Novel Transaminase Enzymes in Metagenomes

Gregory Ledderboge-Vucinic

Primary supervisor: Prof. Christine Orengo
(Institute of Structural and Molecular Biology, UCL)

Secondary supervisor: Prof. John Ward
(Department of Biochemical Engineering, UCL)

A thesis submitted for the degree of
MSc Bioinformatics with Systems Biology

Institute of Structural and Molecular Biology
Biological Sciences, Birkbeck, University of London
Malet St, Bloomsbury, London WC1E 7HX

Abstract

Transaminases (TAms) are a functionally diverse class of enzymes with many current and potential industrial uses, particularly in the pharmaceutical industry, due to their stereoselectivity and high reaction turnover under mild reaction conditions. There is a demand for TAms with new or improved substrate specificities; to this end, a variety of approaches have been employed, including mining of metagenomic sequences, rational mutagenesis, and directed evolution. These methods generally involve a lot of trial and error and are heavily reliant on the existing structural and functional characterisation of the enzyme in question. Recent advances in sequencing technology have enabled an explosion of metagenomic data – DNA sequences recovered from environmental samples. However, this abundance of raw sequence data has yet to be fully exploited, mostly because of the bottleneck that is accurately inferring the function of sequences that have no well-characterised close homologues. This project has aimed to address ways of making use of distant homologues in metagenome sequences to gain insights into functional divergence within the class III subgroup of TAms, often synonymous with omega-TAms or amine TAms. The CATH FunFams classification was used to define specificity groups and to provide a library of profile hidden Markov models with which to scan two salt evaporation pond metagenomes. Specificity determining positions (SDPs) were predicted in a structurally similar group of FunFams and novel mutations were then identified in these positions in distant homologues in the metagenomes. Mutations in the highest scoring SDP, a residue at the back of the binding pocket, were modelled *in silico*, showing one way in which the steric exclusion of bulky substrates could be removed. As well as investigating class III TAms in detail, this project has presented a workflow that can be applied to any enzyme family to identify candidate residues for rational mutagenesis or mutant libraries for directed evolution.

Contents

List of Tables and Figures

List of Abbreviations

ATA – Amine Transaminase
BLAST – Basic Local Alignment Search Tool

CATH – Class, Architecture, Topology, Homologous Superfamily

crh – cath-resolve-hits

DAPA – Diaminopelargonic Acid

DOPS – Diversity of Position Score

EC – Enzyme Commission

E-value – Expectation value

FD – Functional Determinant

FunFam – Functional Family

GABA – gamma-Aminobutyric Acid

HMM – Hidden Markov Model

KAP(A) – 7-keto-8-aminopelargonic acid

M-CSA – Mechanism and Catalytic Site Atlas

MDA – Multi-domain Architecture

MSA – Multiple Sequence Alignment

ORF – Open Reading Frame

PDB – Protein Data Bank

pHMM – profile HMM

PLP – Pyridoxal-5’-phosphate

PMP – Pyridoxamine-5’-phosphate

PSI-BLAST – Position-Specific Iterative BLAST

RMSD – Root Mean Squared Deviations

SC – Structural Cluster

SDP – Specificity Determining Position

SFLD – Structure Function Linkage Database

TAm – Transaminase

UniProtKB – UniProt Knowledge Base

zDOPE – Normalised Discrete Optimised Protein Energy

Acknowledgements

I would like to thank my supervisor, Prof. Christine Orengo, for allowing me the opportunity to experience her exciting research group and for her valuable guidance and immense patience. I am grateful for the help of Dr. Natalie Dawson, Dr. Sayoni Das, and Dr. Ian Sillitoe, and to all others in the group for their warm welcome. I also want to thank Prof. John Ward and his group for entrusting me with the metagenome sequence data. Finally, I would like to thank all of those involved in the MSc Bioinformatics course at Birkbeck for their teaching and support, in particular my personal tutor, Dr. Adrian Shepherd, and Dr. Irilenia Nobeli for their patience and understanding.

1 INTRODUCTION

1.1 Transaminases

Transaminases (TAms, EC 2.6.1.x), also known as aminotransferases (ATs) are pyridoxal-5'-phosphate (PLP) dependent enzymes that catalyse the stereoselective transfer of an amino group to the carbonyl position of a prochiral alpha-keto acid, ketone, or aldehyde. In nature, they are mostly involved in amino acid metabolism and in a wider sense, nitrogen metabolism. They have a broad substrate specificity compared to most other enzyme families (107 EC entries under 2.6.1). The need for some kind of TAm enzyme activity is an ancient one and so divergent evolution has had time to vastly expand the repertoire of TAm activities.

1.1.1 Potential for Industrial Uses

Due to their fast reaction turnover, broad substrate specificity, remarkable stereoselectivity, and lack of a requirement for external cofactor regeneration, TAms have become valuable enzymes for biocatalytic preparation of optically pure amine precursors for the pharmaceutical and agrochemical industries among others (Fuchs, Farnberger and Kroutil, 2015). As of 2013, 1 in 4 of the top 200 selling drugs contain a chiral amine moiety (Weber and Sedelmeier, 2014), which highlights the importance of this functional group. TAms provide high regiospecificity, which is necessary when the substrate is a functionally dense compound with several reactive ketone and/or aldehyde groups,

and stereospecificity, which is often difficult or impossible to achieve by chemical means. Compared to commonly used metal-catalysed chemical reductive aminations, they can also offer higher atom and step economy, with fewer environmentally damaging waste products, and without the need for energy-intensive pressure or heat conditions.

Another factor that makes TAms good candidates for biocatalysis is their high substrate promiscuity as a group (although individually most TAms have a narrow substrate scope). The structural basis for this promiscuity can be exploited in directed evolution bioengineering approaches such as substrate walking, which can dramatically increase the synthetic repertoire of TAms. In particular, this can allow the active site to accommodate larger ring structures, which are present in over 90% of marketed drug molecules but are not in most natural substrates of TAms. Several cases of rational engineering of point mutations have demonstrated significant alteration of substrate specificity, which means a 'brute force' high throughput approach is not always necessary.

A solution of optically pure amines can be generated by two different methods: kinetic resolution and direct asymmetric synthesis. In kinetic resolution, a stereospecific TAm acts on the undesired enantiomer in a racemic mixture, leaving only the desired enantiomer. This reduces the yield by half, so asymmetric synthesis from the prochiral ketone is preferable where possible. The substrates and products of TAms exist in equilibrium with each other, so the ketone co-product must be removed in order to shift the reaction towards the desired amine product.

In addition to unfavourable yields and thermodynamic equilibria, a challenge of using TAms in biocatalysis is product inhibition (due to the dual substrate recognition). Enzyme cascades or other methods may be used to remove co-products, thereby shifting the equilibrium in the desired direction. Much progress has been made in developing biocatalytic strategies with viable enantiomeric purities and yields (Fuchs, Farnberger and Kroutil, 2015). This provides a framework of knowledge into which newly discovered or engineered TAms can be plugged into.

1.1.2 Classification Systems and Nomenclature

The first major effort to classify TAms was carried out by Mehta et al. (1993): 51 sequences of 14 different aminotransferase enzymes were classified into 4 subgroups using sequence-based secondary structure predictions. This was expanded to 5 classes based on multiple sequence alignments of sequences available in the Pfam database (Hwang *et al.*, 2005), with a sixth class

being added later. Class III (highlighted in table 1.1) has received growing attention over the past 10-15 years and is the focus of this project.

