry for short time before placing the plate into the staining reservoir containing a PMA mix. The plate was then heated with the airgun to activate the stain and develop the plate.

4.2.10 HPLC Assays performed using a 300 mm Aminex HPX-87H ion-exchange column

General HPLC methods are described in Section 2.5.12. HPLC assays performed for *Eco*TK and the ancestral TK mutants for the β-HPA + GA reaction were conducted as described in Section 2.5.12.3a. *Eco*TK and N58TK were assayed on HPLC for the potential pyruvate + GA reaction as described in Section 2.5.12.3b.

4.2.10.1 The β-HPA + erythrose reaction

The HPLC configuration was exactly the same as described in Section 2.5.12.3a. The retention times of β-HPA, erythrose and fructose using this method are 8.5, 10.05 and 11.92 minutes, respectively. The progress of the reaction was followed by the appearance of the fructose product, the peak area of which was used in subsequent analysis.

Figure 4.3 shows the calibration curve for fructose on the 300 mm Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories). The relationship between injection concentration and peak area is linear up to 15 mM concentration of fructose.

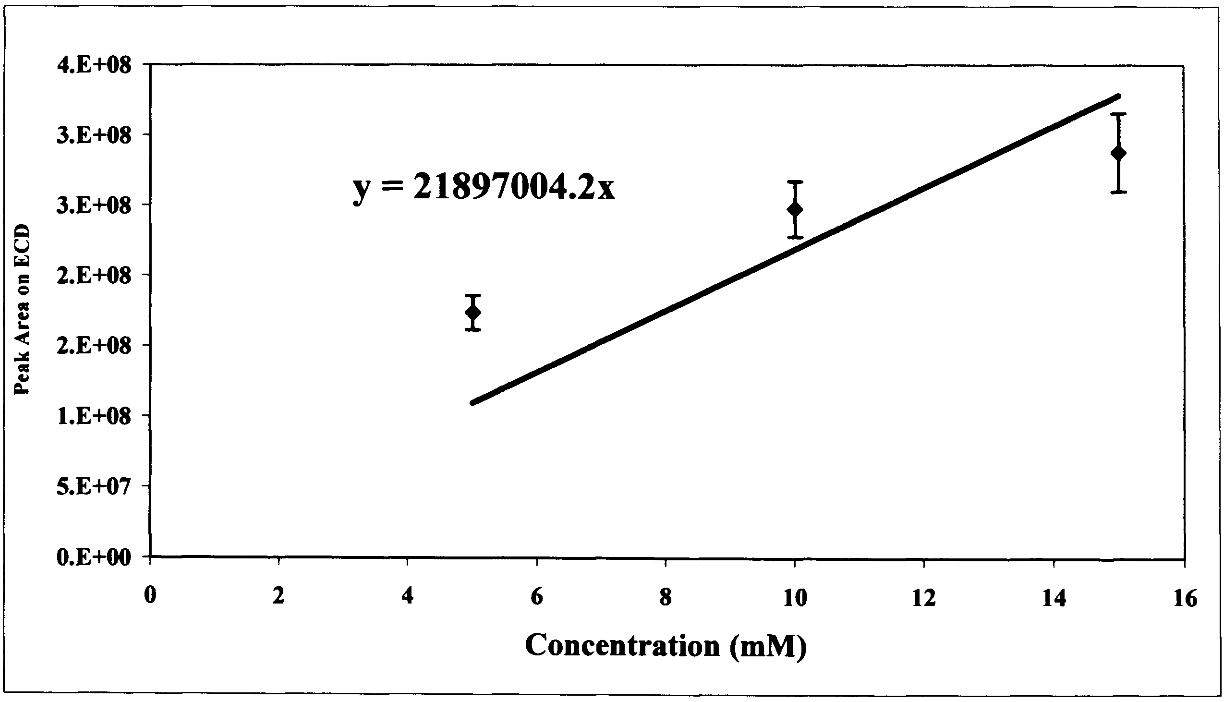


Figure 4.3: Calibration curve for fructose on a 300 mm Aminex HPX-87H ion-exclusion column. Samples were prepared in the following conditions: 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 °C. Samples were then diluted 1:2 in 0.1% TFA and analysed on an Aminex HPX-87H ion-exclusion column.

4.2.10.2 The GA + fructose reaction

The HPLC configuration was exactly the same as described in Section 2.5.12.3a.

The retention times of L-erythrulose, fructose, erythrose and GA are 6.98, 10.05, 11.92 and 12.88 minutes, respectively. The progress of the reaction was followed by the appearance of the erythrose product, the peak area of which was used in subsequent analysis.

Figure 4.4 shows the calibration curve for erythrose on the 300 mm Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories). The relationship between injection concentration and peak area is linear up to 25 mM concentration of erythrose.

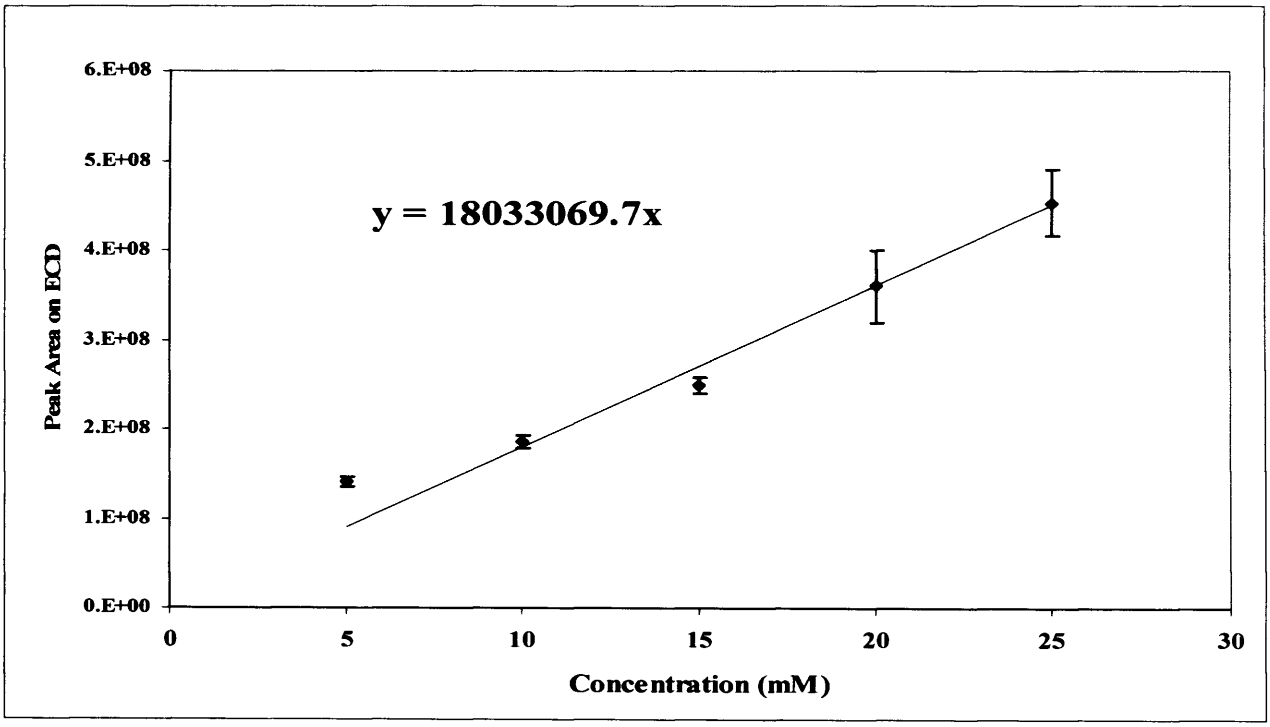


Figure 4.4: Calibration curve for erythrose on a 300 mm Aminex HPX 87H ion-exclusion column. Samples were prepared in the following conditions: 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 °C. Samples were then diluted 1:2 in 0.1% TFA and analysed on an Aminex HPX-87H ion-exclusion column.

4.2.11 Assays performed using a SphereClone 5 µm SAX-N(CH3)3 column

This HPLC assay protocol is for the following reactions:

β-HPA + glyceraldehyde, β-HPA + G3P, β-HPA + E4P, β-HPA + ribose, β-HPA + R5P, β-HPA + arabinose, β-HPA + A5P, β-HPA + glucose, β-HPA + G6P, GA + xylulose, GA + sedoheptulose, GA + F6P.

The mobile phase was 0.045 M KH_2PO_4 with a flow rate was 0.5 $ml.min^{-1}$. Separation of the components was attempted using a SphereClone 5 µm SAX-N(CH3)3 strong anion exchange column (Phenominex). The temperature of the column was maintained at 25 °C by the chromatography oven. The ECD was set up as described in Section 2.5.12.2.

In reactions where β-HPA was used as the ketol donor, β-HPA depletion was monitored. Where GA was the aldol acceptor, the depletion of GA was monitored. The retention times of β-HPA and GA are 17.8 and 6.88 minutes, respectively.

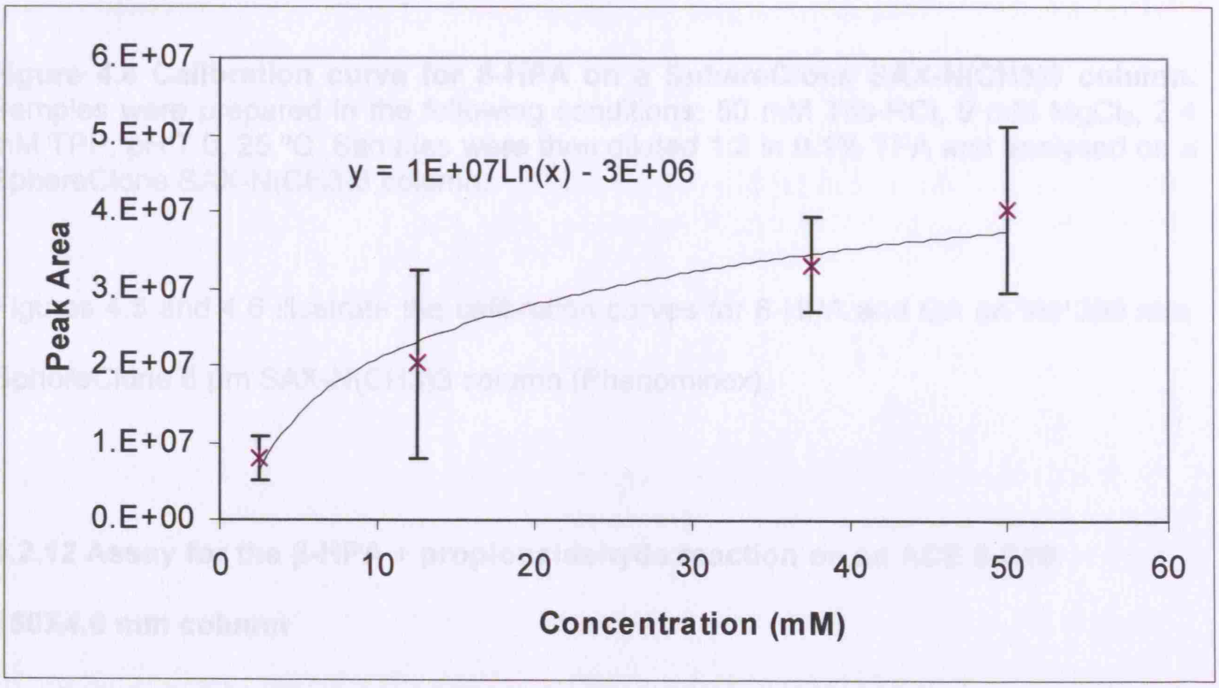


Figure 4.5 Calibration curve for GA on a SphereClone SAX-N(CH3)3 column. Samples were prepared in the following conditions: 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 ºC. Samples were then diluted 1:2 in 0.1% TFA and analysed on a SphereClone SAX-N(CH3)3 column.

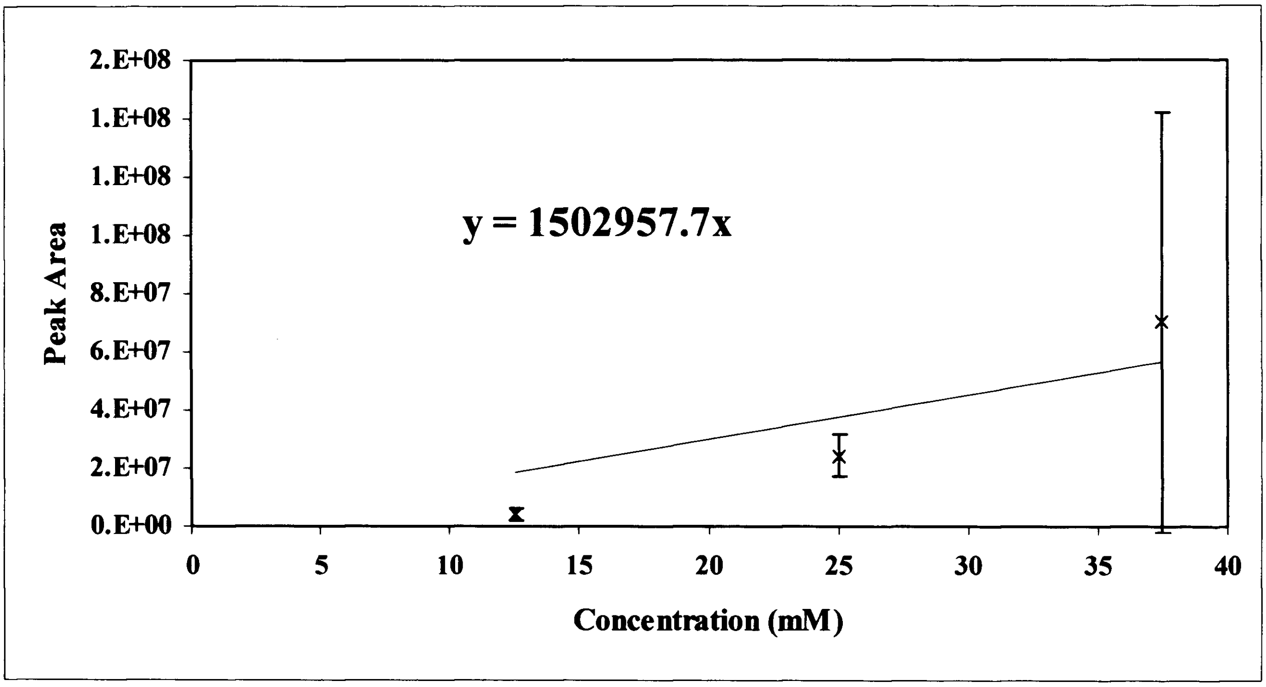


Figure 4.6 Calibration curve for β-HPA on a SphereClone SAX-N(CH3)3 column. Samples were prepared in the following conditions: 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 °C. Samples were then diluted 1:2 in 0.1% TFA and analysed on a SphereClone SAX-N(CH3)3 column.

Figures 4.5 and 4.6 illustrate the calibration curves for β-HPA and GA on the 300 mm SphereClone 5 µm SAX-N(CH3)3 column (Phenominex).

4.2.12 Assay for the β-HPA + propionaldehyde reaction on an ACE 5 C18 150X4.6 mm column

The mobile phase was 0.1 % (v/v) TFA and the flow rate was 0.5 $ml.min^{-1}$. Separation of the components was achieved using an ACE 5 C18 column. The temperature of the column was maintained at 30 °C by the chromatography oven. Concentrated NaOH was mixed with output stream from the column prior to reaching the detector. This NaOH was kept pressurised at 50 psi. The ECD monitored the nC of the output stream. The injection volume was 10 µl for all samples.

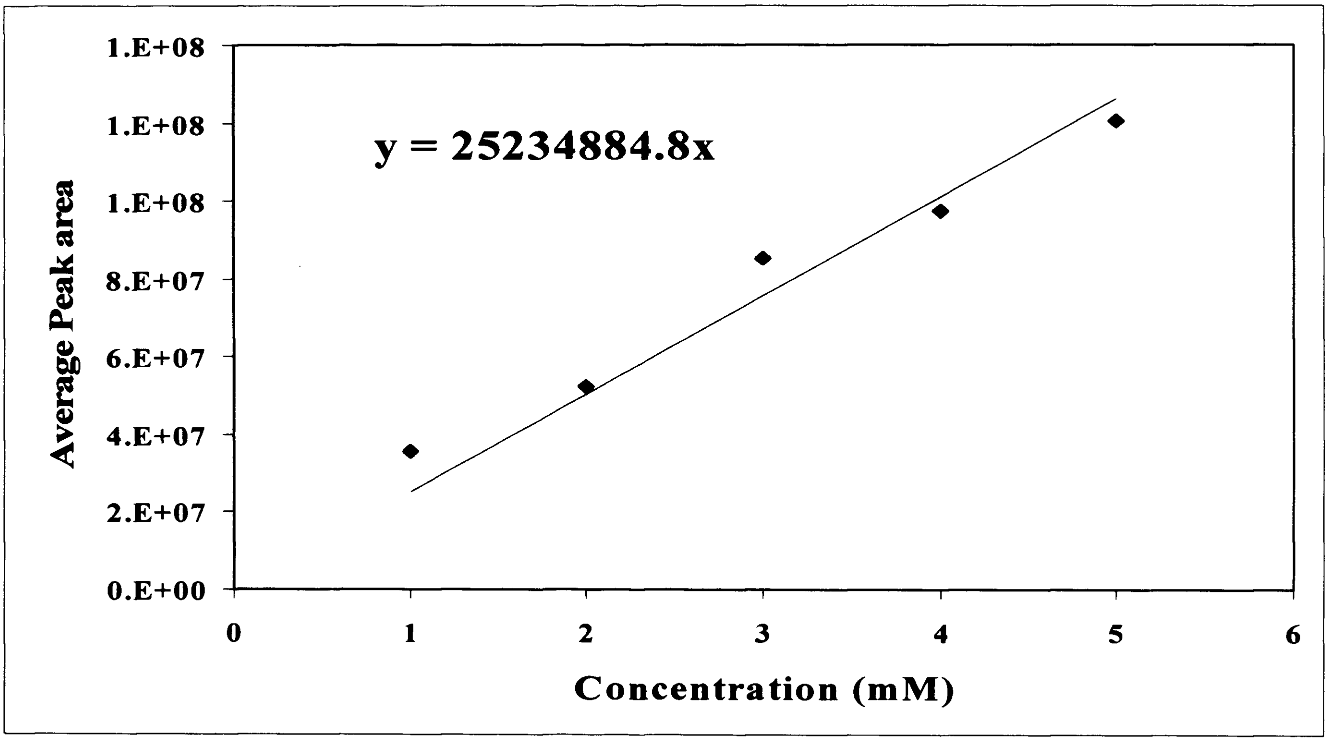


Figure 4.7: Calibration curve for propionaldehyde ketodiol on an ACE 5 C18 (150X4.6 mm) column. Samples were prepared in the following conditions: 50 mM Gly-Gly, 9 mM $MgCl_2$, 2.4 mM TPP, pH 7.0, 25 °C. Samples were then diluted 1:2 in 0.1% TFA and analysed on an ACE 5 C18 (150X4.6mm) column.

The retention times of β-HPA, propionaldehyde and PDK using this method are 3.17, 6.82 and 5.7 minutes, respectively. The progress of the reaction was followed by the appearance of PDK product, the peak area of which was used in subsequent analysis.

Figure 4.7 shows the calibration curve for PDK on the ACE 5 C18 column and was provided by Tarik Senussi of the department of Biochemical Engineering, UCL. The relationship between injection concentration and peak area is linearly proportional up to 5 mM.

4.3 Results

4.3.1 Ancestral reconstruction

Ancestral reconstruction with *PAML* returned the Alignment 4.2 on the CD-ROM, where node numbers correspond with those on the ML trees shown in Figure 4.10. In all, 52 nodes were reconstructed. The *PAML* program returns a score of percentage accuracy at a given site each reconstructed sequence, the values for which are summarised in Figure 4.8.

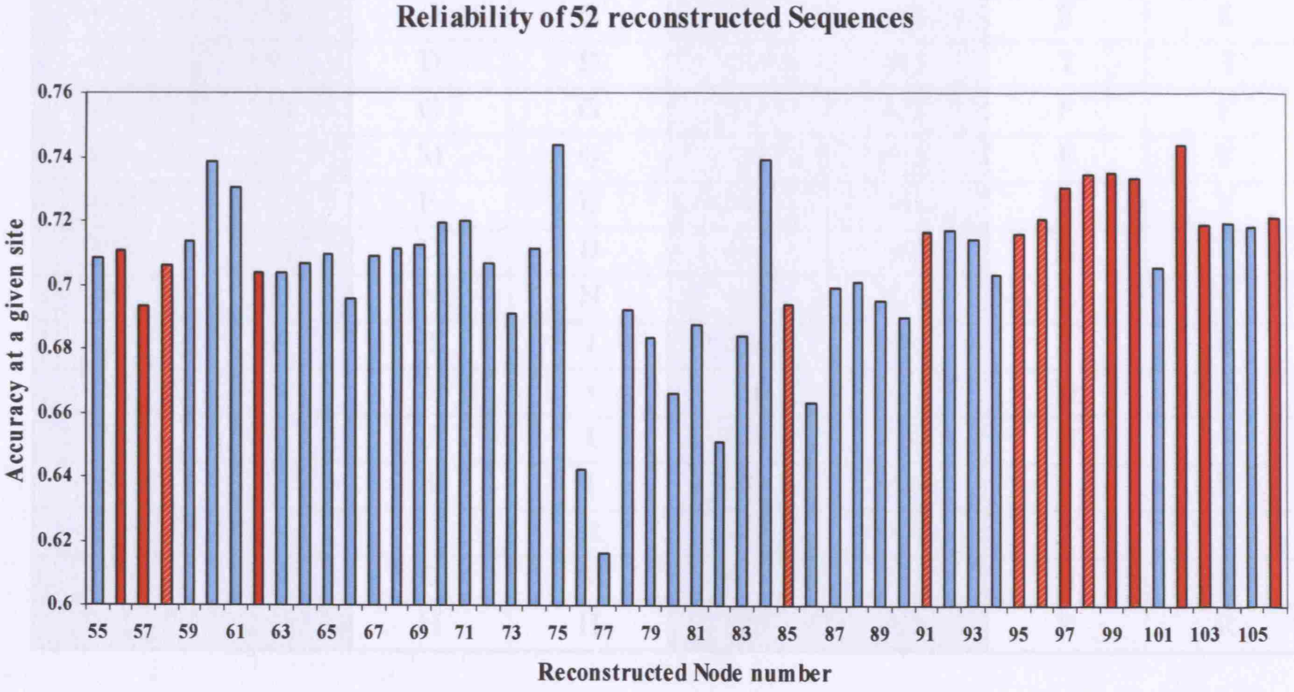


Figure 4.8: The percentage accuracy of reconstructed sites as calculated by PAML. Bars in red show the nodes that link *Eco*TK and *Sce*TK in the reconstructed ML tree (Figure 4.10), a subset of which (coloured in red and white stripes) involve mutations at the active-site and are used in Section 4.2.3 for generating ancestral TK enzymes.

*Eco*TK	*Sce*TK	*Eco*TK	*Sce*TK	*Eco*TK	*Sce*TK	*Eco*TK	*Sce*TK
23	27	K	N	262	264	G	G
24	28	S	S	263	265	A	A
25	29	G	G	264	266	P	P
26	30	H	H	358	359	R	R
29	33	A	A	380	381	A	A
64	67	N	N	381	382	D	D
66	69	H	H	382	383	L	L
91	94	R	R	383	384	A	T
100	103	H	H	384	385	P	P
114	116	G	G	385	386	S	S
115	117	P	P	387	388	L	L
116	118	L	L	409	416	V	I
154	156	G	G	411	418	E	E
155	157	D	D	433	441	T	T
156	158	G	G	434	432	F	F
159	161	M	Q	437	435	F	F
160	162	E	E	440	448	Y	Y
183	185	D	D	461	469	H	H
185	187	N	N	466	474	L	V
187	189	I	I	468	476	E	E
188	190	S	T	469	477	D	D
189	191	I	I	470	478	G	G
247	250	I	I	471	479	P	P
258	260	H	H	472	480	T	T
259	261	D	S	473	481	H	H
261	263	H	H	520	528	R	R

Table 4.1: The 52 Active-site residues picked for study. All residues were found within 10 Å of the TPP molecule. Residues were chosen using the *Sce*TK crystal structure (1NGS.pdb). The numbering is given for both the *Eco*TK and *Sce*TK structures. Residues in *Sce*TK that differ from *Eco*TK at the corresponding position are coloured in red.

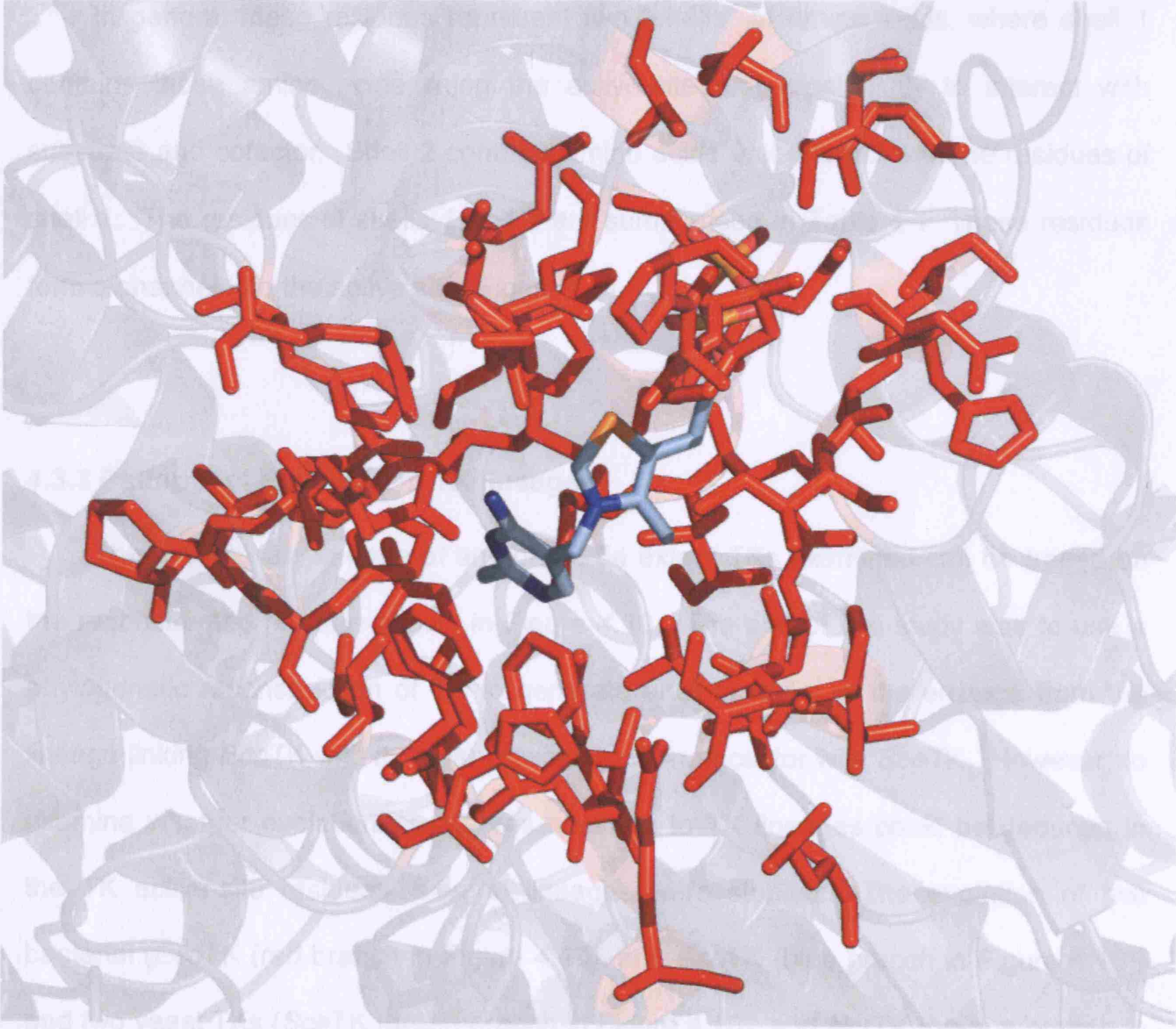


Figure 4.9: The 52 active-site residues within 10 Å of the TPP molecule. 52 active-site residues are found within 10 Å of the TPP cofactor and coloured red. The TPP molecule shown with a CNOS colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

4.3.2 Choice of active-site residues to study

The 52 active-site residues found to be within 10 Å of the TPP moiety are shown in Figure 4.9. These residues are likely to be most proximal to TK substrates and are closest to the TPP group. Such residues are expected to have a strong influence on the determination of TK substrate specificity (Section 1.6).

In general these residues represent two "shells" of amino acids, where shell 1 contains those amino acids lining the active-site and most likely to interact with substrate and cofactor. Shell 2 contains amino acids that interact with the residues of shell 1. The residues of shells 1 and 2 are summarised in Table 4.1. These residues form a channel into the active site (Figure 4.9).

4.3.3 Patterns of Evolution in TK lineages

The evolutionary lineage of any of the 54 extant TKs examined can be traced on the reconstructed ML tree shown in Figure 4.10. The aim of this study was to use a phylogenetic reconstruction of TK to generate ancient forms of the enzyme from the lineage linking *Eco*TK with its most recent common ancestor with *Sce*TK. However, to examine whether evolutionary patterns common to TK lineages could be deduced in the TK active-site residues, several lineages were studied. The evolution of two bacterial (*Eco*TK (red branch in Figure 4.10), and *Bsu*TK (blue branch in Figure 4.10)) and two yeast TKs (*Sce*TK (green branch in Figure 4.10), and *Ncr*TK (orange branch in Figure 4.10) were studied. Since the yeast groups join the rest of the clades on the tree at node 58, it works effectively as an outgroup and thus node 58 is deemed the most likely common ancestor node for the 54 TK species examined. For the lineage describing the evolution of *Eco*TK from the N58TK common ancestor, the order of nodes encountered is as follows:

N58-N82-N85-N91-N95-N96-N97-N98N-N99-N100-N102-*Eco*TK, highlighted by a red branch in Figure 4.10.

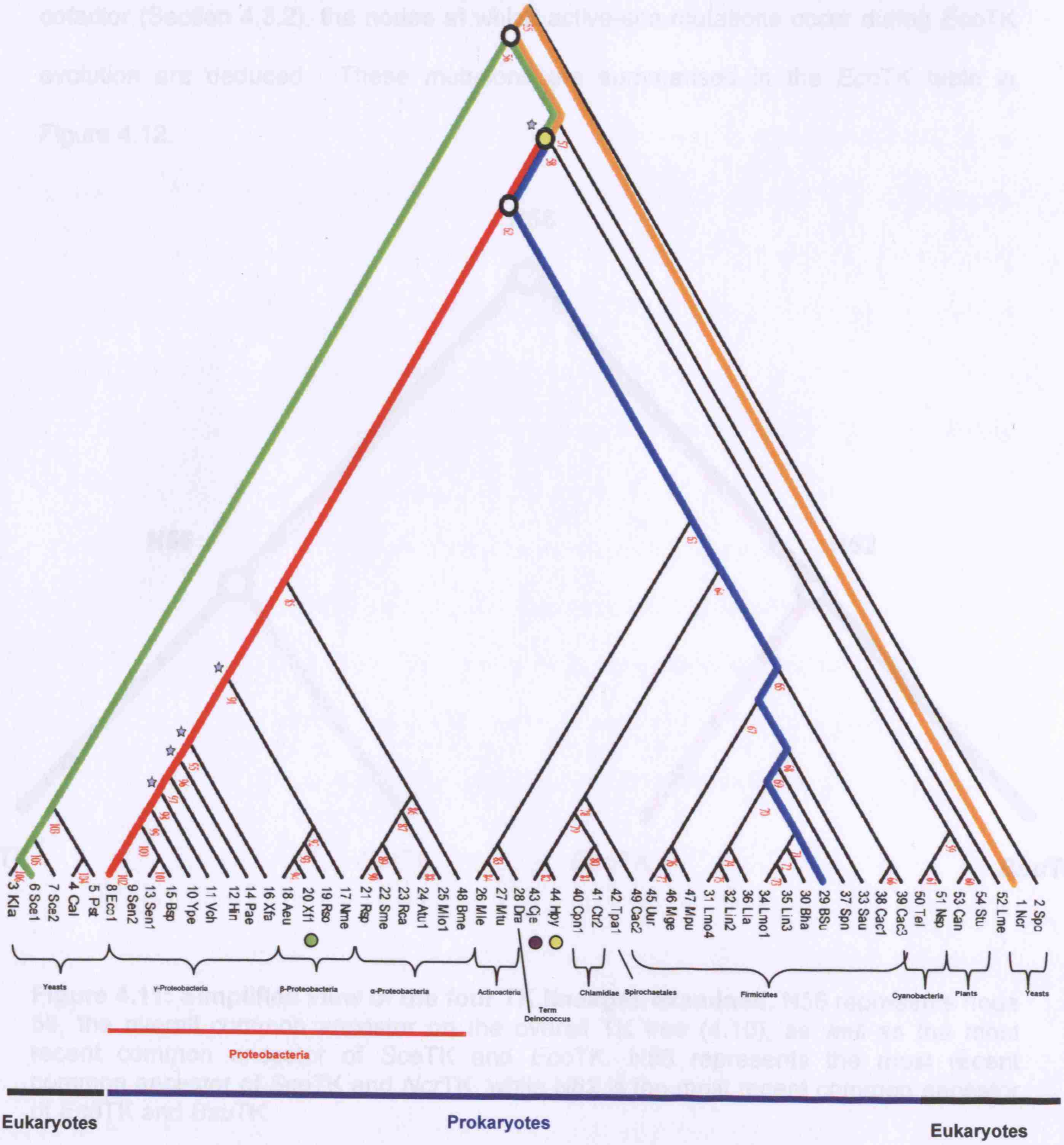


Figure 4.10: Maximum Likelihood tree for transketolase. Node numbers are those assigned by the *PAML* program. Branches are coloured to represent lineages described in Section 4.3.4. Outliers Xfl, Cje and Hpy are indicated with coloured circled. Nodes chosen for reconstruction in Section 4.2.2 are indicated with stars in the diagram.

As mentioned in Section 4.2.2, the alignment of these nodes is shown in Alignment 4.1 on the CD-ROM. Focussing on the 52 active-site residues within 10 Å of the TPP cofactor (Section 4.3.2), the nodes at which active-site mutations occur during *Eco*TK evolution are deduced. These mutations are summarised in the *Eco*TK table in Figure 4.12.

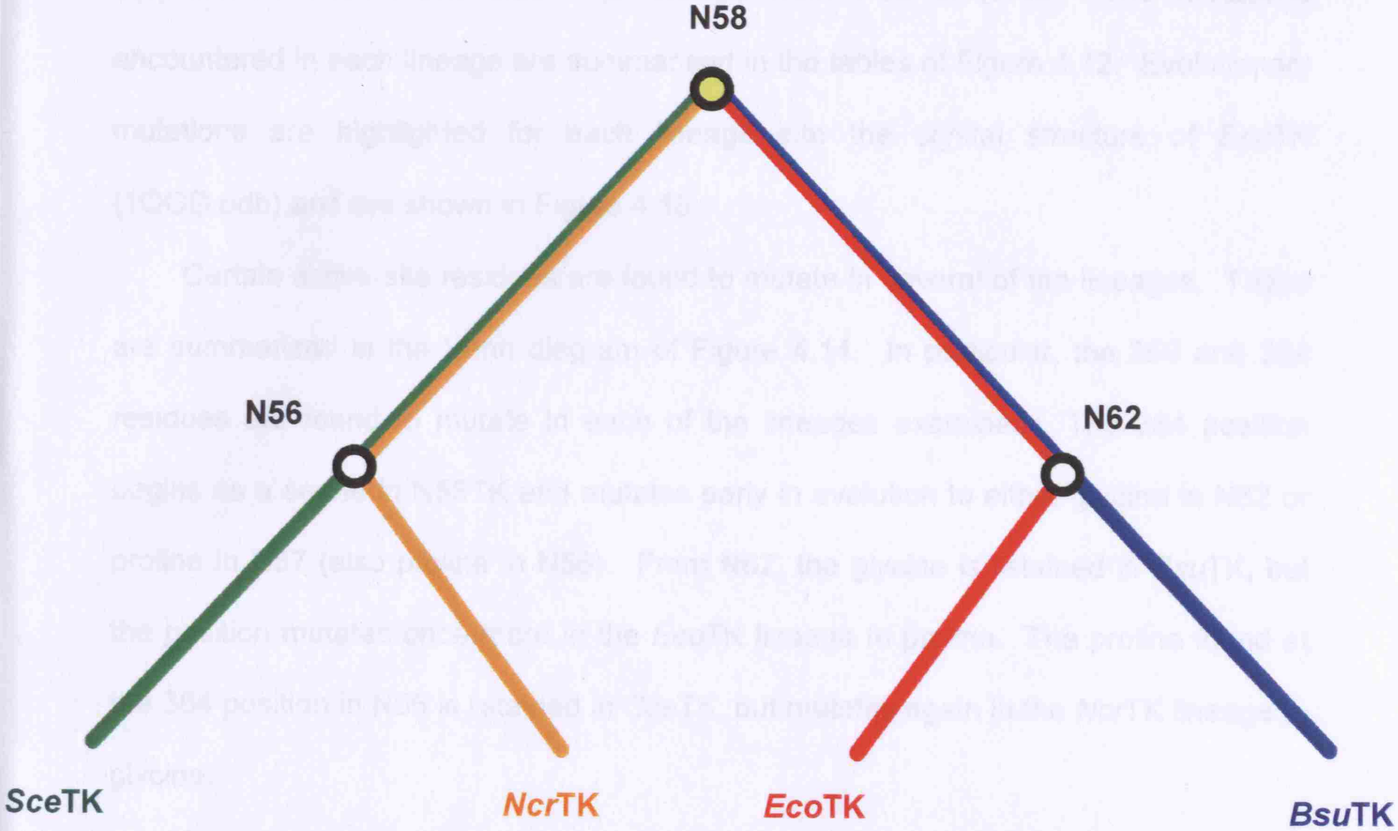


Figure 4.11: Simplified view of the four TK lineages examined. N58 represents node 58, the overall common ancestor on the overall TK tree (4.10), as well as the most recent common ancestor of *Sce*TK and *Eco*TK. N56 represents the most recent common ancestor of *Sce*TK and *Ncr*TK, while N62 is the most recent common ancestor of *Eco*TK and *Bsu*TK

The same procedure is repeated for the evolutionary lineages of *Sce*TK, *Ncr*TK and *Bsu*TK. *Eco*TK and *Bsu*TK share their most recent common ancestor at node 62, while the yeasts *Sce*TK and *Ncr*TK have their most recent common ancestor at node 56. Thus to simplify the study and avoid repetition when summarising the evolutionary mutations, the lineages linking the four TK species are considered in the simplified form illustrated in Figure 4.11.