Pfam Family (Accession)	**CATH Superfamilies (major + minor domain)**	**Positional Specificities[a] (regiospecificity)**	**Archetypal enzyme names**	**Archetypal substrates[b] (amino donor:acceptor)**
Aminotransferase Class I/II (PF00155)	3.40.640.10 + 3.90.1150.10	alpha	aspartate TAm	aspartate:alpha-ketoglutarate
			alanine TAm	alanine:alpha-ketoglutarate
			tyrosine TAm	tyrosine:alpha-ketoglutarate
			histidinol-phosphate TAm	histidinol-P:alpha-ketoglutarate
			phenylalanine TAm	phenylalanine:pyruvate
Aminotransferase Class III (PF00202)	3.40.640.10 + 3.90.1150.10	omega (==delta)	(N-acetyl)ornithine TAm	(acetyl)ornithine:alpha-ketoglutarate
		omega (==delta)	omega-amino acid TAm	beta-alanine:pyruvate
		gamma (==omega)	gamma-aminobutyrate TAm	gamma-aminobutyrate:alpha-ketoglutarate
		omega	diaminopelargonate TAm	SAM:7-keto-8-aminopelargonate
		beta	beta-aminocarboxylic acid TAm	beta-homoleucine:pyruvate
Aminotransferase Class IV (PF01063)	3.20.10.10 + 3.30.470.10	alpha	D-alanine TAm	D-alanine:alpha-ketoglutarate
			branched chain amino acid TAm	leucine:alpha-ketoglutarate
Aminotransferase Class V (PF00266)	3.40.640.10 + 3.90.1150.10	alpha	serine TAm	serine:pyruvate
			phosphoserine TAm	phosphoserine:alpha-ketoglutarate
			cysteine desulfarase	(not a transaminase)
DegT/DnrJ/EryC1/StrS aminotransferase family[c] (PF01041)	3.40.640.10 + 3.90.1150.10	(ring structure)	TDP-4-amino-4,6-dideoxy-D-glucose TAm	TDP-4-amino-4,6-dideoxy-D-glucose:alpha-ketoglutarate
Alanine-glyoxylate aminotransferase[d] (PF12897)	3.40.640.10 + 3.90.1150.10	alpha	alanine-glyoxylate TAm	alanine:glyoxylate

Table 1.1 *(Previous page) Structural and functional classification of the 6 TAm families in Pfam. The Pfam to CATH mapping was obtained by scanning the Pfam seed sequences for each family with cath-genomescan. The Pfam classification considers the whole chain to be one domain, while CATH chops it into two – a major/large domain and a minor/small domain. All amino acids are L-stereoisomers unless otherwise stated. Class IV is the only group that naturally contains R-stereoselective enzymes (D-amino acids).*
[a] *The position of the amino group of the donor, relative to the carboxyl or other major group. In this context, omega simply refers to any non-alpha position for the amino donor in either the forward or reverse reaction*
[b] *The main donors are amino acids, but many other amino compounds can take their place, for instance isopropylamine, beta-alanine, and (S)-alpha-methylbenzylamine (MBA)*
[c] *Also called Transaminase Class-VI or Sugar Aminotransferases*
[d] *Also known as the MocR subfamily of the Helix-Turn-Helix GntR family of transcriptional regulators*

TAms have several complicated and overlapping naming conventions. In the original 4-group classification (1993-2005) the current class I/II was named subgroup-I, class III was subroup-II, class IV was subgroup-III, and class V was subgroup-IV. The current 6 Pfam classes can be placed into two distinct evolutionary groups – classes I/II, III, V and VI in one group, and class IV alone in a separate group. The major domain of these groups correspond to CATH homologous superfamilies 3.40.640.10 and 3.20.10.10, and are also known as fold-types I and IV. These evolutionary groups are functionally diverse, each containing over 200 unique Enzyme Commission (EC) annotations.

Class III TAms are also commonly referred to as omega-TAms, due to their diverse regioselectivity, or amine TAms (ATAs), since they do not require a carboxyl group adjacent to the amine. However, there are also some omega-TAms present in class IV. Class III and class IV omega-TAms are often mentioned in the same breath without noting their distinct evolutionary groups. To futher complicate matters, the omega position is often equivalent to the beta, gamma or delta position; for instance, a common substrate of omega-amino acid TAm is beta-alanine.

1.1.3 Reaction Mechanism

In the ground state of the enzyme, a catalytic lysine residue forms a Schiff base with PLP. This is attacked by the amine group of the first substrate in a nucleophilic addition-elimination, forming an external aldimine. The PLP then shuttles electrons around its ring, eventually eliminating the deaminated substrate and leaving behind pyridoxamine-5’-phosphate (PMP) in the enzyme active

site. The same reaction essentially occurs in reverse for the second substrate. An example of a TAm reaction is beta-alanine:pyruvate transamination, shown in figure 1.1.

Both substrates in the ping-pong bi-bi mechanism bind at the same site. In most cases, this dual substrate recognition is achieved through a large-scale rearrangement of the active site hydrogen-bonding network caused by the induced fit (Hirotsu *et al.*, 2005).

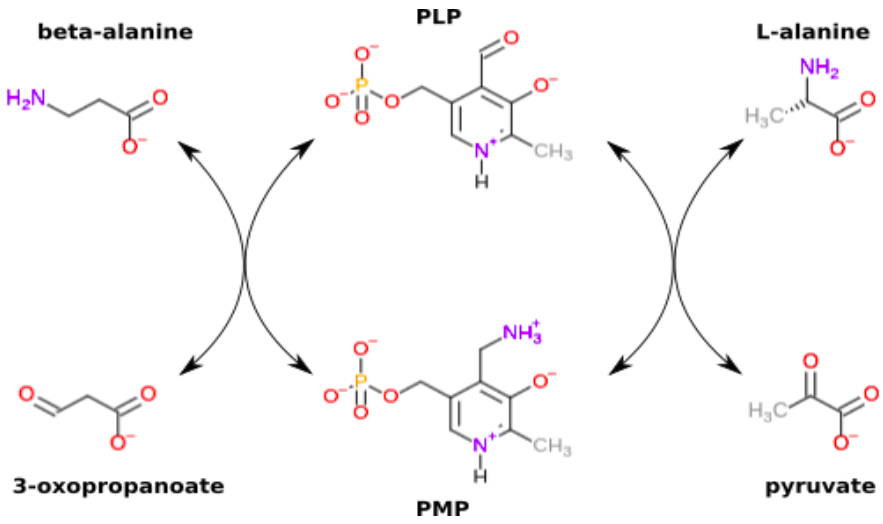

Figure 1.1 *The overall reaction scheme of beta-alanine:pyruvate transaminase (EC 2.6.1.18), in which L-alanine is used to aminate 3-oxopropanoate via the PLP cofactor, forming the products pyruvate and beta-alanine. The reaction is fully reversible and its natural role is often in beta-alanine degradation.*

1.2 The Search for New Functions – Existing Methods

Given the large biocatalytic potential of TAms, finding new functions is of great interest. As well as improving the diversity of the toolkit at our disposal, new sequences can bypass the current patent space. Sequences with improved or novel functions can be arrived at in two ways: finding existing sequences and bioengineering new ones.

It is a sensible strategy to first look at the solutions evolution has come up with over hundreds of millions of years, rather than trying to reinvent the wheel by bioengineering new functions through complicated prediction, trial and error. The conventional approach towards enzyme discovery involves enrichment screening of microorganisms (Mathew and Yun, 2012). For example, Clay et al. (2010) tested over 100 lyophilised whole cell preparations for omega-TAm activity. However, these methods are costly, time-consuming, and limited to culturable organisms. In the past decade, bioinformatic data mining approaches have been employed to make use of the

wealth of sequence information in public databases. BLAST (Altschul *et al.*, 1990) is a powerful tool for identifying related sequences, and the related tool PSI-BLAST (Altschul *et al.*, 1997) has an enhanced sensitivity to weak similarities. Höhne et al. (2010) developed an *in silico* strategy that involved analysis of sequence motifs in (R)-selective omega-TAms versus (S)-selective omega-TAms. This information was used to create an annotation algorithm that proved highly successful in screening a library of 5700 sequences (17/21 predicted sequence showed positive results when the stereoselectivity was tested).

Despite the great variety of the microbial world, the wild type enzymes represent only a tiny fraction of the possible 'functional space' that evolution has explored, which means it is rare to find a wild-type enzyme posessing all the desirable features for industrial applications. This is where bioengineering approaches step in, in the form of site-specific or random mutagenesis.

Enzyme stability, size of preferred substrate and enantiopreference can be modified with as little as a single mutation. For example, an amine:alpha-keto acid TAm was changed to an amine:aldehyde TAm by changing the so-called 'flipping' arginine to a hydrophobic residue (Genz *et al.*, 2015), and the stereoselectivity of an omega-TAm was reversed with a single point mutation (Svedendahl *et al.*, 2010).

Random mutagenesis can be performed in the absence of detailed structural and functional information. It can generate enzymes that perform well under extreme reaction conditions such as high substrate concentration, presence of solvents, and high temperature. These stabilising mutations are harder to predict than mutations that will affect the function. However, high throughput (HTP) screening is costly and time-intensive. The limiting factor in directed evolution is the HTP screening – mutant libraries for each generation are usually larger than the 10^3-10^6 variants that can be screened. Therefore, predicting residues that are both tolerant to mutation and important in determining function would be a valuable step towards reducing the size of mutant libraries.

An example of the application of random mutagenesis towards omega-TAms is the ATA-117 enzyme used for the final step of the synthesis of the type-II diabetes drug sitaglipin. It contains 27 mutations, which were accumulated over 11 rounds of directed evolution (Savile *et al.*, 2010). This allowed the enzyme to accommodate a drastically larger substrate than it's orignal preferred substrate, and serves as a demonstration of the immense potential of this group of enzymes as templates for bioengineering. The success story of ATA-117, an R-selective omega-TAm from Class IV, has not yet been replicated for S-selective class III omega-TAms. Hence, focusing on the S-

selective Class III will help in the efforts to fill this gap of R-selective TAms that can tolerate bulky substrates.

The two approaches – mining natural sequences and bioengineering – can be combined by analysing mutation patterns in natural TAms to predict useful point mutations or to help guide and streamline bioengineering projects.

1.3 Metagenomic Approach to Enzyme Discovery

The vast majority of DNA and protein sequences deposited in public databases is from culturable organisms such as *E. coli*, but this represents only a small fraction of the total genetic information of all organisms on the planet. With the ever-increasing speed, quality and cost-efficiency of sequencing technologies, it has become viable to sequence and assemble DNA from environmental samples with the sole aim of mining the data for novel and potentially useful proteins. Metagenomics is the surveying of microbial communities and their encoded metabolic activities.

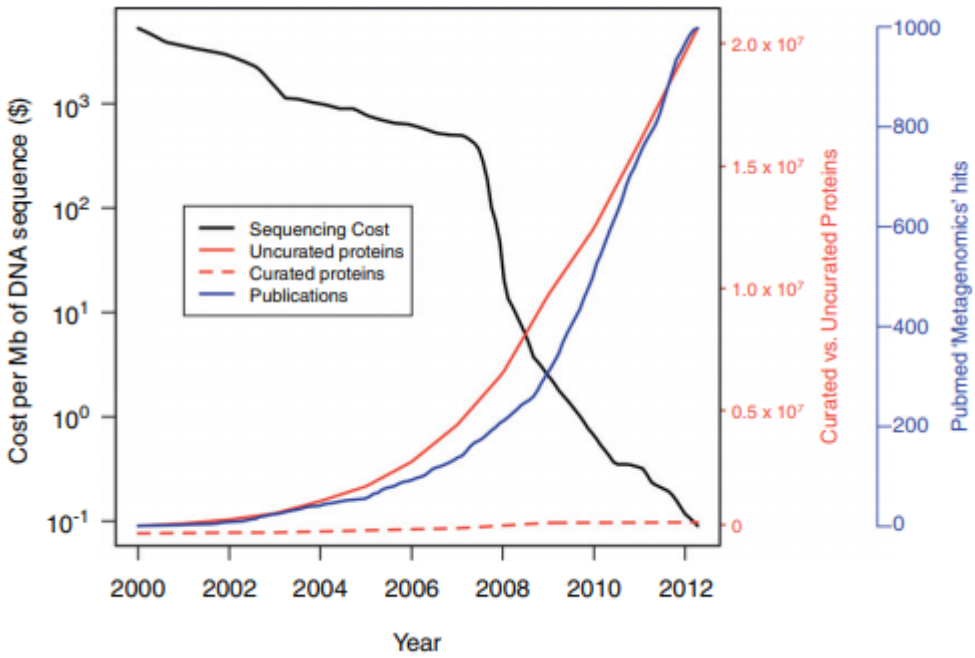

Figure 1.2 *The effect of DNA sequencing cost on number of curated and uncurated proteins, taken from Temperton & Giovannoni (2012). The reduction in cost of DNA sequence has seen an explosion of genomic and metagenomic data, but the number of manually curated proteins lags far behind, with less tha 1% of all proteins in UniProtKB being experimentally characterised. The number of curated vs uncurated proteins refers to the UniProtKB and SwissProt databases.*

While metagenomics has contributed to the growing function annotation gap (see fig. 1.2), it also holds part of the solution to bridging this gap, by allowing exploration of the as yet undiscovered sequence and function diversity. Novel functional repertoires are emerging from metagenomic studies around the world. Baud et al. (2017) used a metagenomics mining strategy to identify class III TAms with activity towards substrates of interest.

1.4 Widening Function Annotation Gap

While the number of sequences in databases have increased exponentially, the costly and time-consuming annotation of function through experimental methods has lagged behind (see fig. 1.2). To make use of the sequence data, homology-based techniques have been employed for function prediction and annotation. This means that most of the function annotations in public databases are simply inferred through homology. While this has provided a valuable tool for researchers, it is not difficult to imagine how incorrect annotations can be propagated as the databases grow. Even with the recent rise in metagenomic sequences, the sequences that receive confident homology-based function annotation are biased towards the homologues of well characterised proteins from culturable organisms. There are major gaps in function annotation that can ultimately only be addressed through experimental validation of sequence-based predictions. Moreover, a large degree of functional diversity is observed among homologous proteins, which means global sequence and structure homology alone cannot be used to infer functional similarity. Processes such as gene duplication followed by subfunctionalisation or neofunctionalisation of the resultant paralogues have been posited to account for the observed functional divergence. It is important to be able to identify the key driver mutations that account for this functional divergence, so that a combination of overall homology and ‘signature’ residues may be used to more accurately infer the function of a sequence.

1.5 The CATH Database

CATH-Gene3D is a domain-based protein structure classification database that organises structures from the Protein Data Bank (PDB) into a hierarchical classification of Class, Architecture, Topology and Homologous superfamily (Dawson *et al.*, 2017), where proteins in a homologous superfamily are strongly predicted to share a common evolutionary ancestry. The sequences of each

homologous superfamily are used to seed profile hidden Markov models (pHMMs) (Eddy, 1998), which are probabilistic models of sequence families. All known protein sequences with no associated structure are pulled in from UniProtKB and Ensembl, scanned against the pHMMs, and classified within Gene3D, the sister resource of CATH (Lewis *et al.*, 2018). Unless otherwise stated, version 4.1 of CATH (*http://wiki.cathdb.info/wiki-beta/doku.php?id=release_notes#cath-plus_version_41*) is used for all sequence alignments and statistics discussed here.

1.5.1 Domains as the Fundamental Units of Evolution

Domains are independently stable sections of proteins which can be thought of as evolutionary modules. A domain-based system is arguably the most useful way to classify proteins in order to discover homology relationships between them.