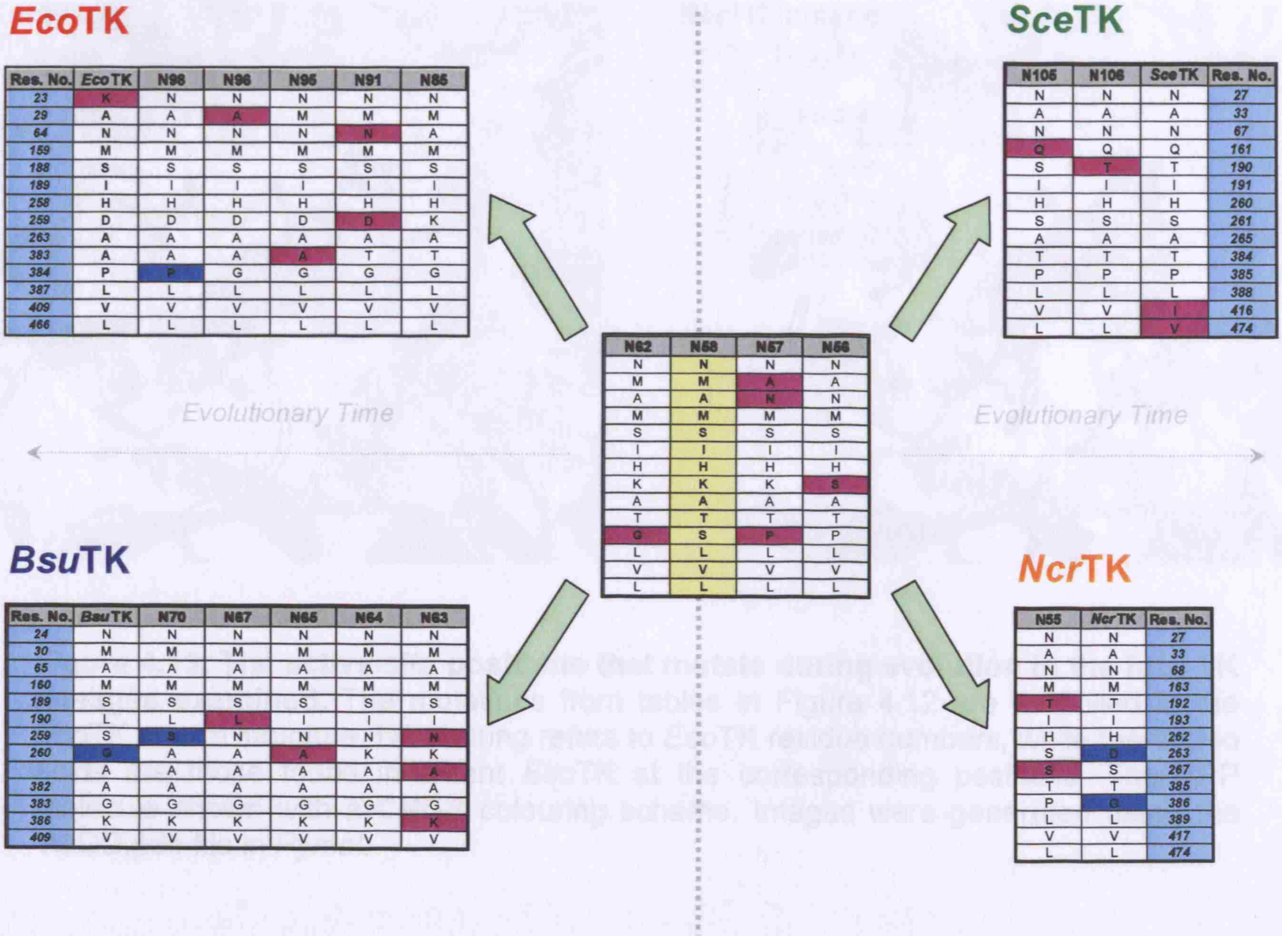
The evolution of N62 and N56 (both indicated as white circles in Figures 4.10 and 4.11) from N58 (yellow circle in Figures 4.10 and 4.11) is studied (Figure 4.12). The evolution of *Sce*TK and *Ncr*TK is studied from their common ancestor N56, while evolution of *Eco*TK and *Bsu*TK is studied from N62 onwards. The mutations encountered in each lineage are summarised in the tables of Figure 4.12. Evolutionary mutations are highlighted for each lineage into the crystal structure of *Eco*TK (1QGD.pdb) and are shown in Figure 4.13.

Certain active-site residues are found to mutate in several of the lineages. These are summarised in the Venn diagram of Figure 4.14. In particular, the 259 and 384 residues are found to mutate in each of the lineages examined. The 384 position begins as a serine in N58TK and mutates early in evolution to either glycine in N62 or proline in N57 (also proline in N56). From N62, the glycine is retained in *Bsu*TK, but the position mutates once more in the *Eco*TK lineage to proline. The proline found at the 384 position in N56 is retained in *Sce*TK, but mutates again in the *Ncr*TK lineage to glycine.

The 259 position is lysine at N58 and switches to serine in the branch to N56. The serine is retained in the *Sce*TK lineage leading from N56, whereas in the *Ncr*TK lineage, the position mutates to aspartate. The lysine is still retained at this position in N62. At node 91 of the *Eco*TK lineage, the lysine to aspartate mutation occurs whereas in the *Bsu*TK lineage, the position mutates twice at N65 and upon the emergence of extant *Bsu*TK.

Thus position 384 and 259 are likely to be of great importance in determining the substrate specificity of TKs in general.

The positions of the mutations were examined. Previous studies have suggested that mutations in the active-site can have a synergistic effect. The orientation of the residues may be important here [193]. Mutating residues that were found to be opposite each other when viewed through either the TPP or the E4P molecules in the *Sce*TK crystal structure were noted. Since *Sce*TK and *Eco*TK structures can be fitted resulting in a RMS of 30.362 Å (data not shown), we can be confident that residues that are opposing couples through E4P in the *Sce*TK structure are also opposing in the *Eco*TK structure. Residues were sometimes found very close to one another in the active-site. The aim of these observations was to identify instances where mutating positions have influenced the evolution of their neighbours in the active-site.



*Eco*TK

Res. No.	*Eco*TK	N98	N96	N95	N91	N85
23	K	N	N	N	N	N
29	A	A	A	M	M	M
64	N	N	N	N	N	A
159	M	M	M	M	M	M
188	S	S	S	S	S	S
189	I	I	I	I	I	I
258	H	H	H	H	H	H
259	D	D	D	D	D	K
263	A	A	A	A	A	A
383	A	A	A	A	T	T
384	P	P	G	G	G	G
387	L	L	L	L	L	L
409	V	V	V	V	V	V
466	L	L	L	L	L	L

*Sce*TK

N105	N106	*Sce*TK	Res. No.
N	N	N	27
A	A	A	33
N	N	N	67
Q	Q	Q	161
S	T	T	190
I	I	I	191
H	H	H	260
S	S	S	261
A	A	A	265
T	T	T	384
P	P	P	385
L	L	L	388
V	V	I	416
L	L	V	474

N62	N58	N57	N56
N	N	N	N
M	M	A	A
A	A	N	N
M	M	M	M
S	S	S	S
I	I	I	I
H	H	H	H
K	K	K	S
A	A	A	A
T	T	T	T
G	S	P	P
L	L	L	L
V	V	V	V
L	L	L	L

*Bsu*TK

Res. No.	*Bsu*TK	N70	N67	N65	N64	N63
24	N	N	N	N	N	N
30	M	M	M	M	M	M
65	A	A	A	A	A	A
160	M	M	M	M	M	M
189	S	S	S	S	S	S
190	L	L	L	I	I	I
259	S	S	N	N	N	N
260	G	A	A	A	K	K
	A	A	A	A	A	A
382	A	A	A	A	A	A
383	G	G	G	G	G	G
386	K	K	K	K	K	K
409	V	V	V	V	V	V

*Ncr*TK

N55	*Ncr*TK	Res. No.
N	N	27
A	A	33
N	N	68
M	M	163
T	T	192
I	I	193
H	H	262
S	D	263
S	S	267
T	T	385
P	G	386
L	L	389
V	V	417
L	L	474

Figure 4.12: Summary of the active-site mutations occurring during evolution in the four different TK lineages examined. Cells are coloured pink to indicate that the position mutates to a different amino acid at this point. Where a second mutation occurs at a givenposition, the cell is coloured blue.

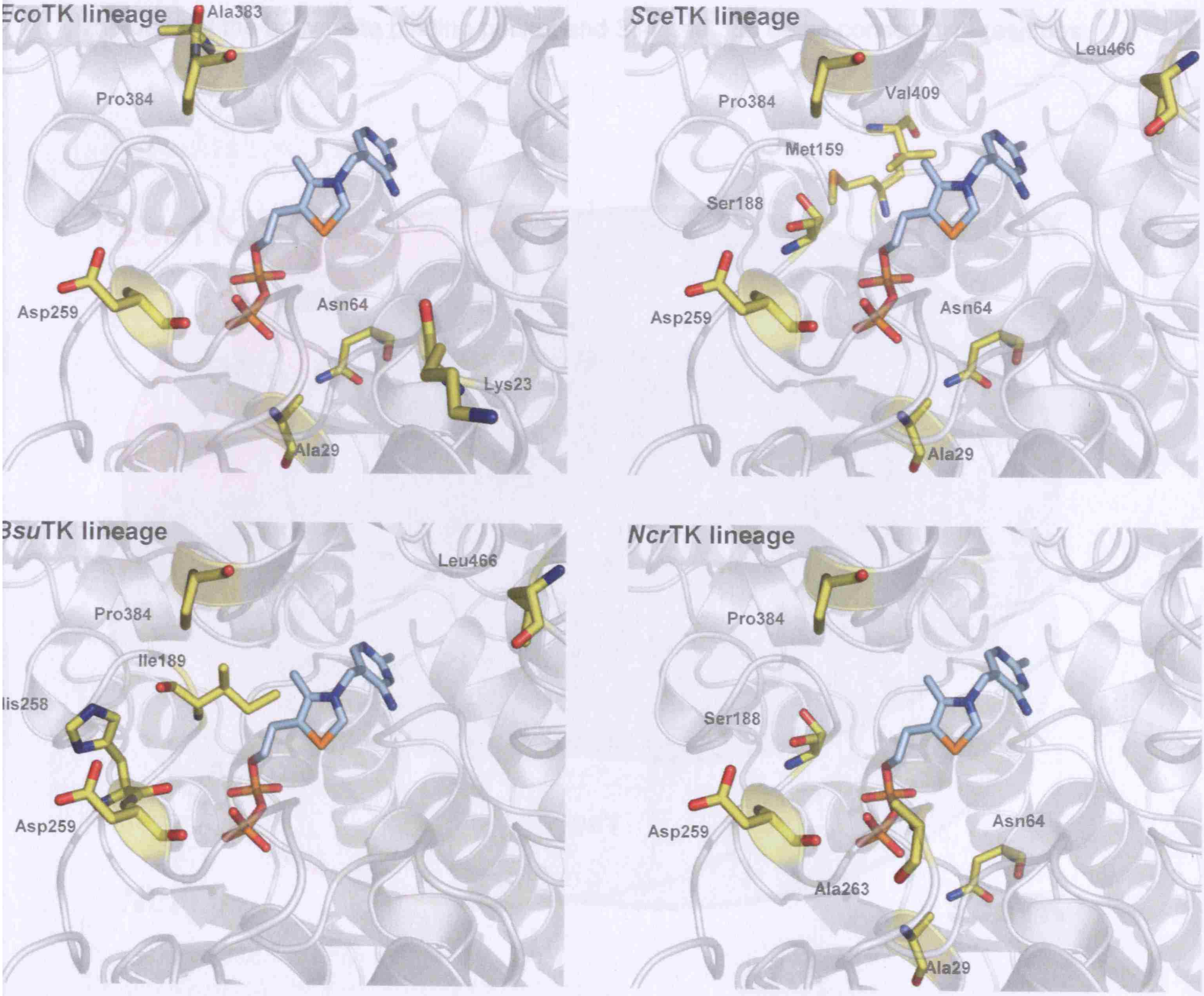


Figure 4.13: The active-site positions that mutate during evolution in the four TK lineages examined. The mutations from tables in Figure 4.12 are modelled in the *Eco*TK crystal structure. Numbering refers to *Eco*TK residue numbers, while the amino acids are those found in extant *Eco*TK at the corresponding positions. The TPP molecule shown with a **CNOS** colouring scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

In the *Eco*TK lineage, two residues are found to be opposite each other when viewed through the E4P molecule, namely 23 and 384. The consecutive residues 383 and 384 are both found to be opposite residue 64, when viewed through the TPP cofactor. Two clusters are noted in the *Eco*TK lineage, with 29 and 64 clustering near the base of the active-site binding pocket and 384 and 385 being consecutive residues.

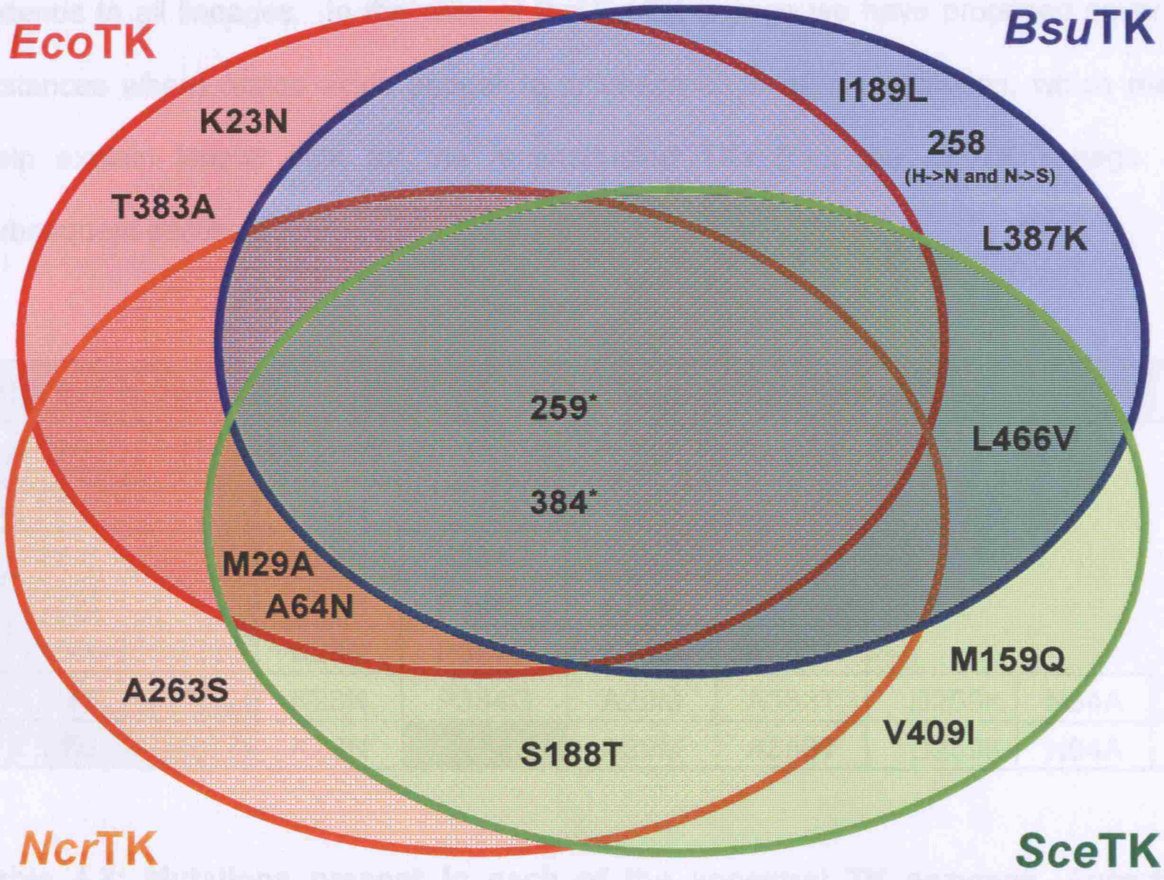


Figure 4.14: Venn diagram summarising the residues which mutate in the evolutionary lineages of *Eco*TK, *Sce*TK, *Ncr*TK and *Bsu*TK. Residues marked with an asterisk undergo different mutations in different lineages as described in the text. The residue corresponding with position 258 in *Bsu*TK mutates twice during the evolution of *Bsu*TK from N58TK. Both of these mutations (H to N, and N to S) are highlighted.

Changes in polarity occur at all residues that are members of opposing couples, whether through the E4P or TPP molecules, suggesting that a change in the polarity of one residue may illicit a response from its opposing partner. The mutations in the consecutive amino acids 383 and 384 both involve a change from hydrophobic to hydrophilic residues during evolution. Perhaps the switch in 383 prompts the same switch in 384.

Evolutionary patterns in the other three lineages studied are less informative. It would appear that there is no clear pattern of evolution in the TK active-site that extends to all lineages. In the case of the *Eco*TK lineage we have proposed several instances where amino acids appear to influence each others evolution, which may help explain kinetic data for the reconstructed TKs from the *Eco*TK lineage in subsequent sections.

TK mutant name	Mutations Present						
*Eco*TK							
N98	K23N						
N96	K23N	P384G					
N95	K23N	P384G	A29M				
N91	K23N	P384G	A29M	A383T			
N85	K23N	P384G	A29M	A383T	D259K	N64A	
N58	K23N		A29M	A383T	D259K	N64A	G384

Table 4.2: Mutations present in each of the ancestral TK enzymes. Ancestral mutants are named after the nodes at which they occur on our reconstructed ML tree for TK (Figure 4.10), where "N" stands for "node". For example, the common ancestor of *Eco*TK and *Sce*TK, N58TK is found at node 58 of the reconstructed tree.

4.3.4 Mutants Generated

Based on the *Eco*TK table in Figure 4.12, six ancestral TK mutants were generated, as described in Section 4.2.3. These mutants were named according to the nodes at which their mutations occurred (Figure 4.12 for details). In the data eventually used for generating ancestral TK sequences for the N58TK to *Eco*TK lineage, simultaneous mutations are seen in the active-site residues in one instance, where the A64N and the K259D mutations occur at the transition from N85TK to

N81TK. Such instances may reflect a real evolutionary event or may be due to a lack of sequences which are informative at the positions under study (Section A1.1).

All mutations were confirmed by DNA analysis, grown, lysed and quantified on a Bioanalyser as detailed in Section 2.5. Figure 4.15 shows a sample gel for a successful site directed mutagenesis experiment, in this case, the K23N mutation. Figure 4.16 is an SDS PAGE gel showing K23NTK levels in cell lysate.

4.3.5 Enzyme assays

In each case, lysate was added to ensure a TK concentration of 0.5 mg.mL^{-1}. Since none of the mutations caused a change in the length of the enzyme, concentrations could be calculated using the calibration curve for commercial TK (Figure 2.5).

4.3.5.1 *Eco*TK and ancestral TK activity for the β-HPA + GA model reaction

Mutations will be discussed in the order in which they would have occurred during evolution i.e. N58 (Common Ancestor) → N85 → N91 → N95 → N96 → N98 → *Eco*TK. Consequently, a mutant generated by mutation of *Eco*TK K23 to N experimentally to yield N98TK, will be discussed in the context that the N23 residue mutated to K during the course of evolution.

Figure 4.17 shows the initial velocities for reactions carried out for *Eco*TK and the ancestral TKs with the substrates β-HPA and GA.

Each of the ancestral mutants has a higher initial velocity for the β-HPA + GA reaction than *Eco*TK, except for N98TK, which shows large error bars, suggesting its activity may not be significantly different to *Eco*TK. Thus, the effect of the K23N mutation is unclear, but compared with the other mutants, its activity appears low. The P384G mutation which yields N96TK increases activity by 1145 % when compared

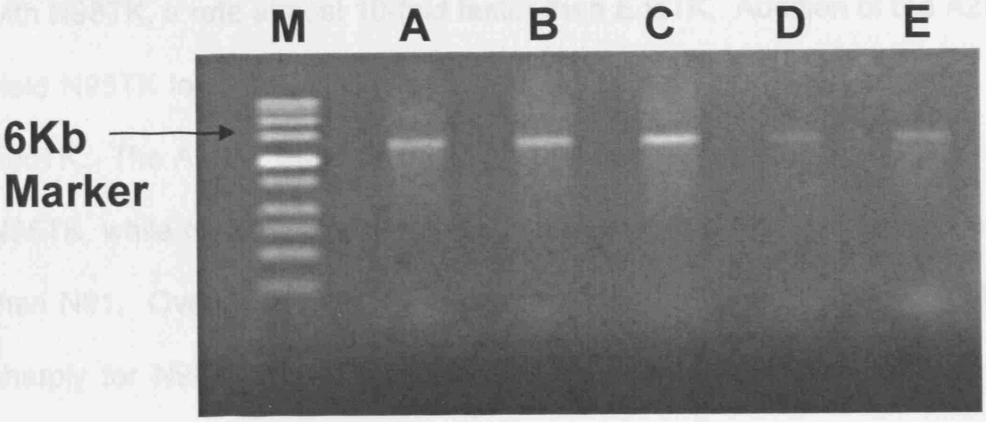


Figure 4.15: Agarose gel showing the product of the Quickchange^{TM} reaction for the K23N mutation. 0.7 % (w/v) agarose was dissolved in buffer (40 mM Tris-acetate and 1 mM EDTA in pure water), to which 0.5 mg.L-1 ethidium bromide was added. Once set, the gel was placed in buffer containing Tris-acetate and EDTA (TAE buffer). Samples were added to 6× loading buffer (Bromophenol blue (0.05% w/v), sucrose (40% w/v), EDTA (0.1 M, pH 8.0), and SDS (0.5% w/v) (Sigma-Aldrich Company Ltd.). Lanes A, B, C, D and E show, respectively, the Quickchage products where 50, 37.5, 25, 12.5 and 5 ng.μL^{-1} of template were used. 1.5 μl of Novagen 0.5–12.0 kbp Perfect DNA Markers (EMD Biosciences Inc.) are shown in lane M. Electrophoresis was performed at 60 V for 1 hour. The gel was visualised and photographed under UV light using a Gel Doc 2000 system (Bio-Rad Laboratories).

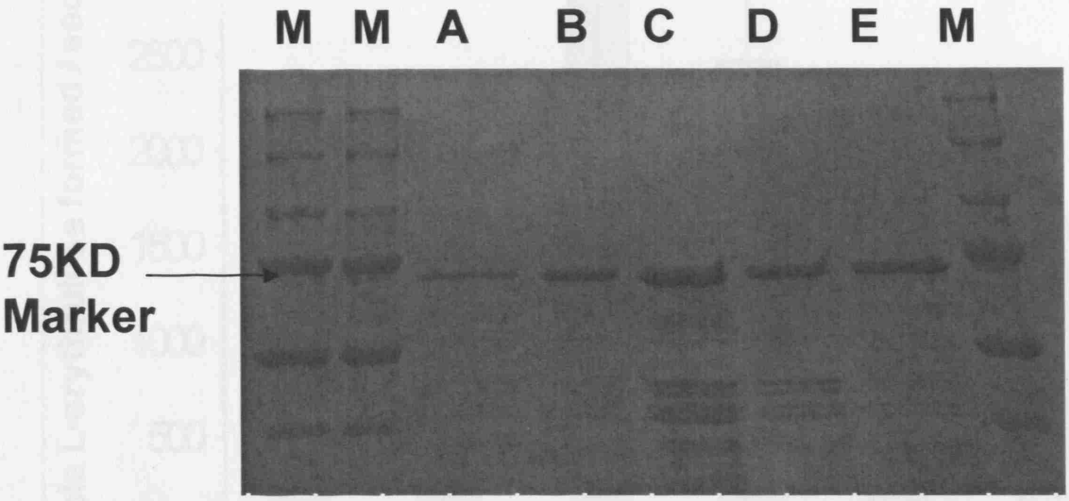


Figure 4.16: SDS PAGE gel for cell lysate containing the K23NTK protein. 12.5 % (w/v) separating and 6 % (w/v) stacking gels were used. Samples were mixed with 2× Laemmli Sample Buffer (Bio-Rad Laboratories), containing 62.5 mM Tris-HCl, pH 6.8, 25% glycerol, 2% SDS and 0.01% Bromophenol. Lanes A to E each contain 20 μL of sample + sample buffer. The wells of marker lanes are labelled "M" and were loaded with 5 μl of Precision Plus Protein Standards (Bio-Rad Laboratories). Protein bands were visualised by staining in an aqueous staining solution containing 0.1% (w/v) Coomassie Blue R-250, 40% (v/v) methanol and 10% (v/v) acetic acid, then destained by boiling for 10 minutes in 1 litre of pure water. The gel was photographed using a Gel Doc 2000 system (Bio-Rad Laboratories).

with N98TK, a rate almost 10-fold faster than *Eco*TK. Addition of the A29M mutation to yield N95TK lowered activity by 6.5 %, but velocity remained ≥7 times faster than for *Eco*TK. The A383T switch to yield N91TK decreased activity by 43 % compared with N95TK, while N85TK, with the additional D259K and N64A mutations was 17 % faster than N91. Overall, TK activity increased from the common ancestor, then decreased sharply for N98TK and *Eco*TK, perhaps due to other evolutionary constraints. This data also shows the potential of natural evolution mutants for obtaining higher activity.

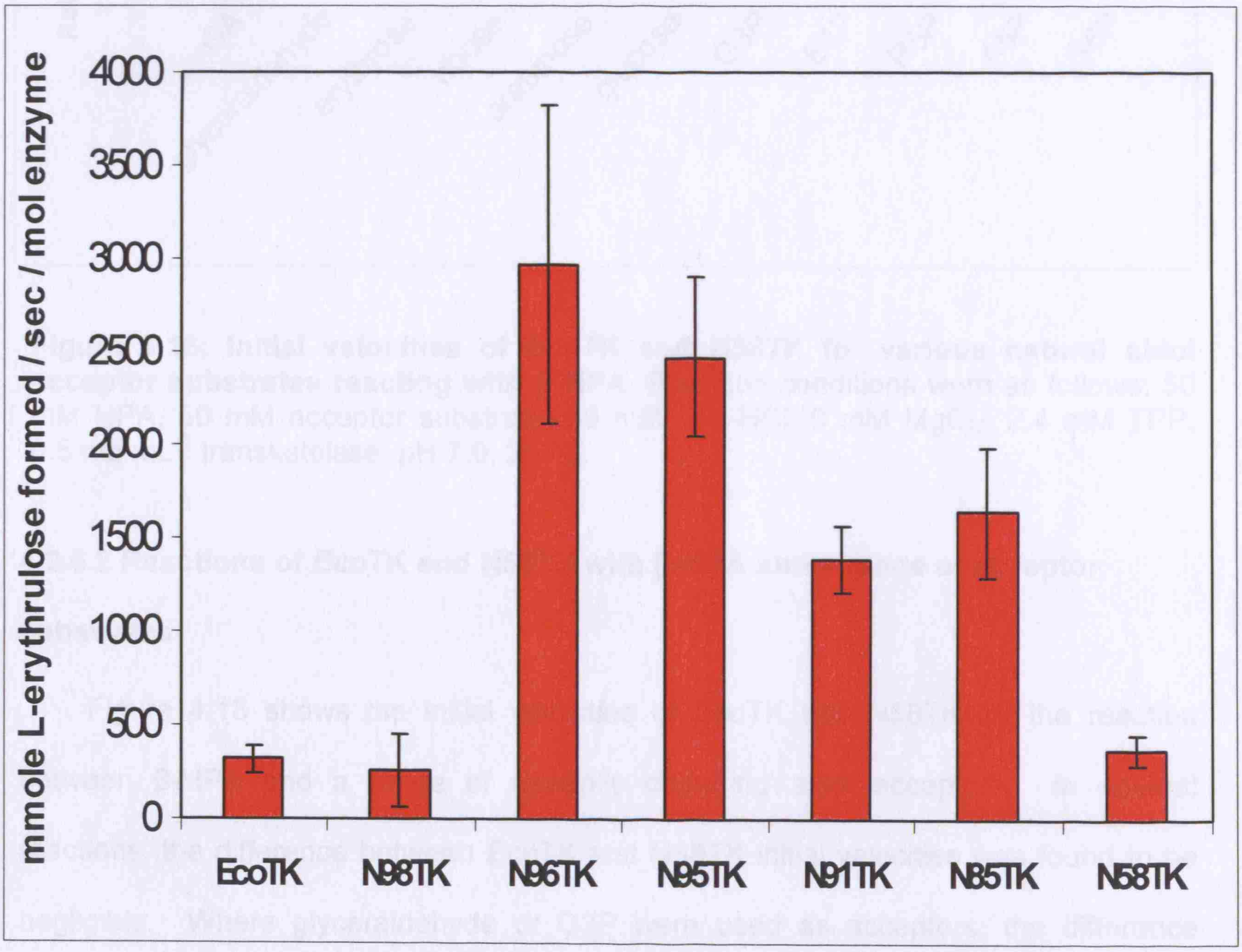


Figure 4.17: The initial velocities of *Eco*TK and ancestral TKs for the β-HPA + GA reaction. Reaction conditions were as follows: 50 mM HPA, 50 mM GA, 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, 0.5 $mg.mL^{-1}$ transketolase, pH 7.0, 25 ºC.

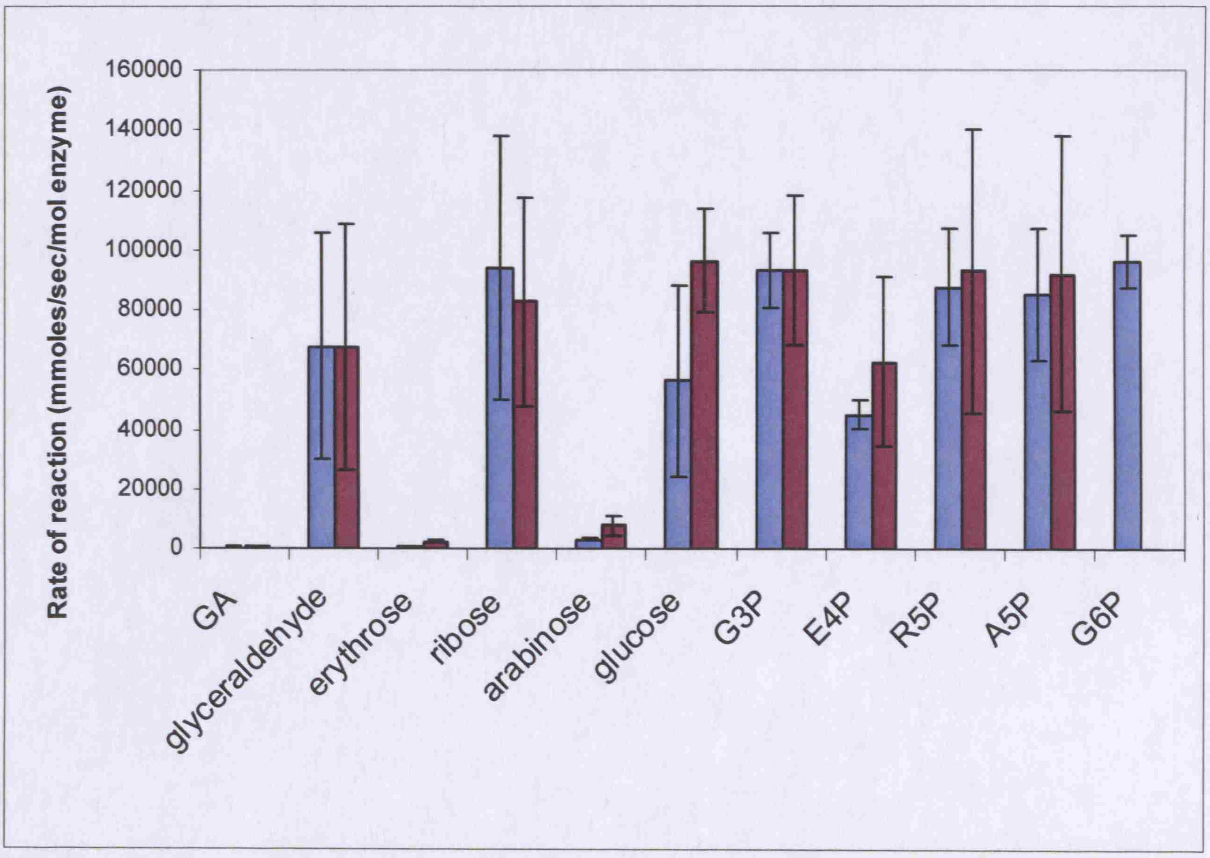


Figure 4.18: Initial velocities of *Eco*TK and N58TK for various natural aldol acceptor substrates reacting with β-HPA. Reaction conditions were as follows: 50 mM HPA, 50 mM acceptor substrate, 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, 0.5 $mg.mL^{-1}$ transketolase, pH 7.0, 25 °C.

4.3.5.2 Reactions of *Eco*TK and N58TK with β-HPA and a range of acceptor substrates

Figure 4.18 shows the initial velocities of *Eco*TK and N58TK for the reaction between β-HPA and a range of naturally occurring aldol acceptors. In several reactions, the difference between *Eco*TK and N58TK initial velocities was found to be negligible. Where glyceraldehyde or G3P were used as acceptors, the difference between *Eco*TK and N58TK activities were found to be 1 % and <1 % respectively. In the cases of R5P, A5P, arabinose and ribose, differences in activity between *Eco*TK and N58TK were 6, 8, 151 and 14 % respectively, which, taking error bars into account shows there is essentially no difference between *Eco*TK and N58TK activities.

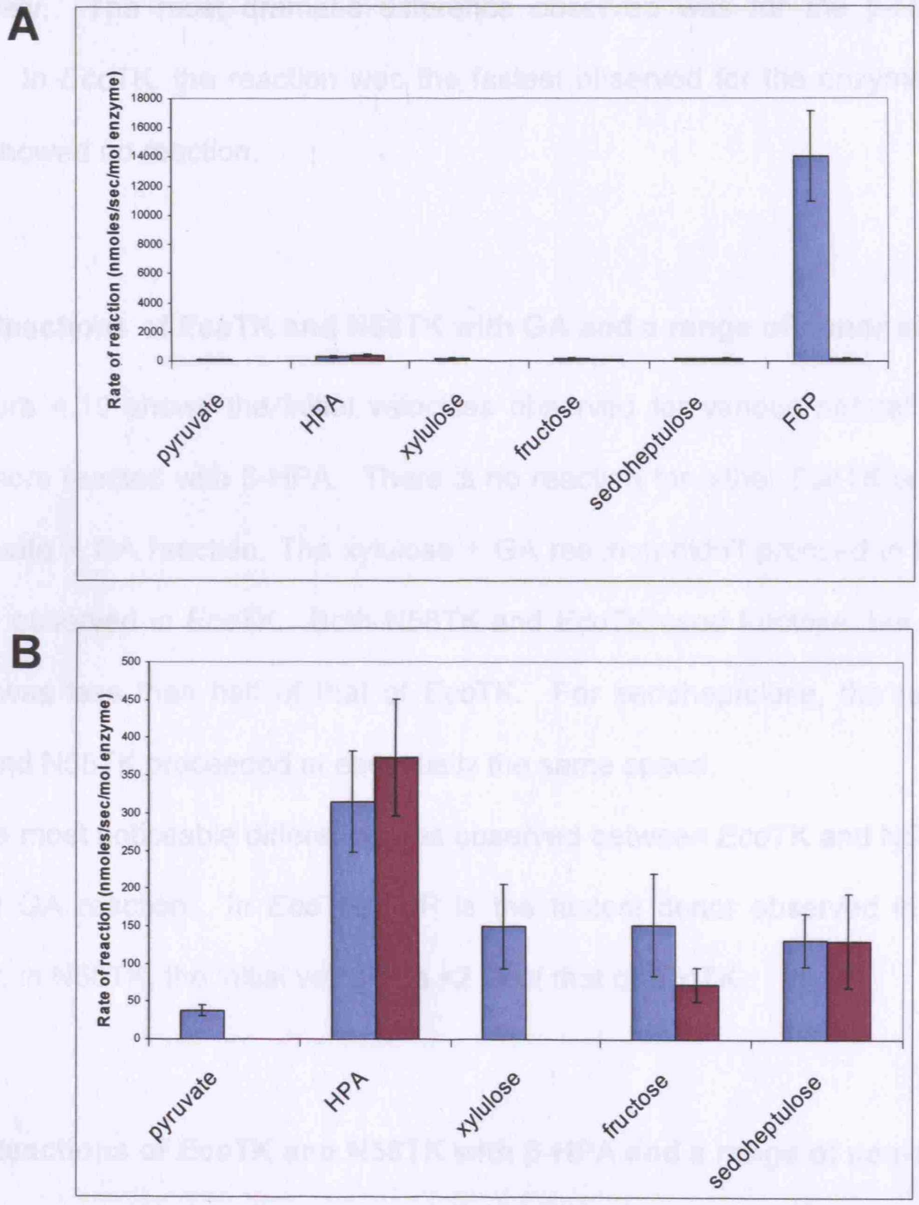


Figure 4.19: Initial velocities of *Eco*TK and N58TK for various natural ketol donor substrates reacted with GA. Reaction conditions were as follows: 50 mM GA, 50 mM donor substrate, 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, 0.5 mg.mL^{-1} TK, pH 7.0, 25 °C. F6P is by far the fastest donor observed and so in order that the remaining donor substrates can be compared graphically, the Graph A shows the F6P reaction rate, while in Graph B, F6P is removed.

Significant differences are observed in the other reactions. In the cases where E4P, glucose or erythrose were the acceptor substrates, the N58TK initial velocity was significantly higher than in the *Eco*TK reaction, being 39, 71, and 521 % faster respectively. The most dramatic difference observed was for the β-HPA + G6P reaction. In *Eco*TK, the reaction was the fastest observed for the enzyme, while the N58TK showed no reaction.

4.3.5.3 Reactions of *Eco*TK and N58TK with GA and a range of donor substrates

Figure 4.19 shows the initial velocities observed for various naturally occurring ketol donors reacted with β-HPA. There is no reaction for either *Eco*TK or N58TK for the pyruvate + GA reaction. The xylulose + GA reaction didn't proceed in N58TK, but could be observed in *Eco*TK. Both N58TK and *Eco*TK used fructose, but the N58TK velocity was less than half of that of *Eco*TK. For sedoheptulose, the reactions for *Eco*TK and N58TK proceeded at essentially the same speed.

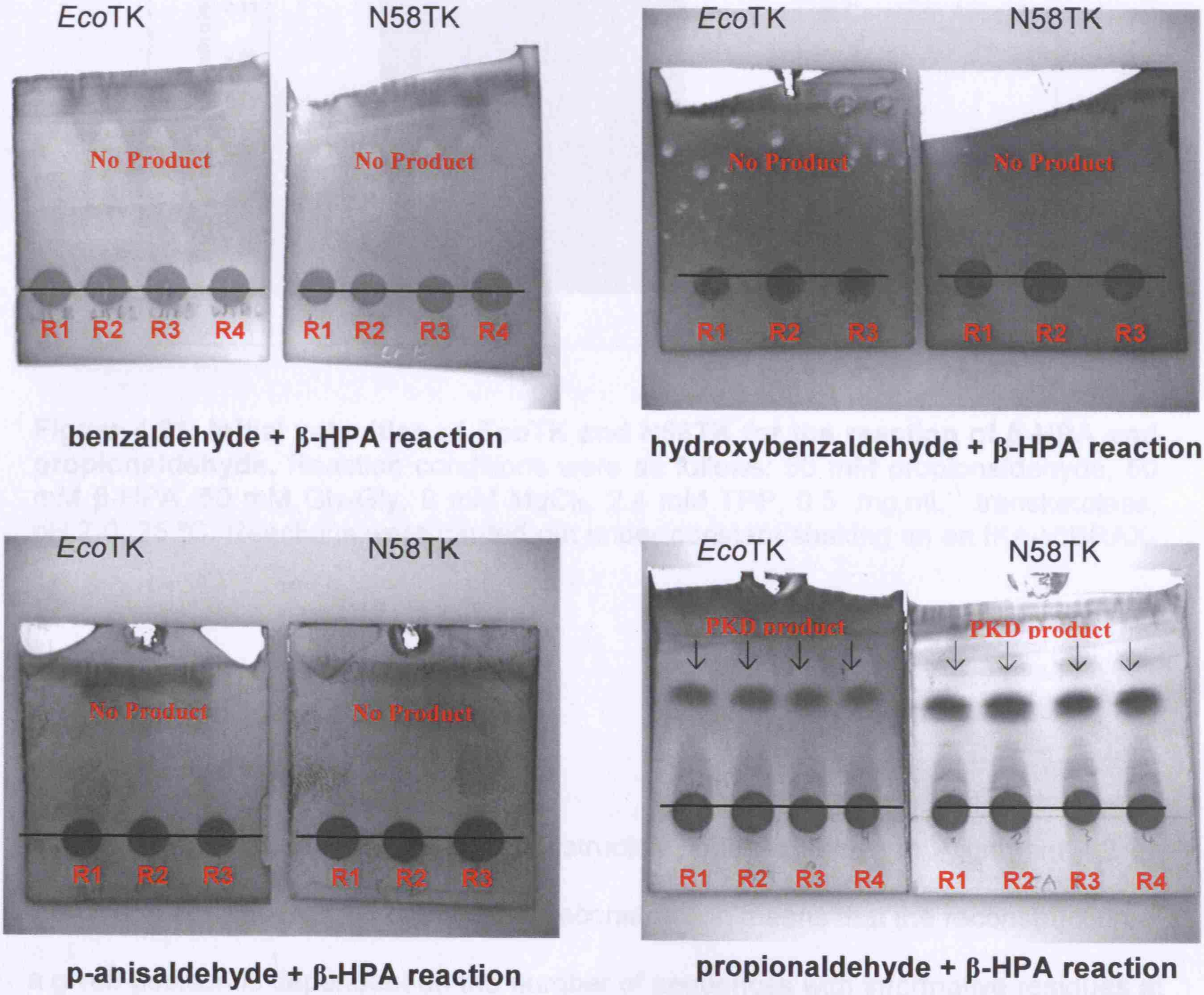
The most noticeable difference was observed between *Eco*TK and N58TK for the F6P and GA reaction. In *Eco*TK, F6P is the fastest donor observed in the study. However, in N58TK, the initial velocity is <2 % of that of *Eco*TK.

4.3.5.4 Reactions of *Eco*TK and N58TK with β-HPA and a range of non-natural acceptor substrates

Figure 4.20 shows the TLC assay to detect product in the reactions where non-natural substrates benzaldehyde, hydroxybenzaldehyde, p-anisaldehyde and propionaldehyde were used as acceptor substrates.

For benzaldehyde, hydroxybenzaldehyde and p-anisaldehyde there was no observable product after 17 hours 40 minutes for reaction catalysed by either the *Eco*TK or N58TK enzymes. In the case of propionaldehyde, product is observed for both the *Eco*TK and N58TK catalysed reactions.

To determine any difference in reaction rates between the *Eco*TK and N58TK enzymes for the propionaldehyde + β-HPA reaction, the reaction was repeated, with samples taken for HPLC analysis, as described in Section 4.2.8.3. A 9 % decrease in initial velocity is observed in the N58TK compared with the *Eco*TK. This difference is not significant, as shown in Figure 4.21.



benzaldehyde + β-HPA reaction

hydroxybenzaldehyde + β-HPA reaction

p-anisaldehyde + β-HPA reaction

propionaldehyde + β-HPA reaction

Figure 4.20: Thin layer chromatography results for non-natural acceptor substrates. Samples were run in 100% ethyl acetate, left to dry for short time before being stained with a PMA mix. The plate was then heat-dried.

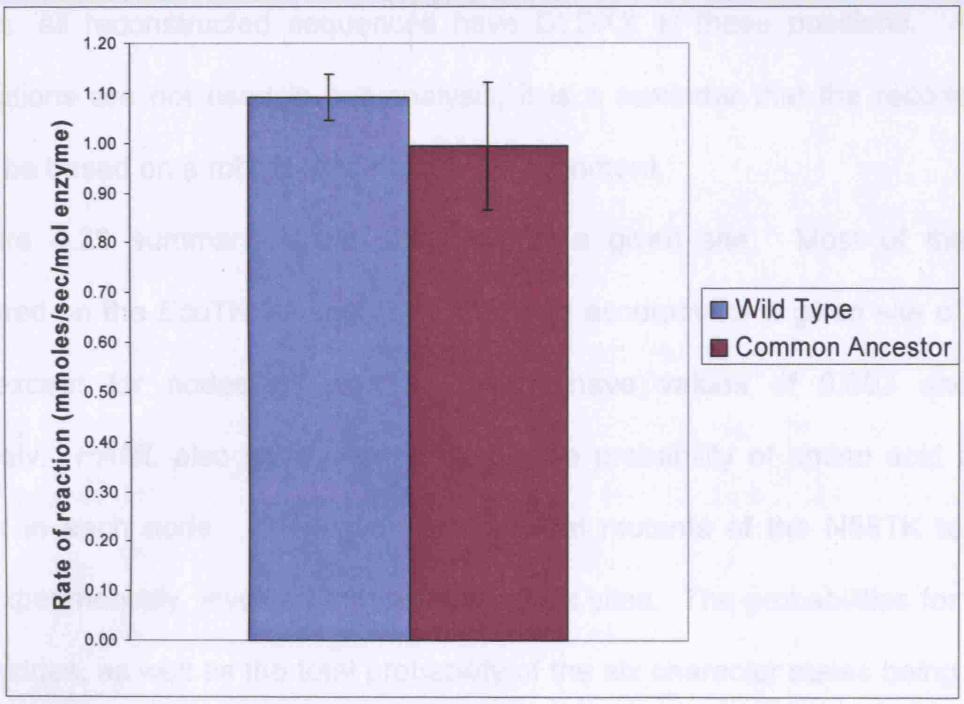


Figure 4.21: Initial velocities of *Eco*TK and N58TK for the reaction of β-HPA and propionaldehyde. Reaction conditions were as follows: 50 mM propionaldehyde, 50 mM β-HPA, 50 mM Gly-Gly, 9 mM $MgCl_2$, 2.4 mM TPP, 0.5 $mg.mL^{-1}$ transketolase, pH 7.0, 25 ºC. Reactions were carried out under constant shaking on an IKA-VIBRAX-

4.4 Discussion

4.4.1 Ancestral reconstruction

The Alignment for the *PAML* reconstructed nodes is shown in Alignment 4.2 on the CD-ROM. The method of ancestral reconstruction means that the reconstruction of a given position is dependent on the number of sequences with informative residues at a given amino acid position. In an alignment, gaps are added to ensure that the maximum number of equivalent residues align. For example, in Alignment A4.2, alignment of the *Cac*2TK sequence required 5 gaps to be introduced in all other sequences, so that the DLDMI run of amino acids (beginning at position 73 in Alignment A4.2) could be accommodated. Upon reconstruction, *PAML* uses only the information provided by the *Cac*2TK (the only information available at these positions) to generate the characters at these positions in the ancestral sequences.

Thus, all reconstructed sequences have DLDXX at these positions. Although such positions are not used in our analysis, it is a reminder that the reconstruction needs to be based on a robust and informative alignment.

Figure 4.23 summarises the accuracy for a given site. Most of the nodes encountered on the *Eco*TK lineage have a level of accuracy for a given site of 70% or higher, except for nodes 57 and 85, which have values of 0.693 and 0.694 respectively. *PAML* also gave information on the probability of amino acid being at each site in each node. Generation of ancestral mutants of the N58TK to *Eco*TK branch experimentally, involved the mutation of six sites. The probabilities for each of these residues, as well as the total probability of the six character states being present is summarised for each of the ancestral mutants to be generated in Section 4.2.3.

Mutant name	Positions						
	23	29	64	259	383	384	Probability of all states
N58TK	(N) 1	(M) 0.586	(A) 0.967	(K) 0.555	(T) 0.988	(S) 0.342	0.106
N85TK	(N) 1	(M) 0.858	(A) 0.934	(K) 0.948	(T) 0.971	(G) 0.988	0.729
N91TK	(N) 1	(M) 0.614	(N) 0.999	(D) 0.685	(T) 0.955	(G) 0.981	0.394
N95TK	(N) 0.995	(M) 0.611	(N) 1	(D) 0.712	(A) 0.9	(G) 0.964	0.376
N96TK	(N) 0.987	(A) 0.989	(N) 1	(D) 0.781	(A) 0.996	(G) 0.954	0.724
N98TK	(N) 0.59	(A) 1	(N) 1	(D) 0.999	(A) 1	(P) 0.992	0.585

Table 4.3: Accuracy of phylogenetic reconstruction at the sites mutated in this study.

Thus, the most reliable mutant is N85TK. The probability of all of the six mutating active-site residues calculated by *PAML* being correct is 0.729. The N96 mutant is similarly reliable. Least probable of all the combinations is that calculated for N58, the common ancestor, 0.106. This is in part due to the low probability of serine being at position 384, although it is the most probable character state. The mutation of Gly384 to Ser at the N85 to N58 transition is accompanied by a dramatic decrease in activity

(Figure 4.18). This should be borne in mind when analysing catalytic data for the N58 mutant.

4.4.2 Choice of active-site residues to study

Across the length of an alignment, there will be many positions at which sequences will vary. Therefore, unless the entire ancestral gene is to be constructed, a focus is needed for the study to enable practical mutagenesis. The ideas of Morley and Kazlauskas [138] and Dalby [130] suggest that residues closest to the active-site are most responsible for determining substrate specificity. The 52 amino acid residues found to be within 10 Å of the TPP molecule are summarised in Figure 4.9 as well as in Table 4.1.

4.4.3 Tracing Lineages

4.4.3.1 Resolution of mutations in different TK lineages

As described in Section 4.3.3, the lineage between any two extant species can be traced on the tree of Figure 4.10. However, since the aim of the study was to focus on the *Eco*TK lineage, the 54 TK sequences were chosen on the basis that they resolve the evolutionary mutations occurring in this lineage well (Section A1.1). However, had the objective been to resolve lineages other than the *Eco*TK lineage, it is plausible that different TK sequences would have been chosen for the alignment and subsequent phylogenetic analysis described in Chapter 3.

Section 4.3.3 describes how five different lineages on the ML tree were studied to examine whether patterns in the evolution of active-site residues could be deduced. While the resolution of the *Eco*TK lineage was optimal by design, the *Ncr*TK and *Sce*TK lineage has a similar level of resolution of mutations. Further resolution of this branch would require additional yeast TK sequences, particularly of the *Schizosaccharomycetes* and *Pezizomycotina* type, which are at present unavailable.

In the *Bsu*TK lineage, evolution occurs from node 62, with good resolution, each mutation occurring singly from the N63TK node to extant *Ncr*TK. At the N62TK to N63TK transition, three mutations occur simultaneously. This transition represents the switch between proteobacteria and other bacteria in the analysis, as well as being in the region where prokaryotic and eukaryotic ancestors meet. Since two of these mutations occur only in the N62TK to *Bsu*TK branch, it is concluded that they are important only in the most recent bacteria and not important in proteobacterial evolution.

Thus, overall, the TK tree and reconstruction yield optimal lineages and well resolved mutations. One could confidently generate the ancestral TK mutants for each of the examined lineages experimentally, and by inference, for any lineage linking bacterial or fungal sequences in Figure 4.10.

4.4.4 Enzyme assays

4.4.4.1 *Eco*TK and Ancestral TK activity for the β-HPA + GA model reaction

It is clear from Figure 4.17, there are, in general, three levels of activity observed by TK variants for the β-HPA + GA reaction. *Eco*TK, N98TK and N58TK have similar and relatively low activity (258–374 mmoles L-erythrulose.sec^{-1}.mol enzyme^{-1}). N91TK and N85TK show similar, significantly higher levels of activity (1391 - 1641 mmoles L-erythrulose.sec^{-1}.mol enzyme^{-1}), while the highest levels of activity are observed for the N95TK and N96TK mutants at 2474 and 2963 mmoles L-erythrulose.sec^{-1}.mol^{-1} enzyme respectively.

In terms of the affects of mutations, the K23N mutation on its own appears to have no observable affect. The mutation involves a decrease in side-chain length, accompanied by a change from basic to neutral and polar character.

The P384G mutation causes a dramatic affect, with activity increasing over ten-fold. The mutation involves a change from the proline ring side-chain, to the much smaller hydrogen side-chain of glycine. The proline to glycine transition represents a switch from the most constrained amino acid residue to the least constrained in terms of conformational flexibility. If one traces the direction N58TK to N85TK, which involves the S384G mutation, the increase in activity is also dramatic. Thus, in two cases it can be observed that mutants have activities dramatically increased by mutation of position 384 to glycine.

The A29M mutation has a small affect, causing a slight decrease in activity in N95TK (M29) when compared with the N96TK mutant (A29).

When the A383 of N95TK is mutated to T383 as in N91TK, there is a marked decrease in activity, although the N91TK mutant still has an activity >4 times that of *Eco*TK. It is plausible that the decrease in activity caused by the A383T mutation is mediated by the residues affect on G384 next to it.

The double mutation D259K and N64A which occur simultaneously between N91TK and N85TK have a slight affect on activity, increasing it by 17 %. It is unclear which residue is most responsible for this increase.

To summarise, the two positions which most affect activity are 383 and 384. Having glycine at position 384 increases activity dramatically compared to either proline or serine at the same position, while having threonine at position 383 causes a decrease when compared with having alanine at the same position. The functions of residues at other positions are less obvious. Positions 23, 29, 64 and 259 alter activity slightly. Such positions could be required to stabilise the active-site or change the polar or hydrophobic/hydrophilic nature of the active-site, each of which may create a context for mutations to evolve which dramatically alter substrate specificity.

A hypothetical description of the evolution of the active-site of *Eco*TK from N58TK might read as follows:

The N58TK exists in an organism that is similar to the common ancestor organism for bacteria, yeast and plants (the "progenote" [146]). Bacteria then branch away from the eukaryotes, leading to the common ancestor of the proteobacteria, similar to N58TK. This organism required a faster TK reaction, perhaps to increase the production of ribose for RNA synthesis and increase the host growth rate (discussed in Section 1.1.11, in the context of tumour proliferation). This faster TK reaction rate was primarily achieved by mutation of S384 to G, which has a smaller side-chain. Within the proteobacteria, the α-proteobacteria diverge from the common ancestor of the β- and γ-proteobacteria (this clade includes the outlier *Xff*TK and lacks the outlier *Hpy*TK). During this period two mutations occur, the K259D and A64N, causing a slight decrease in activity for the TK reaction, perhaps by the interaction of N64TK with Gly384, which oppose each other on opposite sides of the TPP molecule.

Next the γ- and β-proteobacteria lineages diverge, and the common ancestor of the γ-proteobacteria, N95TK, is arrived at. This is accompanied by an increase in activity, caused by the T383A mutation, possibly due to its affect on or with the Gly384 residue. It is also possible that the Gly384 and Asn64 residues, which oppose each other through the TPP molecule, affect each other in some way.

*Xff*TK then diverges from the N96TK, which is the ancestor of the rest of the γ-proteobacteria. N96TK has the highest activity for any TK mutant observed in the N58TK to *Eco*TK branch. The increase in activity is achieved by mutation of the Met29 to Ala. It is unclear how this is achieved, although 29 and 64 cluster close to each other near the base of the TPP binding pocket, perhaps having an affect on one another.

Next, the N98TK node is reached, which is close to the time at which *Sen*TK evolved, approximately 140 m.y.a [194]. The glycine at 384 mutates to proline, causing a ten-fold loss of activity. Next the N23K mutation yields extant *Eco*TK. This mutation has little effect on activity. Since both the 384 and 23 positions are on opposite sides of the E4P substrate molecule from each other in the *Sce*TK crystal structure, it is possible they affect one another. Perhaps Asn is required at position 23 for the Gly384 state to affect reaction kinetics (Figure 4.17). Perhaps, once the glycine evolved to proline at position 384, the Asn was no longer required at position 23 and mutated to lysine. This would be consistent with modern organisms adopting a slower rate of replication. In such a context a slower TK would be adequate for metabolism.

4.4.5.2 Reactions of *Eco*TK and N58 TK with a range of acceptor substrates

As summarised in Section 4.3.5.2, there are essentially no differences in activity between *Eco*TK and N58TK for the reactions of β-HPA with glyceraldehyde, G3P, R5P, ribose and A5P (Figure 4.18).

It is unsurprising that no significant differences between N58TK and *Eco*TK are observed for G3P or R5P, since these are physiological substrates of TK. The non-phosphorylated versions of these substrates, glyceraldehyde and ribose are used at a high rate with little difference between N58TK and *Eco*TK. The observation that A5P is used at a level comparable with the physiological substrate R5P is perhaps surprising, considering it has previously been suggested to inhibit *Eco*TK [31]. It is interesting to note that E4P is used at a significantly faster rate by N58TK than *Eco*TK, since it is a physiological substrate of TK. The G6P + β-HPA result is very surprising. The fact that no activity is observed for N58TK suggests that perhaps the G6P doesn't bind it at all, while the rate for *Eco*TK is comparable with activity for the physiological substrates even though G6P is not normally considered a natural substrate of TK.

4.4.5.3 Reactions of *Eco*TK and N58TK with a range of donor substrates

The most remarkable observation is how low the rate for the β-HPA + GA reaction is, when compared with β-HPA reactions with other donor substrates in Section 4.3.5.3 (Figure 4.19). This suggests that GA is a poor acceptor substrate in both *Eco*TK and N58TK. In the cases of most donor substrates reacted with GA, activity is very low, for xylulose, fructose and sedoheptulose. For pyruvate, the reaction rate is zero. This is unsurprising since there is no hydroxyl group at the C-2 of pyruvate (Section 5.1). However, the physiological donor substrate, F6P shows the highest reaction rates for both *Eco*TK and N58TK observed when reacted with GA. The rate in N58TK for GA + F6P, although the highest observed for the enzyme where GA is the acceptor, is still relatively low. The *Eco*TK GA + F6P reaction however, is over 50 times faster than the model β-HPA + GA reaction. This observation is discussed in Section 4.4.5.5c.

4.4.5.4 Theoretically fastest reactions for *Eco*TK and N58TK

The theoretically fastest reaction for *Eco*TK, given that F6P was found to be the fastest donor and G6P the fastest acceptor, would be F6P + G6P. Since the β-HPA + G6P and the β-HPA + G3P reactions are very similar in rate, the F6P + G3P reaction is also a theoretically fast reaction. The β-HPA + ribose reaction is also very fast, but has larger error bars than the β-HPA + G6P and the β-HPA + G3P reactions. F6P + G3P is the backwards reaction of the E4P + X5P reaction of TK in the PPP (Section 1.1) so the observation that it is so favoured by *Eco*TK is perhaps unsurprising.

It is unusual that the F6P + G6P reaction is so fast, since it is has not been previously described as a physiological reaction of *Eco*TK. An alternative version of the PPP has been previously described by Williams *et al.* [195], based on ^{14}C labelling experiments. The "L-type" PPP was found to operate in rat hepatocytes. In the L-type PPP, G6P was found to act as an acceptor substrate for TK, resulting in generation of

O8P [195-197]. Perhaps catalysis with O8P can be considered a moonlighting function of *Eco*TK, which has been present since the time when *Eco*TK and TK from rat liver shared a common ancestor. Upon the emergence of rat hepatocytes, the O8P function may have become important *in vivo*.

The theoretically fastest reactions for N58TK are also the fastest observed reactions, of β-HPA + glucose and β-HPA + G3P. The R5P + A5P reaction for N58TK is of a similar speed, but error bars are large (Figure 4.18).

4.4.5.5 Reactions of *Eco*TK and N58 with β-HPA and a range of non-natural acceptor substrate

The observation that *Eco*TK and N58TK have no activity for the non-natural substrates benzaldehyde, hydroxybenzaldehyde and p-anisaldehyde is unsurprising. Since these substrates would not have been encountered during the course of evolution of *Eco*TK and they represent a considerable change in size and hydrophobicity of substrate. In order to generate mutants capable of using such substrates, it is likely that the experimenter will need to look outside the set of mutations observed in this lineage. Since *Eco*TK can use propionaldehyde it was perhaps unsurprising that N58TK uses it also. The HPLC assay showed no appreciable difference in activity.

Work conducted by Tarik Senussi in this laboratory produced saturation mutagenesis libraries based on the mutating active-site residues suggested by the *Eco*TK table in Figure 4.12. Using NNS primers, libraries were generated for the positions 23, 29, 64, 159, 188, 259, 383, 384, 409 and 466 in *Eco*TK. Certain mutants of position 259 were shown to have activity towards benzaldehye in the Senussi study. As proposed earlier in this section, the benzaldehyde using mutants contained residues at position 259 outside the natural range of amino acids found amongst TKs at this position. These mutants are currently being examined in this laboratory. This observation underlines the influence of position 259, anticipated in Section 4.3.3, where it was observed to mutate early in the evolution of all lineages examined. In the same study mutation of 29 to either glutamate of aspartate (neither of which is the natural range in TK) produced mutants with higher activity than *wild-type* for the β-HPA + GA reaction.

This observation suggests that the variable active-site positions in TK are more flexible than suggested by their character states and conservation in extant species. Manipulation of the amino acid residues in these variable active-site positions could lead to substrate repertoires far broader than those observed within modern TK species.

4.4.5.5 Substrate specificity of EcoTK and ancestral TKs

4.4.5.5a Broadening / narrowing of substrate specificity during evolution

There are two reactions in this study which proceeded in *Eco*TK and not in N58TK, namely the G6P + β-HPA and the xylulose + GA reactions, while N58TK activity for the F6P + GA reaction is reliable compared with the *Eco*TK reaction. This could support the notion that substrate specificity has broadened in the course of evolution, although such a hypothesis is difficult to quantify. A greater repertoire of substrates needs to be tested to explore this matter further.

In the three instances where significant differences are seen between *Eco*TK and N58TK for different acceptors erythrose, glucose and E4P, the N58TK is the more active. Where different donors are tested with GA, in two instances the *Eco*TK activity is higher than N58TK, fructose + GA (2X) and F6P + GA (70X). These results suggest that for donor substrates, N58TK has slightly higher activity, while the acceptors, in general have higher reaction rates in *Eco*TK. It should be borne in mind however, that in the fourteen cases (including propionaldehyde + β-HPA) where substrates are used to some extent by both the *Eco*TK and N58TK, 9 show essentially the same activity in both enzymes.

Thus, making a definitive statement on the broadening or narrowing of substrate specificity in TK during the course of evolution is not possible on the basis of this study.

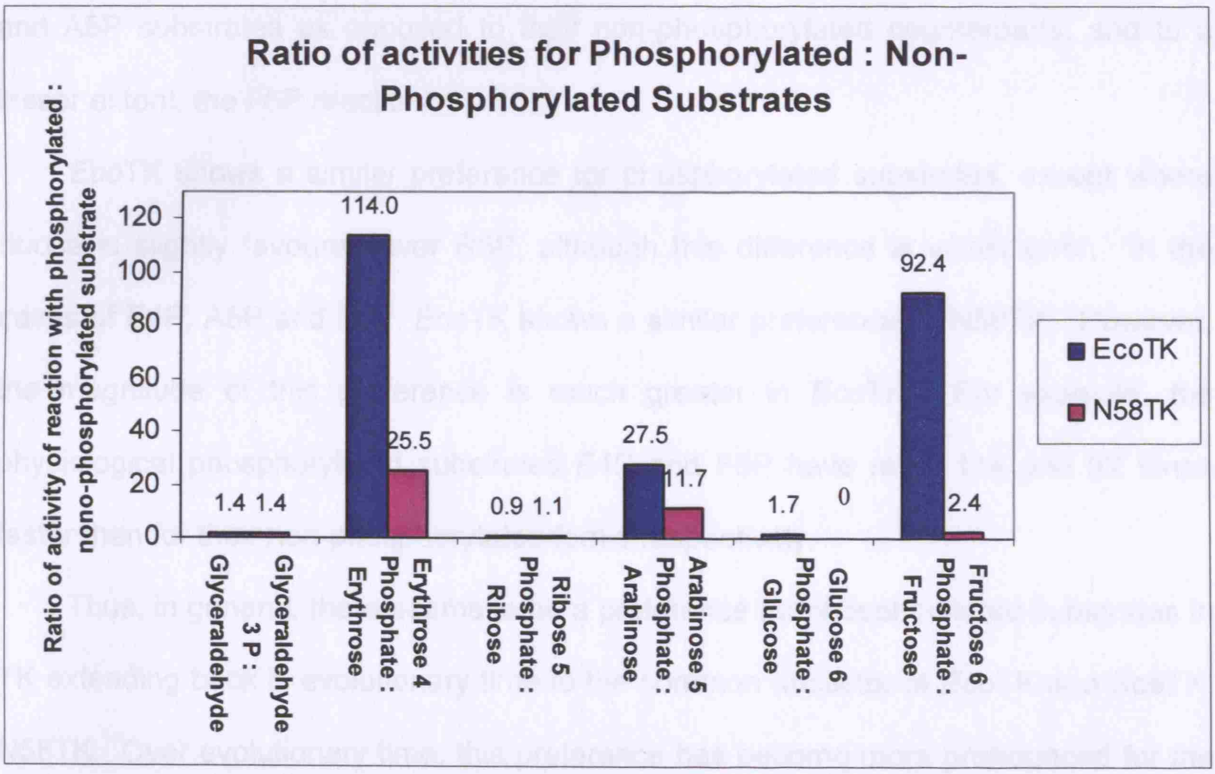


Figure 4.22: Comparison of the preferences of *Eco*TK and N58TK for phosphorylated substrates.

4.4.5.5b: Activity for phosphorylated versus non-phosphorylated substrates

There are six cases where activity for nonphosphorylated substrates can be compared with activity for the phosphorylated counterpart. These are: glyceraldehyde and G3P, erythrose and E4P, ribose and R5P, arabinose and A5P, glucose and G6P as well as fructose and F6P.

In Figure 4.22, the ratio of activity for the phosphorylated substrate to the non-phosphorylated substrate is shown for different substrate pairs for *Eco*TK and N58TK. Apart from G6P which isn't used at all by N58TK (although N58TK uses glucose). Phosphorylated substrates are favoured in general by both enzymes.

In N58TK, in every case, except G6P, the reaction with the phosphorylated substrate is higher than with the non-phosphorylated substrate, particularly for the E4P and A5P substrates as opposed to their non-phosphorylated counterparts, and to a lesser extent, the F6P reaction.

*Eco*TK shows a similar preference for phosphorylated substrates, except where ribose is slightly favoured over R5P, although this difference is within error. In the cases of E4P, A5P and F6P, *Eco*TK shows a similar preference to N58TK. However, the magnitude of this preference is much greater in *Eco*TK. For example, the physiological phosphorylated substrates E4P and F6P have rates 114 and 92 times faster than for their non-phosphorylated forms respectively.

Thus, in general, there seems to be a preference for phosphorylated substrates in TK extending back in evolutionary time to the common ancestor of *Eco*TK and *Sce*TK, N58TK. Over evolutionary time, this preference has become more pronounced for the physiological (and similar) substrates, emphasising the importance of TK for primary metabolism.

4.4.5.5c: Physiological reactions

Since TK catalyses two reactions *in vivo*, it is assumed that both are equally important to central metabolism. The physiological substrates for *Eco*TK are G3P, E4P, R5P and F6P. For G3P and R5P, reactions with β-HPA there is virtually no difference in activity between N58TK and *Eco*TK. This suggests that the R5P + X5P to G3P + S7P reaction occurs at virtually the same rate in *Eco*TK and N58TK.

In the case of the E4P + β-HPA, and F6P + GA reactions, differences are observed between *Eco*TK and N58TK. N58TK uses E4P at a slightly faster rate than *Eco*TK. However, for the F6P + GA, the difference between the *Eco*TK and N58TK is highly significant. The *Eco*TK reaction is over 70 times faster than the N58TK reaction. This could suggest that the E4P + X5P to G3P + F6P reaction is much slower in N58TK.

It may be that the R5P + X5P reaction was favoured over the E4P + X5P reaction in the N58TK and that during the course of evolution, the slower E4P + X5P reaction became equally important to central metabolism in the PPP. This would certainly explain why the reverse reaction of E4P + X5P, F6P + G3P is one of the theoretically fastest reactions proposed. The F6P + G3P TK reaction is also seen in plants, where it is the first reaction of the Calvin cycle, participating in the regeneration of ribulose 1,5-bisphosphate. Less easy to explain is the observation that F6P + G6P is theoretically the fastest reaction we can suggest for *Eco*TK. A decrease in the R5P + X5P reaction would support the idea, discussed in Section 4.4.6.1 that modern organisms may replicate slower than their ancestors and so can afford a low rate of synthesis of RNA from ribose.

4.5 Conclusions

Phylogenetic reconstruction uncovers "hidden" evolutionary mutations. Comparing *Eco*TK and *Sce*TK sequences suggests seven active-site differences. Phylogenetic reconstruction suggests that ten sites have mutated during their evolution from a common ancestor TK. One such "hidden" position is 384, shown to have a dramatic impact on TK activity in the N96TK mutant.

Focussing on the active site positions that vary during the evolution of *Eco*TK from the overall common ancestor TK reduced the number of targets for mutagenesis from fifty-two active-site residues to seven sites. These sites were shown to impact on enzyme activity as well as substrate specificity.

*Eco*TK and N58TK have similar activity for the model β-HPA + GA reaction. Similarly, when glyceraldehyde, G3P, R5P, A5P, arabinose, ribose or propionaldehyde is reacted with β-HPA , reaction rates are essentially the same in *Eco*TK and N58TK, while neither can utilise the aromatic substrates benzaldehyde, hydroxybenzaldehyde or p-anisaldehyde.

Where E4P, glucose or erythrose are reacted with β-HPA, reaction rate with N58TK is significantly faster than *Eco*TK for the same respective reaction. The xylulose + GA reaction proceeds in N58TK, but not *Eco*TK.

*Eco*TK can catalyse the G6P + β-HPA, and the xylulose + GA reactions, neither of which proceed with N58TK. *Eco*TK also exhibits a preference for the fructose + GA ,which it performs twice as fast as N58TK and the F6P + GA reaction, which proceeds at a rate >50-fold faster than in N58TK.

Both *Eco*TK and N58TK prefer phosphorylated substrates. This preference is greater in *Eco*TK, perhaps reflecting the important role TK has come to occupy during the course of evolution.

No definite pattern can be observed for either the broadening or narrowing of substrate specificity during the evolution of *Eco*TK from N58TK.

Chapter 5: Engineering a pyruvate utilizing TK enzyme

5.1 Introduction

In studies of TK, β-HPA is often used as the donor substrate, yielding CO_2 as one of the products formed (Figure 1.12) [21,199]. Thus TK reactions with β-HPA are essentially irreversible, which drives reactions to completion. From an industrial point of view, the irreversibility of such reactions is of great interest. However, the lack of readily available commercial β-HPA and its prohibitive price (> £116.g^{-1}) mean that if TK is ever to fulfil its industrial potential, a different, cheaper donor is needed. Pyruvate differs from β-HPA by the absence of a single hydroxyl group and is far cheaper (≈ £0.46.g^{-1}). However, no TK has yet been reported which can use pyruvate as a ketol donor. The aim of this study is to generate TK mutants, which will catalyse the reaction between pyruvate and GA to produce (*S*)-3,4-dihydroxybutan-2-one (Figure 5.1).

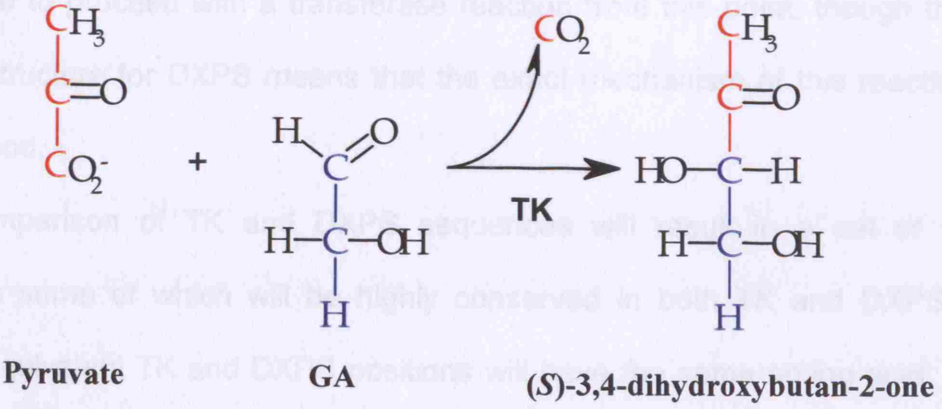


Figure 5.1: The hypothetical TK catalysed reaction of pyruvate and GA to form (*S*)-3,4-dihydroxybutan-2-one and CO_2.

The study of Meshalkina [36] suggests that the intermediate formed when TPP reacts with pyruvate, HETPP (Section 1.1.4) is not conducive to the transferase reaction. HETPP differs from the natural TK catalytic intermediate DHETPP by a single hydroxyl group (Section 1.1.4) perhaps explaining the preference of TK for substrates with a hydroxyl group at the C-1 atom. Where preformed HETPP was bound to TK, the result was the release of free aldehyde in the same manner as the PDC reaction. Thus, for TK to utilise pyruvate in a tranferase reaction, it is necessary to remove this acetaldehyde release step. It is also necessary for TK to form HETPP from pyruvate and TPP, which it cannot do.

As the hypothetical reaction between pyruvate and GA requires a mechanism not found in any TK examined, a study such as that of Chapter 4 where the variable active-site residues of TK were examined, is unlikely to yield a pyruvate utilising TK.

An alternative technique would be to compare the active-site residues of TK with the corresponding residues of D-xylulose-5-phosphate synthase (DXPS). DXPS is similar in domain architecture to TK, being a member of the TK-like group of TPP-dependent enzymes (Section 1.2.2), yet it is evolutionarily distinct (Section 3.3.5) and, importantly can catalyse a transferase reaction involving pyruvate. Thus, DXPS has, during evolution, acquired the ability to both generate HETPP from pyruvate and TPP, and to proceed with a transferase reaction from this point, though the lack of a crystal structure for DXPS means that the exact mechanism of this reaction is poorly understood.

Comparison of TK and DXPS sequences will result in a set of amino acid positions some of which will be highly conserved in both TK and DXPS. In some cases, equivalent TK and DXPS positions will have the same amino acid. Such sites are ignored since their function is likely to be the same in both enzymes. The highly conserved use of pyruvate in DXPS enzymes is likely to be mediated by conservation at the amino acid level. It is also likely that certain highly conserved residues may prevent DXPS from utilising some of the same substrates as TK. DXPS cannot, for example catalyse the pyruvate + GA reaction [200], suggesting that GA cannot bind as it can in TK.

This study will focus on equivalent residue positions that are highly conserved in DXPS, yet contain different amino acids to those found at equivalent positions in TK. Such residue positions suggest targets in the TK active-site for rational design. To select within this set of potential active-site targets and reduce the number of targets to a practical level, the analysis is refined using several approaches.

The first approach involves comparing the target residues suggested by the TK structure with equivalent residues in DXPS and PDC, as described in Chapter 3 (Section 3.3.3). PDC uses pyruvate, but has a different domain architecture and catalytic mechanism to TK and DXPS, as described in (Sections 1.2.2 and 1.1.4). Cross referencing of the highly conserved residues for TK and DXPS with the highly conserved PDC residue in the equivalent positions allows the user to ascertain which positions are highly conserved in all three enzymes, but with similar amino acid residues only in the pyruvate using enzymes PDC and DXPS, and a different amino acid residue in TK. Such residues are potentially responsible for the activities common to DXPS and PDC, but absent from TK, such as the formation of HETPP from TPP and pyruvate. Such sites would be are ideal targets for rational mutagenesis.

The second approach to narrow the number of targets for mutagenesis involves examining the residues that may interact with the hydroxyl group of the TK substrate erythrulose, recently modelled into the crystal structure of *Eco*TK in this laboratory. Since the hypothetical product of the pyruvate + GA reaction, (*S*)-3,4-dihydroxybutan-2-one differs from erythrulose by a single hydroxyl group at the C-1 atom, such residues may be important in discriminating between a TK reaction with a DHETPP or a HETPP intermediate.

Two TK mutants are generated using SDM. Along with a mutant previously generated in this laboratory, the two mutants are characterised with the model β-HPA + GA reaction as well as the hypothetical pyruvate + GA reaction.

This study demonstrates how using the analysis described here, a TK mutant is generated with activity towards pyruvate.

5.2 Methods

5.2.1 Sequence Alignments

Here an alignment of TK, DXPS and PDC sequences is compiled for study. The TK sequences used are exactly the same as those used for TK Alignment 3.1 in Chapter 3 for 54 TK sequences, except that the species *Cac2*TK and *Bme*TK have been removed, while the 89 DXPS sequences are the same as those used in the phylogenetic analysis of DXPS in Chapter 3 (Alignment 3.2). TK and DXPS sequences were aligned together prior to their inclusion in the overall PP and Pyr domain alignments of Chapter 3 (Alignments 3.19 and 3.21). Where comparable the sequences align the same in the alignment used in this chapter (Alignment 5.1) as in Alignments 3.19 and 3.21. Since the TKC domain is unlikely to affect the modes of catalysis in either TK or DXPS, the alignment was trimmed at the N and C termini so that residues aligning with *Eco*TK residue numbers 10 – 540 were chosen, a region defining the PP and Pyr domains in TK [2].

The 27 PDC sequences used in Chapter 3 were also included in Alignment 5.1, aligning exactly as in Alignments 3.19 and 3.21.

5.2.2 Analysis of highly conserved amino acid residues in TK and DXPS

TK and DXPS were analysed individually as they align in Alignment 5.1. Individual subalignments for TK and DXPS were input into the *Scorecons* server (see Section 2.2 for URL), where the "vadar01" parameter [201] was used. *Scorecons* then gave a score of the level of conservation at each position in the alignments.

Scorecons data for the TK and DXPS alignments were converted in *Excel*, to give the percentage occurrence of each character state at each amino acid position. Using the same rationale as in Chapter 4, the study focussed on the active-site (Section 1.6). The same 52 amino-acid positions found within 10 Å of the TPP moiety (Section 4.3.2) were picked from the *Scorecons* data and tabulated in *Excel.* In this way, the positions equivalent to the 52 active-site residues selected in the TK structure could be compared in TK and DXPS.