1.5.2 Functional Families (FunFams)

Due to the functional diversity within CATH superfamilies, they have been sub-classified into functionally coherent sequence groups called FunFams (Das, Lee, *et al.*, 2015). The FunFHMMer algorithm supervises an agglomerative clustering of the CATH-Gene3D sequences in each homologous superfamily. As it works its way up a tree, a novel Functional Coherence index is calculated for each pair of nodes (representing multiple sequence alignments), to decide if they should be merged or not. This is based on the differential conservation of mechanistically important residues, which is more powerful than relying on overall homology. The resulting FunFam groupings are used to create pHMMs. These are available for use through the CATH FunFHMMer web server (Das, Sillitoe, *et al.*, 2015) and as part of genomescan (*https://github.com/UCLOrengoGroup/cath-tools-genomescan*), which is a powerful tool for annotating large sequence datasets with FunFam assignments. FunFam groupings can be used as a framework for analysis of the sequence-function relationship between similar groups of proteins. FunFHMMer predicted residues have been successfully used in the prediction of functional determinants (FDs) and rational classification of different beta-lactamase classes (Lee *et al.*, 2016). In this project, the FunFam profiles of TAms are used to detect and analyse distant homologues in metagenome sequences. This allows access to a sweet spot between similarity and novelty i.e. just enough novelty to affect substrate specificity, while still remaining a TAm.

Each FunFam has an associated diversity of position score (DOPS), which signifies the information content of a multiple sequence alignment (MSA). DOPS = 100 when no two positions have the same conservation score, and DOPS = 0 when all positions have the same conservation score (this is seen in an alignment of identical sequences). The larger the evolutionary distance captured by the FunFam – while staying functionally coherent – the more informative it is. This means a conserved residue in a high DOPS MSA carries more weight than in a low DOPS MSA.

Many superfamilies are structurally diverse enough to prohibit confident comparisons between FunFams. To get around this issue, each superfamily is sub-classified into distinct structural clusters (SCs), where all intra-cluster normalised root-mean-square deviation (RMSD) scores are less than 5Å.

1.6 Transaminases in CATH

The 6 Pfam TAm families map onto two pairs of homologous superfamilies in CATH (see table 1.1) – this was confirmed by scanning the seed sequences for each Pfam family with the CATH FunFam profiles (see section 2.5). The large domain – 3.40.640.10 (Type I PLP-dependent aspartate aminotransferase-like (Major domain)) – which contains the PLP-coordinating residues, was the focus of this project. It belongs to class 3 (Alpha Beta), architecture 3.40 (3-layer (aba) Sandwich), and topology 3.40.640 (Aspartate Aminotransferase; domain 2). In version 4.1, it is sub-classified into 134 FunFams.

1.6.1 Domain Structure of Homologous Superfamily 3.40.640.10

The domain exhibits a 3-layer (aba) sandwich, one of the most abundant and functionally diverse folds in nature. The strong conservation of the structural core, with one or two exceptions, is apparent in figure 1.3.

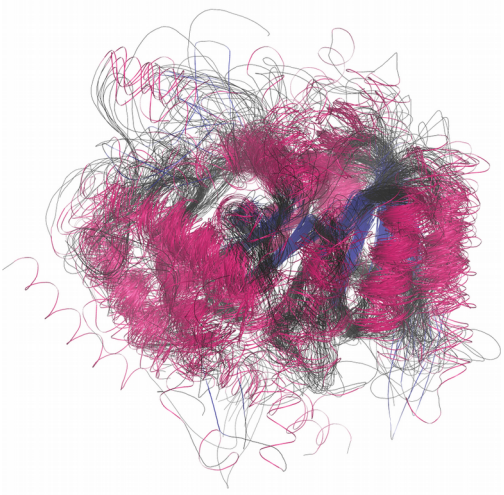

Figure 1.3 *Superposition of the 134 representative domains (one for each FunFam) within the 3.40.640.10 superfamily (taken from the CATH website).*

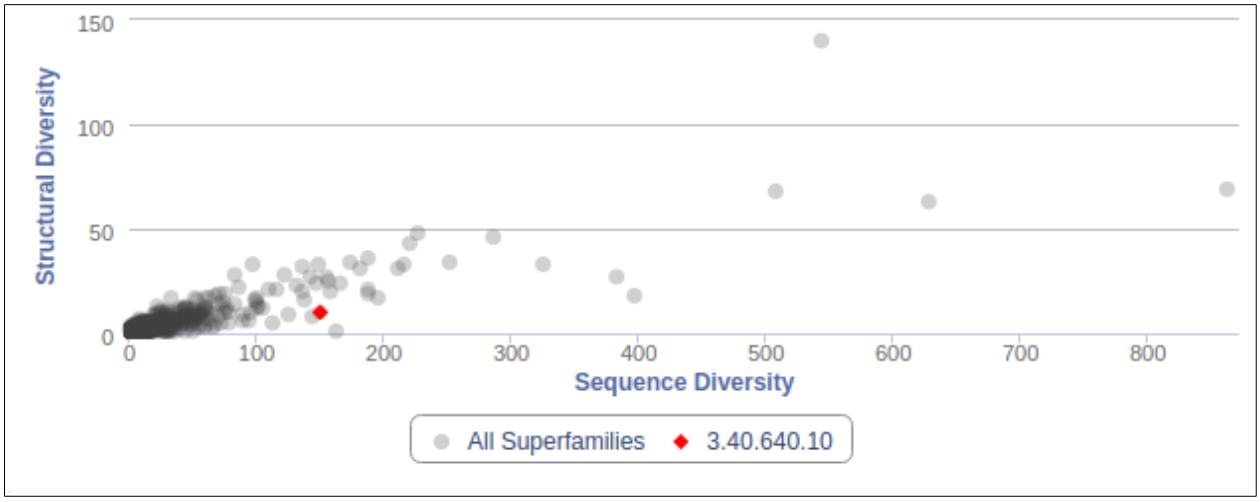

Figure 1.4 *Sequence diversity and structural diversity of 3.40.640.10 relative to all other homologous superfamilies in CATH (taken from the CATH website). Each point represents one of the 6119 superfamilies in CATH v4.2. The sequence diversity is the number of FunFams in the superfamily and the structural diversity is the number of structural clusters (normalised RMSD < 5Å between members of the same cluster).*

As seen in figure 1.4, the superfamily 3.40.640.10 has a high sequence diversity relative to its structural diversity – 151 FunFams grouped into just 10 structural clusters. This means the same structural core can carry out many different functions, which is useful for the analysis of SDPs in this project. In contrast, a higher structural diversity relative to sequence diversity would translate to fewer FunFams per structural cluser, which may make the SDP analysis less informative.

1.6.2 Multi-domain Architectures of 3.40.640.10

Figure 1.5 *The distribution of most frequent domain architectures of 3.40.640.10 sequences in Gene3D, taken from http://gene3d.biochem.ucl.ac.uk/search?sterm=3.40.640.10&mode=family. The small domain is discontinuous – it makes up the N-terminus and the C-terminus of the sequence.*

The most common domain architecture of TAms is 3.90.1150.10—3.40.640.10 (fig. 1.5). In this arrangement, the 3.40.640.10 domain is called the large/major domain while the 3.90.1150.10 domain is the small/minor domain. Typically, these TAms are homodimers (sometimes homotetramers) in their native states, with a large buried surface area between the large domains of the two monomers. Figure 1.6 depicts the quarternary structure of a representative TAm.