It was decided that highly conserved amino acid positions within an alignment would be defined by *Scorecons* returning a score of 0.8 or higher and would suggest that the residue is performing an important functional role in the enzyme(s).

TK active-site positions lacking a corresponding position in DXPS were removed. Positions found to have the same amino acid residue highly conserved in both TK and DXPS were removed, since these residues probably perform the same function in both enzymes. In such instances, amino acids with similar characteristics, for example Ile and Leu, or Lys and Arg, were considered to be equivalent. Residues not highly conserved in the DXPS active-site were eliminated from the analysis, since they are unlikely to be responsible for pyruvate usage. Only highly conserved DXPS residues found to be different from the corresponding residue in TK were considered.

Highly conserved DXPS residues found to occur in specific TK species, known to lack function towards pyruvate are removed, since they are unlikely to confer pyruvate activity to TK.

5.2.3 Choosing targets by comparison with PDC

The candidate sites for rational design that remain after the initial round of elimination described in Section 5.2.2 were compared with the corresponding PDC residues, where structurally equivalent (Section 3.3.3). The region described in Alignment 3.20 is included in this analysis. This region was excluded form the phylogenetic analysis, but is homologous for the TK and PDC structures. Positions found to be the same in TK and PDC but different from DXPS were examined in the crystal structure to determine if they are likely to perform the same function. Where a position is occupied by a highly conserved amino acid residue in DXPS and PDC, which is highly conserved as a different residue in TK, the relative position in the *Eco*TK and *Sce*PDC structures are examined to determine their structural equivalence.

5.2.4 Choosing targets by proximity to the hydroxyl group of the ketose substrate

A recent study was performed by John Strafford at this laboratory where erythrulose was modelled into the active-site of *Eco*TK (1QGD.pdb). Using this model, the candidate sites remaining after Section 5.2.3 are examined to see if any are proximal to the hydroxyl group of erythrulose.

5.2.5 Primer designs for D183N and S385G

The D381A mutant had already been generated by Jean Aucamp.

Primers were designed using *AnnHyb* and ordered from Qiagen Ltd., salt free at 50 nmol scale. Primers designed to generate the D183NTK and S385GTK mutants (Abbreviations section for nomenclature), with mutagenic codons in bold were as follows:

D183N F: GATTGCATTCTAC**TAA**GACAACGGTATTTCTATCG

D183N R: CGATAGAAATACCGTTGTC**TTA**GTAGAATGCAATC

S385G F: GCTGACCTGGCGCCG**GGT**AACCTGACCC

S385G R: GGGTCAGGTT**ACC**CGGCGCCAGGTCAGC

5.2.6 Site Directed Mutagenesis, transformation and storage of mutants

The protocol for generating mutants D183NTK and S385GTK is the same as Protocol 2 (Section 4.2.5.2). Parental DNA was digested as described in Section 4.2.6, followed by gel visualisation and transformation into XL1 Blue competent cells, as detailed in Section 2.5. Miniprepped mutant DNA was sequenced at the Wolfson Institute. Mutant plasmids were transformed into XL10 Gold cells, using the protocol in Section 2.5.9. Glycerol stocks were made with 40 % sterile glycerol and stored at -80 °C.

5.2.7 Enzyme Reactions

β-HPA + GA and pyruvate + GA reactions were performed as described in Sections 2.5.11.2 and 2.5.11.3 and reactions were monitored by HPLC as described in Sections 2.5.12.3a and 2.5.12.3b.

5.2.8 Visualisation of enzyme crystal structures

In all cases, enzymes were visualised in *Pymol.* For TK the following crystal structure were used (*E. coli* TK, 1QGD.pdb; *E. coli* TK with erythrulose modelled into the active-site, 1QGD_TK_2_Ery.pdb and the *S.cerevisiae* TK, 1GPU.pdb), while for PDC the *S.cerevisiae* structure was used (1PVD.pdb). Where a mutant enzyme is presented, the mutagenesis is modelled in *Deepview* without energy minimisation and viewed in *Pymol*.

5.3 Results

5.3.1 Sequence Alignments

The full final alignment of 52 TK, 89 DXPS and 27 PDC sequences is presented in the Phylip 4 format in Alignment 5.1 - CD-ROM. A simplified version of the alignment, containing five representative sequences from each of the enzyme types, TK, DXPS and PDC is presented in Alignment 5.2.

5.3.2 Analysis of highly conserved amino acid residues in TK, DXPS and PDC

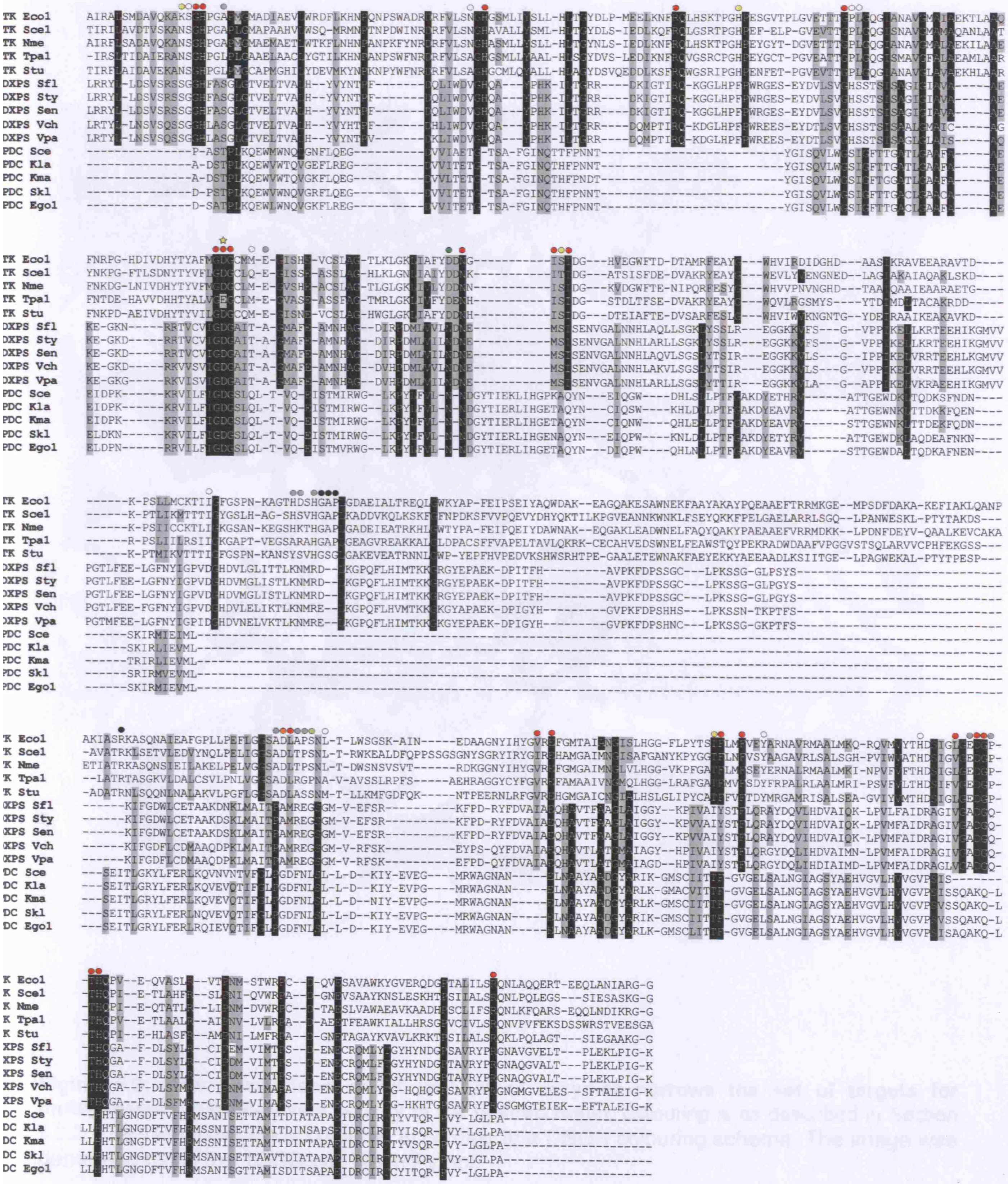
Figure 5.2 summarises how beginning with the 52 TK active-site residues, the methods described in Sections 5.2.2 and 5.2.3 enabled the selection of two residues for rational design. Table 5.1 summarises the *Scorecons* score for all of the 52 active-site residues considered for mutagenesis. In this section, reference to colouring applies to the markers in Alignment 5.2, residue colouring in Figure 5.2 and colouring of cells in Table 5.1, where applicable. Of the 52 TK active-site residues referenced in Alignment 5.1, four failed to align with equivalent residues in DXPS (coloured black). Of the remaining 48 residues, 21 were highly conserved as the same residues in both TK and DXPS. As described in Section 5.2.2, these residues are excluded from the study (coloured red). Residues responsible for the pyruvate utilising activity of DXPS will be highly conserved in DXPS. Ten positions, where the *Scorecons* score was less than 0.8 were removed from the analysis (coloured grey), leaving fifteen candidate sites. At this stage TK residues Lys23 and Ser188 were also removed from the analysis.

Position	TK	DXPS	PDC
23	0.613	0.802 (G)	N A.
24	0.857 (S)	1.000 (G)	N A.
25	1.000 (G)	1.000 (G)	0.430
26	1.000 (H)	1.000 (H)	N A.
29	0.514	0.614	0.433
64	0.657	0.982 (V)	0.826 (T)
66	1.000 (H)	1.000 (H)	0.003
91	0.970 (R)	0.975 (R)	N A.
100	1.000 (H)	0.939 (F)	N A.
114	1.000 (G)	0.825 (G)	0.869 (G)
115	1.000 (P)	1.000 (H)	0.881 (S)
116	1.000 (L)	0.831 (S)	1.000 (I)
154	0.922 (G)	1.000 (G)	1.000 (G)
155	0.983 (D)	1.000 (D)	1.000 (D)
156	1.000 (G)	1.000 (G)	1.000 (G)
159	0.637	0.808 (T)	0.938 (Q)
160	1.000 (E)	0.724	N A.
183	1.000 (D)	1.000 (N)	0.825 (N)
185	1.000 (N)	1.000 (N)	0.942 (N)
187	0.888 (I)	0.847 (M)	0.202
188	0.641	0.817 (S)	0.218
189	0.754	0.860 (I)	0.483
247	1.000 (I)	0.951 (D)	N A.
258	0.434	0.545	N A.
259	0.445	0.505	N A.
261	1.000 (H)	0.708	N A.
262	0.904 (G)	0.017	N A.
263	0.662	0.000	N A.
264	0.750	0.006	N A.
358	1.000 (R)	N. A.	0.466
380	1.000 (A)	0.698	0.841(P)
381	1.000 (D)	1.000 (A)	0.923 (G)
382	0.960 (L)	1.000 (M)	1.000 (D)
383	0.612	0.475	0.843 (F)
384	0.457	0.520	1.000 (N)
385	0.974 (S)	1.000 (G)	0.931 (L)
387	0.426	0.810 (G)	0.683
409	0.849 (V)	1.000 (I)	N A.
411	1.000 (E)	1.000 (E)	1.000 (E)
433	0.971 (T)	0.929 (I)	1.000 (T)
434	1.000 (F)	1.000 (Y)	0.942 (F)
437	1.000 (F)	1.000 (F)	1.000 (V)
440	1.000 (Y)	1.000 (R)	0.914 (L)
461	1.000 (H)	1.000 (D)	0.418
466	0.706	0.983 (V)	0.445
468	0.920 (E)	0.608	0.468
469	1.000 (D)	1.000 (D)	0.439
470	1.000 (G)	1.000 (G)	0.435
471	0.973 (P)	0.548	N A.
472	1.000 (T)	1.000 (T)	0.826 (L)
473	1.000 (H)	1.000 (H)	1.000 (H)
520	1.000 (R)	0.979 (R)	0.631

Table 5.1: Summary of the *Scorecons* scores for TK, DXPS and PDC. Positions are referenced according to their *Eco*TK numbering. Scores are given for the corresponding positions in TK, DXPS and PDC. Where a position shows a high level of conservation (≥ 80%) the most common amino acid at that position is shown in brackets. Colouring of cells is as described in Section 5.3.2. Where an enzyme was found to have no corresponding amino acid at a given position, the cell is labelled N.A.

TK positions aligning with *Eco*TK Lys23 are variable as discussed in Chapter 4, while the equivalent position in DXPS is highly conserved.

The majority of the DXPS species examined have Gly at this position, with Ser occurring also, albeit at a much lower frequency. Both Ser and Gly are found in species of TK at this position, suggesting that the site does not mediate the pyruvate activity of DXPS. *Eco*TK residue Ser188 is found at a position not highly conserved in TK, but highly conserved as Ser in DXPS. Since it is known that *Eco*TK cannot catalyse a reaction involving pyruvate, this Ser is unlikely to be conserved in DXPS for such a reason and so is removed from the analysis. Both Lys23 and Ser188 are coloured in yellow.



Alignment 5.2: Simplified version of Alignment 5.1. Five representative species from the TK, DXPS and PDC sequences are shown. Positions are marked with coloured circles as described in the text. The star symbol indicates residues discussed in relation to *Tpa*TK (Section 5.4.3.2).

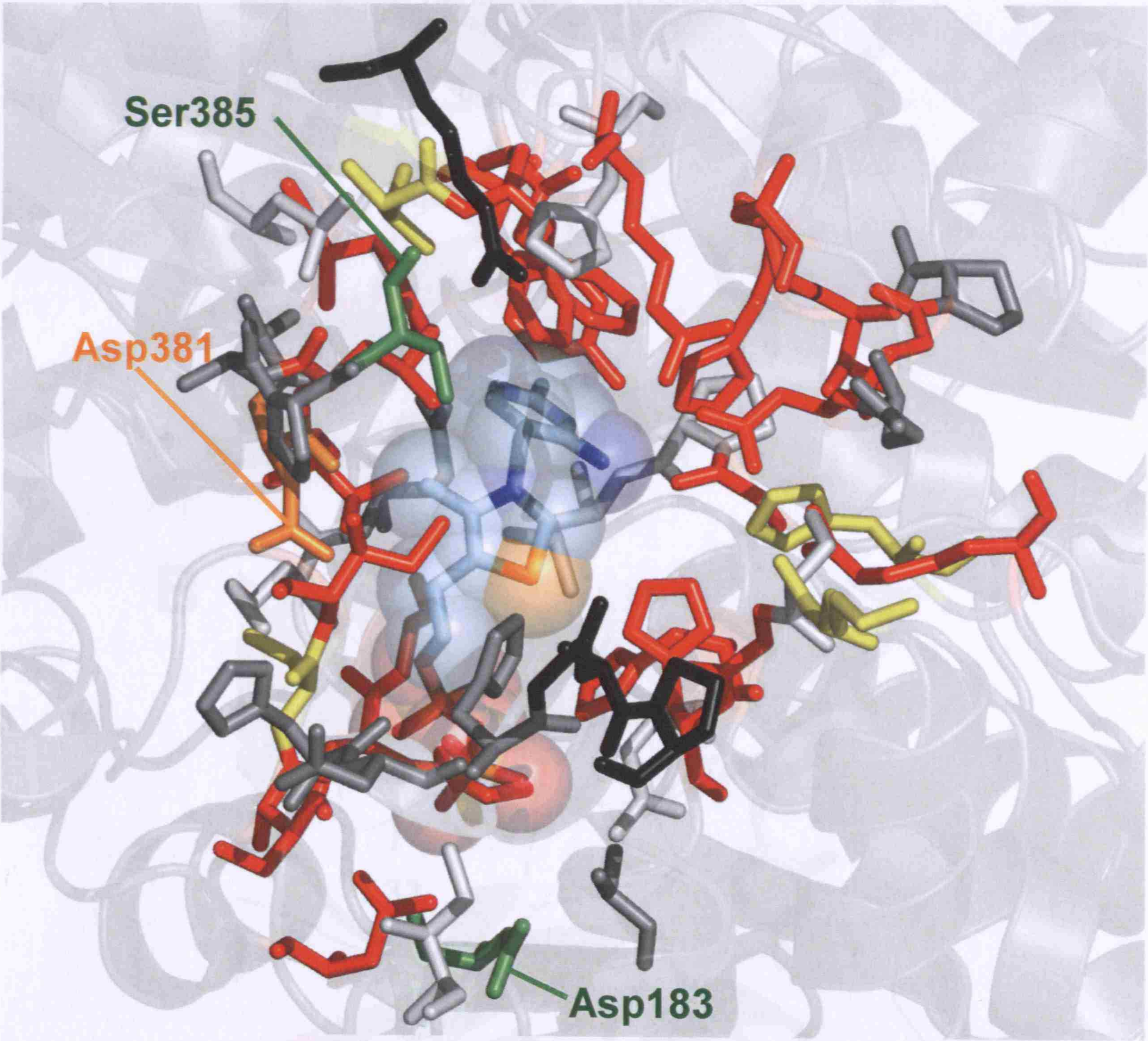


Figure 5.2: How comparison of TK with other enzymes narrows the set of targets for mutagenesis, where pyruvate activity is the desired result. Colouring is as described in Section 5.3.2. The TPP molecule is coloured according to a CNOS colouring scheme. The image was generated using the 1QGD.pdb file in *Pymol.*

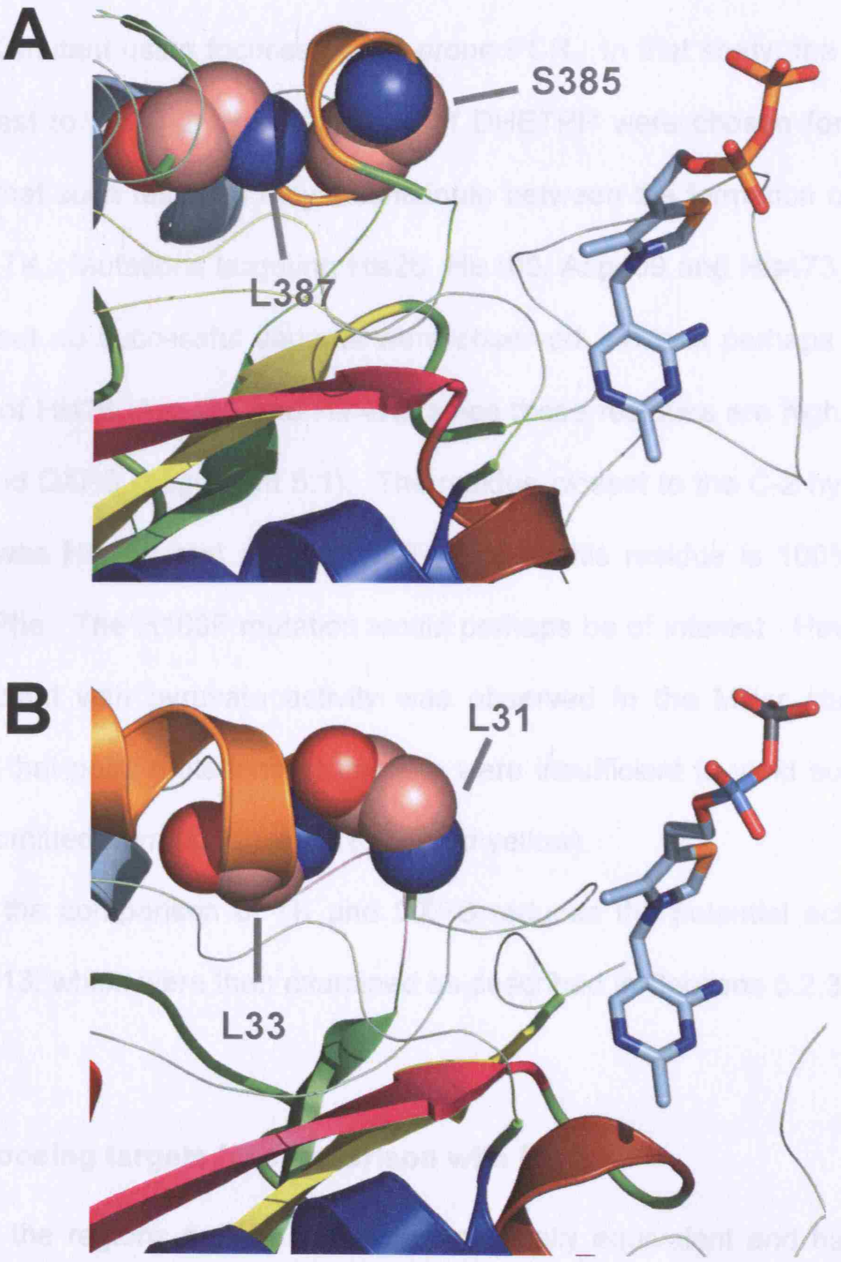


Figure 5.3: Structural comparison of TK residues 385 and 387 with the corresponding PDC positions in Alignment 5.1. Image A shows the TK structure (1QGD.pdb) with residues Leu387 and Ser385 as spheres. Image B highlights as sphere residues Leu31 and Leu33 in the PDC structure (1PVD.pdb), which align with TK positions 385 and 387 respectively. Equivalent regions of secondary structure are coloured as in Figure 3.5 of Chapter 3. Nonequivalent secondary structure regions are coloured light yellow for β-sheets and light green for α-helices. Spheres are coloured according to the scheme CHNOS, while the TPP molecule is shown with a CNOS colouring scheme. Images were generated in *Pymol.*

A previous study by Oliver Miller in this department attempted to generate a pyruvate utilising TK mutant using focussed error-prone PCR. In that study, the residues found to be closest to the C-2 hydroxyl group of DHETPP were chosen for study. It was reasoned that such residues may discriminate between the formation of DHETPP and HETPP in TK. Mutations targeting His26, His100, Asp469 and His473 in *Eco*TK were produced but no successful variants were observed. This is perhaps unsurprising in the cases of His26, Asp469 and His473, since these residues are highly conserved in both TK and DXPS (Alignment 5.1). The residue closest to the C-2 hydroxyl group of DHETPP was His100 and as seen in Table 5.1, this residue is 100% conserved in DXPS as Phe. The H100F mutation would perhaps be of interest. However, since no His100 mutant with pyruvate activity was observed in the Miller study and it was suggested that point mutations at this site were insufficient to yield such activity, this residue is omitted from our analysis (coloured yellow).

Thus the comparison of TK and DXPS reduces the potential active-site targets from 52 to 13, which were then examined as described in Sections 5.2.3.

5.3.2.1 Choosing targets by comparison with PDC

Using the regions defined as being structurally equivalent and having sequence homology in Section 3.3.3 (including the extra region discussed in Section 5.2.1), six of the thirteen remaining target sites from Section 5.3.2 align reliably with sections of the PDC sequences. Although residues corresponding with *Eco*TK Ser385 and Leu387 lie in regions that were not considered equivalent for the purposes of the phylogenetic study of TPP-dependent enzymes in Chapter 3, they are considered here for the reasons discussed below:

TK residues Ser385 and Leu387 are found on, or immediately adjacent to, an equivalent helix (the orange helix in Figure 5.3). This helix varies in length among the

TPP-dependent enzymes, so it was difficult to infer structural equivalence across all of the TPP-dependent enzyme types examined in Chapter 3 (Section 1.2.2). Thus the region was omitted from the phylogenetic analysis. However, as can be seen in Figure 5.3 Ser385 and Leu387 are structurally equivalent to *Sce*PDC residues Leu31 and Leu33, with which they align, respectively, in Alignment 5.1. Thus, eight candidate residues are compared with the equivalent positions in PDC, namely those positions aligning with *Eco*TK residues Pro115, Leu116, Met159, Asp183, Ser385, Leu387, Thr433 and His461. In the cases of Leu116 and Thr433 the residue conserved in TK is very similar to the residue conserved in PDC. For Leu116, the Leu residue is 100 % conserved in TK, while the very similar Ile is 100% conserved in all PDC species examined. In DXPS, the position is highly conserved as Ser. For the position corresponding to TK Thr433, thr is highly conserved in TK and PDC, but conserved as Ser in DXPS. In the cases of Leu116 and Thr433, it could be argued that since the residues are conserved in both TK and PDC, they may be responsible for the common functions of TK and PDC, such as the release of free acetaldehyde where HETPP is the reaction intermediate (Section 5.1). Since we wish to remove this function, perhaps these residues are useful targets. To investigate further, the positions of Leu116 and Thr433 and their corresponding PDC residues were examined in the *Eco*TK and *Sce*PDC crystal structures. As seen in Figure 5.4, Thr433 is immediately adjacent to Phe434, which with the neighbouring Phe437 forms part of the hydrophobic pocket within which the MT ring of TPP binds (Section 1.1.3). In PDC the residue that aligns with *Eco*TK Thr433, Thr73 is also found next to a Phe residue, Phe74, although a phenylalanine corresponding with *Eco*TK Phe437 is missing.

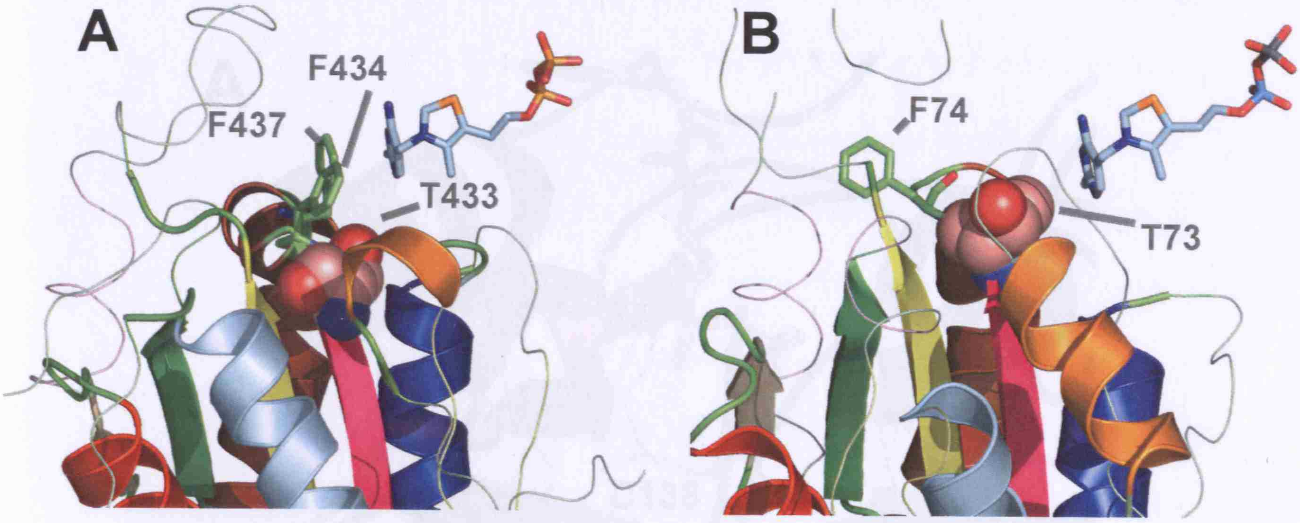


Figure 5.4: Structural comparison of *Eco*TK residue thr433 with the corresponding PDC positions in Alignment 5.1. Image A shows the TK structure (1QGD.pdb) with residue Thr433 shown as spheres. As discussed in the text, Thr433 is next to the important residue Phe434 and close to Phe437 (highlighted as sticks). Image B highlights as spheres residue Thr73 from PDC (1PVD.pdb) which aligns with *Eco*TK Thr433, as well as Phe74, which aligns with EcoTK Phe434 as discussed. Equivalent regions of secondary structure are coloured as in Figure 3.6 of Chapter 3. Nonequivalent secondary structure regions are coloured light yellow for β-sheets and light green for α-helices. Spheres are coloured according to the scheme CNOS, sticks according to CNOS, while the TPP molecule is shown with a CNOS colouring scheme. Images were generated in *Pymol*.

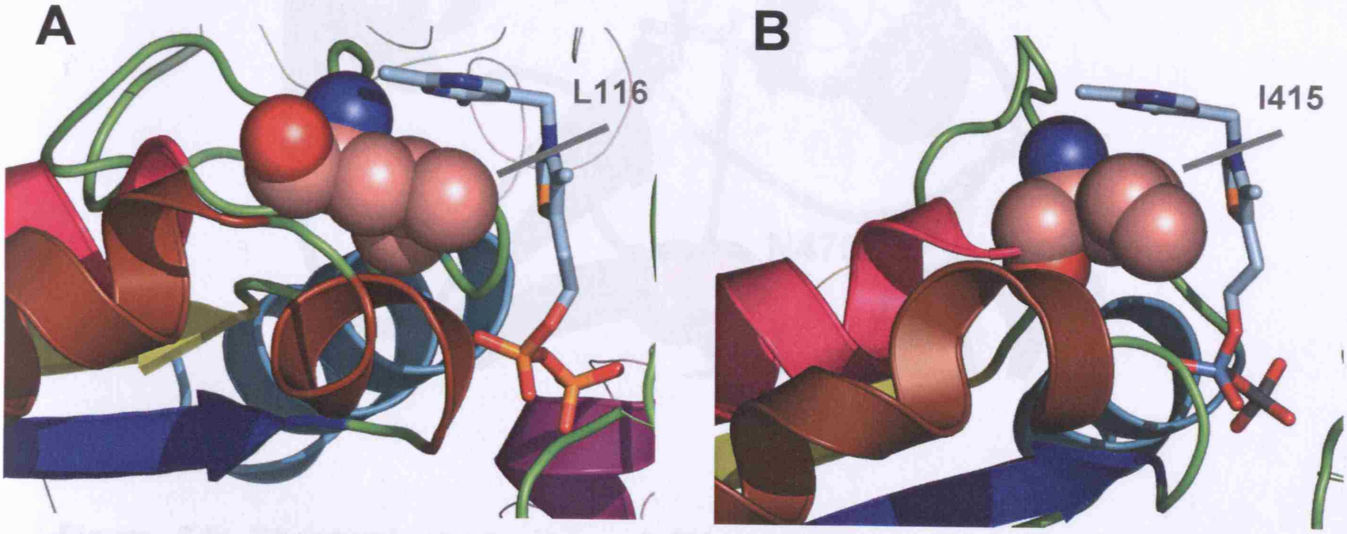


Figure 5.5: Structural comparison of *Eco*TK residue L116 with the corresponding PDC positions in Alignment 5.1. Image A shows the TK structure (1QGD.pdb) with residue Leu116 shown as spheres. Image B highlights as spheres residue Ile415 from PDC (1PVD.pdb) which aligns with *Eco*TK Leu116, Equivalent regions of secondary structure are coloured as in Figure 3.3 of Chapter 3. Spheres are coloured according to the scheme CNOS, while the TPP molecule is shown with a CNOS colouring scheme. Images were generated in *Pymol*

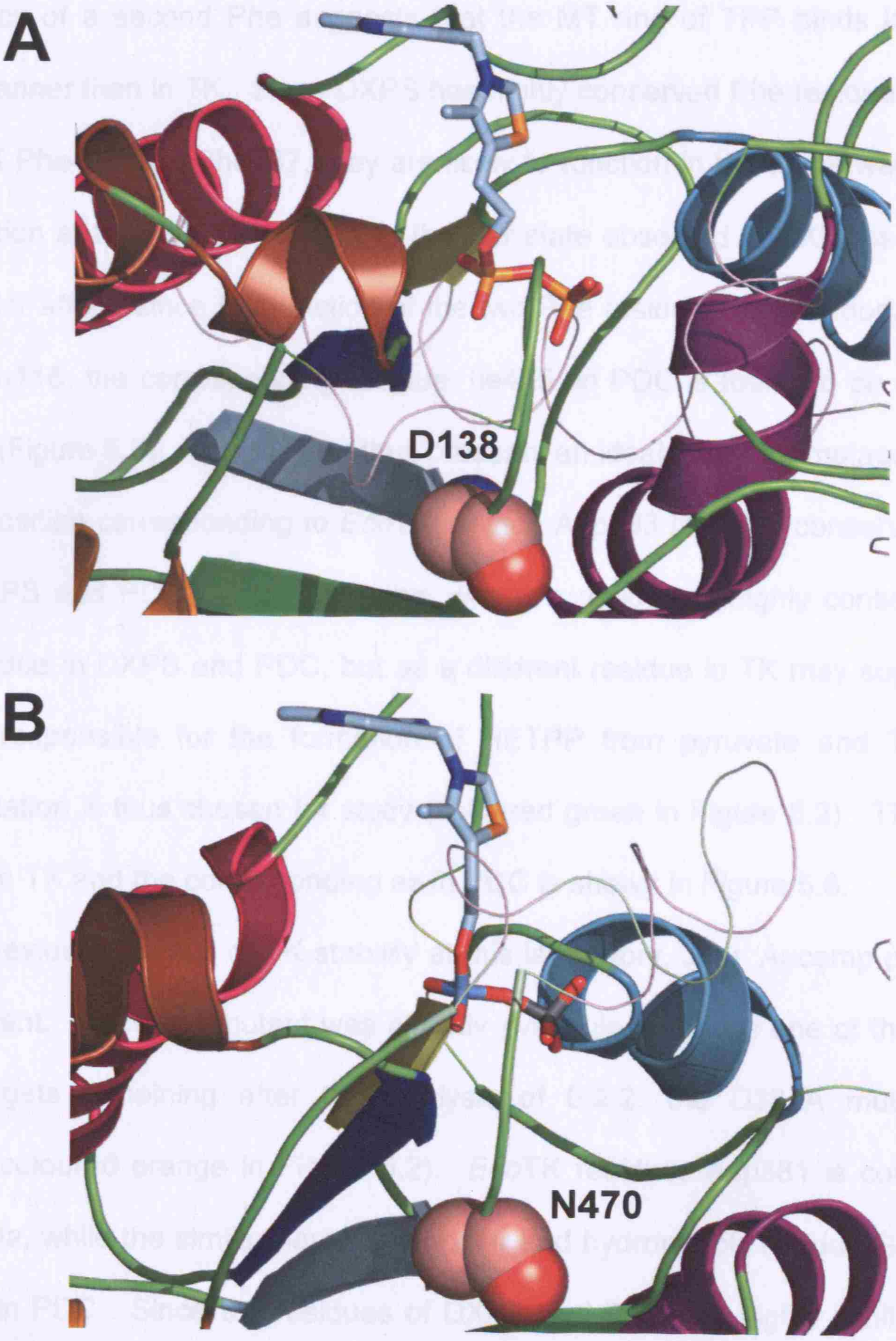


Figure 5.6: Structural comparison of TK residue Asp183 with PDC residue Asn470. Image A shows Asp138 in TK (1QGD.pdb), while image B shows the corresponding position in PDC (1PVD.pdb). Equivalent regions of secondary structure are as backbone spheres. Nonequivalent secondary structure regions are coloured light yellow for β-sheets and light green for α-helices. Spheres are coloured according to the scheme CNOS, while the TPP molecule is shown with a CNOS colouring scheme. Images were generated in *Pymol*

As seen in Figure 5.4, Phe74 faces away from the TPP moiety in PDC. This, along with the lack of a second Phe suggests that the MT ring of TPP binds in a slightly different manner than in TK. Since DXPS has highly conserved Phe residues that align with *Eco*TK Phe434 and Phe437, they are likely to function in the same way as in TK. Thus mutation at the Thr433 position to the Ser state observed in DXPS is unlikely to have a major affect, since the function of the two Phe residues is so important. In the case of Leu116, the corresponding residue, Ile415, in PDC is found to be structurally equivalent (Figure 5.5). This residue thus presents an ideal target for mutagenesis.

The position corresponding to *Eco*TK residue Asp183 is highly conserved as Asp in both DXPS and PDC. Such positions, where a residue is highly conserved as a certain residue in DXPS and PDC, but as a different residue in TK may suggest sites potentially responsible for the formation of HETPP from pyruvate and TPP. The D183N mutation is thus chosen for study (coloured green in Figure 5.2). The position of Asp183 in TK and the corresponding as in PDC is shown in Figure 5.6.

In a previous analysis of TK stability at this laboratory, Jean Aucamp produced a D381A mutant. Since the mutant was already available, and was one of the potential thirteen targets remaining after the analysis of 5.2.2, the D381A mutation was examined (coloured orange in Figure 5.2). *Eco*TK residues Asp381 is conserved in DXPS as Ala, while the similar neutral, nonpolar and hydrophobic residue Gly is found conserved in PDC. Since the residues of DXPS and PDC are highly similar to each other and different from the residue found in TK, this site may be interesting for study.

Finally the *Eco*TK polar residue Ser385 was found to align with nonpolar residues, Gly in DXPS and either Leu or Met in PDC. Thus the S385G mutation was chosen for study (coloured green in Figure 5.2).

The L116S mutation would be interesting to study, but the mutant is not examined in this study. Here, the study focuses on positions where similar residues are found to be highly conserved in DXPS and PDC as a certain character state different to that found at the corresponding position in TK. Thus the point mutations D183N and S385G were generated in TK and examined along with already available D381A in the subsequent study.

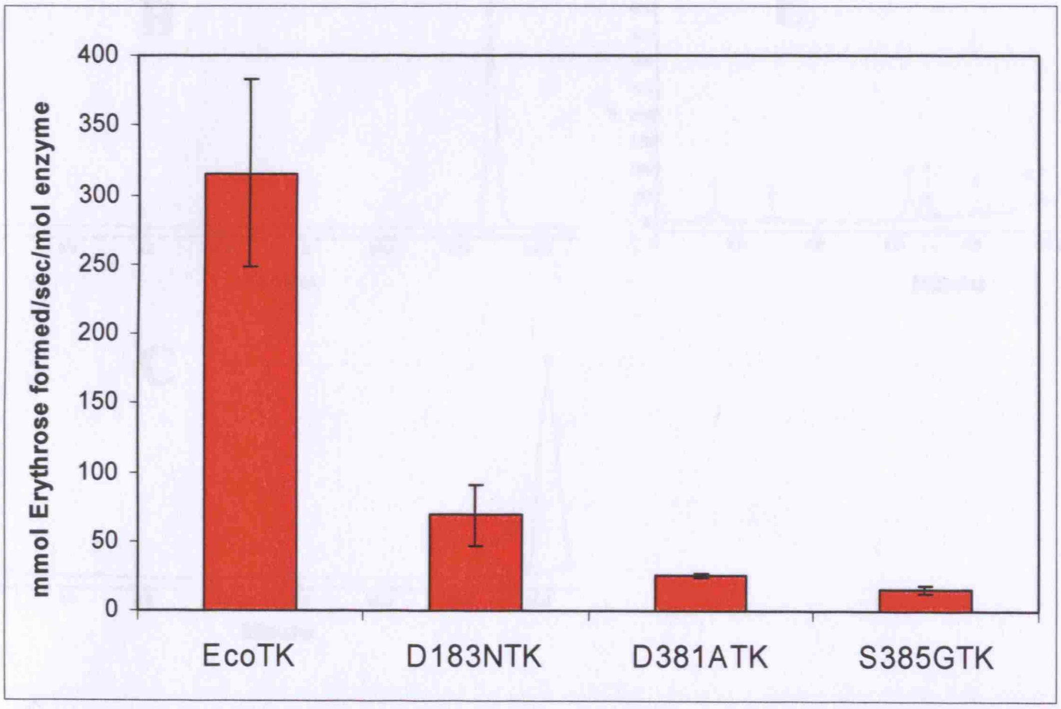


Figure 5.7: Initial velocities for the β-HPA + GA reaction with *Eco*TK and mutant TKs. Reaction conditions were as follows: 50 mM GA, 50 mM β-HPA, 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, 0.5 mg.mL^{-1} enzyme, pH 7.0, 25 ºC.