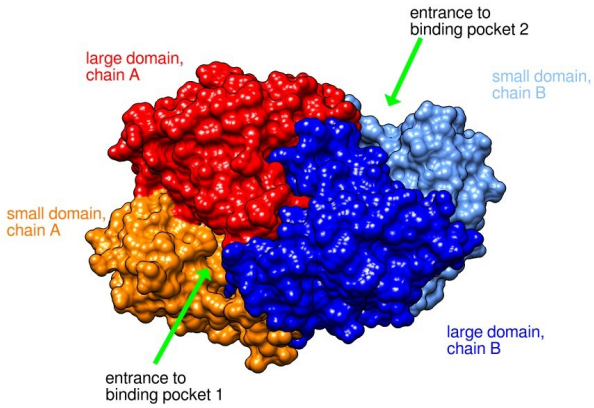

Figure 1.6a *TAm homodimer structure. The large domains (CATH domains 1dtyA02 and 1dtyB02) are in superfamily 3.40.640.10, and the small domains (CATH domains 1dtyA01 and 1dtyB01) are in superfamily 3.90.1150.10.*

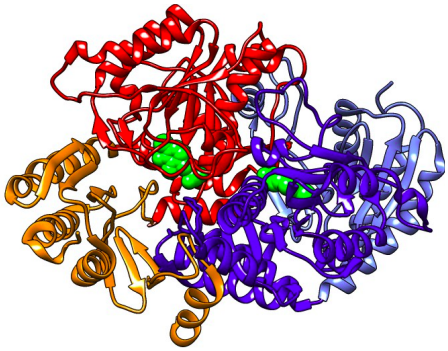

Figure 1.6b *A ribbon diagram from the same angle and with the same colour scheme as above. The PLP cofactor atoms are shown as green spheres.*

1.6.3 Focus of this Project

The four FunFams in table 1.2 make up the bulk of class III TAms and are the primary focus of this project. Figure 1.7 is a visualisation of the sequence similarity between all 134 FunFams in 3.40.640.10, with the FunFams of interest highlighted. These four structurally similar FunFams were analysed to identify potential specificity determining positions (SDPs) – residues that display strong intra-FunFam conservation but vary between FunFams. The metagenomic sequences were scannned with all the FunFam profile HMMs in the CATH database to identify matches to the four FunFams of interest. Particular attention was paid to weak hits that are in potentially novel FunFams. The predicted SDPs in these sequences were analysed to identify novel mutations, and the predicted effects of these mutations are discussed in the context of known structural and functional information. SDPs are often found in the secondary shell of residues around the catalytic residues and are likely to play an important role in the functional divergence of the different groups. Mutating a catalytic residue is almost guaranteed to abrogate the function of the protein, whereas mutating an SDP is likely to have a more subtle effect, often improving the enzyme's kinetic constants towards some substrates at the expense of other substrates.

FunFam code	**Name**	**Structural representative**
3.40.640.10/FF/***62757***	Related to Adenosylmethionine—8-amino-7-oxononanoate aminotransferase*	1dtyA02
3.40.640.10/FF/***63097***	4-aminobutyrate aminotransferase GabT	1sffC02
3.40.640.10/FF/***63148***	Aminotransferase class-III family protein+	4ba5B02
3.40.640.10/FF/***63154***	Probable acetylornithine aminotransferase, mitochondrial	4addD02

Table 1.2 *(Previous page) The four TAm FunFams that are the primary focus of this project. The FunFam names are automatically generated based on the most common terms in the UniProtKB descriptions (Dawson et al., 2017). The structural representative is the CATH domain ID (4-character PDB code followed by chain letter and domain number).*
** Also known as DAPA transaminase*
+ The other 3 FunFams are also in class-III

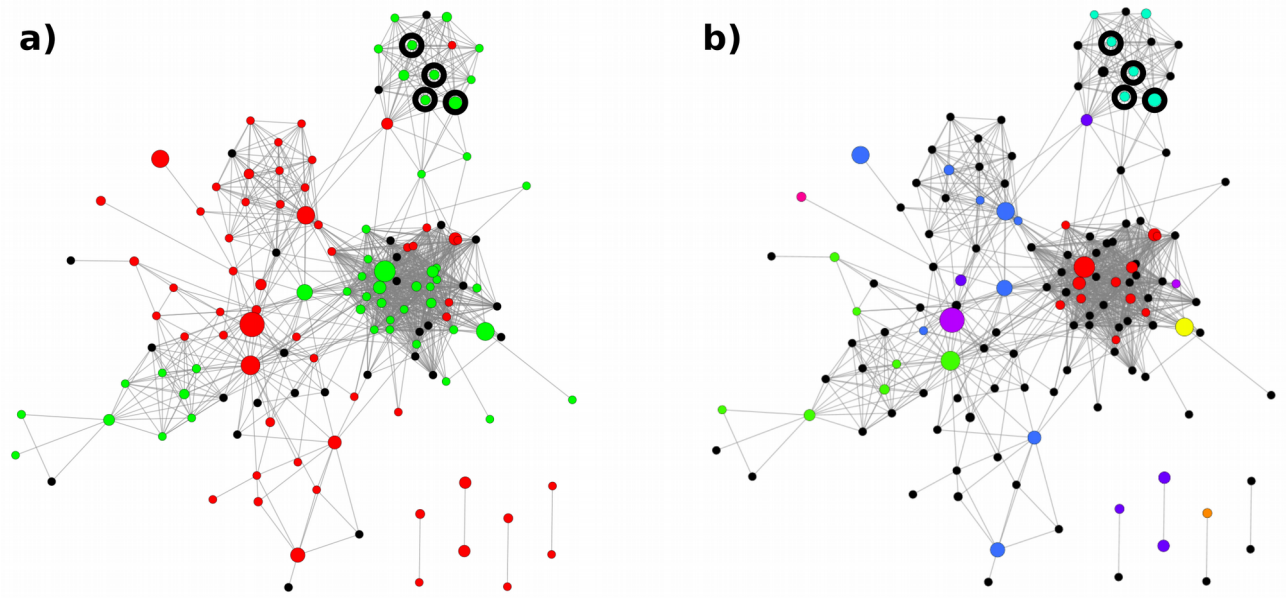

Figure 1.7 *Network diagrams of 3.40.640.10 FunFam sequence profiles. Nodes represent FunFams and are linked by edges if their pHMM similarity score is above 15, as calculated by Profile Comparer (Madera, 2008). The size of the nodes are proportional to the number of sequences in the FunFam. The top right cluster of FunFams essentially represents class-III TAms. The FunFams highlighted with bold circles are the focus of this project. (a) The FunFams are coloured green if >50% of EC annotations are 2.6.1.-, red if not, and black if there are no EC annotations associated with sequences in this FunFam. (b) The FunFams are coloured according to 5Å structural clusters. The black nodes are FunFams without a structural representative in CATH v4.1. The sequences of FunFams within the same structural cluster group together fairly neatly, so the sequence profile network can be used to confidently predict the structural cluster of many of the as yet structurally uncharacterised FunFams.*

2 MATERIALS AND METHODS

2.1 Input Data: Assembled Contigs from Sequenced Metagenomes

The environmental samples were collected by John Ward's group at UCL (Dept. of Biochemical Engineering). DNA was obtained from a number of different locations in Peru, including two salt evaporation pond samples – pilluana and maras3 – which were analysed in this project. Two samples were used to increase the chance of finding novel sequences and mutations. They were

sequenced with the Illumina MiSeq platform. The resulting sequence reads were quality-filtered by Dragana Dobrijevic with trimmomatic (Bolger, Lohse and Usadel, 2014). MEGAHIT (Li *et al.*, 2016) was used by Natalie Dawson (Orengo group) to assemble the quality-filtered reads into contigs. A kmer size of 141 was used. The final assembly has a 500 base cut-off. The distribution of contig lengths for each sample is shown in figure 2.1.