5.3.2.1 Choosing target by proximity to the C-2 hydroxyl group of erythrulose

Examination of the crystal structure of *Eco*TK with erythrulose modelled into the active-site shows clearly that the residue most proximal to the C-2 hydroxyl group of the erythrulose substrate is Ser385. This hydroxyl group may be a discriminating factor for pyruvate usage. Thus, Ser385, which was already suggested by the analysis of Section 5.2.3 is an interesting target for mutagenesis.

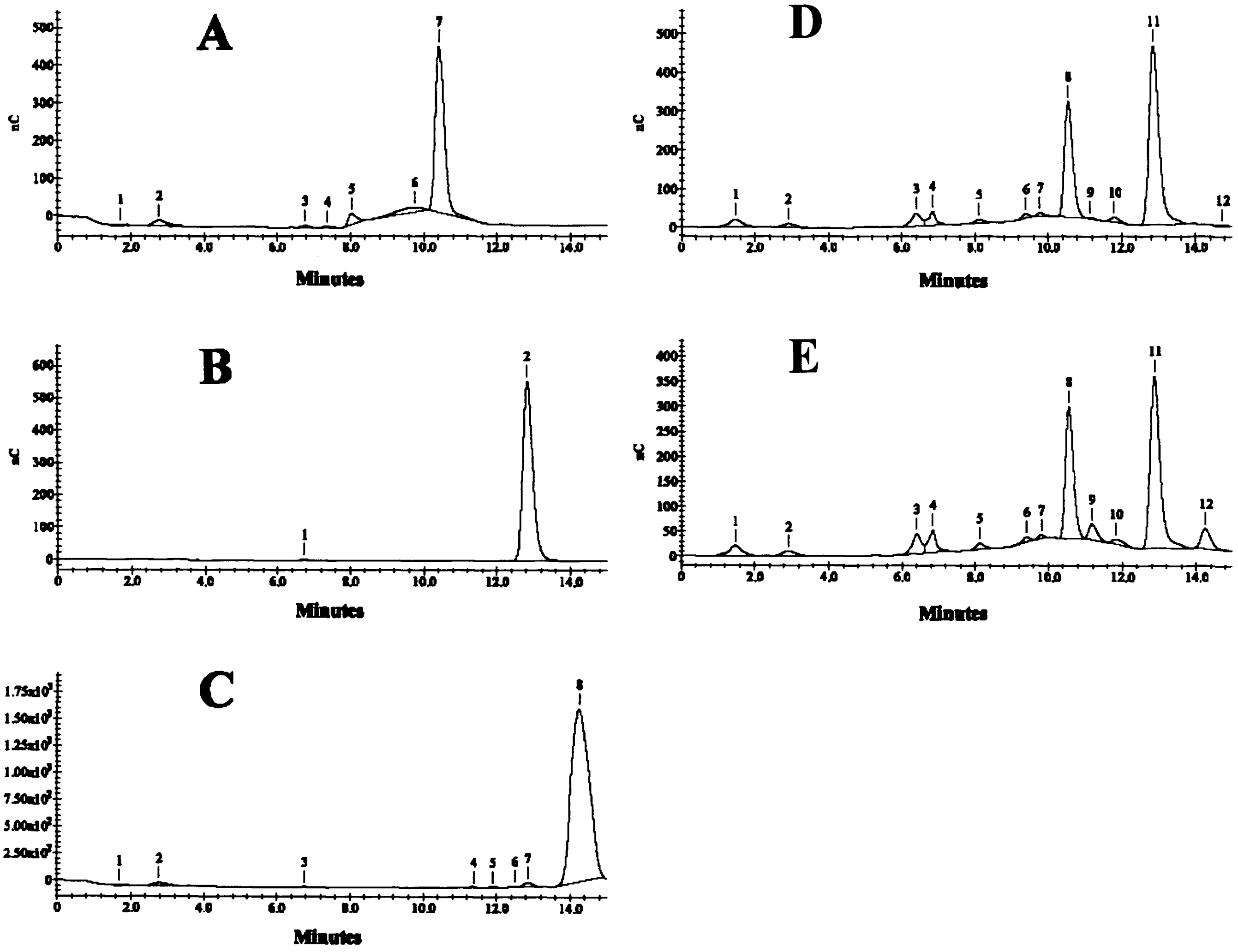


Figure 5.8: HPLC traces for the product standards and assay of pyruvate + GA for the S385GTK mutant. Figure A shows the product standard for Pyruvate. Figures B and C show the product standards for GA and (*S*)-3,4-dihydroxybutan-2-one respectively. Figure D shows the reaction profile of pyruvate + GA with the S385GTK mutant at time 5 minutes, for one of the replicates. Figure E shows the reaction profile for the same reaction at time 24 hours (1440 minutes). Reactions were conducted as described in section 5.2.4.2.

5.3.3 Enzyme Reactions

5.3.3.1 The β-HPA + GA reactions

As seen in Figure 5.7, each of the 3 TK mutants show significantly lower rates of L-erythrulose production than *Eco*TK. The D183N, D381A and S385G mutations caused 80.1 %, 92 % and 95 % decreases in activity relative to *Eco*TK respectively.

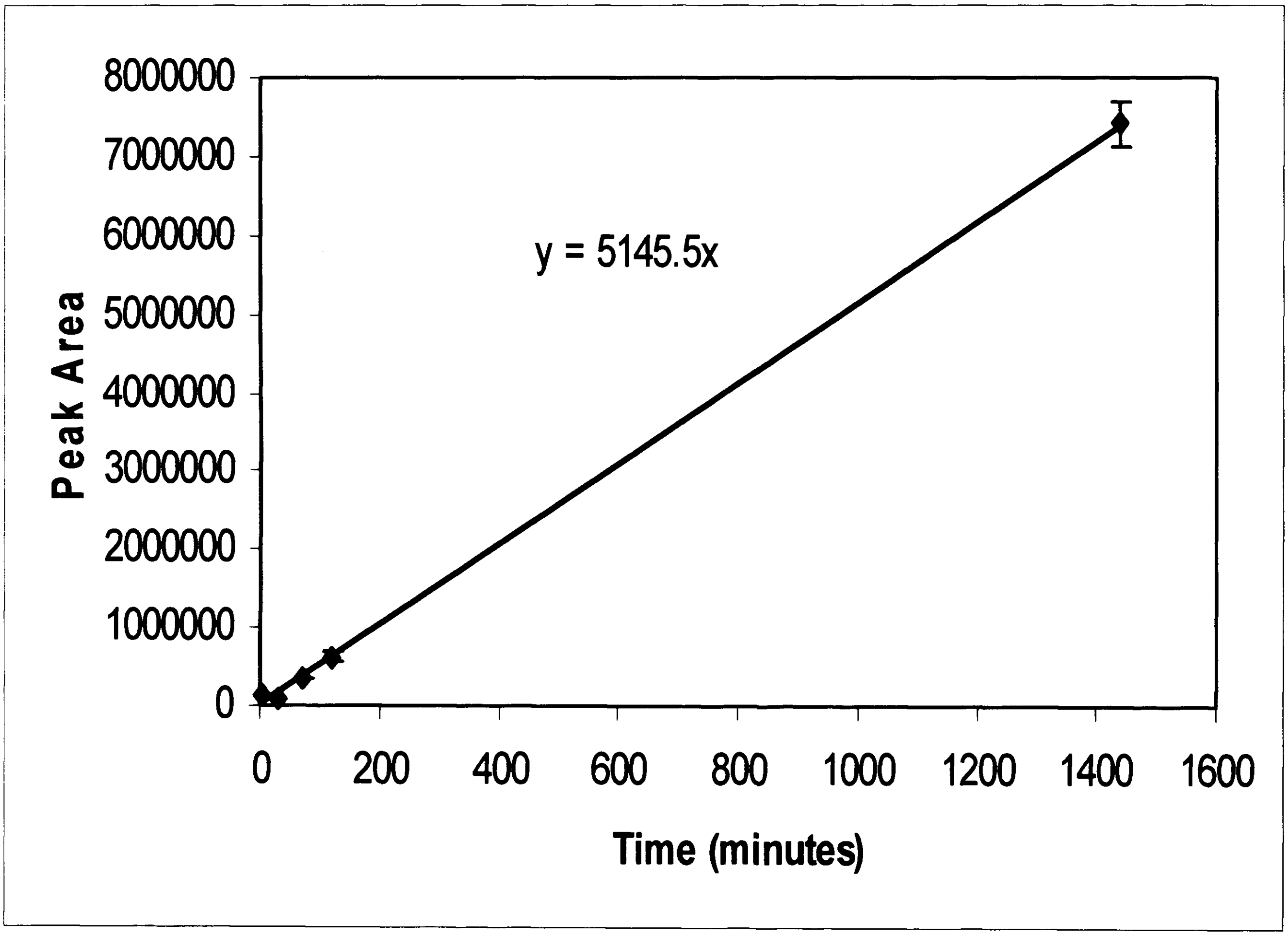


Figure 5.9: The appearance of (*S*)-3,4-dihydrohybutan-2-one for the S385GTK catalysed reaction of pyruvate with GA. (*S*)-3,4-dihydrohybutan-2-one was provided by Mark Smith of the Chemistry Department at UCL. However, only a small quantity was available for standards. Thus, no calibration curve has been generated as of yet. Hence, the use of peak area in this graph.

5.3.3.2 Pyruvate + GA reactions

Over the course of the 24-hour reaction (*S*)-3,4-dihydroxybutan-2-one production was not observed for *Eco*TK, D183NTK or D381ATK. In the case of the S385G mutant, (*S*)-3,4-dihydroxybutan-2-one product was observed. Figure 5.8 shows the HPLC traces for the substrate standards and the t_{5mins} and t_{24hr} time points (Images DE and E, respectively, in Figure 5.8).

As can be seen, peak number 12 (as designated by the *Peaknet* software) increases over time and corresponds exactly with the retention time of (*S*)-3,4-dihydroxybutan-2-one, the standard for which is shown in C in Figure 5.8. The appearance of (*S*)-3,4-dihydroxybutan-2-one was monitored at time points 5, 30, 71, 120 and 1440 minutes. This data is shown in Figure 5.9. As can be seen, the reaction profile is still in the linear range after 24 hours.

Figures 5.10 to 5.12 show, respectively, the differences in the active-site of TK made by mutation of Asp183 to asparagine, Asp381 to alanine and Ser385 to glycine.

In Figure 5.10, the difference between aspartate and asparagine can be seen. Both amino acids are hydrophilic. However, the aspartate is acidic, while the asparagine is neutral and polar. The mutation is located close to the pyrophosphate region of the TPP molecule. The 183 position in TK is found at the end of a β-strand region (residues 178-184). The mutation of aspartate to asparagine does not disrupt the polar contacts made with Met153, Asn185 or Thr245.

In Figure 5.11, the difference between aspartate and alanine at the 381 position is shown. The acidic and hydrophilic aspartate residue is replaced by the smaller, neutral and hydrophobic alanine. This results in a loss of polar contacts with Ser188, Ile189 and Asp190.

As seen in Figure 5.12, the 385 position is at the surface of the TK molecule, close to the TPP binding pocket. The difference between having Ser and Gly at this position is illustrated. In the S385G mutants, the hydrophilic neutral and polar serine is replaced with the smaller, hydrophobic, neutral and nonpolar glycine. The mutation disrupts none of the polar contacts with other residues (Arg358, Leu382 and Pro384).

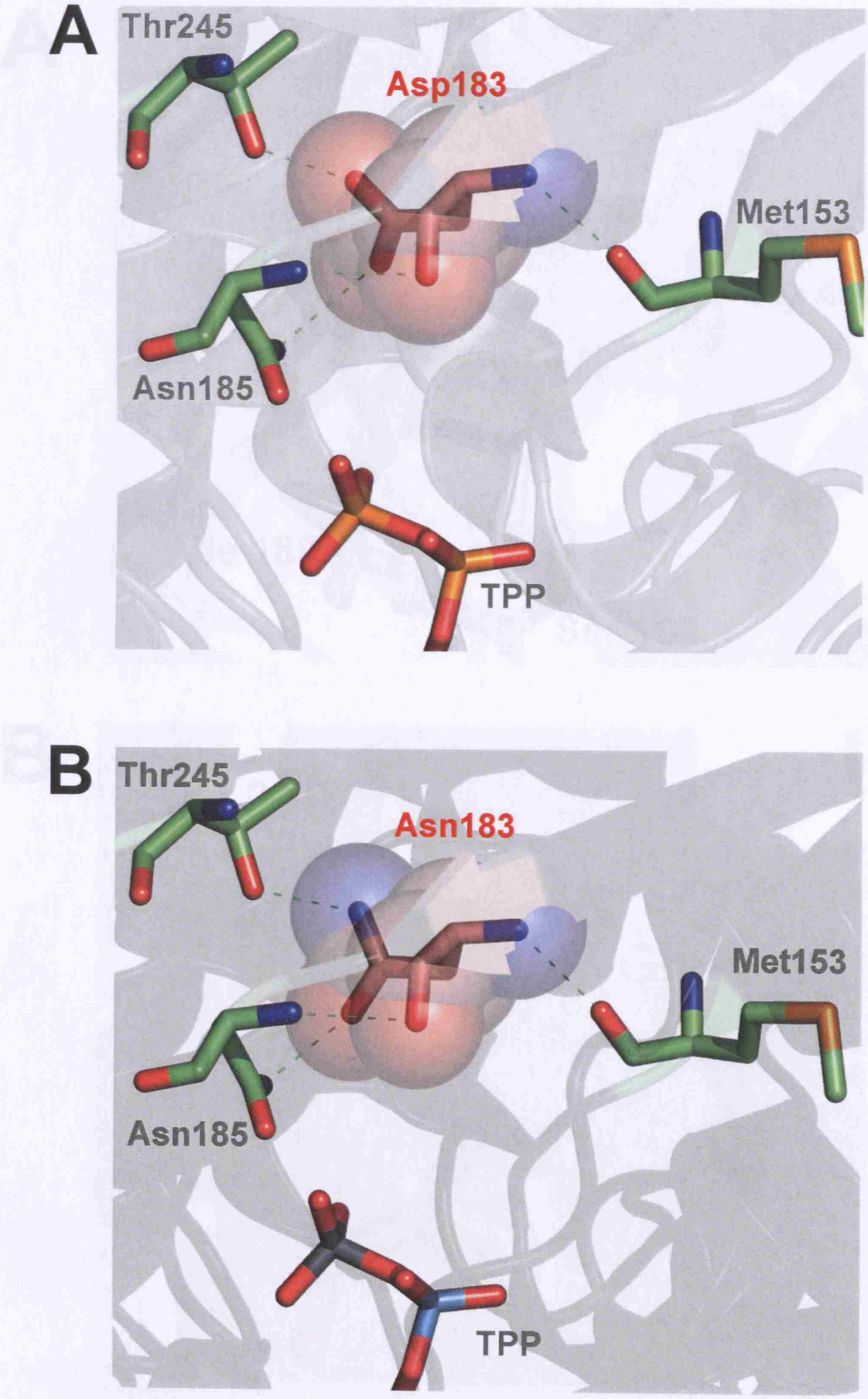


Figure 5.10: Structural view of position 183 in (A) *Eco*TK and (B) the D183NTK mutant. In both cases, the 183 position is coloured according to the CHNOS scheme, while the Met153, Asn185 and Thr245 residues which form polar contacts with the 183 residue adopt the CHNOS scheme. Images were generated using the 1QGD.pdb file in *Pymol.*

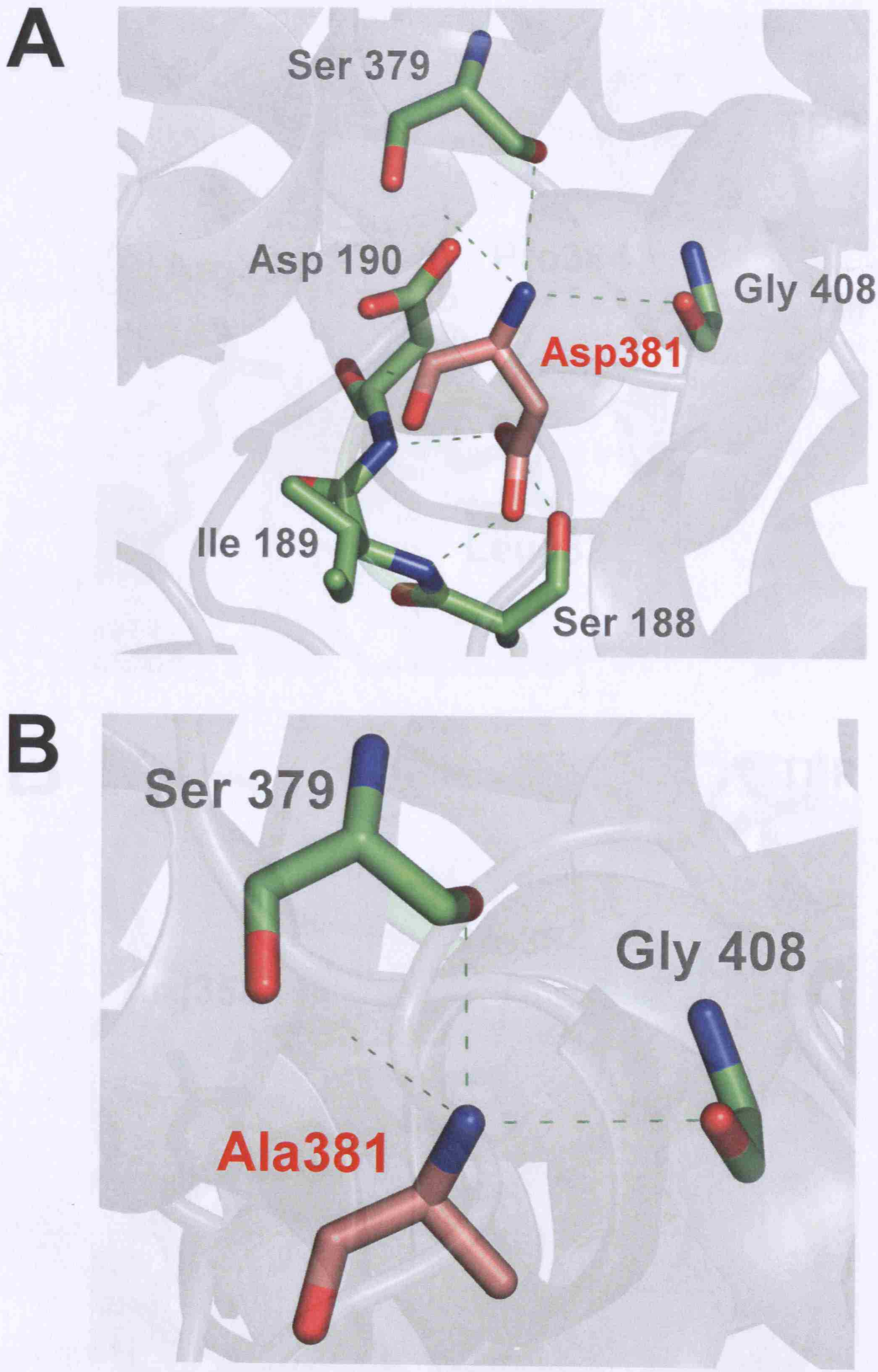


Figure 5.11: Structural view of position 381 in (A) *Eco*TK and (B) the D381ATK mutant. In both cases, the 381 position is coloured according to the CHNOS scheme, while the Ser188, Ile189, Asp190, Ser379 and Gly408 residues adopt the CHNOS scheme. Images were generated using the 1QGD.pdb file in *Pymol*.

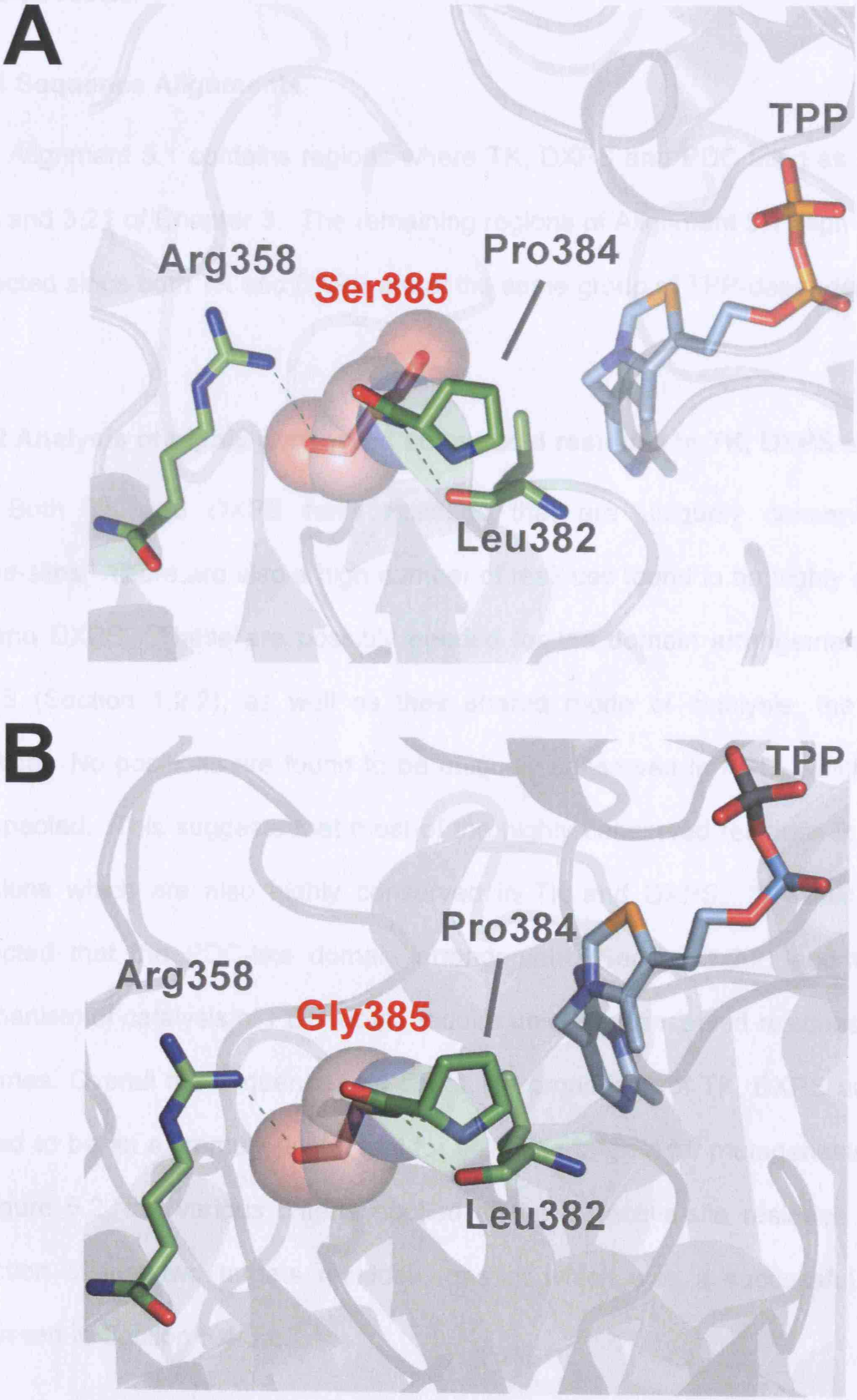


Figure 5.12: Structural view of the 385 position in *Eco*TK and the S385GTK mutant. In both cases, the 385 position is coloured according to the CHNOS scheme, while the Arg358, Leu382 and Pro384, residues adopt the CHNOS scheme. Images were generated using the 1QGD.pdb file in *Pymol.*

5.4 Discussion

5.4.1 Sequence Alignments

Alignment 5.1 contains regions where TK, DXPS and PDC align as in Alignment 3.19 and 3.21 of Chapter 3. The remaining regions of Alignment 5.1 align very well, as expected since both TK and DXPS are of the same group of TPP-dependent enzymes.

5.4.2 Analysis of highly conserved amino acid residues in TK, DXPS and PDC

Both TK and DXPS have residues that are uniquely conserved in their active-sites. There are also a high number of residues found to be highly conserved in TK and DXPS. These are possibly needed for the domain arrangement of TK and DXPS (Section 1.2.2), as well as their shared mode of catalysis, the transferase reaction. No positions are found to be uniquely conserved in PDC, which is perhaps unexpected. This suggests that most of the highly conserved residues in PDC are at positions which are also highly conserved in TK and DXPS. It would have been expected that the PDC-like domain arrangement (Section 1.2.2) and the different mechanism of catalysis of PDC would require uniquely conserved residues in the PDC enzymes. Overall the sequence and structure comparison of TK, DXPS and PDC has proved to be an extremely useful tool for selecting targets for mutagenesis. As shown in Figure 5.2, the various criteria applied to the 52 active-site residues allowed the selection of just two targets residues, one of which was a successful variant, as discussed in Section 5.4.3.2.

5.4.3 Enzyme Reaction

5.4.3.1 The β-HPA + GA reactions

All of the mutations generated significantly reduce the rate of L-erythrose formation in the TK catalysed β-HPA + GA reaction.

In the D183NTK mutant, the polar character changes, from acidic to neutral polar. We can also suggest that the orientation of the side-chain changes, although energy minimisation would need to be performed to establish this. Perhaps the β-strand (residues 178 – 184) is disturbed by the change at 183, but without proper energy minimisation, this is difficult to deduce.

The D381A mutation causes a 92% decrease in activity. Mutation of the equivalent residue in yeast, Asp382, which is buried upon the binding of TPP, to either Asn or Ala results in TK mutants with severe impairments in TPP binding and catalytic efficiency [202]. The D381A mutation has previously been shown to cause a decrease in TK monomer stability [203], resulting in a mutant with activity decreased by 98.3 % (56-fold decrease) for an enzyme linked TK assay. This corresponds within error to the observed 92 % decrease in activity for the β-HPA + GA reaction for the D381ATK mutant. Examination of the structure shows that mutation of aspartate to alanine at this position results in the loss of three hydrogen bond interactions with Ser188, Ile189 and Asp190 (Figure 5.11). In the Aucamp study, it was suggested that the removal of interactions with the interfacial Ile189 and Asp190 may cause the observed instability. The 381 position is also likely to be important in stabilising the adjacent loop region (382-392).

The S385G mutation decreases the formation of L-erythrulose by 95%. Perhaps this is explained by the switch in amino acid character from hydrophilic to neutral, although the marked decrease in the length of the side observed for the S385G switch may cause a steric effect, whereby the G385 may be optimally distant from the pyruvate molecul to form a H-bond. Modelling studies conducted in this laboratory by John Strafford suggest that the Ser385 residue is hydrogen bonded to L-erythrulose, the product of the β-HPA + GA reaction (Figure 5.13). Perhaps loss of

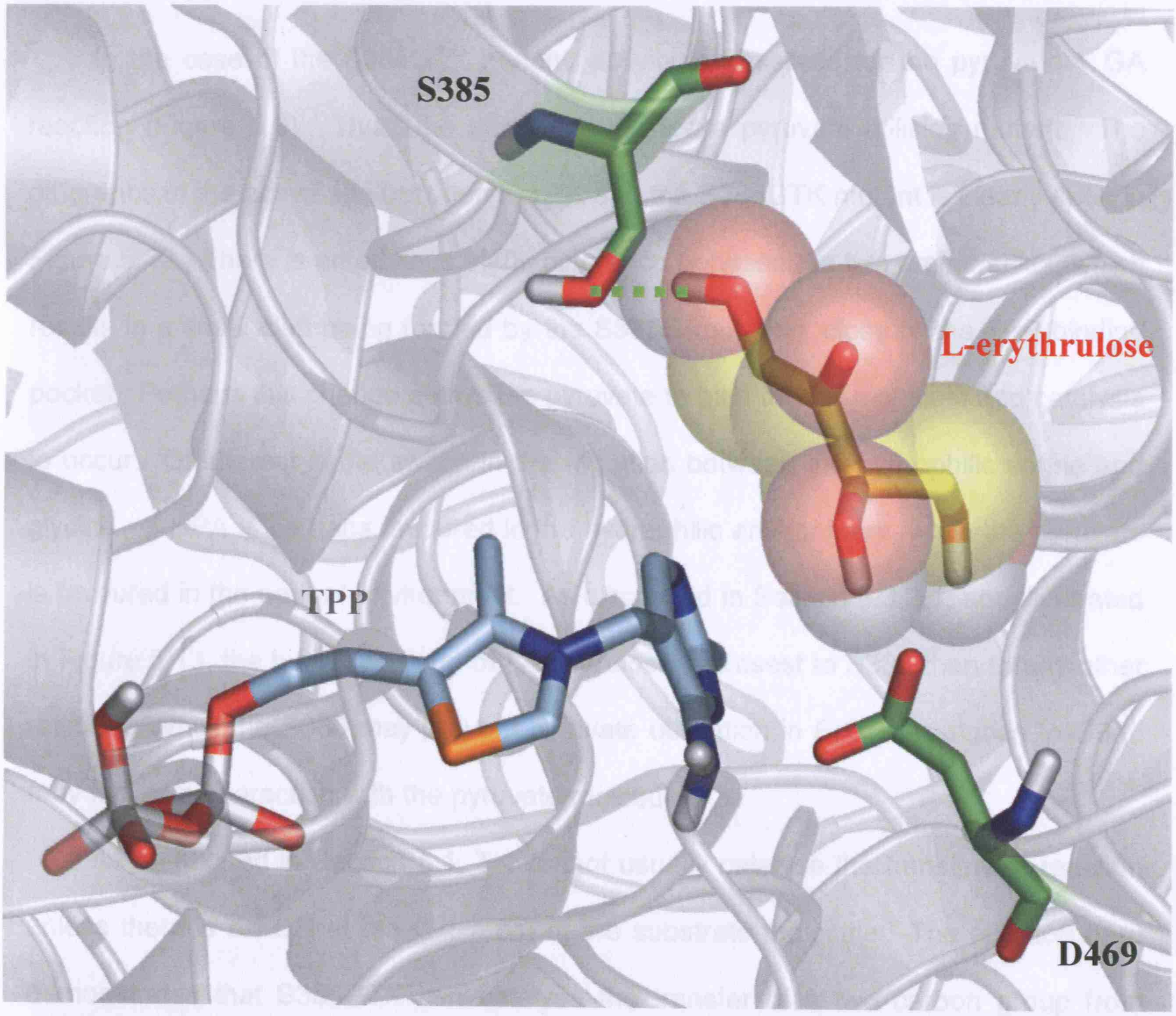


Figure 5.13: The proximity of Ser385 to the hydroxyl group of bound L-erythrulose in the *Eco*TK active-site. It would appear that L-erythrulose is bound at opposite ends by S385 and D469. L-erythrulose was modelled into the active-site by John Strafford, department of Biochemical Engineering, UCL. The TPP molecule is coloured according to the CNOS scheme, residues Ser385 and Asp469 are coloured as CNOS scheme, while L-erythrulose adopts the CNOS colouring. The proposed interaction between the Ser385 amino acid and the C1 atom of erythrulose is shown in green. Images were generated using the 1QGD.pdb file in *Pymol*.

this H-bond in the S385G mutant is responsible for the observed decrease in activity for the β-HPA + GA reaction.

5.4.3.2 The pyruvate + GA reactions

In the case of the S385GTK mutant, activity is observed for the pyruvate + GA reaction (Figure 5.9). Thus, the study has yielded a pyruvate utilising mutant. The difference in the active-site between *Eco*TK and the S385GTK mutant is clearly seen in Figure 5.12. There is an obvious steric difference between the two amino acids, which results in a small cleft being formed by the S385G mutation, close to the TPP binding pocket. Perhaps this change allows the pyruvate to bind the TK molecule and catalysis to occur. Of interest is the switch in the mutation between the hydrophilic serine and glycine. β-HPA is perhaps favoured in the hydrophilic environment, while the pyruvate is favoured in the neutral environment. As discussed in Section 5.3.2.1, and illustrated in Figure 5.13, the hydroxyl group of L-erythrulose is closest to S385 than to any other residue. While the S385 may prevent pyruvate utilisation in *Eco*TK, mutation to G385 may facilitate interaction with the pyruvate molecule.

As mentioned in Section 5.1, TK cannot usually catalyse the transferase reaction unless there is an OH at the C-1 atom of the substrate molecule. The reaction here demonstrates that S385GTK can catalyse the transfer of a two carbon group from pyruvate to GA. This reaction must proceed through a HETPP reaction intermediate (Section 1.1.4), from where TK normally yields free aldehyde, in a similar manner to the natural PDC reaction, thus preventing the transferase reaction from occurring. Thus, the S385G mutation must allow TK to form the HETPP intermediate from pyruvate and TPP, while also allowing the transferase reaction to occur from the HETPP intermediate. In general, we hypothesise that the 385 position in TK is responsible for the acceptance of substrates with an OH group at their C-1 atom, as backed up by the modelling studies described.

The 385 position may also have an effect on the 384 position, which is shown to have a dramatic effect on the β-HPA + GA reaction in Chapter 4 (Section 4.3.5.1). This could help explain the decrease in activity for the β-HPA + GA reaction in S385G.

The formation of a new product for the pyruvate + GA reaction is clearly observed for the S385GTK mutant. However, it should be noted that the rate of formation is still in the linear range after 24 hours. The reaction is thus comparatively slow. On reflection, this may be due to the use of GA as an acceptor substrate. The hypothesis is that Ser385 interacts with the OH group of the donor substrate, β-HPA (Image A, Figure 5.14). Mutation to Gly385 allows interaction with pyruvate (Image B, Figure 5.13). However, this discrimination is likely to extend to the acceptor substrate aswell. Since GA has an OH at the C-1 position, perhaps it is a poor substrate for the S385GTK mutant, as is the case in DXPS (Section 5.1). Mutants should therefore be tested for pyruvate activity with a different acceptor substrate, one lacking the C-1 OH group.

Based on the activities observed for the pyruvate + GA reaction, in order for TK to use pyruvate in industrially useful quantities, either lots of enzyme is needed or a form of TK with a faster reaction rate for pyruvate than S385G is needed. To generate such a mutant, the S385GTK makes an ideal starting point for directed evolution or further point mutations. Saturation mutagenesis of the 385 position may yield mutants with higher activity towards pyruvate than when S385G. An interesting mutation to investigate would be an S385P mutant. In the analysis, only 1 TK sequence did not have S at position 385. This was the TK from *Treponema pallidum*, the causative agent of syphilis (indicated with a star in Alignment 5.2). TK from *Tpa* has a proline at position 385. Interestingly, in an analysis of the *Tpa* metabolic system [204] two forms of

TK were found. However, no accompanying transaldolase (TA) enzyme could be detected. Since TK is accompanied by TA in all other organisms examined (due to the central metabolic roles of both enzymes in the PPP), this suggests an unusual function for TK in this species.

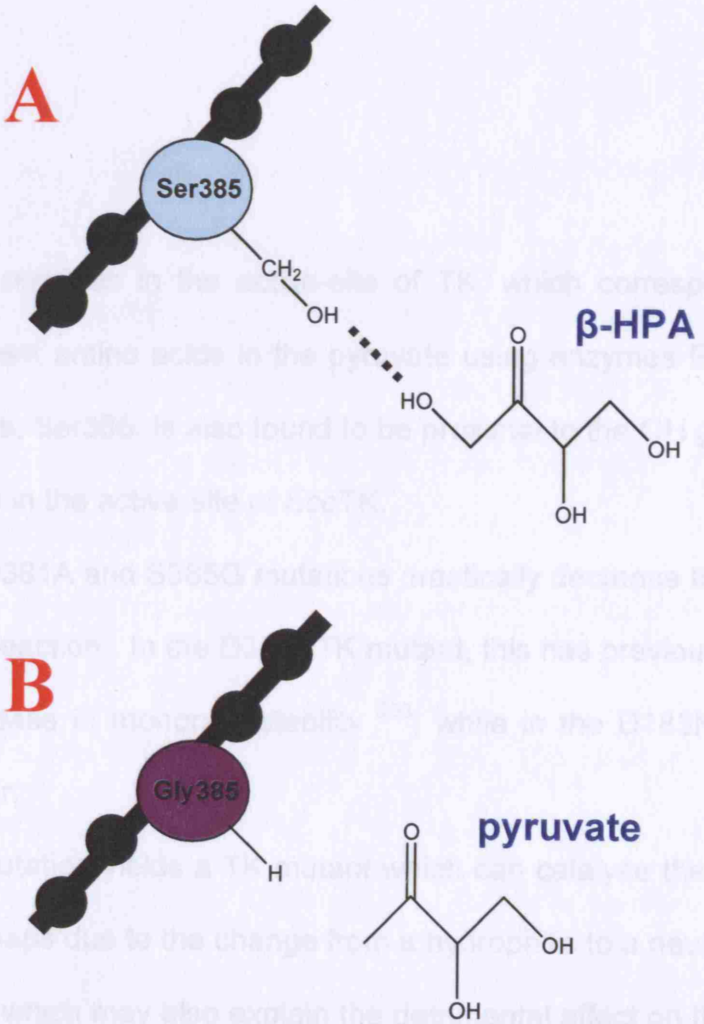


Figure 5.14: A model of how β-HPA and pyruvate may interact, respectively with Ser385 of EcoTK and Gly385 of S385GTK.

This analysis examined the effects of each mutation on pyruvate usage individually. It is possible that there are synergistic effects between the positions. It is plausible, for example that the appearance of pyruvate activity in S385GTK is due to the hydrophilic to neural switch at position 385. An accompanying switch of 381 from aspartate to alanine, may increase the hydrophobic character of the active-site and lead to a TK mutant with higher activity towards pyruvate than with the S385G mutation alone. Thus the generation of double and triple mutants of the identified residues would be of interest.

5.5 Conclusions

There are two residues in the active-site of TK, which correspond with highly conserved, but different amino acids in the pyruvate using enzymes PDC and DXPS. One of these residues, Ser385, is also found to be proximal to the OH group of the C-2 atom of L-erythrulose in the active-site of *Eco*TK.

The D183N, D381A and S385G mutations drastically decrease the activity of TK for the β-HPA + GA reaction. In the D381ATK mutant, this has previously been shown to be due to a decrease in monomer stability [203], while in the D183NTK mutant the reasons are less clear.

The S385G mutation yields a TK mutant which can catalyse the pyruvate + GA reaction. This is perhaps due to the change from a hydrophilic to a neutral residue with the S385G mutation, which may also explain the detrimental affect on the β-HPA + GA reaction. Position 385 is likely to have an important role in discriminating between substrate with hydroxyl groups at their C-1 atoms.