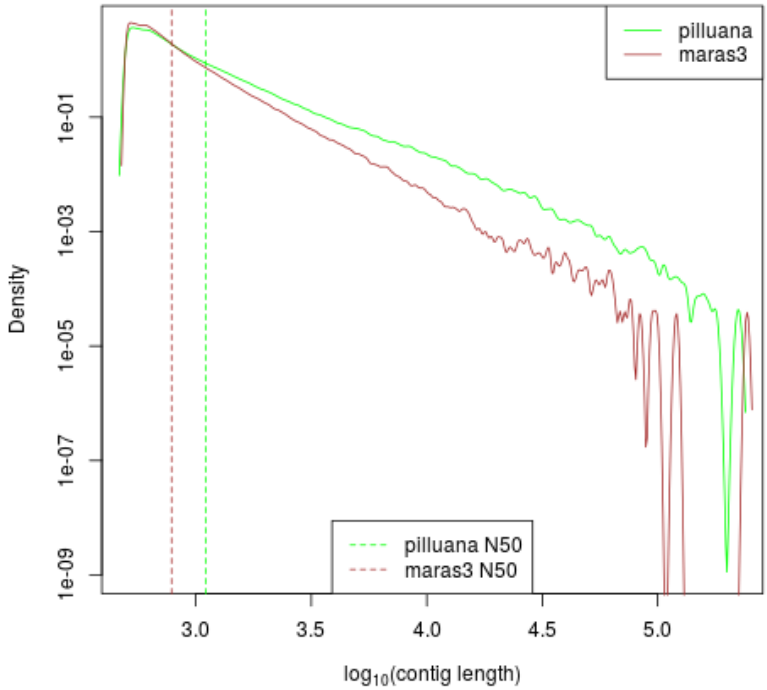

Figure 2.1 *Log-log density plot of the distribution of contig lengths in the two samples. The density is relative to the overall amount of DNA in each sample. N50 is a weighted median, such that 50% of all nucleotides are contained in contigs longer than this value.*

Sample	**Pilluana**	**Maras3**
Total length	933 Mbp	1,214 Mbp
Number of contigs	857,886	1,437,380
Min. contig length (due to cut-off)	500 bp	500 bp
Average contig length	1,087 bp	844 bp
Max. contig length	223,333 bp	243,827 bp
N50	1,104 bp	785 bp
Contigs > 5 kbp	14,700 (total: 159 Mbp)	6,365 (total: 56 Mbp)
Average fold coverage	5.25	4.55

Table 2.1 *Key statistics relating to the assembly quality.*

More DNA was recovered from the maras3 sample (table 2.1), but the quality of the contig assembly is better for the pilluana metagenome. This translates into a marginally higher information

content for the pilluana metagenome in terms of the number of full-length genes that were predicted and, more specific to this project, the number of domain sequences in the CATH superfamily 3.40.640.10 (see table 3.1).

2.2 ORF Prediction

There are many different programs available for open reading frame (ORF) prediction. FragGeneScan, which combines codon usage patterns and sequencing error models into an HMM, was used here as it has been shown to be superior for shorter sequence reads (Rho, Tang and Ye, 2010). In addition, it finds fragmented genes as well as full-length genes. Whole domains can be recognised within these fragments, which adds to the overall information content extracted from the sample (see fig. 2.2). While these additional domains cannot be physically cloned and expressed as part of a whole gene from the sample, the sequences can still be analysed for the identification of novel mutations in SDPs. They can also give a more reliable estimate of the gene composition helping to guide future sampling locations (assuming a location with similar conditions is likely to contain a similar proportion of sequence families).

Illumina MiSeq has an error rate of ~ 0.1% (Ross *et al.*, 2013), so a 0.1% error model file was downloaded from *https://github.com/hallamlab/FragGeneScanPlus/blob/master/train/illumina_1* and placed in the 'train' directory of FragGeneScan. The program was run with the following options:
-complete=0
-train=illumina_1
-threads=1.
It took 58 minutes to run FragGeneScan on the pilluana contigs (~ 16.1 Mbp/min), and 83 minutes for the maras3 contigs (~ 14.7 Mbp/min). The number of ORFs predicted for each sample is shown in table 3.1.

2.3 Redundant Sequences

Fastx_collapser (Gordon and Hannon, 2010) was used to remove duplicates from the fasta files of predicted ORFs.

2.4 Discriminating Between Full-length and Fragmented Gene Sequences

Fragmented gene sequences are missing either the C- or N-terminal section, which means primers cannot be designed to clone the whole gene for expression and functional assays. Therefore it is important to know which genes are full-length and which are just fragments being used for domain sequence analysis. A python script (get_full_gene_ids.py) was written to work out which of the predicted ORFs were full genes, and write a list of their gene identifiers (fasta headers). This list was imported and used in other scripts such as the stats tables of TAm hits.

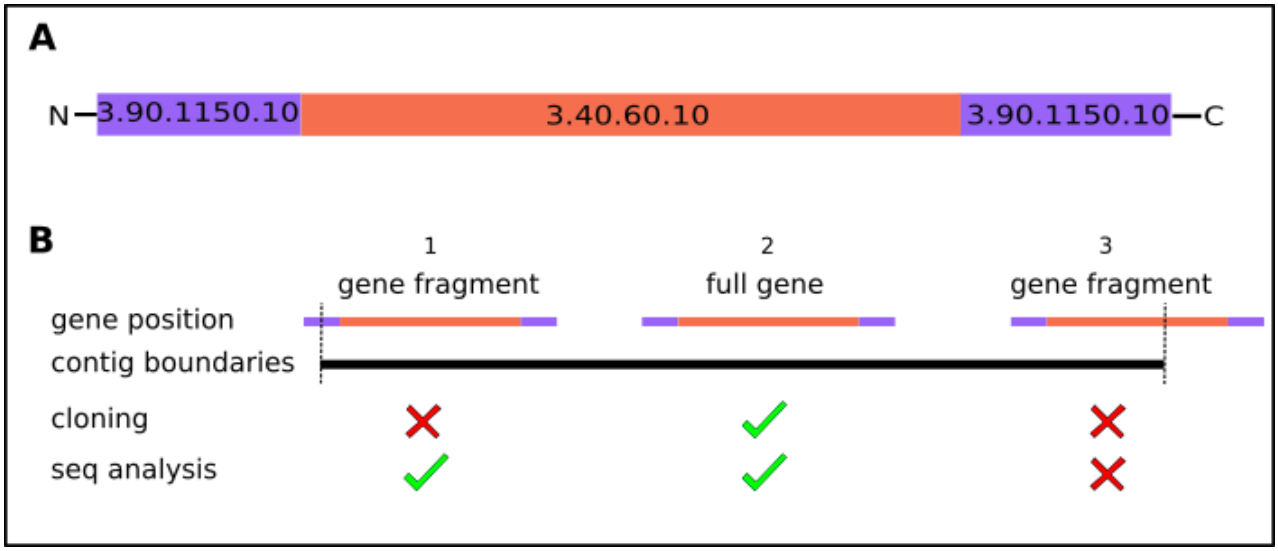

Figure 2.2 *The rationale behind scanning gene fragments as well as full-length genes. (a) In the most common multidomain architecture of TAms, a discontinuous small domain flanks the major domain on both sides. (b) 3 possible positions for a TAm gene in a contig sequence. Many of the incomplete gene fragments found at the ends of contigs still include the full major domain, as in position 1. Gene fragments in position 3 have a truncated major domain. They may still weakly match a FunFam profile, but were discarded from the analysis (using a domain length cut-off of 200, which is ~ 75% of the average domain length) since the match score is distorted by the missing sequence.*

The get_full_gene_ids.py script sacrifices some sensitivity for specificity – about 0.1-0.2% of full-length genes will have been labelled as fragments – as it cannot be determined with certainty whether genes that start at exactly the first position of a contig actually start here or are fragmented. This is because translation initiation is known to occur from 47 out of the 64 codons (Hecht *et al.*, 2017), all of which also occur in the coding sequence. The 3 stop codons however cannot occur in the coding sequence (except in selenoproteins and some methyltransferases where TGA and TAG

code for selenocysteine and pyrrolysine respectively), so full-length genes that end right on the contig boundary can be distinguished from fragmented genes that extend beyond the sequence that has been captured in the contig.

2.5 Pfam-CATH Mapping

The 6 TAm families in Pfam were used as a starting point for sequences with known or strongly predicted TAm function. Pfam is a semi-manually curated database where proteins are grouped into families based on sequence similarity (Finn *et al.*, 2016). The size/diversity of a Pfam family is somewhere between a CATH superfamily and FunFam. Pfam version 29.0 contains 16295 families, while CATH v4.1 groups proteins into 2737 superfamilies and 92,882 FunFams in Gene3D. The domains may also be chopped differently – as is the case for TAms.

The Pfam seed sequences (v29.0) were downloaded for each of the 6 TAm families (PF00155, PF00202, PF01063, PF00266, PF01041, PF12897) and run through genomescan. Five of the six families map onto a 3.40.640.10—3.90.1150.10 multi-domain architecture, with the remaining one (PF01063, a.k.a. Amino-transferase class IV) mapping onto a 3.30.470.10—3.20.10.10 multi-domain architecture. Amino-transferase class IV was not studied here. The major domain (3.40.640.10) was chosen as the subject of this study as it contains the main catalytic residues – this was confirmed by looking at the M-CSA entries under 2.6.1.

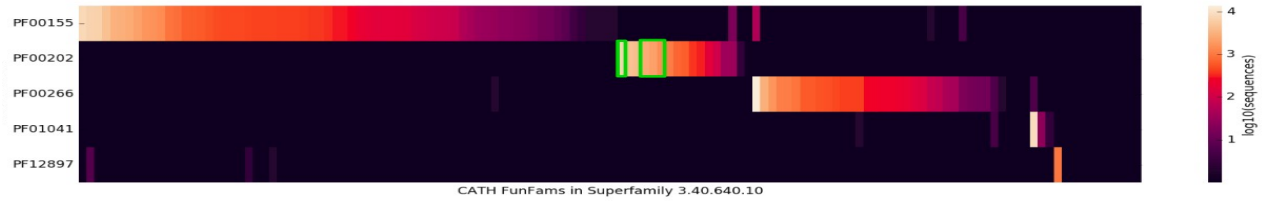

Figure 2.3 *Heatmap of the Pfam-FunFam mapping. Genomescan was used to scan all sequences in the 5 Pfam families (version 29.0) that were identified as TAms belonging to the 3.40.640.10 homologous superfamily in CATH. The FunFams are sorted along the x-axis in such a way that makes it easy to identify 'fuzzy' areas of the mapping, where a FunFam overlaps from one Pfam to another. As the mapping shows, there are very few cases of this, so there is near enough to a one-to-many mapping for each of the Pfam families. Most of the 134 FunFams have at least one match in the 5 Pfam families, but some are sparsely represented and 10 FunFams (right side of graph) have no matches.*

The full set of sequences for the 5 Pfam families (v29.0) with a 3.40.640.10 structure were then downloaded and run through genomescan to give a detailed Pfam-FunFam mapping (fig. 2.3). The superfamily 'purity' of the Pfam families was 98.2%; the 1.8% of sequences that didn't contain a

match to a 3.40.640.10 FunFam could be anomalous fragments. The purpose of the mapping was partly to demonstrate that all of the 4 FunFams of interest, highlighted in green in fig 2.3, map onto the same Pfam family – PF00202, a.k.a. Amino-transferase class III.

2.6 Genomescan

Genomescan is a tool developed by the Orengo group at UCL and is freely available at *https://github.com/UCLOrengoGroup/cath-tools-genomescan*. It uses HMMER3 (Eddy 2009) and a library of FunFam sequence profiles to generate CATH assignments for large datasets of query sequences. Here, genomescan was used on the pilluana and maras3 genes and gene fragments. The FunFam library was from the v4.1 release of the CATH database. The genomescan program generates a 'cath-resolve-hits' (.crh) file as one of its outputs. This is a flat text file where each row denotes a domain assignment and associated information such as the sequence range and E-value score.

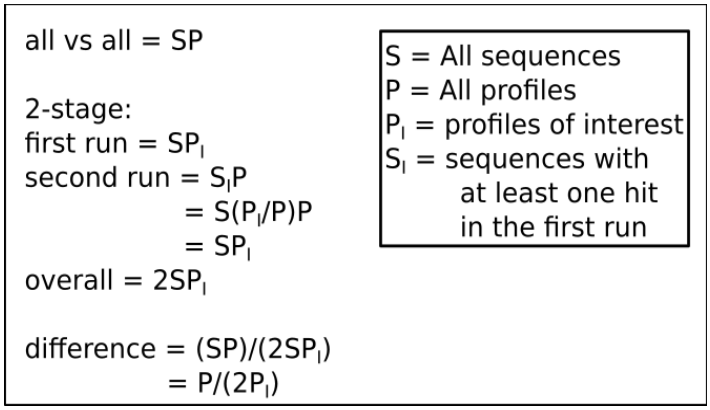

Figure 2.4 *Equations for number of scanning operations required for an all vs all genomescan compared to a 2-stage scan, where the time taken for a scan is assumed to be proportional to the number of sequences multiplied by the number of profiles being used. Assuming the subset of profiles of interest matches a proportional subset of the total sequences, S_I is equal to $S(P_I/P)$. The final result means that a 2-stage scan would be faster than an all vs all scan provided the number of FunFams of interest is less than half of the total library of FunFams. In this project, there are 134 FunFams of interest out of a total of 92882, so a 2-stage scan is theoretically 92882/(2*134) = 346.6 times faster.*

Running an all vs all scan (~ 2 million genes per sample vs 92,882 profile HMMs) locally on an average laptop (8GB RAM, intel i3 processor) would have been too inefficient. Using a 2-stage genomescan was predicted to be up to 350 times quicker (see fig. 2.4) – this is the difference between 1 hour and over 2 weeks. While this method is convenient for large datasets when time and

computing power is limited, an all vs all genomescan is recommended for a more comprehensive analysis of more than one CATH superfamily.

The first stage was a scan of the sample genes with only the 134 FunFams of interest. This required a python script (extract_funfamHMMs.py) to extract the desired subset of profile HMMs from the full library; this subset was then indexed using the hmmpress function in HMMER3. The subset of genes that hit these FunFams (~ 0.5%) was then scanned against all 92,882 funfam profile HMMs. Another python script (extract_run1hits_genes.py) was written to separate the subset of genes to be used as input. The second stage was necessary because very weak hits below the FunFam inclusion threshold are picked up in the first run. The weak hits that matched a FunFam in a different (non-TAm) CATH superfamily more strongly in the second run were ignored for the further analysis. The second genomescan run was also necessary to get more accurate resolved domain boundaries, since all of the adjacent domains are assigned.

2.7 Generating Tables of Hits and Choosing Structural Cluster

A python script (create_funfam_table.py) was written to create a flat text table of FunFams. Web scraping was used to extract information from *http://www.cathdb.info/version/v4_1_0/superfamily/3.40.640.10/alignments*. Then, the list of FunFam IDs was used to download stockholm files for each FunFam, and EC annotations were extracted and counted to determine the prevalent functions within each FunFam. The full list of Swiss-Prot IDs was used by the script to determine which EC annotations are associated with manually annotated and reviewed sequences. A FunFam-SC mapping, provided by Sayoni Das, was then merged with the FunFam table using pandas.

A python script (genomescan_stats.py) was then written to read in the genomescan output (.crh file) and generate several reports. This included tables of hits per FunFam – separated by sample, strength of hits, and whether the hit is in a full-length gene or not – and subsets of the crh file that correspond to each table. Weak hits are domain assignments that do not meet the inclusion threshold for any existing FunFam; these are of particular interest since they represent putative novel FunFams while being similar enough to an existing FunFam to assume structural homology. The reporting E-value threshold used by HMMER3 is 10.0, so hits with 0.001<E-value<10.0 were disregarded from the analysis as they are considered too distant from any known FunFam.