Pyruvate use is slow with the S385GTK mutant, and the reaction is still linear after 24hours, perhaps due to GA being a poor acceptor substrate. The S385GTK mutant makes an ideal starting point for future development of an industrially viable pyruvate-using TK mutant. Here it is shown that analysis of enzymes within an enzyme family such as the TPP-dependent enzymes, can suggest excellent targets for mutagenesis to yield a desired enzyme function.

Chapter 6: Investigation of the role of the C-terminal domain of transketolase

6.1 Introduction

TK is composed of 3 domains, which are, the PP and Pyr domains are common to all TPP-dependent enzymes and a C-terminal domain, designated the TKC domain. The TKC domain is also found in other TPP-dependent enzymes of the TK-like group, the 2OXO-like group as well as in the PFRD enzymes (Section 1.2.2). While the PP and Pyr domains are known to be essential for TPP binding and catalysis, the role of the TKC domain remains undefined.

Despite the lack of information on the function of the TKC domain, it is highly conserved both at the sequence and structure levels.

To investigate the role of the TKC domain, a stop codon was introduced into the TK gene, yielding a truncated TK enzyme, without a TKC domain. When this mutant was found to be active for the β-HPA + GA reaction, further truncations of the TK gene were performed yielding four more mutants with varying activity for the β-HPA + GA reaction. Modelling studies were then used to try and explain the observed activities of the truncated TK enzymes.

6.2 Methods

6.2.1 Conservation of the TK C-terminus

Alignment 3.1 from Chapter 3 contains the TKC domains of 54 TK sequences, as they align with *Eco*TK. In that alignment, due to the variability of the lengths of some of the C-termini, the sequences were trimmed to those residues aligning with *Eco*TK residues 10-627. Since the TKC domain of *Eco*TK is defined as beginning at residue

540 ^{7}, this alignment contains the TKC domains of the 54 sequences as they align with the *Eco*TK TKC domain residues 540-627. When discussing the alignment of *Eco*TK and *Sce*TK in this chapter, as in Alignment 6.1, the residues of the two enzymes align exactly as in Alignment 3.1. To examine the conservation of the TKC domain structure in TK, the two structures for *Eco*TK (1QGD.pdb) and *Sce*TK (1NGS.pdb) were overlaid against each other in *Pymol.* The residues which are found to be the same in *Eco*TK and *Sce*TK in Alignment 6.1 were modelled as sticks in the TKC domain.

6.2.2 *BLAST* search using *Eco*TK residues 540-627 as the query

To examine whether the TKC domain of TK shares homology with any another protein domain of known function, the *Eco*TK C-terminal sequence of residues 540-627 was input into a protein-protein *BLAST* search (Section 2.2 for URL).

6.2.3 Choice of target amino acid residues for mutagenesis

Initially, since the TKC domain is defined as beginning at residue 540 in *Eco*TK, residue 540 was chosen as a target for insertion of a stop codon by SDM. The *Eco*TK crystal structure was examined in *Pymol* to ensure that no regions of secondary structure were disturbed by this mutation.

Additional sites were picked in TK for the introduction of stop codons. In each case, the TK enzyme was increasingly truncated. Each mutation site was chosen by examining the *Eco*TK crystal structure to ensure stop codons were being inserted into loop regions and not in areas of secondary structure. It was thought that such mutations were less likely to cause structural disruptions that may inactivate the enzyme for other reasons, for example, protein misfolding or aggregation.

6.2.4 Insertion of stop codons into TK

TK mutants in subsequent sections will be named according to the mutations they contain as described in the Abbreviations section.

6.2.4.1 Primers for the G540stop, E527stop, R492stop, H461stop and Q453stop mutations

Primers were designed using *Annhyb* and ordered from Qiagen Ltd., salt free and at the 50 nmol scale. Primers used for mutagenesis were as follows (mutagenic codons are in bold):

G540stopF: CGCGCGCGGT**TGA**TATGTGCTG

G540stopR: CAGCACATA**TCA**ACCGCGCGCG

E527stopF: GGCGCAGCAG**TAA**CGAACTGAAG

E527stopR: CTTCAGTTCG**TTA**CTGCTGCGCC

R492stopF: GTCTACATGG**TGA**CCGTGTGAC

R492stopR: GTCACACGG**TCA**CCATGTAGAC

H461stopF: GGTTTACACC**TAG**GACTCCATCGG

H461stopR: CCGATGGAGTC**CTA**GGTGTAAACC

Q453stopF: GCTGATGAAA**TAG**CGTCAGGTGATG

Q453stopR: CATCACCTGACG**CTA**TTTCATCAGC

6.2.4.2 Site directed mutagenesis, transformation and storage of mutants

The protocol for generating mutants G540stopTK, E527stopTK, R492stopTK, H461stopTK and Q453stopTK is the same as Protocol 2, described in Section 4.2.5.2.

Parental DNA was digested as described in Section 4.2.6, followed by gel visualisation and transformation into XL1 Blue competent cells, as detailed in

Section 2.5. Miniprepped mutant DNA was sequenced by the Wolfson Institute. Mutant plasmids were transformed into XL10 Gold cells, again using the protocol in Section 2.5. Glycerol stocks were made with 40 % sterile glycerol and stored at -80 °C.

6.2.5 Enzyme reactions of the truncated TK mutants for the model β-HPA + GA reaction

The reactions of *Eco*TK, G540stopTK, E527stopTK, R492stopTK, H461stopTK and Q453stopTK exactly as described in Section 2.5.11.2, with purified lysate bioanalysed and used in the reactions to ensure an enzyme concentration of 0.5 mg.mL^{-1}.

6.2.6 HPLC assays

Assays were performed for *Eco*TK, G540stop, E527stop, R492stop, H461stop and Q453stop exactly as described in Section 2.5.12.3a.

6.2.7 Structural models of mutant TKs

In all cases, enzymes were manipulated and visualised in *Pymol*, using the *Eco*TK crystal structure, 1QGD.pdb [187].

6.3 Results

6.3.1 Conservation of the TKC domain

In the TKC domain region of Alignment 3.1, of the 87 residues, there are 40 at which *Eco*TK and *Sce*TK have the same amino acid and a further 11 where the amino acids are very similar (using a Blossum62 matrix).

Figure 6.1 illustrates the fitting of the *Eco*TK and *Sce*TK structures that was performed in *Pymol.* The RMSD after fitting the entire *Eco*TK structure to the entire *Sce*TK structure was 30.262 Å. In terms of regions of secondary structure and the location of the 51 equivalent residues, discussed above, the level of structural conservation in the TKC domains of *Eco*TK and *Sce*TK is high.

```
EcoTK    GYVLKDCAGQPELIFIATGSEVELAVAAYEKLTA-EGVKARVVSMSSTDAFDKQDAAYRESVLPKAVTARVAVEAGIADYWYKYVGLNG
SceTK    GYVLQDVAN-PDIILVATGSEVSLSVEAAKTLA-AKNIKARVVSLPDFFTFDKQPLEYRLSVLPDNVP-IMSVEVLATTCWGKYAH---
```

Alignment 6.1: How the *Eco*TK and *Sce*TK TKC domain sequences align in Alignment 3.1. Residues highlighted in black are those found to be 100 % conserved between *Eco*TK and *Sce*TK by identity, while those shaded grey are 100 % conserved by similarity. These conserved residues are modelled into the *Eco*TK and *Sce*TK crystal structures in Figure 6.1.

6.3.2 *BLAST* search using *Eco*TK residues 540-628 as the query

The *BLAST* search returned many TK sequences and other enzymes containing the TKC domain, but no other homologous domains, the characteristics for which may have been used to hypothesise the function of the TKC domain.

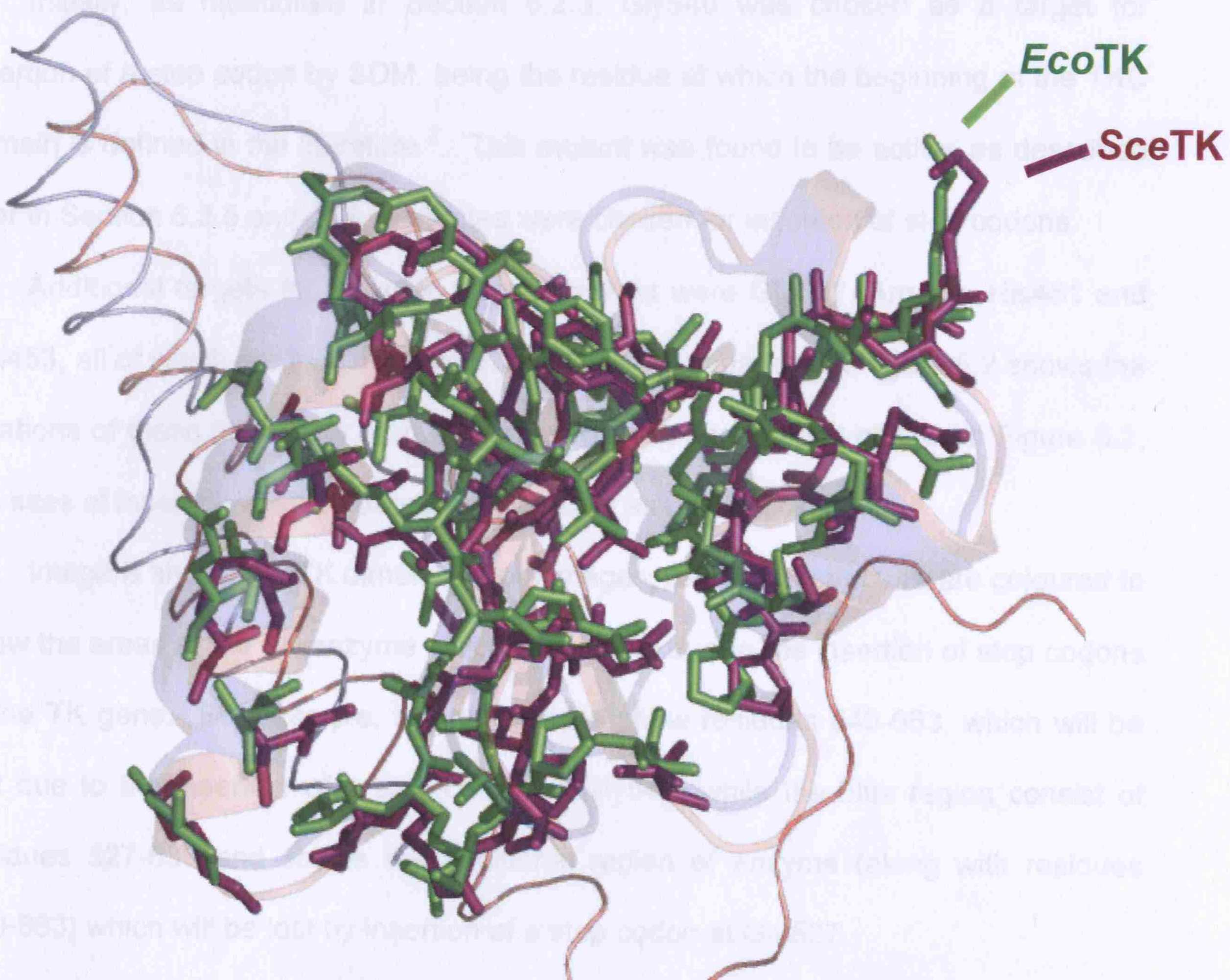


Figure 6.1: The TKC domain structures for *Eco*TK and *Sce*TK fitted together. In the regions where the sequences of *Sce*TK and *Eco*TK align (Alignment 6.1), the secondary structures are shown. Secondary structure appears in blue for the *Eco*TK and red for the *Sce*TK. Those residues in the *Eco*TK and *Sce*TK sequences that are 100 % conserved either by sequence identity or similarity using a Blossum62 matrix (Alignment 6.1), are shown as sticks, in green in the *Eco*TK (1QGD.pdb) structure and purple in the *Sce*TK (1NGS.pdb) structure. The regions of structure which do not align sequentially, at the extreme C-termini of *Eco*TK and *Sce*TK are shown in ribbon form, where *Eco*TK is blue and *Sce*TK is coloured red . The image was generated using *Pymol.*

6.3.3 Choice of target amino acid residues for mutagenesis

Initially, as mentioned in Section 6.2.3, Gly540 was chosen as a target for insertion of a stop codon by SDM, being the residue at which the beginning of the TKC domain is defined in the literature [2]. This mutant was found to be active, as described later in Section 6.3.5 and so further sites were chosen for insertion of stop codons.

Additional targets for insertion of stop codons were Glu527, Arg492, His461 and Gln453, all of which are located in loop regions in TK. Image A in Figure 6.2 shows the locations of these targets for stop codon insertion in one subunit of TK. In Figure 6.2, the sites of insertion of stop codons are depicted as black sticks.

Image B shows the TK dimer. In both images, regions of structure are coloured to show the areas of the TK enzyme which will be lost due to the insertion of stop codons in the TK gene. For example, the red regions show residues 540-663, which will be lost due to the insertion of a stop codon at Gly540, while the blue region consist of residues 527-539 and shows the additional region of enzyme (along with residues 540-663) which will be lost by insertion of a stop codon at Glu527.

6.3.4 Enzyme reactions and HPLC analysis of the truncated TK mutants for the model β-HPA + GA reaction

Figure 6.3 shows the initial velocities for *Eco*TK, G540StopTK, E527StopTK, R492StopTK, H461StopTK and H453StopTK. *Eco*TK and E527StopTK have very similar initial velocities, while the H461StopTK mutant has a slightly lower activity. The R492StopTK mutants has an initial velocity that is 12% that of *Eco*TK.

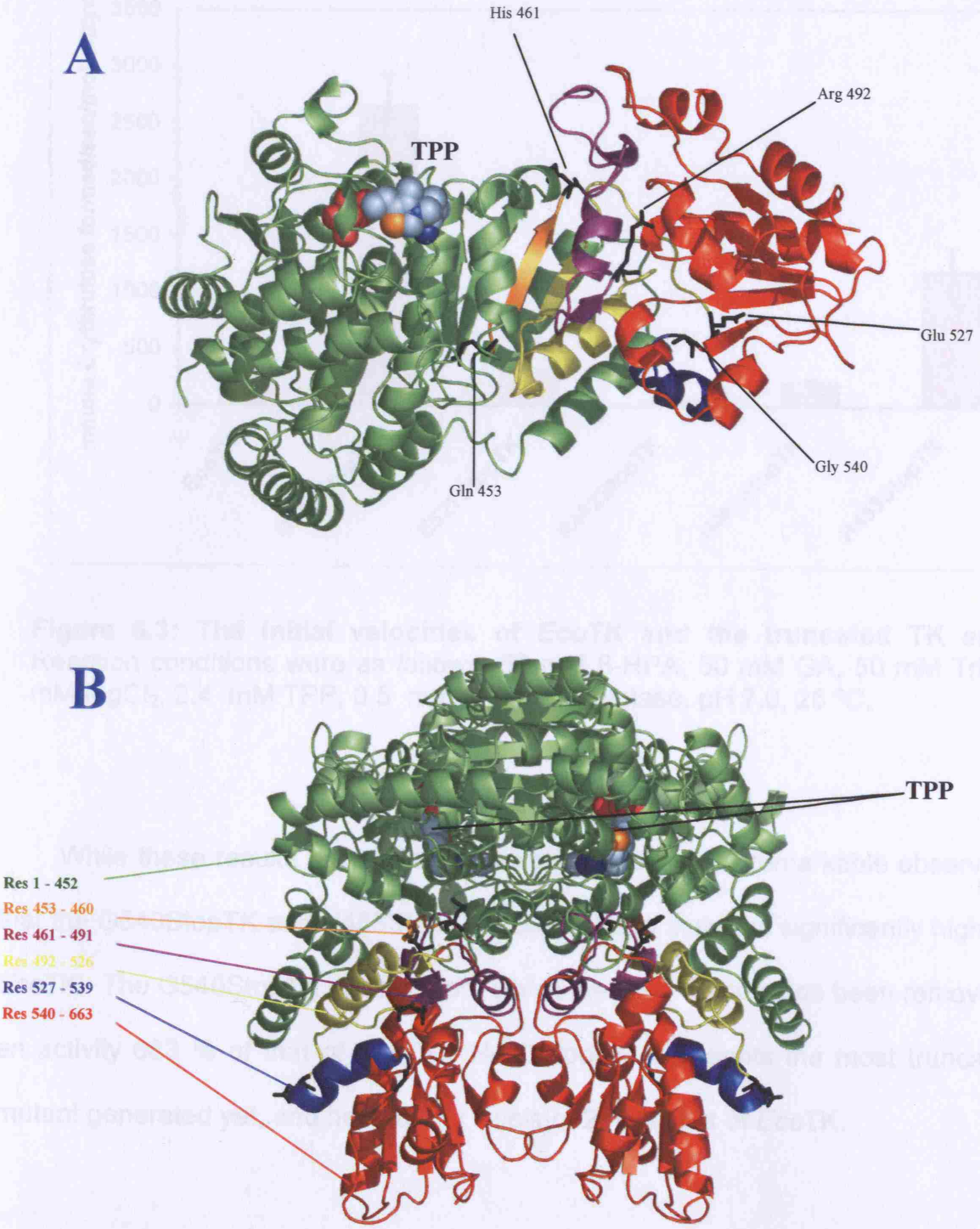


Figure 6.2: Targets chosen for insertion of stop codons by SDM. Image A shows a monomer of TK, with the sites chosen for stop codon insertion labelled, while image B shows the TK dimer, illustrating how the enzyme will be truncated by insertion of the stop codons of image A into the TK gene. The unaffected region of the TK molecule is coloured green, while the red region indicates the area that will be lost due to insertion of a stop codon at G540. Further truncation of the TK enzyme will result in the other coloured regions being lost, as suggested by the numbering in image B and in the text. Residues chosen as targets for SDM are depicted as black sticks. The TPP molecule is coloured according to a CNOS colouring scheme. The image was generated using the 1QGD.pdb file in *Pymol*.

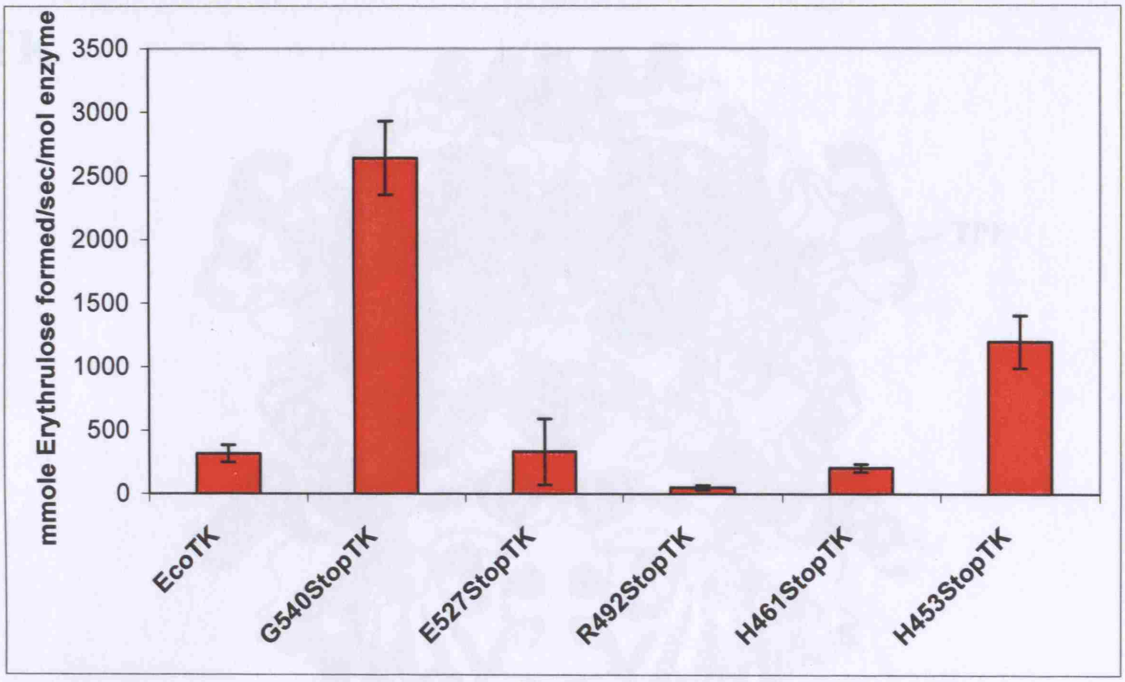


Figure 6.3: The initial velocities of *Eco*TK and the truncated TK enzymes. Reaction conditions were as follows: 50 mM β-HPA, 50 mM GA, 50 mM Tris-HCl, 9 mM $MgCl_2$, 2.4 mM TPP, 0.5 $mg.mL^{-1}$ transketolase, pH 7.0, 25 °C.

While these results are perhaps unexpected, the most remarkable observation is that the G540StopTK and H453StopTK mutants have activities significantly higher than *Eco*TK. The G540StopTK mutant, from which the TKC domain has been removed, has an activity 683 % of that of *Eco*TK. H453StopTK represents the most truncated TK mutant generated yet, and has activity levels of 260 % that of *Eco*TK.

6.3.5 Structural models of mutant TKs

Figure 6.4 shows structural models for *Eco*TK and the five truncated TK mutants generated. These models show how gradual truncation of the TK gene produced successively shorter enzymes. These models are used in the discussion to attempt to explain the kinetic data obtained for the mutant TK enzymes.

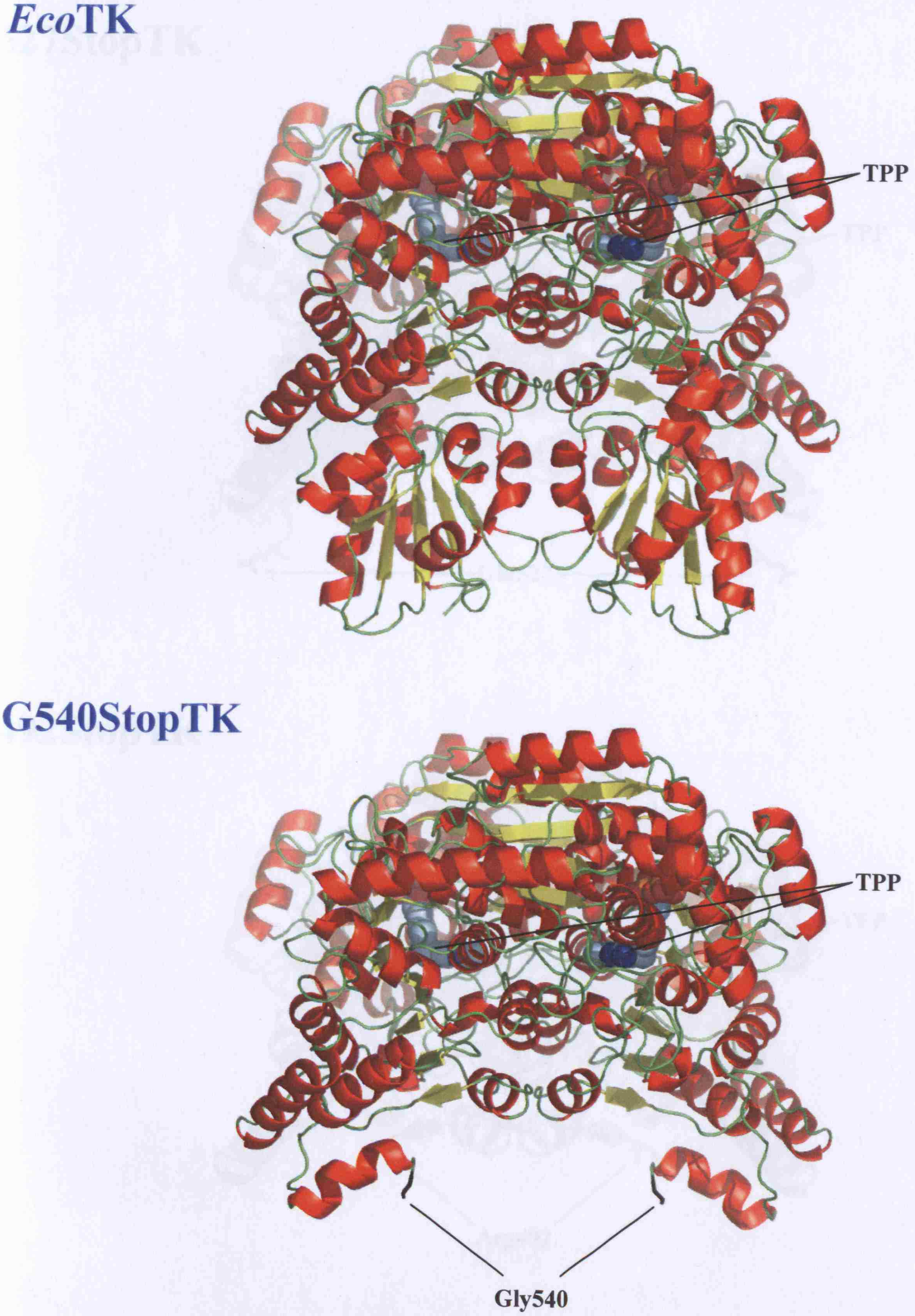
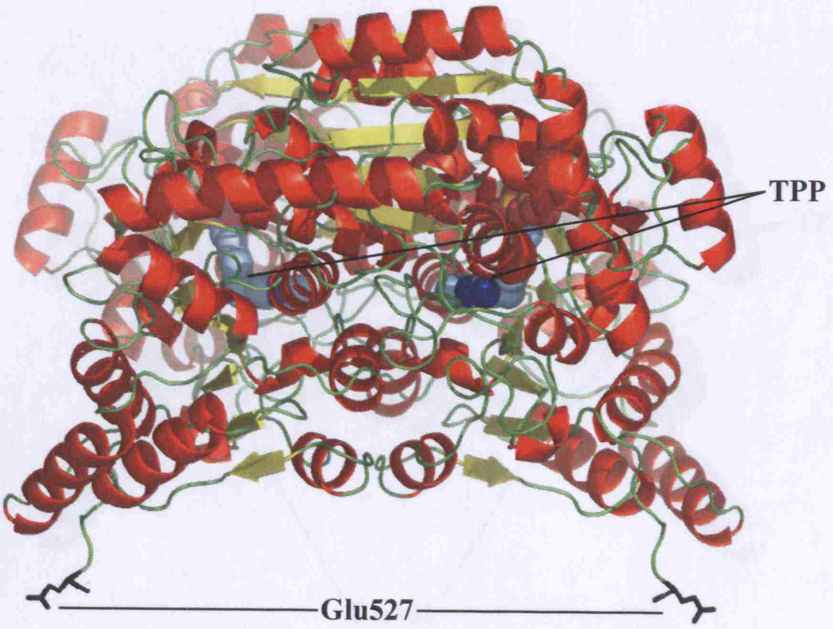


Figure 6.4: Structural models of *Eco*TK and the truncated TK mutants. Loop regions are coloured green, while α-helices and β-strands are coloured red and yellow respectively. The TPP molecule is coloured according to the CNOS scheme. Sites which were targeted for mutagenesis are displayed as black sticks. Images were generated using the 1QGD.pdb file in *Pymol*.

E527StopTK



R492StopTK

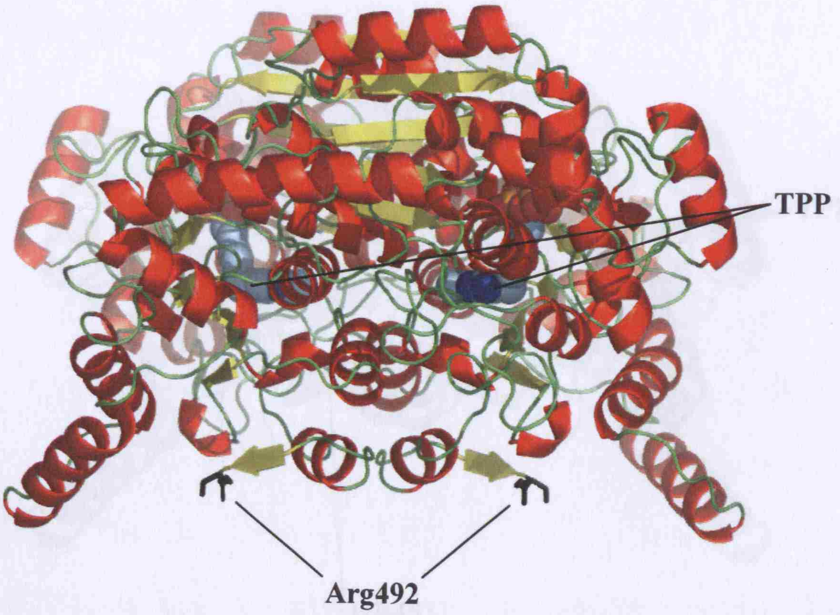
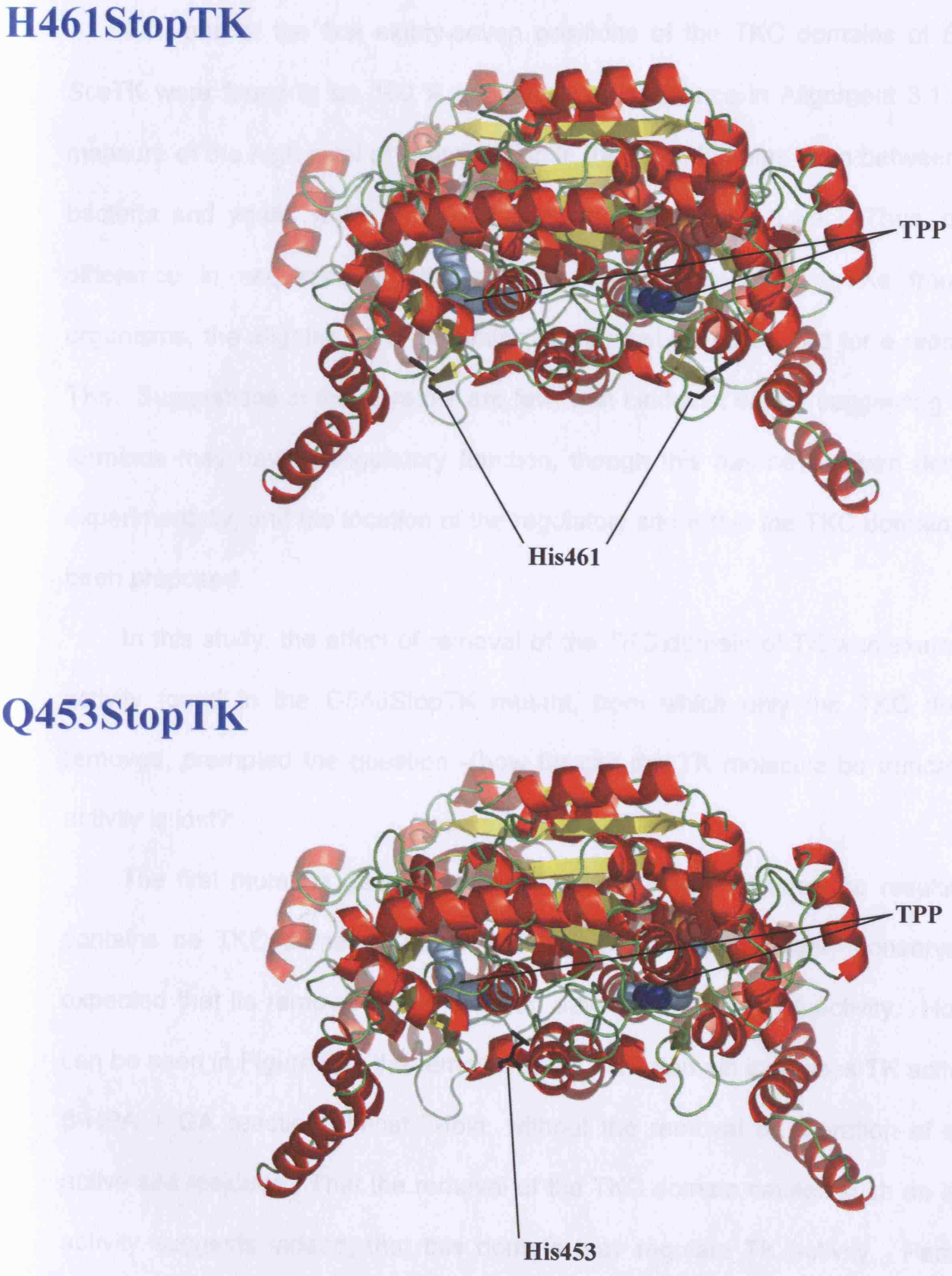


Figure 6.4 continued.



H461StopTK

Q453StopTK

Figure 6.4 continued.

6.4 Discussion

Fifty-one of the first eighty-seven positions of the TKC domains of *Eco*TK and *Sce*TK were found to be 100 % conserved by sequence in Alignment 3.1. This is a measure of the high level of conservation in the TKC domains even between TKs from bacteria and yeast, which are likely to have diverged ~2 b.y.a.. Thus, despite the difference in sequence lengths at the extreme C-termini of TKs from different organisms, the aligning 89 or so amino acids must be conserved for a reason among TKs. Suggestions in the literature are few, with Lindqvist *et al.* [2] suggesting that the C-terminus may have a regulatory function, though this has never been demonstrated experimentally, and the location of the regulatory site within the TKC domain has never been proposed.

In this study, the affect of removal of the TKC domain of TK was examined. The activity found in the G540StopTK mutant, from which only the TKC domain was removed, prompted the question - how far can the TK molecule be truncated before activity is lost?

The first mutation generated was the G540StopTK, where the resulting mutant contains no TKC domain. Since the TKC domain is so highly conserved, it was expected that its removal would have an adverse affect on TK activity. However, as can be seen in Figure 6.3, the removal of the TKC domain increases TK activity for the β-HPA + GA reaction almost 7-fold, without the removal or alteration of any of the active-site residues. That the removal of the TKC domain causes such an increase in activity suggests indeed, that this domain may regulate TK activity. Perhaps such regulation is essential *in vivo*, but from an industrial point of view, the more active, shorter form of TK would be of significant interest. It is also possible that removal of the TKC domain causes a conformational shift in the PP and Pyr domains, leading to an increase in the TK reaction rate. Figure 6.4 notionally shows the G540StopTK structure. Energy minimisation of such a structure is beyond the scope of our computational resources, so the α-helices at the C-terminal ends of the G540StopTK mutant may adopt a different conformation to that shown. The presence of these helices seems to be important to the high level of activity of the G540StopTK, since their removal with the E527Stop mutation has a dramatic effect on TK activity, as shown in Figure 6.3. In instances where removal of helices ia accompanied by a decrease in activity, this may be due to the exposure of hydrophopbic surfaces. Such hydrophobic regions may cause the protein to misfold and have a detrimental affect on activity.

The E527stopTK has an activity roughly the same as for *Eco*TK. Removal of the 527-539 region involves the loss of eight residues, which either by sequence identity or similarity were shown to be 100% conserved in an alignment of fifty-two TK sequences (Alignment 3.1, less *Cac2*TK and *Bme*TK) in Chapter 3 (Table 3.2). These were Tyr541, Thr557, Gly558, Ser559, Glu560, Val561, Ser582 and Glu612. Residues 541, 558, 559 and 560 were found at the surface. The functions of these residues are undefined, but collectively, they have a marked influence on TK activity.

The R492Stop mutation decreases activity to the lowest level observed for any of the truncated TK mutants, with an initial velocity 12 % that of *Eco*TK.

The R492Stop mutation involves the loss of six additional residues 100 % conserved in TK, namely Arg492, Asp495, gGlu498, Ser519, Arg520 and Gln521. Of these, Gln522 is found at the surface and along with Glu498 and Ser519, its function is unknown. Highly conserved residues Arg492 and Asp495, as well as Pro493 and Cys494, which are much less conserved, form part of the TK motif, discussed in Chapter 3 (Section 3.3.2.3). Interestingly, the Asp495 residue is invariant in all TK sequences examined here or elsewhere and is even conserved at the equivalent position of the NADH-binding like motif, which shares homology with the TK motif (Section 3.3.2.3).

The function of active-site residue Arg520 is believed to be the binding of substrate. In the crystal structure of *Sce*TK, the corresponding yeast residue, R528, is shown to be involved in the binding of E4P (Figure 1.8), by direct interaction with the phosphate.

The effect on TK activity, of removing residues 493-527, is dramatic. This is perhaps due to the combined removal of D495 and R520. A recent library screen in this laboratory demonstrated a R520V mutant with increased activity for the β-HPA + GA reaction [205]. It would be interesting to test this R520V mutant for activity with E4P. Since manipulation of the R520 position can affect the rate of reaction of β-HPA + GA, neither of which is phosphorylated, the complete removal of this amino acid position may have an even more dramatic affect on the rate of TK reaction with phosphorylated substrates than that observed for our model TK reaction. Also of note is that a R520Stop mutant generated in the Hibbert library screen was found to be as active as the *Eco*TK.

Further truncation of the TK molecule by introduction of the H461Stop mutants involves the loss of most of the TK motif. Eleven 100 % conserved residues (Section 3.3.2.2) are lost, six of which (His461, Asp469, Gly470, Thr472, His473 and Gln474) as well as 3 less conserved residues, Pro471, Glu468 and Leu466 are found in the TK active-site region. The active-site Gly470 and surface residue Gly467 are highly conserved in the TK motif, being found at this position in all TK enzymes as well as at the equivalent position in the NADH-binding like domain. Other active-site residues, the functions for which have been proposed are His461, Asp469 and His473. Residue His469 in *Sce*TK corresponds with *Eco*TK His461 and is shown to be involved in the binding of E4P in the *Sce*TK crystal structure (Figure 1.8).

Asp469 is thought to be essential for proper substrate binding in TK, where it is proposed to bind the hydroxyl group of the donor substrate. Figure 1.8 shows how the corresponding *Sce*TK residue Asp477 is involved in binding E4P. Interestingly, the D477A mutant of the *Sce*TK generated by Nilsson *et al* [29] was shown to change the stereospecificity of the enzyme and a D469Stop mutant also generated in the Hibbert library was found to be active for the model TK reaction.

His473 is conserved at the corresponding positions in all non-mammalian TKs examined. It has been proposed that this residue has a role in abstracting protons from the imino group during TK catalysis. This proposal is, however, contentious [11].

Several of the highly conserved residues in the 461-491 region are found at the interface of the TK dimer. These are Asp469, Thr472, His473, Gln474 and Glu477. The loss of all of these residues results in an almost 4-fold increase in activity relative to the R492StopTK.

The H453Stop mutation produced the most truncated form of TK, with 210 of the 663 TK residues removed from the C-terminal. Apart from removal of the entire TKC domain, 87 of the 220 residues comprising the Pyr domains have been removed in the H453StopTK. However, removal of the 453-460 region, which contains the active-site residue Thr460 produces an enzyme which has 572% the activity of R461StopTK. This mutant, as described in Section 6.3.5, has 260% the activity of *Eco*TK. Thus, the Thr460 residue may be important for TK activity in this context.

Overall, the activities of the truncated TKs are surprising and raise many questions. Firstly, what is the function of the TKC domain? It would appear that it regulates the activity of TK in some way. Such regulation may be very important *in vivo*. Perhaps the TKC domain has a stabilising effect on the TK molecule. Perhaps a compromise between a more stable TK dimer and a less active enzyme has been achieved in TK and conserved during evolution in all TKs. Stability work on the truncated TK mutants described here are currently being performed in this laboratory.

There are implications for some other TPP-dependent enzymes. This study shows that essentially only the PP and Pyr domains are required for catalysis, a feature reinforcing the idea that the most ancient TPP-dependent enzymes (similar to extant PPDC and SPDC enzymes) had only PP and Pyr domains (Section 3.4.3).

Primarily, the increase in the rate of reaction observed by removal of the TKC domain is of greatest interest from an industrial point of view. As discussed in Chapter 5, for TK to be industrially important, a cheap, readily available donor substrate such as pyruvate would be desirable. The S385G mutant of TK, described in Chapter 5 uses pyruvate, but only slowly. Perhaps removal of the TK C-terminal domain of the S385GTK, could lead to a faster, more industrially viable pyruvate using TK mutant.

There are many residues that could be responsible for the activities observed for the other TK mutants, E527StopTK, R492StopTK, H461StopTK and Q453StopTK. In general it seems that when regions of high conservation are disturbed, as those residues that make up the TK motif, one of the most conserved regions among TK enzymes, activity decreases, as observed for the R492StopTK. Residues in such motifs have coevolved over long periods of time and their correct function is likely to require the topology to be conserved. Removing some of the residues could upset the function of the motif as a whole and cause the observed decrease in TK activity.

However, when motifs are almost entirely removed, as in the H461StopTK, activity is recovered to *Eco*TK levels. When the final, highly conserved active-site residue in the region, T460, is removed, activity increases significantly. Thus, activity, it would seem is higher if the TK motif is either fully removed or fully present, while activity is drastically reduced if it is merely disturbed. The observation that so many regions of secondary structure can be removed while still retaining TK activity may have interesting evolutionary implications. The study conducted in Chapter 3 shows how the various TPP-dependent enzymes have coevolved. Section 1.1.2 of Chapter 1 described how the PP and Pyr domains are very similar structurally having both evolved from a similar ancestral domain type (Section 3.4.3). Since PP and Pyr domains are present in all TPP-dependent enzymes, they are likely to have diverged from each other before the emergence of distinct TPP-dependent enzymes. 40% of the primary structure of the Pyr domain can be removed while activity remains higher than in *Eco*TK. It could thus be hypothesised that the PP domain is the more essential domain for catalysis. Perhaps the progenitor, single domain TPP-dependent enzyme resembled the PP domain more than the Pyr domain. Expression and characterisation of a single PP domain from TK could test the plausibility of such an assertion. It is important to note that none of the highly conserved Pyr domain residues thought to interact with the MAP ring of TPP have thus far been removed from the Pyr domain. It is thus possible that the region of Pyr domain remaining in the H453StopTK is the most important region of the Pyr domain. Truncation of TK so as to disrupt these residues would thus be of interest. During evolution of the TPP-dependent enzymes, the Pyr domain may have provided more specificity and enantioselectivity while addition of the TKC may have provided a regulatory mechanism for the enzyme.

The high activity of H453Stop with the substrate β-HPA and GA in particular fits with structure seen in Figure 6.4. Although energy minimisation has not been performed for the images in Figure 6.4, and it is likely that the regions of secondary structure will orient differently once exposed, it is clear that the active-site has become more exposed in H453StopTK. This may have allowed greater access to substrates in general and removed some of the enzyme architecture that has evolved to increase the selectivity of TK towards phosphorylated substrates (Section 4.4.5.5b). Removal of these regions may allow TK to catalyse at a faster rate the reaction involving smaller, non-phosphorylated substrates, such as β-HPA and GA.

It would be interesting to investigate if removal of the D469 residue removes the enantioselectivity of TK. As mentioned, the D477A mutant of *Sce*TK had different enantioselectivity than in the *Eco*TK.

6.5 Conclusions

The TKC domain of TK is highly conserved at the sequence and structure levels, yet it isn't necessary for TK activity. In fact, its removal increases TK activity more than 8-fold. In addition to the TKC domain, 40 % of the Pyr domain, including the TK motif and 10 active-site residues can be removed and activity retained at a level 4 -fold higher than activity in *Eco*TK. Removal of the TKC domain may have industrial implications with faster rates for small non-phosphorylated substrates and less cost in preparing the enzyme. However, the *in vivo* role of the TKC domain remains unresolved.

Chapter 7: General discussion

The focus of this thesis has been on the industrially important TPP-dependent enzyme transketolase (TK). TK is an ideal biocatalyst given its broad substrate repertoire and the fact that a small number of variable active-site residues result in the diverse substrate specificities seen amongst TK species. Several approaches were used in the studies of Chapter 3 through 6.

Firstly, the evolution of TK in the context of the TPP-dependent enzyme family was examined. The combined phylogenetic study of 17 TPP using enzymes shows a family of enzymes for which most members diverged before the emergence of the five kingdoms of life, i.e. in the progenote population. In the case of GXC, the enzyme seems to have evolved more recently, from an ALS-like ancestor. Evidence of this evolutionary relationship survives in the ability of GXC to catalyse the ALS reaction. All of the GXC sequences examined are from bacteria, but the poor resolution of the ALS phylogeny in both the individual ALS tree and the overall TPP-dependent enzyme tree makes it impossible to infer any more details about the divergence event. The strong signal observed in the phylogenetic analysis suggests that the structurally equivalent regions analysed have become refined during evolution towards specific chemistries in the extant enzymes in which they are found. Thus, these regions of comparable structure can confidently be used in the analysis of Chapter 5, where the TK and PDC structures are compared in detail. In particular, the phylogenetic study highlights the usefulness of structural data. Use of a simple *ClustalW* alignment would have produced an incorrect and confusing topology. Since sequence homology between certain enzymes examined is in the "twilight-zone", construction of the TPP-dependent enzyme phylogeny would have proved impossible without the available crystal structures which allowed alignment by eye.

In the future, wherever it is necessary to compare regions of secondary structure between TPP-dependent enzymes, the structurally equivalent regions defined in Chapter 3 can be used as a starting point. To investigate relationships between positions in the alignments of the PP and Pyr domains of the TPP-dependent enzyme family, cluster analysis is currently being performed at this laboratory. In this way, functionally related residues may be elucidated, suggesting where synergistic effects occur and perhaps prompting mutagenesis targets which may compliment the mutations discussed in Chapter 4 and 5.

The lack of certain regions of secondary structure in the *Tma*-like PFRD enzymes and their consequent exclusion from the phylogenetic study highlights an important issue: that as the number of distinct enzymes examined increases, the regions of comparable structure will be likely to decrease, lowering the number of informative positions in a given alignment and hindering phylogenetic reconstruction. What is also clear from construction of the TPP-dependent phylogeny is the inadequacies of the various tree-viewing software available. These limitations mean that cladograms are often adjusted by hand, since the graphics produced for large trees are often skewed, with sequence names not corresponding well with branch positions [206].

An interesting observation in the overall TPP phylogeny is that the SPDC and PPDC enzymes diverged before the emergence of the other enzymes examined. PPDC can catalyse the decarboxylation of pyruvate, sulfopyruvate as well as its physiological substrate, phosphopyruvate. Since PPDC may resemble some of the earliest TPP-dependent enzymes, as discussed in Chapter 3, this catalytic promiscuity may support the idea that ancient proteins had broader substrate specificities than modern enzymes, a hypothesis explored in Chapter 4.

The in-depth analysis of TK was necessary since its phylogeny would be subject to ancestral reconstruction. What is shown, unsurprisingly is that TK is highly conserved and this conservation extends to each of the three TK domains. As discussed in Chapters 1 and 4, such high homology allows the diverse substrate specificities in the TK species to be mediated by a small number of variable residues. Since the substrate specificities of *Eco*TK and *Sce*TK are considerably different and their crystal structures and sequences are so highly homologous, the choice was made to study the evolution from their common ancestor to modern *Eco*TK, focussing on the active-site. For the ancestral TKs, the 384 position had the most dramatic effect on activity. The presence of glycine at this position, as in the N96TK mutant, showed activity 10-fold higher than *Eco*TK for the β-HPA +GA reaction. During evolution, the initial velocity for the model TK reaction increases in general, with a sharp fall ~140 MYA. Since the β-HPA + GA reaction is not a natural TK reaction, it would not have been encountered during evolution. Thus, screening of the ancestor TK mutants with the physiological substrates of TK may shed light on the true metabolic responses to evolutionary stress.

The *Eco*TK and N58TK (common ancestor) enzymes were examined for a much broader range of donor and acceptor substrates. No definitive pattern could be observed for the broadening or narrowing of the substrate repertoire during evolution. This question itself is subjective, since it is difficult to suggest which substrates define a broad repertoire. What can be stated, however, is that the amino acids targeted have a significant effect on substrate specificity, with certain substrates favoured in one enzyme to a greater degree than in the other. The most unexpected result is the use of A5P by both *Eco*TK and N58TK. A5P has previously been proposed to inhibit *Eco*TK [31], although, it must be pointed out that in that study, β-HPA was not used as the donor substrate. Similarly, the use of G6P by *Eco*TK was unexpected, given that it has previously been reported as not binding to *Eco*TK, although both G6P and A5P are known to be used by yeast TK [30]. Such observations are difficult to explain. In the future, it would be interesting to develop more refined assays to measure the kinetics of the G6P + β-HPA, and A5P + β-HPA reactions for both enzymes.

In terms of the non-natural substrates examined, only propionaldehyde was used by either *Eco*TK or N58TK and the initial velocities were essentially the same for the two enzymes. However, recent saturation mutagenesis experiments in this laboratory have shown that the evolutionarily variable active-site residues in TK can modulate the use of non-natural substrates where positions are mutated to amino acid states outside the range found at a given position in nature. An example of this is the TK variants that can use benzaldehyde, all of which contained amino acids at the 259 position not found in any of the sequences examined in our studies. Thus, examination of evolutionarily variable active-site positions can be used as an excellent starting point for modulating the substrate specificity by rational evolutionary design.

Another observation was the fact that N58TK cannot use pyruvate. Since TK was shown in Chapter 3 to have diverged long ago from other pyruvate using TPP-dependent enzymes such as PDC, or slightly more recently from DXPS, it is unlikely that the variable residues within the active-sites of the TKs examined could potentially be responsible for pyruvate usage. Such a change in substrate specificity also requires mechanistic change as discussed in Chapter 5.

The S385GTK represents the first reported pyruvate utilising TK. Such an enzyme has great industrial potential, since pyruvate is ~230 times cheaper than β-HPA, as discussed in Chapter 5. The fact that a single mutation was sufficient to overcome the twin hurdles of producing HETPP from pyruvate and TPP, and eliminating the release of free acetaldehyde from TK-bound HETPP, is surprising. However, the proximity of Ser385 to the C-1 hydroxyl group of L-erythrulose in the *Eco*TK active-site shows how pyruvate usage may be a matter of substrate interaction with the 385 position. It would appear that Gly385 allows the HETPP to undergo the transferase reaction. If the C-1 hydroxyl group is the key to pyruvate discrimination however, it does not explain why previous studies, where the His100 was mutated by Oliver Miller in this laboratory were unsuccessful (the corresponding His103 is most proximal to the C-1 hydroxyl of the DHETPP intermediate in *Sce*TK). As discussed in Chapter 5, this may be due to GA being a poor acceptor substrate in TK mutants where the C-1-hydroxyl interaction is altered. Perhaps the use of an alternative acceptor substrate, lacking the C-1 hydroxyl group may lead to faster reaction rates for S385G.

The S385GTK would make an excellent starting point for production of a more efficient pyruvate–using TK enzyme. In general, Chapter 5 demonstrates how using the structurally equivalent regions and alignments defined in Chapter 3, along with modelling studies, a successful variant is produced, with catalytic properties outside the usual TK spectrum.

While Chapters 3, 4 and 5, by nature of their focus on the active-site were concerned with the PP and Pyr domains, Chapter 6 explores the TKC domain of TK. Removal of this domain results in a mutant with elevated activity for the β-HPA + GA reaction. This is important since it supports the previous notion that the TKC domain may be regulatory [7]. If the TKC domain is slowing the rates of biotransformations and if a shorter TK is cheaper to express on larger scales, then its removal is of great industrial interest.

Characterisation of the role of the TKC domain, which has no homology with any other domains in the sequence databases, would be of interest. It is likely that detailed structural studies will be necessary to elucidate its function.

Since the activity of G540StopTK was so high, it was decided to further truncate the TK molecule. The characterisation of these variants showed how the Pyr domain can be truncated without loss of activity, although effects on stereospecificity and preference for phosphorylated substrates were not explored. These characteristics would make interesting functional studies. Currently, our laboratory, in conjunction with the UCL chemistry department, is testing variants for stereoselectivity. The H461StopTK will be included in this study, since it is active and yet the Asp469, proposed as responsible for stereoselectivity has been removed. Since a significant portion of the Pyr domain can be removed without losing activity, it is tempting to hypothesise that the role of the Pyr domain is in controlling catalytic subtleties such as substrate specificity and stereoselectivity rather than the core TK catalytic mechanism. Further truncation of the Pyr domain would result in the loss of residues known to be involved in the binding of TPP. Should such mutants retain activity, then perhaps it could be deduced that the PP domain is the most important for TK activity and is by extension, similar to the progenitor of the PP and Pyr domains. Future studies on the TK stop mutants in this laboratory will focus on the stability of the truncated TK variants. In the case of the G540StopTK, stability is perhaps unlikely to be an issue, due to the relatively few points of contact between the TKC domains in the TK dimer. For the mutants where regions of the Pyr domain are removed, the affect on stability is more likely to be an issue.

Overall, the studies in this thesis show that a combination of sequence alignments, phylogenetic studies and reconstruction, as well as the use of structural information can be employed to successfully alter substrate specificity in enzymes such as TK. The use of such studies within the TK phylogeny proposes residues highly important for substrate specificity. While mutating these residues to amino acids within the natural range observed for the enzyme can alter substrate specificity, their potential is fully realised where saturation mutagenesis is performed. To alter the mechanism of TK, it was necessary to mutate residues that were highly conserved in the active-site, which showed the usefulness of comparing between enzymes with different domain arrangements and modes of catalysis. While focussing on the catalytic centre is useful for modulating substrate specificity and increasing activity in some cases, the non-catalytic domains cannot be ignored as illustrated for the TKC domain. Such domains can have dramatic effects on the rate of catalysis and can potentially be manipulated in biotransformations.

Eventually, as information from these as well as other TK studies increases, a "molecular toolbox" will be assembled, where the knowledge of evolutionary relationships, catalytically important residues and the effect of domain arrangements will allow the user to engineer enzymes with a high degree of predictability.

References

1. Racker,E., de la Haba,G. & Leder,J.G. Thiaminpyrophosphate, a coenzyme of transketolase. *J. Am. Chem. Soc.* **75**, 1010-1011 (1953).

2. Lindqvist,Y., Schneider,G., Ermler,U. & Sundstrom,M. Three-dimensional structure of transketolase, a thiamine diphosphate dependent enzyme, at 2.5 Å resolution. *EMBO J.* **11**, 2373-2379 (1992).

3. Nikkola,M., Lindqvist,Y. & Schneider,G. Refined structure of transketolase from *Saccharomyces cerevisiae* at 2.0 Å resolution. *J. Mol. Biol.* **238**, 387-404 (1994).

4. Cavaliere,S.W., Neet,K.E. & Sable,H.Z. Enzymes of pentose biosynthesis. The quaternary structure and reacting form of transketolase from baker's yeast. *Arch. Biochem. Biophys.* **171**, 527-532 (1975).

5. Sundstrom,M., Lindqvist,Y., Schneider,G., Hellman,U. & Ronne,H. Yeast TKL1 gene encodes a transketolase that is required for efficient glycolysis and biosynthesis of aromatic amino acids. *J. Biol. Chem.* **268**, 24346-24352 (1993).

6. Muller,Y.A. & Schulz,G.E. Structure of the thiamine and flavin dependent enzyme pyruvate oxidase. *Science* **259**, 965-967 (1993).

7. Schneider,G. & Lindqvist,Y. Crystallography and mutagenesis of transketolase: mechanistic implications for enzymatic thiamin catalysis. *Biochim. Biophys. Acta* **1385**, 387-398 (1998).

8. Meshalkina,L., Nilsson,U., Wikner,C., Kostikowa,T. & Schneider,G. Examination of the thiamin diphosphate binding site in yeast transketolase by site-directed mutagenesis. *Eur. J. Biochem.* **244**, 646-652 (1997).

9. Esakova,O.A., Meshalkina,L.E., Golbik,R., Hubner,G. & Kochetov,G.A. Donor substrate regulation of transketolase. *Eur. J. Biochem.* **271**, 4189-4194 (2004).

10. Sprenger,G. & Pohl,M. Synthetic potential of thiamin diphosphate-dependent enzymes. *Journal of Molecular Catalysis B: Enzymatic* **6**, 145-159 (1998).

11. Schneider,G. & Lindqvist,Y. Enzymatic Thiamine catalysis: Mechanistic implications from the three dimensional structure of transketolase. *Bioorganic Chemistry* **21**, 109-117 (1993).

12. Muller,Y.A., Lindqvist,Y., Furey,W., Schulz,G.E., Jordan,F. & Schneider,G. A thiamin diphosphate binding fold revealed by comparison of the crystal structures of transketolase, pyruvate oxidase and pyruvate decarboxylase. *Structure* **1**, 95-103 (1993).

13. Dyda,F., Furey,W., Swaminathan,S., Sax,M., Farrenkopf,B. & Jordan,F. Catalytic centers in the thiamin diphosphate dependent enzyme pyruvate decarboxylase at 2.4-A resolution. *Biochemistry* **32**, 6165-6170 (1993).

14. Jordan,F., Nemeria,N.S., Zhang,S., Yan,Y., Arjunan,P. & Furey,W. Dual catalytic apparatus of the thiamin diphosphate coenzyme: acid-base via the 1',4'-

iminopyrimidine tautomer along with its electrophilic role. *J. Am. Chem. Soc.* **125**, 12732-12738 (2003).

15. Hawkins,C.F., Borges,A. & Perham,R.N. A common structural motif in thiamin pyrophosphate-binding enzymes. *FEBS Lett.* **255**, 77-82 (1989).

16. Schenk,G., Layfield,R., Candy,J.M., Duggleby,R.G. & Nixon,P.F. Molecular evolutionary analysis of the thiamine-diphosphate-dependent enzyme, transketolase. *J. Mol. Evol.* **44**, 552-572 (1997).

17. Arjunan,P., Umland,T., Dyda,F., Swaminathan,S., Furey,W., Sax,M., Farrenkopf,B., Gao,Y., Zhang,D. & Jordan,F. Crystal structure of the thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast *Saccharomyces cerevisiae* at 2.3 Å resolution. *J. Mol. Biol.* **256**, 590-600 (1996).

18. Hasson,M.S., Muscate,A., McLeish,M.J., Polovnikova,L.S., Gerlt,J.A., Kenyon,G.L., Petsko,G.A. & Ringe,D. The crystal structure of benzoylformate decarboxylase at 1.6 Å resolution: diversity of catalytic residues in thiamin diphosphate-dependent enzymes. *Biochemistry* **37**, 9918-9930 (1998).

19. Candy,J.M. & Duggleby,R.G. Structure and properties of pyruvate decarboxylase and site-directed mutagenesis of the *Zymomonas mobilis* enzyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1385**, 323-338 (1998).

20. Diefenbach,R.J., Candy,J.M., Mattick,J.S. & Duggleby,R.G. Effects of substitution of aspartate-440 and tryptophan-487 in the thiamin diphosphate binding region of pyruvate decarboxylase from *Zymomonas mobilis*. *FEBS Lett.* **296**, 95-98 (1992).

21. Bolte,J., Demuynck,C. & Samaki,H. Utilization of enzymes in organic chemistry: Transketolase catalyzed synthesis of ketoses. *Tetrahedron Letters* **28**, 5525-5528 (1987).

22. Hecquet,L., Demuynck,C., Schneider,G. & Bolte,J. Enzymatic synthesis of ketoses: study and mosification of the substrate specificity of transketolase from *Saccharomyces cerevisiae. Journal of Molecular Catalysis B: Enzymatic* **11**, 771-776 (2001).

23. Kobori,Y., Myles,D.C. & Whitesides,G.M. Substrate specificity and carbohydrate synthesis using transketolase. *J. Org. Chem.* **57**, 5899-5907 (1992).

24. Myles,D.C., AndrulisIII,P.J. & George,M. A transketolase-based synthesis of (+)-exo-brevicomin. *Tetrahedron Letters* **32**, 4835-4838 (1991).

25. Kochetov,G.A. Transketolase: structure and mechanism of action. *Biokhimiia.* **51**, 2010-2029 (1986).

26. Heinrich,P.C., Steffen,H., Janser,P. & Wiss,O. Studies on the reconstitution of apotransketolase with thiamine pyrophosphate and analogs of the coenzyme. *Eur. J. Biochem.* **30**, 533-541 (1972).

27. Wikner,C., Meshalkina,L., Nilsson, U., Backstrom,S., Lindqvist,Y. & Schneider,G. His103 in yeast transketolase is required for substrate recognition and catalysis. *Eur. J. Biochem.* **233**, 750-755 (1995).

28. Wikner,C., Nilsson,U., Meshalkina,L., Udekwu,C., Lindqvist,Y. & Schneider,G. Identification of catalytically important residues in yeast transketolase. *Biochemistry* **36**, 15643-15649 (1997).

29. Nilsson,U., Meshalkina,L., Lindqvist,Y. & Schneider,G. Examination of substrate binding in thiamin diphosphate-dependent transketolase by protein crystallography and site-directed mutagenesis. *J. Biol. Chem.* **272**, 1864-1869 (1997).

30. Nilsson,U., Hecquet,L., Gefflaut,T., Guerard,C. & Schneider,G. Asp477 is a determinant of the enantioselectivity in yeast transketolase. *FEBS Lett.* **424**, 49-52 (1998).

31. Sprenger,G.A., Schorken,U., Sprenger,G. & Sahm,H. Transketolase A of *Escherichia coli* K12. Purification and properties of the enzyme from recombinant strains. *Eur. J. Biochem.* **230**, 525-532 (1995).

32. Gyamerah,M. & Willetts,A.J. Kinetics of overexpressed transketolase from *Escherichia coli* JM 107/pQR 700. *Enzyme Microb. Technol.* **20**, 127-134 (1997).

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

34. Singleton,C.K., Wang,J.J., Shan,L. & Martin,P.R. Conserved residues are functionally distinct within transketolases of different species. *Biochemistry* **35**, 15865-15869 (1996).

35. Tittmann,K., Golbik,R., Uhlemann,K., Khailova,L., Schneider,G., Patel,M., Jordan,F., Chipman,D.M., Duggleby,R.G. & Hübner,G. NMR Analysis of Covalent Intermediates in Thiamin Diphosphate Enzymes. *Biochemistry* **42**, 7885-7891 (2003).

36. Meshalkina,L.E., Neef,H., Tjaglo,M.V., Schellenberger,A. & Kochetov,G.A. The presence of a hydroxyl group at the C-1 atom of the transketolase substrate molecule is necessary for the enzyme to perform the transferase reaction. *FEBS Letters* **375**, 220-222 (1995).

37. Waltham,M. Studies on dihydrofolate reductase and transketolase. PhD Thesis, The University of Queensland, Brisbane, Australia. (1990).

38. Villafranca,J.J. & Axelrod,B. Heptulose synthesis from nonphosphorylated aldoses and ketoses by spinach transketolase. *J. Biol. Chem.* **246**, 3126-3131 (1971).

39. Usmanov,R.A. & Kochetov,G.A. [Function of the arginine residue in the active center of baker's yeast transketolase]. *Biokhimiia.* **48**, 772-781 (1983).

40. Mocali,A. & Paoletti,F. Transketolase from human leukocytes. Isolation, properties and induction of polyclonal antibodies. *Eur. J. Biochem.* **180**, 213-219 (1989).

41. Paoletti,F. Purification and properties of transketolase from fresh rat liver. *Arch. Biochem Biophys.* **222**, 489-496 (1983).

42. Patel,R.N. *Stereoselective Biocatalysis.*, pp. 87-130 (Marcel Dekker Inc., New York, 2000).

43. FDA. Policy statement for the development of new stereoisomeric drugs. http://www.fda.gov/cder/guidance/stereo.htm. (1992)
Electronic Citation

44. Stinson,S.C. Chiral Drugs. *Chem. Eng. News* **78**, 55-78 (2000).

45. Comprehensive organic synthesis. Pergamon Press , New York (1991).

46. Seyden-Penne,J. Chiral Auxiliaries and Ligands in Asymmetric Synthesis. Wiley-Interscience, (1995).

47. Turner,J. Applications of transketolases in organic synthesis. *Curr. Opin. in Biotech.* **11**, 527-531 (2000).

48. Datta,A.G. & Racker,E. Mechanism of action of transketolase. I. Properties of the crystalline yeast enzyme. *J. Biol. Chem.* **236**, 617-623 (1961).

49. Effenberger,F., Null,V. & Ziegler,T. Preparation of optically pure L-2-hydroxyaldehydes with yeast transketolase. *Tetrahedron Letters* **33**, 5157-5160 (1992).

50. Effenberger,F., Straub,G. & Null,V. Enzym-katalysierte Reaktionen, 14. Stereoselektive Darstellung von Thiozuckern aus achiralen Vorstufen mittels Enzymen. *Liebigs Ann. Chem.* 1297-1301 (1992).

51. Demuynck,C., Bolte,J., Hecquet,L. & Dalmas,V. Enzyme-catalyzed synthesis of carbohydrates: synthetic potential of transketolase. *Tetrahedron Letters* **32**, 5085-5088 (1991).

52. Hobbs,G.R., Lilly, M.D., Turner, N.J., Ward, J.M., Willets, A.J., Woodley, J.M. Enzyme-catalysed carbon-carbon bond formation: use of transketolase from *Escherichia coli. J. Chem. Soc Perkin Trans* **1**, 165-166 (1993).

53. Morris,K.G., Smith, M.E.B., Turner N.J., Lilly M.D., Mitra R.K., Woodley J.M. Transketolase from *Escherichia coli*: A practical procedure for using the biocatalyst for asymmetric carbon-carbon bond synthesis. *Tetrahedron: Asymmetry* **7**, 2185-2188 (1996).

54. Schorken,U., Sahm,H. & Sprenger,G.A. Biochemistry and physiology of thiamin diphosphate enzymes. *A&C Intemann Verlag*, Prien, Germany (1996).

55. International Union of Biochemistry. *Enzyme Nomenclature: Recommendations 1964 of the International Union of Biochemistry.* Elsevier, Amsterdam. Chapter 4. (1965).

56. Rohmer,M., Knani,M., Simonin,P., Sutter,B. & Sahm,H. Isoprenoid biosynthesis in bacteria: a novel pathway for the early steps leading to isopentenyl diphosphate. *The Biochemical Journal* **295**, 517-524 (1993).

57. Demuynck,C., Bolte,J., Hecquet,L. & Samaki,H. Enzymes as reagents in organic chemistry: transketolase-catalysed synthesis of -[1,2-13C2]xylulose. *Carbohydrate Research* **206**, 79-85 (1990).

58. Effenberger,F. & Null,V. Enzym-katalysierte Reaktionen, 13. Eine neue, effiziente Synthese von Fagomin. *Liebigs Ann. Chem.* 1211-1212 (1992).

59. Ziegler,T., Straub,A. & Effenberger,F. Enzyme-Catalyzed Synthesis of 1-Deoxymannojirimycin, 1-Deoxynojirimycin, and 1,4-Dideoxy-1,4-imino-D-arabinitol. *Angewandte Chemie International Edition in English* **27**, 716-717 (1988).

60. Hecquet,L., Bolte,J. & Demuynck,C. Enzymatic synthesis of "natural-labeled" 6-deoxy-L-sorbose precursor of an important food flavor. *Tetrahedron* **52**, 8223-8232 (1996).

61. Uerard,C. Alphand,V. Archelas,A. Demuynck,C. Hecquet,L. Furstoss,R. Bolte,J. Transketolase mediated synthesis of 4-deoxy--fructose 6-phosphate by epoxide hydrolase-catalysed resolution of 1,1-diethoxy-3,4-epoxybutane. *Eur. J. Org. Chem.* 3399-3402 (1999).

62. Zimmermann,F.T., Schneider,A., Schorken,U., Sprenger,G.A. & Fessner,W.D. Efficient multi-enzymatic synthesis of D-xylulose 5-phosphate. *Tetrahedron: Asymmetry* **10**, 1643-1646 (1999).

63. Humphrey,A.J., Parsons,S.F., Smith,M.E.B. & Turner,N.J. Synthesis of a novel N-hydroxypyrrolidine using enzyme catalysed asymmetric carbon-carbon bond synthesis. *Tetrahedron Letters* **41**, 4481-4485 (2000).

64. Draths,K.M. & Frost,J.W. Synthesis using plasmid-based biocatalysis: plasmid assembly and 3-deoxy-D-arabino-heptulosonate production. *J. Am. Chem. Soc.* **112**, 1657-1659 (1990).

65. French,C. & Ward,J.M. Improved production and stability of *E. coli* recombinants expressing transketolase for large scale biotransformation. *Biotechnology Letters* **17**, 247-252 (1995).

66. Hobbs,G.R., Mitra,R.K., Chauhan,R.P., Woodley,J.M. & Lilly,M.D. Enzyme-catalysed carbon-carbon bond formation: Large-scale production of Escherichia coli transketolase. *Journal of Biotechnology* **45**, 173-179 (1996).

67. Mitra,R.K., Woodley,J.M. & Lilly,M.D. *Escherichia coli* transketolase-catalyzed carbon-carbon bond formation: biotransformation characterization for reactor evaluation and selection. *Enzyme Microb. Tech.* **22**, 64-70 (1998).

68. Chauhan,R.P., Woodley,J.M. & Powell,L.W. *In situ* product removal from *E. coli* transketolase-catalyzed biotransformations. *Ann. N. Y. Acad. Sci.* **799**, 545-554 (1996).

69. Vasic-Racki,D., Bongs,J., Schorken,U., Sprenger,G.A. & Liese,A. Modeling of reaction kinetics for reactor selection in the case of L-erythrulose synthesis. *Bioprocess. Biosyst. Eng.* **25**, 285-290 (2003).

70. R.B.Sterzel, M.Semar, E.T.Lonergan, G.Treser & K.Lange. Relationship of nervous tissue transketolase to the neuropathy in chronic uremia. *The Journal of Clinical Investigation* 50(11): 2295–2304 (1971).

71. Tombaccini,D., Mocali,A. & Paoletti,F. Characteristic transketolase alterations in dermal fibroblasts of Alzheimer patients are modulated by culture conditions. *Exp. Mol. Pathol.* **60**, 140-146 (1994).

72. Cascante M, Centelles JJ, Veech RL, Lee WN & Boros LG. Role of thiamin (vitamin B-1) and transketolase in tumor cell proliferation. *Nutrition and Cancer* **36**, 150-154 (2000).

73. Arigoni,D., Sagner,S., Latzel,C., Eisenreich,W., Bacher,A., Zenk,M.H... Terpenoid biosynthesis from 1-deoxy-D-xylulose in higher plants by intramolecular skeletal rearrangement. *PNAS* **94**, 10600-10605 (1997).

74. Broers,S.T.J. (1994). Eidgenössiche Technische Hochschule, Zürich, Switzerland.
Thesis/Dissertation

75. Rohmer,M., Seemann,M., Horbach,S., Bringer-Meyer,S. & Sahm,H. Glyceraldehyde 3-Phosphate and Pyruvate as Precursors of Isoprenic Units in an Alternative Non-mevalonate Pathway for Terpenoid Biosynthesis. *J. Am. Chem. Soc.* **118**, 2564-2566 (1996).

76. Zeidler,J.G., Lichtenthaler,H.K., May,U. & Lichtenthaler,F.W. *Zeitung Naturforsch.* **52c**, 15-23 (1997).

77. Yanase,H. Okuda,M., Kita,K., Sato,Y., Shibata,K., Sakai,Y. & Kato,N.. Enzymatic preparation of [1,3-13C]dihydroxyacetone phosphate from [13C]methanol and hydroxypyruvate using the methanol-assimilating system of methylotrophic yeasts. *Applied Microbiology and Biotechnology* **43**, 228-234 (1995).

78. Gottschalk,G. Bacterial Metabolism. Springer-Verlag, New York, (1986).

79. Thompson,J.D., Higgins,D.G. & Gibson,T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680 (1994).

80. Schorken,U. & Sprenger,G.A. Thiamin-dependent enzymes as catalysts in chemoenzymatic syntheses. *Biochim. Biophys. Acta - Protein Structure and Molecular Enzymology* **1385**, 229-243 (1998).

81. Goldberg,M., Fessenden,J.M. & Racker,E. Phosphoketolase. *Methods Enzymol.* **9**, 515-520 (1966).

82. Namba,Y., Yoshizawa,K., Ejima,A., Hayashi,T. & Kaneda,T. Coenzyme A- and Nicotinamide Adenine Dinucleotide-dependent Branched Chain α-Keto Acid Dehydrogenase. *J. Biol. Chem.* **244**(16), 4437-4447 (1969).

83. Konig,S. Subunit structure, function and organisation of pyruvate decarboxylases from various organisms. *Biochim. Biophys Acta - Protein Structure and Molecular Enzymology* **1385**, 271-286 (1998).

84. Iding,H., Siegert,P., Mesch,K. & Pohl,M. Application of alpha-keto acid decarboxylases in biotransformations. *Biochim. Biophys. Acta* **1385**, 307-322 (1998).

85. Pohl,M. Protein design on pyruvate decarboxylase (PDC) by site-directed mutagenesis. Application to mechanistical investigations, and tailoring PDC for the use in organic synthesis. *Adv. Biochem. Eng. Biotechnol.* **58**, 16 (1997).

86. Bringer-Meyer,S. & Sahm,H. Acetoin and phenylacetylcarbinol formation by the pyruvate decarboxylase of *Zymomonas mobilis* and *Saccharomyces carlsbergensis*. *Biocatalysis* **1**, 321-331 (1988).

87. Bruhn,H., Pohl,M., Gratzinger,J. & Kula,M.R. The replacement of Trp392 by alanine influences the decarboxylase/carboligase activity and stability of pyruvate decarboxylase from *Zymomonas mobilis*. *European Journal of Biochemistry* **234**, 650-655 (1995).

88. Bruhn,H., Pohl,M., Mesch,K. & Kula,M.R. Verfahren zur Gewinnung von Acyloinen, dafür geeignete Pyruvatdecarboxylase sowie deren Herstellung und DNA-Sequenz des für diese kodierenden PDC-Gens. *German patent number:* 195 23 269, 0-41. (1995). Germany.

89. Bournemann,S. Crout,D.H.G., Dalton,H., Hutchinson,D.W., Dean,G., Thomspon,N., Turner,M.M. Stereochemistry of the formation of lactaldehyde and acetoin produced by the pyruvate decarboxylases of yeast (*Saccharomyces* sp.) and *Zymomonas mobilis* : different Boltzmann distributions between bound forms of the electrophile, acetaldehyde, in the two enzymatic reactions. *J. Chem. Soc. Perkin Trans* **1**, 309 (1993).

90. Chen,G.C. & Jordan,F. Brewers' yeast pyruvate decarboxylase produces acetoin from acetaldehyde: a novel tool to study the mechanism of steps subsequent to carbon dioxide loss. *Biochemistry* **23**, 143-149 (1984).

91. Crout,D.H.G., Littlechild,J.A. & Morrey,S.M. Acetoin metabolism: stereochemistry of the acetoin produced by the pyruvate decarboxylase of wheat germ and by the -acetolactate decarboxylase of *Klebsiella aerogenes*. *J. Chem. Soc. Perkin Trans* **1**, 105-108 (1986).

92. Koga,J., Adachi,T. & Hidaka,H. Purification and characterization of indolepyruvate decarboxylase. A novel enzyme for indole-3-acetic acid biosynthesis in *Enterobacter cloacae*. *J. Biol. Chem.* **267**, 15823-15828 (1992).

93. Barak,Z., Calvo,J.M. & Schloss,J.V. Acetolactate synthase isozyme III from *Escherichia coli*. *Methods In Enzymology* **166**, 455-458 (1988).

94. Eoyang,L. & Silverman,P.M. Purification and assays of acetolactate synthase I from *Escherichia coli* K12. *Methods In Enzymology* **166**, 435-445 (1988).

95. Schneider,S., Mohamed,M.E.S. & Fuchs,G. Anaerobic metabolism of *L* - phenylalanine via benzoyl-CoA in the denitrifying bacterium *Thauera aromatica*. *Archives of Microbiology* **168**, 310-320 (1997).

96. Ward,O.P. & Singh,A. Enzymatic asymmetric synthesis by decarboxylases. *Current Opinion in Biotechnology* **11**, 520-526 (2000).

97. Graupner,M., Xu,H. & White,R.H. Identification of the Gene Encoding Sulfopyruvate Decarboxylase, an Enzyme Involved in Biosynthesis of Coenzyme M. *J. Bacteriol.* **182**, 4862-4867 (2000).

98. Zhang,G., Dai,J., Lu,Z. & Dunaway-Mariano,D. The Phosphonopyruvate Decarboxylase from *Bacteroides fragilis*. *The J. Biol. Chem.* **278**, 41302-41308 (2003).

99. Wilcocks,R., Ward,O.P., Collins,S., Dewdnwy,N.J., Hong,Y. & Prosen,E. Acyloin formation by benzoylformate decarboxylase from *Pseudomonas putida*. *Applied And Environmental Microbiology* **58**, 1699-1704 (1992).

100. Iding,H. PhD thesis, University of Düsseldorf (1998).

101. Koland JG,G.RB. Proximity of reactive cysteine residue and flavin in *Escherichia coli* pyruvate oxidase as estimated by fluorescence energy transfer. *Biochemistry*. 21(18), 4438-4442. 1982.

102. Bertagnolli,B.L. & Hager,L.P. Role of Flavin in Acetoin Production by Two Bacterial Pyruvate Oxidases. *Arch. Biochem. Biophys.* **300**, 364-371 (1993).

103. Chang,Y.Y., Wang,A.Y. & Cronan,J.E.,Jr. Expression of *Escherichia coli* pyruvate oxidase (PoxB) depends on the sigma factor encoded by the rpoS(katF) gene. Mol. Microbiol. 11(6), 1019-1028. 1994.

104. Abdel-Hamid,A.M., Attwood,M.M. & Guest,J.R. Pyruvate oxidase contributes to the aerobic growth efficiency of *Escherichia coli*. *Microbiology* **147**, 1483-1498 (2001).

105. Kiuchi K,H.LP. Reconstitution of the lipid-depleted pyruvate oxidase system of *Escherichia coli*: the palmitic acid effect. *Arch. Biochem. Biophys.* **233**, 776-784 (1984).

106. Grabau C,C.J.Jr. *In vivo* function of *Escherichia coli* pyruvate oxidase specifically requires a functional lipid binding site. *Biochemistry* 25(13), 3748-3751. (1986).

107. Bertagnolli,B.L. & Hager,L.P. Activation of *Escherichia coli* pyruvate oxidase enhances the oxidation of hydroxyethylthiamin pyrophosphate. *J. Biol. Chem.* **266**, 10168-10173 (1991).

108. Mather,M.W. & Gennis,R.B. Spectroscopic studies of pyruvate oxidase flavoprotein from *Escherichia coli* trapped in the lipid-activated form by cross-linking. *J. Biol. Chem.* **260**, 10395-10397 (1985).

109. Mather,M.W. & Gennis,R.B. Kinetic studies of the lipid-activated pyruvate oxidase flavoprotein of *Escherichia coli. J. Biol. Chem.* **260**, 16148-16155 (1985).

110. F E Dailey and J E Cronan,J. Acetohydroxy acid synthase I, a required enzyme for isoleucine and valine biosynthesis in *Escherichia coli* K-12 during growth on acetate as the sole carbon source. *J. Bacteriol.* **165**, 453-460 (1986).

111. Chang,Y.Y., Wang,A.Y. & Cronan,J.E., Jr. Molecular cloning, DNA sequencing, and biochemical analyses of *Escherichia coli* glyoxylate carboligase. An enzyme of the acetohydroxy acid synthase-pyruvate oxidase family. *J. Biol. Chem.* **268**, 3911-3919 (1993).

112. Cromartie,T.H. & Walsh,C.T. *Escherichia coli* glyoxalate carboligase. Properties and reconstitution with 5-deazaFAD and 1,5-dihydrodeazaFADH2. *J. Biol. Chem.* **251**, 329-333 (1976).

113. Demir,A.S., Sesenoglu,O., Erin,E., Hosrik,B., Pohl,M., Janzen,E., Kolter,D., Feldmann,R. & Dunkelmann,P.. Enantioselective Synthesis of -Hydroxy Ketones via Benzaldehyde Lyase-Catalyzed CC Bond Formation Reaction. *J. Am. Chem. Soc.* **244**, 96-103 (2001).

114. Meile,L., Rohr,L.M., Geissmann,T.A., Herensperger,M. & Teuber,M. Characterization of the D-Xylulose 5-Phosphate/D-Fructose 6-Phosphate Phosphoketolase Gene (xfp) from *Bifidobacterium lactis. J. Bacteriol.* **183**, 2929-2936 (2001).

115. Cavazza,C., Contreras-Martel,C., Pieulle,L., Chabriere,E., Hatchikian,E.C. & Fontecilla-Camps,J.C. Flexibility of Thiamine Diphosphate Revealed by Kinetic Crystallographic Studies of the Reaction of Pyruvate-Ferredoxin Oxidoreductase with Pyruvate. *Structure* **14**, 217-224 (2006).

116. Todd,A.E., Orengo,C.A. & Thornton,J.M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113-1143 (2001).

117. Kazlauskas,R.J. Enhancing catalytic promiscuity for biocatalysis. *Current Opinion in Chemical Biology* **9**, 195-201 (2005).

118. Petrov,D.A. Evolution of genome size: new approaches to an old problem. *Trends in Genetics* **17**, 23-28 (2001).

119. Jeffery,C.J. Moonlighting proteins: old proteins learning new tricks. *Trends in Genetics* **19**, 415-417 (2003).

120. Jeffery,C.J. Multifunctional proteins: examples of gene sharing. *Annals of Medicine* **35**, 28-35 (2003).

121. Jeffery,C.J. Moonlighting proteins. *Trends in Biochemical Sciences* **24**, 8-11 (1999).

122. Wool,I.G. Extraribosomal functions of ribosomal proteins. *Trends in Biochemical Sciences* **21**, 164-165 (1996).

123. Jeffery,C.J. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Current Opinion in Structural Biology* **14**, 663-668 (2004).

124. Aharoni,A. Gaidukov,L., Khersonsky,O., Gould,S.M., Roodveldt,C. & Tawfik,D.S.. The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73-76 (2005).

125. Branneby,C., Carlqvist,P., Magnusson,A., Hult,K., Brinck,T. & Berglund,P. Carbon-Carbon Bonds by Hydrolytic Enzymes. *J. Am. Chem. Soc.* **125**, 874-875 (2003).

126. Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nat. Genet.* **30**, 406-410 (2002).

127. Ulmer,K.M. Protein Engineering. *Science* **219**, 666-671 (1983).

128. Cedrone,F., Menez,A. & Quemeneur,E. Tailoring new enzyme functions by rational redesign. *Current Opinion in Structural Biology* **10**, 405-410 (2000).

129. Hult,K. & Berglund,P. Engineered enzymes for improved organic synthesis. *Current Opinion in Biotechnology* **14**, 395-400 (2003).

130. Dalby,P.A. Optimising enzyme function by directed evolution. *Current Opinion in Structural Biology* **13**, 500-505 (2003).

131. Shao,Z. & Arnold,F.H. Engineering new functions and altering existing functions. *Current Opinion in Structural Biology* **6**, 513-518 (1996).

132. Lingen,B., Grotzinger,J., Kolter,D., Kula,M.R. & Pohl,M. Improving the carboligase activity of benzoylformate decarboxylase from *Pseudomonas putida* by a combination of directed evolution and site-directed mutagenesis. *Protein Eng.* **15**, 585-593 (2002).

133. Bessler,C., Schmitt,J., Maurer,K.H. & Schmid,R.D. Directed evolution of a bacterial {alpha}-amylase: Toward enhanced pH-performance and higher specific activity. *Protein Sci.* **12**, 2141-2149 (2003).

134. Zaccolo,M., Williams,D.M., Brown,D.M. & Gherardi,E. An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. *J. Mol. Biol.* **255**, 589-603 (1996).

135. Fisher,R.A. The genetical theory of Natural Selection. Oxford University Press , Oxford, UK. (1930).

136. Stemmer,W.P.C. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389-391 (1994).

137. Doolittle,W.F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).

138. Morley,K.L. & Kazlauskas,R.J. Improving enzyme properties: when are closer mutations better? *Trends in Biotechnology* **23**, 231-237 (2005).

139. Page,R. & Holmes,E. Molecular Evolution: A Phylogenetic approach. (Blackwell Science,1998).

140. Nei,M. & Kumar,S. Molecular Evolution and Phylogenetics. Oxford University Press, USA, (2000).

141. Russo,C.A., Takezaki,N. & Nei,M. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13**, 525-536 (1996).

142. Asai,T., Zaporojets,D., Squires,C. & Squires,C.L. An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria.*PNAS U.S.A.* **96**, 1971-1976 (1999).

143. Kurland,C.G. Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.* **26**, 29-50 (1992).

144. Kurland,C.G., Canback,B. & Berg,O.G. Horizontal gene transfer: a critical view. *PNAS U.S.A.* **100**, 9658-9662 (2003).

145. Woese,C.R. Evolutionary questions: the "progenote". *Science* **247**, 789 (1990).

146. Woese,C.R. & Fox,G.E. The concept of cellular evolution. *J. Mol. Evol.* **10**, 1-6 (1977).

147. Woese,C.R. & Gupta,R. Are archaebacteria merely derived 'prokaryotes'? *Nature* **289**, 95-96 (1981).

148. Raymond,J., Zhaxybayeva,O., Gogarten,J.P., Gerdes,S.Y. & Blankenship,R.E. Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**, 1616-1620 (2002).

149. Snel,B., Bork,P. & Huynen,M.A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17-25 (2002).

150. Pauling,L. & Zuckerkandl,E. Chemical Paleogenetics. Molecular "Restoration Studies" of Extinct Forms of Life. *Acta Chemica Scandinavica* **17**, S9-S16 (1963).

151. Stackhouse,J., Presnell,S.R., McGeehan,G.M., Nambiar,K.P. & Benner,S.A. The ribonuclease from an extinct bovid ruminant. *FEBS Letters* **262**, 104-106 (1990).

152. Jermann,T.M., Opitz,J.G., Stackhouse,J. & Benner,S.A. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**, 57-59 (1995).

153. Adey,N.B., Tollefsbol,T.O., Sparks,A.B., Edgell,M.H. & Hutchison CA,I.I.I. Molecular Resurrection of an Extinct Ancestral Promoter for Mouse L1. *PNAS* **91**, 1569-1573 (1994).

154. Zhang,J. & Rosenberg,H.F. From the Cover: Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *PNAS* **99**, 5486-5491 (2002).

155. Chang,B.S., Jonsson,K., Kazmi,M.A., Donoghue,M.J. & Sakmar,T.P. Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19**, 1483-1489 (2002).

156. Gaucher,E.A., Thomson,J.M., Burgan,M.F. & Benner,S.A. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**, 285-288 (2003).

157. Benner,S.A. The past as the key to the present: Resurrection of ancient proteins from eosinophils. *PNAS* **99**, 4760-4761 (2002).

158. Chandrasekharan,U.M., Sanker,S., Glynias,M.J., Karnik,S.S. & Husain,A. Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science* **271**, 502-505 (1996).

159. Dykhuizen,D.E. & Dean,A.M. predicted fitness changes along an environmental gradient. *Evol. Ecol.* **8**, 524-541 (1994).

160. Krishnan,S., Hall,B.G. & Sinnott,M.L. Catalytic consequences of experimental evolution: catalysis by a 'third-generation' evolvant of the second β-galactosidase of *Escerichia coli*, ebg^{abcde,} and by ebg^{abcd}, a 'second-generation' evolvant containing two supposedly 'kinetically silent' mutation. *Biochem. J.* **312**, 971-977 (1995).

161. Rosenzweig,R.F., Sharp,R.R., treves,D.S. & Adams,J. Microbial evolution in a simple unstructured environment : genetic differentiation in *Escherichia coli*. *Genetics* **137**, 903-917 (1994).

162. Kimura,M. & Kimura,M. Population genetics, molecular evolution, and the neutral theory :selected papers. University of Chicago Press, Chicago (1994).

163. Uzzell,T. & Corbin,K.W. Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089-1096 (1971).

164. Johnson,N.L. & Kotz,S. Distribution in statistics: Discrete distributions. Houghton Mifflin, Boston, (1969).

165. Rzhetsky,A. & Nei,M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* **10**, 1073-1095 (1993).

166. Saitou,N. & Nei,M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).

167. Differences and similarities between the Protein 200 assay and SDS-PAGE. (2001). Agilent Technologies.

168. Kuschel,M. Protein sizing and analysis using the Agilent 2100 Bioanalyzer and Protein 200 LabChip® kit. (2000). Agilent Technologies.

169. The Bench Guide. (2001). Qiagen Ltd.

170. Brocklebank,S., Woodley,J.M. & Lilly,M.D. Immobilised transketolase for carbon-carbon bond synthesis: biocatalyst stability. *J. Mol. Cat. B - Enzymatic* **7**, 223-231 (1999).

171. Chauhan,R.P., Powell,L.W. & Woodley,J.M. Boron based separations for *in situ* recovery of L-erythrulose from transketolase-catalyzed condensation. *Biotechnol. Bioeng.* **56**, 345-351 (1997).

172. Mitra,R.K. & Woodley,J.M. A useful assay for transketolase in asymmetric syntheses. *Biotechnology Techniques* **10**, 167-172 (1996).

173. Woodley,J.M., Mitra,R.K. & Lilly,M.D. Carbon-carbon bond synthesis – reactor design and operation for transketolase-catalyzed biotransformations. *Ann. Ny. Acad. Sci.* **799**, 434-445 (1996).

174. Sambrook,J.E., Fristsch,F. & Maniatis,T. Molecular Cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. (1989).

175. Talarico,L.A., Ingram,L.O. & Maupin-Furlow,J.A. Production of the Gram-positive *Sarcina ventriculi* pyruvate decarboxylase in *Escherichia coli. Microbiology* **147**, 2425-2435 (2001).

176. Laemmli,U.K. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature* **227**, 680 – 685. (1970).

177. Corpet,F. Multiple sequence alignment with hierarchical clustering. *Nucl. Acids Res.* **22**, 10881-10890 (1988).

178. Duggleby,R.G. Domain Relationships in Thiamine Diphosphate-Dependent Enzymes. *Acc. Chem. Res.* ***39(8)**, 550-557* (2006).

179. Green,J.B.A. Pyruvate decarboxylase is like acetolactate synthase (ILV2) and not like the pyruvate dehydrogenase E1 subunit. *FEBS Letters* **246**, 1-5 (1989).

180. Rost,B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85-94 (1999).

181. Chung,S.Y. & Subbiah,S. A structural explanation for the twilight zone of protein sequence homology. *Structure.* **4**, 1123-1127 (1996).

182. Felsenstein,J. PHYLIP (Phylogeny Inference Package) version 3.5c. (1993).

183. Hall,T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.* 41, 95-98. (1999).

184. Kumar,S., Tamura,K., Jakobsen,I.B. & Nei,M. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244-1245 (2001).

185. Page,R.D. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357-358 (1996).

186. Chang,Y.Y. & Cronan,J.E., Jr. Conversion of *Escherichia coli* pyruvate oxidase to an 'alpha-ketobutyrate oxidase'. *Biochem. J.* **352 Pt 3**, 717-724 (2000).

187. Isupov,M.N., Rupprecht,M.P., Wilson,K.S., Dauter,Z. & Littlechild,J.A. Crystal Structure of *Escherichia coli* Transketolase. To be published (2003).

188. Schenk,G., Duggleby,R.G. & Nixon,P.F. Properties and functions of the thiamin diphosphate dependent enzyme transketolase. *The international Journal of Biochemistry and Cell Biology* **30**, 1297-1318 (1998).

189. Veitch,N.J. & Barrett,M.P. The Characterisation of a *Leishmania mexicana* Transketolase. PhD thesis, Department of Ibls, University of Glasgow, Glasgow, United Kingdom. (2004)

190. Kochetov,G.A. Transketolase from yeast, rat liver, and pig liver. *Methods Enzymol.* **90 Pt E**, 209-223 (1982).

191. Yang,Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).

192. DeLano,W.D. The PyMOL Molecular Graphics System. (2004).

193. Wang,L. & Schultz,P.G. Expanding the genetic code. *Angewandte Chemie (International Ed. In English)* **44**, 34-66 (2004).

194. Ochman,H. & Wilson,A.C. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *Journal of Molecular Evolution* **26**, 377 (1987).

195. Williams,J.F., Clark,M.G. & Blackmore,P.F. The fate of 14C in glucose 6-phosphate synthesized from [1-14C]Ribose 5-phosphate by enzymes of rat liver. *The Biochemical Journal* **176**, 241-256 (1978).

196. Horecker,B.L., Paoletti,F. & Williams,J.F. Occurrence and significance of octulose phosphates in liver. *Annals of The New York Academy of Sciences* **378**, 215-224 (1982).

197. Williams,J.F., Arora,K.K. & Longenecker,J.P. The pentose pathway: a random harvest. Impediments which oppose acceptance of the classical (F-type) pentose cycle for liver, some neoplasms and photosynthetic tissue. The case for the L-type pentose pathway. *The International Journal of Biochemistry* **19**, 749-817 (1987).

198. Wood,T. *The Pentose Phosphate Pathway.* Academic Press, Orlando, FL., USA (1985).

199. Datta,A.G. & Racker,E. Mechanism of action of transketolase. I. Properties of the crystalline yeast enzyme. *J. Biol. Chem.* **236**, 617-623 (1961).

200. Ward,J.M. Personal Communication. (2006).

201. Valdar,W.S. & Thornton,J.M. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313**, 399-416 (2001).

202. Meshalkina,L., Nilsson,U., Wikner,C., Kostikowa,T. & Schneider,G. Examination of the thiamin diphosphate binding site in yeast transketolase by site-directed mutagenesis. *Eur. J. Biochem.* **244**, 646-652 (1997).

203. Aucamp,J. PhD thesis, Department of Biochemical Engineering, University College London.(2005).

204. Fraser,C.M., Fraser,C.M., Norris,S.J., Weinstock,G.M., White,O., Sutton,G.G., Dodson,R., Gwinn,M., Hickey,E.K., Clayton,R., Ketchum,K.A., Sodergren,E., Hardham,J.M., McLeod,M.P., Salzberg,S., Peterson,J., Khalak,H., Richardson,D., Howell,J.K., Chidambaram,M., Utterback,T., McDonald,L., Artiach,P., Bowman,C., Cotton,M.D. & Venter,J.C.. . Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375-388 (1998).

205. Hibbert,E.G. Personal Communication. (2006).

206. Carrizo,S.F. Phylogenetic Trees: An Information Visualisation Perspective. *Proceedings of the second conference on Asia-Pacific bioinformatics* **29**, 315-320. (2004).

Appendix 1: Appendix for Chapter 3

A1.1 Choosing the TK sequences to study

During the initial choosing of TK sequences for analysis, the goal was not to choose between methods of phylogenetic inference, but rather to use several tree building methods to deduce the optimum number of amino acid sequences that should make up the TK alignment in order to resolve the mutations that have occurred during the course of evolution. In order to make mutagenesis possible, mutations must be resolved, preferably occurring one at a time between nodes. Even when one focussed on a subset of the total amino acid residues, mutations are regularly found to occur simultaneously. Thus, the optimal set of TK sequences was assembled in order to minimise the occurrence of such multiple simultaneous mutations.

A1.1.1 Round one – phylogenetic analysis and reconstruction of 45 TK sequences

Initially 66 TK sequences retrieved using the *BLAST* search were subject to phylogenetic analysis and reconstruction using the Phylip [182] program *ProML*, by Paul Dalby (N.B. in subsequent analyses the *ProML* program was used to generate phylogenetic trees only. In order to generate this tree, the *ProML* program needed to run for over 5 weeks. This analysis was run on a desktop computer with an Intel Pentium X processor.)

In order to reduce computational time, a subset of 45 sequences was chosen in such a way as to eliminate all putative TK sequences. An alignment of these 45 TK sequences was input into the Phylip programs *Protpars*, *ProML* and *Protdist* to generate, respectively, a parsimony tree, a ML tree and a distance matrix for analysis with the *Neighbour* program, which in turn produced a phylogenetic tree using the NJ method. These 3 trees are shown in Figures A1.1, A1.2 and A1.3.

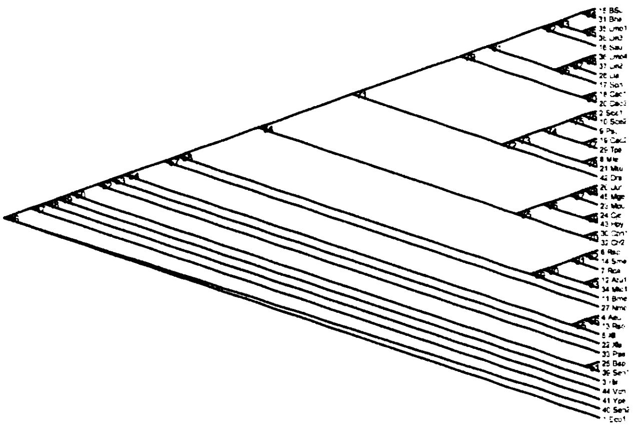


Figure A1.1 : Neighbour joining tree for 45 TK sequences, as described in section A1.1.1

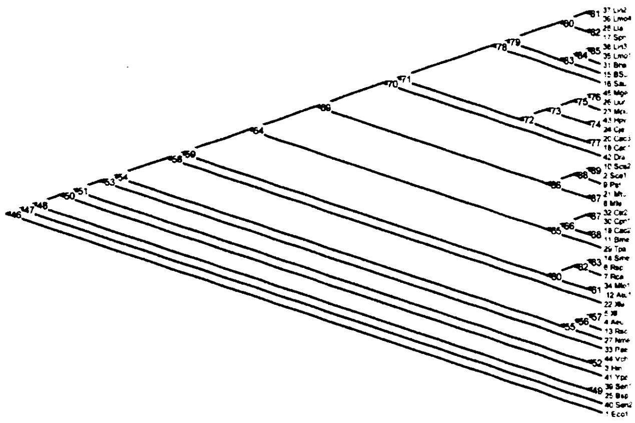


Figure A1.2 : Parsimony tree for 45 TK sequences, as described in section A1.1.1

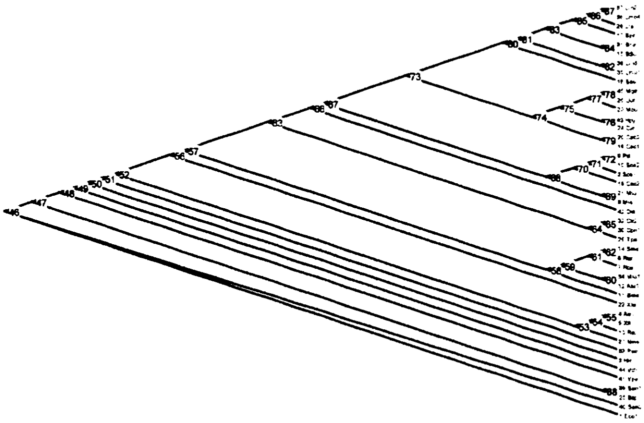


Figure A1.3 : Maximum Likelihood tree for 45 TK sequences, as described in section A1.1.1.

To determine how each of these trees differed in terms of their ancestral sequences, each was input, along with the original sequence file into the *PAML* program. Reconstructions yielded the Tables A1.1, A1.2 and A1.3. The ultimate use of these

AS res no	E.coli	48	52	54	57	58	64	72	75	S. cerevisiae
23	K	N	N	N	N	N	N	N	N	N
29	A	A	A	A	A	M	M	M	A	A
64	N	N	N	N	N	A	A	A	N	N
159	M	M	M	M	M	M	M	M	Q	Q
188	S	S	S	S	S	S	S	S	S	T
259	D	D	D	D	K	K	K	K	G	S
383	A	A	A	T	T	T	A	A	T	T
384	P	P	G	G	G	G	G	G	P	P
387	L	L	L	L	L	L	K	K	L	L
409	V	V	V	V	V	V	V	V	V	I
466	L	L	L	L	L	L	V	L	L	V

Table A1.1 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Neighbour joining tree for 45 TK (Figure A1.1) sequences, as described in section A1.1.1.

AS res no	E.coli	51	53	58	59	64	88	89	S. cerevisiae
23	K	N	N	N	N	N	N	N	N
29	A	A	A	M	M	M	A	A	A
64	N	N	N	N	A	A	N	N	N
159	M	M	M	M	M	M	M	Q	Q
188	S	S	S	S	S	S	S	S	T
259	D	D	D	D	K	K	G	G	S
383	A	A	A	A	A	A	T	T	T
384	P	P	G	G	G	G	P	P	P
387	L	L	L	L	L	K	L	L	L
409	V	V	V	V	V	V	V	V	I
466	L	L	L	L	L	L	L	L	V

Table A1.2 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Parsimony tree for 45 TK sequences, as described in section A1.1.1.

AS res no	E.coli	49	51	56	57	63	68	70	71	S. cerevisiae
23	K	N	N	N	N	N	N	N	N	N
29	A	A	A	M	M	M	M	A	A	A
64	N	N	N	N	A	A	A	A	N	N
159	M	M	M	M	M	M	Q	Q	Q	Q
188	S	S	S	S	S	S	S	S	S	T
259	D	D	D	D	K	K	K	K	S	S
383	A	A	A	A	A	A	A	A	T	T
384	P	P	G	G	G	G	G	G	P	P
387	L	L	L	L	L	K	K	K	L	L
409	V	V	V	V	V	V	V	V	V	I
466	L	L	L	L	L	L	L	L	L	V

Table A1.3 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Maximum Likelihood tree for 45 TK sequences, as described in section A1.1.1.

phylogenetic reconstructions was the generation of ancestral TKs using SDM. For this purpose, it would have been most useful if each mutation encountered along an evolutionary lineage were to occur one at a time. Across the three reconstructions it was found that not all of the mutations occurred singly and in some instances occurred six or seven at a time (Tables A1.1 to A1.3). It was decided that more sequences were needed in order to better resolve the order in which these mutations had evolved.

A1.1.2 Round two – phylogenetic analysis and reconstruction of 49 TK sequences

It was decided that more yeast TK sequences should be added to the alignment described in Section A1.1.1. The following 4 yeast sequences were obtained from a *BLAST* search using *Sce*TK as the query:

Kluyveromyces lactis

Candida albicans

Neurospora crassa

Schizosaccharomyces pombe

It was felt that more yeast sequences were needed since there were considerably more bacterial sequences in the analysis than yeast and the mutations of the bacterial clades were better resolved than those on the yeast clades of the tree. The resulting 49 TK sequences were aligned using *ClustalW*.

To ascertain the effect of these 4 new yeast sequences on the overall ancestral reconstruction, a dendrogram was generated using the *Propars* program. This tree was subject to reconstruction using the *PAML* program as before. The reconstructed tree as well as the distribution of mutations in the active-site can be seen in Figure A1.4 and Table A1.4.

Adding the yeast sequences to the analysis reduced the overall number of mutations seen in the active-site of this lineage. The resolution of mutations was also sharper, although at node no. 56, 5 simultaneous mutations are found where *Pst*TK joins the tree (marked with a red asterisk in Figure A1.4). More sequences were added to the analysis to try and resolve the mutations at node number 56.

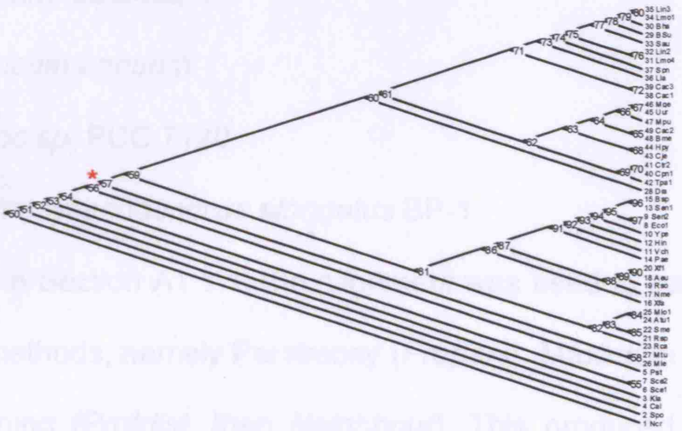


Figure A1.4 : Parsimony tree for 49 TK sequences, as described in section A1.1.2. The part of the tree where *Pichia stipitis* joins is marked with a red asterisk.

AS res no	*E.coli*	92	91	86	81	56	54	55	*S. cerevisiae*
23	K	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**N**
29	A	**A**	**A**	**M**	**M**	**A**	**A**	**A**	**A**
64	N	N	N	N	**A**	**N**	**N**	**N**	**N**
159	M	M	M	M	M	M	M	**Q**	**Q**
188	S	S	S	S	S	S	**T**	**T**	**T**
259	D	D	D	D	**K**	**S**	**S**	**S**	**S**
383	A	A	A	A	A	**T**	**T**	**T**	**T**
384	P	P	**G**	**G**	**G**	**P**	**P**	**P**	**P**
409	V	V	V	V	V	V	V	V	**I**
466	L	L	L	L	L	L	L	L	**V**

Table A1.4 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Parsimony tree for 49 TK sequences (Figure A1.4, above), as described in section A1.1.2.

1.1.3 Round three – phylogenetic analysis and reconstruction of 49 TK sequences.

Using the *Pichia stipitis* TK as the query in a *BLAST* search, 5 additional TK sequences were chosen and aligned with the previous 49 TKs from Sections A1.1.1 and A1.1.2, yielding a 54 TK alignment. The additional TKs added were :

Leishmania mexicana mexicana

Solanum tuberosum

Capsicum annuum

Nostoc sp. PCC 7120

Thermosynechococcus elongatus BP-1

As described in Section A1.1.1, the alignment was used to generate phylogenies by three different methods, namely Parsimony (*Propars*), Maximum Likelihood (*ProML*) and Neighbour joining (*Protdist*, then *Neighbour*). This produced 4 trees, since the *Propars* method produced two "tie-trees", both of which were equally parsimonious Reconstruction of these trees with *PAML* yielded the best resolution of mutations yet (see Tables A1.5 to A1.7 and Table 1.6 in Section 1.3.3 for the ML reconstruction). It was decided that this 54 sequences alignment should be used in the analysis of TK. The ProML tree became a candidate tree as described in Sections A1.3 and 1.3.2.1, while the Neighbour joining tree (which was generated using the default parameters of the *Neighbour* program (JTT matrix)) and both Parsimony tree were rejected in the overall analysis. The ProML tree was eventually chosen as the best overall tree, as described in Section 1.3.2.1, not before the 54 TK sequences were subject to phylogenetic study by a host of methods (see Section A1.2).

AS res no	*E.coli*	74	70	68	67	61	56	58	59	*S.cerevisiae*
23	K	N	N	N	N	N	N	N	N	N
29	A	A	A	A	M	A	A	A	A	A
64	N	N	N	N	A	N	N	N	N	N
160	M	M	M	M	M	M	M	Q	Q	Q
189	S	S	S	S	S	S	S	S	T	T
260	D	D	D	K	K	K	S	S	S	S
383	A	A	T	T	T	T	T	T	T	T
409	V	V	V	V	V	V	V	V	V	I
466	L	L	L	L	L	L	L	L	L	V

Table A1.5 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Neighbour joining tree for 54 TK sequences, as described in Section 1.1.3.

AS res no	*E.coli*	98	97	93	92	87	59	58	106	107	*S. cerevisiae*
23	K	N	N	N	N	N	N	N	N	N	N
29	A	A	A	A	M	M	M	A	A	A	A
64	N	N	N	N	N	A	A	N	N	N	N
159	M	M	M	M	M	M	M	M	Q	Q	Q
188	S	S	S	S	S	S	S	S	S	T	T
259	D	D	D	D	D	K	S	S	S	S	S
383	A	A	A	T	T	T	T	T	T	T	T
384	P	P	G	G	G	G	S	P	P	P	P
409	V	V	V	V	V	V	V	V	V	V	I
466	L	L	L	L	L	L	L	L	L	L	V

Table A1.6 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Parsimony tie tree 1 for 54 TK sequences, as described in Section 1.1.3.

AS res no	*E.coli*	98	97	93	92	87	58	57	106	107	*S. cerevisiae*
23	K	N	N	N	N	N	N	N	N	N	N
29	A	A	A	A	M	M	M	A	A	A	A
64	N	N	N	N	N	A	A	N	N	N	N
159	M	M	M	M	M	M	M	M	Q	Q	Q
188	S	S	S	S	S	S	S	S	S	T	T
259	D	D	D	D	D	K	S	S	S	S	S
383	A	A	A	T	T	T	T	T	T	T	T
384	P	P	G	G	G	G	P	P	P	P	P
409	V	V	V	V	V	V	V	V	V	V	I
466	L	L	L	L	L	L	L	L	L	L	V

Table A1.7 : Active-site mutations occurring during the evolution of *E. coli* and *S.cerevisiae* from their common ancestor in the Parsimony tie tree 2 for 54 TK sequences, as described in Section 1.1.3.

A1.2 Choice of most suitable phylogenetic tree for transketolase

As described in Section 3.2.2.2a, 29 phylogenetic trees were generated in all for the 54 TK sequences under study. 26 of these trees were rejected on the basis that they failed to conform with our general knowledge of species evolution. Since TK is found in all organisms and is involved in primary metabolism, it is likely to be a very ancient gene. The evolutionary lineage for such an enzyme is likely to be similar to the organism tree for the species involved.

For example, the UPGMA tree generated using the categories model with a gamma distribution (see Figure A4.5), was rejected straight away. In this case it was obvious that the method was producing a tree where plant, bacterial and yeast sequences were not distributed in a manner accurately reflecting the organism tree for the subjects involved. Upon closer inspection, trees such as the UPGMA tree generated using the categories model, without the gamma distribution (Figure A4.6) were also found to contain anomalies, such as where a clade of bacteria appear to have a common ancestor with the yeast and plant species (marked with an asterisk in Figure A4.6) more recently than with the rest of the bacteria in the tree.

The remaining trees were examined in detail. The 3 candidate trees (Figures 3.8, 3.9 and 4.10) were chosen on the basis that species grouped tightly within the trees. This is described in further detail in Section 3.3.4.1.

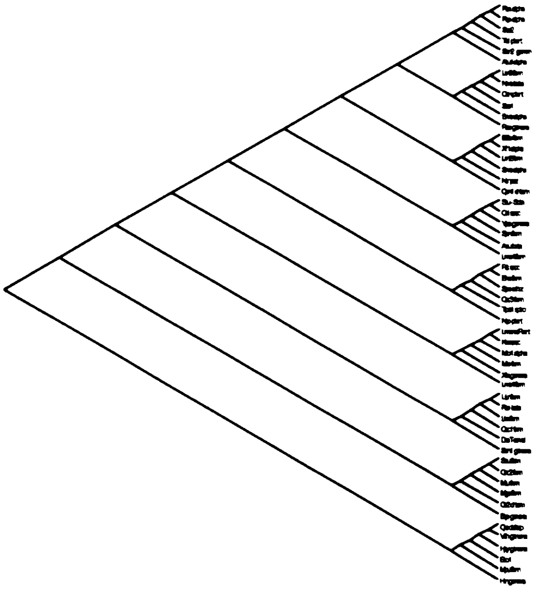


Figure A1.5 : UPGMA tree generated using the Categories model with the gamma parameter

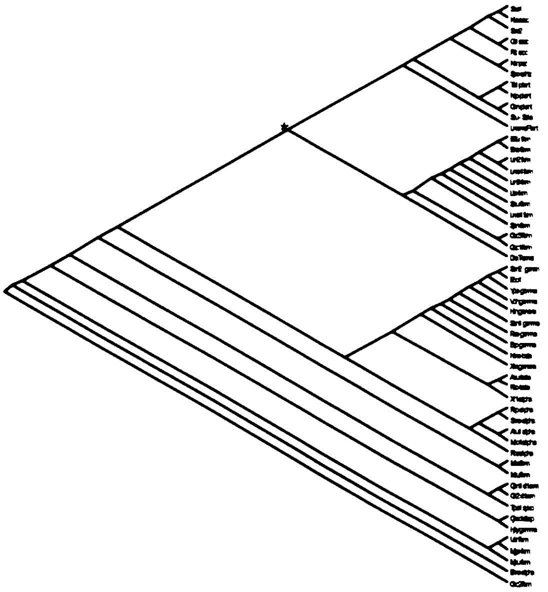


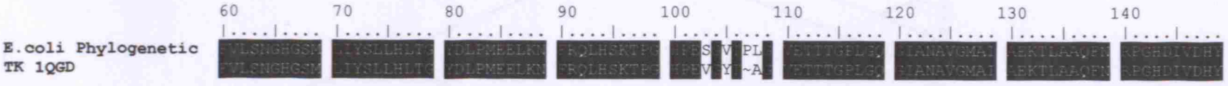
Figure A1.6 : UPGMA tree generated using the Categories model without the gamma parameter

A1.3 Differences between *E.coli* TK sequences used in Phylogenetic and structural analyses.

As can be seen from Table 4.2 in Section 4.3.1, the *E.coli* TK sequence used for phylogenetic analyses is *BLAST* accession number X68025. However, the sequence of the *E.coli* TK used in the majority of our TK structural studies, 1QGD.pdb differs slightly from X68025 version. This difference is likely to be due to an error in the sequencing of X68025. As can be seen in Alignment A1.1, residues differ at positions 103 and 105 as well as at position 108 (corresponding to residue 107 in the 1QGD sequence). An extra residue can be seen in the X68025, which could potentially lead to confusion in amino acid numbering when combining phylogenetic and structural analyses. For example, alanine 383 in the TK structure corresponds with alanine 384 in the X68025 sequence.

The differences between the two sequences occur at positions that are not important to our study, being as they are in a poorly aligned area of the TK alignment. Neither are the residues among the active site residues chosen for study (see Section x).The inserted proline at position 107 in is conserved in 21 of 54 TK sequences examined, while the other contentious positions have negligible identity and similarity.

Since the sequence of our experimental TK gene, on the pqr791 plasmid is identical to the 1QGD sequences, in all cases, the structural numbering will be used when referring to *E.coli* TK residues numbers. Only in the actual alignment of 54 TK sequences (Alignment A4.1) is the X68025 numbering is used.


60
70
80
90
100
110
120
130
140
E.coli Phylogenetic
TK 1QGD
FVLSNGHGSM LIYSLLHLTG YDLPMEELKN FRQLHSKTPG HPESGVTPLG VETTTGPLGQ GIANAVGMAI AEKTLAAQFN RPGHDIVDHY
FVLSNGHGSM LIYSLLHLTG YDLPMEELKN FRQLHSKTPG HPEVGYT~AG VETTTGPLGQ GIANAVGMAI AEKTLAAQFN RPGHDIVDHY

Alignment A1.1: The difference between the sequence of *Eco*TK (X68052) used in the phylogenetic studies of Chapters 3 and 4 and the sequence corresponding to the *Eco*TK crystal structure (1QGD.pdb).

Appendix 2: Appendix for Chapter 4

A2.1: Examination of the affect of *Cac2*TK and *Bme*TK on the phylogenetic reconstruction

The *Cac2*TK and *Bme*TK sequences align relatively poorly with the other TK sequences (Alignment 3.1). The concern was that this could have skewed the alignment somewhat and affected the overall accuracy of the TK tree. To assess the affect of *Cac2*TK and *Bme*TK on the phylogeny, they were removed from Alignment 3.1. This alignment of 52 TKs was then realigned and used to generate a maximum likelihood tree using *ProML*. The resulting ProML tree, shown in Figure A2.1, was compared with the Maximum Likelihood tree that was chosen as the "best tree" (Figure 4.11 in Chapter 3). It was found that the removal of *Cac2*TK and *Bme*TK had no effect on how the remaining 52 TK sequences were related. In all subsequent analyses, *Bme*TK and *Cac2*TK were included unless explicitly stated otherwise.

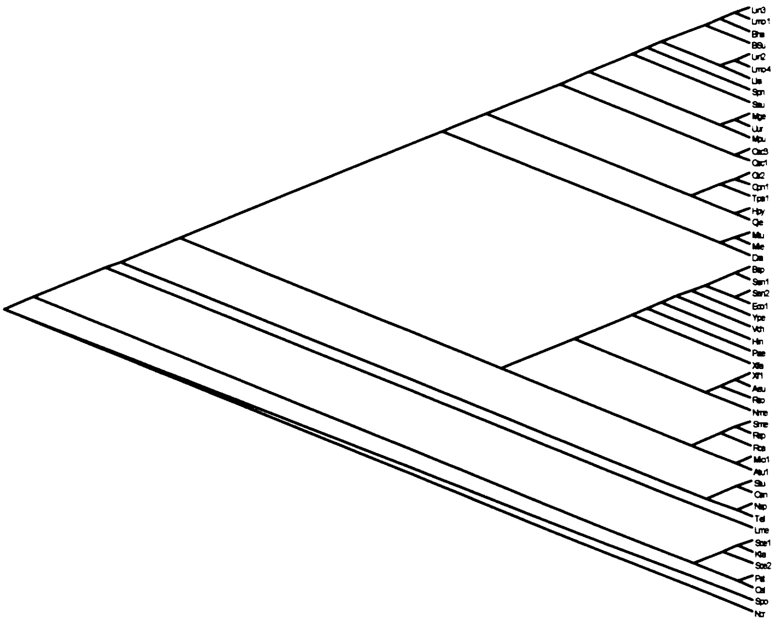


Figure A2.1: Maximum Likelihood tree for 52 TK sequences, examining the effect of removing *Cac2*TK *2* and *BmeTK* from the phylogenetic analysis.