The hits tables were then merged to the FunFam table, which was grouped by SC and filtered by information content (FunFams with DOPS < 70 were removed). The results were analysed to select a suitable cluster to study – this turned out to be SC4, which is functionally pure at the EC3 level and had a good number of weak hits in the metagenome. A python script (write_fasta.py) was written to extract the fasta sequences of all weak hits to SC4 FunFams from the original file of all predicted ORFs in the metagenome. These are the sequences that were carried forward for further analysis.

2.8 Multiple Sequence Alignments and GroupSim

In order to investigate interesting novel mutations in metagenome sequences, the specificity determining positions (SDPs) were first predicted among the FunFams of interest within SC4. When analysing two or more groups of structurally homologous proteins with slightly different functions, an SDP is highly conserved within a group but not between groups.

The following process was carried out for each of the four FunFams (62757, 63097, 63148 and 63154) separately:

1. The stockholm file was downloaded from CATH (*http://www.cathdb.info/version/[version]/superfam/[superfam]/funfam/[funfam]/files/stockholm*).
2. The stockholm file was loaded into Jalview (Waterhouse *et al.*, 2009) and viewed as a multiple sequence alignment.
3. Redundancy was removed at 90% sequence identity, making sure to keep the FunFam structural representative in the alignment. The resulting alignment was saved in fasta format. This step ensures that the MSA isn't biased towards redundant sets of sequences (for instance, there are often identical sets of sequences from different chains of a single homooligomer PDB entry).
4. A python script (groupsim_prepare.py) was written to add the FunFam ID to the end of the fasta header (so it is recognised by the GroupSim program) and to the start of the header (so it is easily apparent when viewing alignments in Jalview).

All four FunFam alignments were then loaded into the same Jalview window and MAFFT (Katoh and Standley, 2013) was run with default settings. The program GroupSim (Capra and Singh,

2008) was then used to calculate SDP scores for each position in the alignment. GroupSim is already used as part of the FunFHMMer algorithm that creates the FunFam classification (Das, Lee, *et al.*, 2015), but there it only compares pairs of groups. Here, a 4-way comparison was used to explicitly show which are the SDPs that were factors in the differential sub-classification of the four FunFams. Standard settings were used (window size = 3, lambda = 0.7). GroupSim calculates a score based on the relative intra-group to inter-group conservation at each position in the MSA. A score less than 0.3 indicates a conserved position, a score greater than 0.7 suggests an SDP, and anything in between indicates a nonconserved position. However, this threshold can be relaxed when comparing more than two groups. The average number of SDPs predicted by the 6 combinations of pairwise GroupSim calculations was 6.67, compared to only 1 SDP for the 4-way comparison. Therefore, the SDP threshold was reduced to 0.6, which resulted in 6 predicted SDPs (close to the pairwise average).

A python script (sort_groupsim_scores.py) was written to read in the GroupSim output and write an output file with a list of the highest scoring positions (user-defined percentage). The script also writes a chimera select command that can be used to quickly select and colour the SDPs in a representative structure. The custom module ‘msa_position_converter.py’ was imported and used to translate from the MSA positions to the representative structure positions.

2.9 Identifying Novel Mutations in SDPs

The fasta file of weak hits to the four FunFams of interest (see section 2.7) was combined with all members of the four FunFams and a new MSA was created with MAFFT (default settings). The python script ‘novel_mutant_finder.py’ was written and used on this MSA. The top 6 predicted SDPs (representative structure numbering) were taken as an input, and the script imported the msa_position_converter module in order to translate to the correct positions in the MSA. In this way, the corresponding positions in the GroupSim input (MSA with only the FunFam members) and the MSA with metagenome sequences were determined by using a reference sequence that is present in both alignments, thus bypassing the need to manually refer to the alignments. The novel_mutant_finder.py script produced two tables: a summary of all the novel mutations in the SDPs and a full list of all the novel mutant sequences. A novel mutation is defined as a new amino acid in an SDP (not seen in any of the existing FunFam sequences), as shown in figure 2.5. The

novel mutation can then be simulated *in silico*, either as a point mutation or homology model of the full sequence.

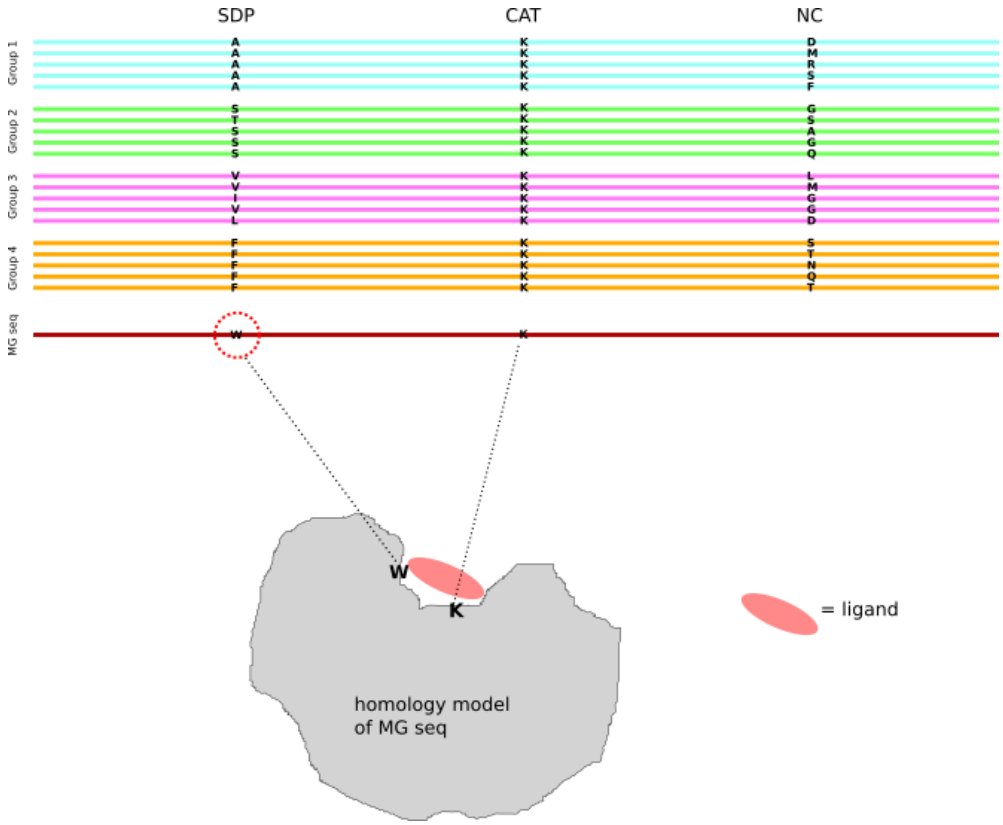

Figure 2.5 *Schematic diagram of the process of SDP prediction and novel mutant identification. CAT = catalytic residue; NC = nonconserved position. Each metagenome sequence (MG seq) is checked in the catalytic residue positions and SDPs. Novel mutants are when the catalytic residue is conserved but a new amino acid is observed in the SDP. A homology model of the MG seq can be built (or just a single SDP mutation simulated), using a representative FunFam structure with known function as the template.*

2.10 Mutation Modelling

Mutations were modelled in UCSF Chimera (Pettersen *et al.*, 2004) and optimised using MODELLER (Webb and Sali, 2016). The idea was to mimic some of the novel SDP mutations observed in the metagenome and predict their effects in isolation. The following steps were carried out for each mutation:

1. Select an appropriate template structure from the nearest matching FunFam.
2. Mutate residue.

3. Select sensible-looking rotamer (as few clashes as possible, and potentially stabilising interactions with neighbouring residues) and check for clashing residues.
4. Model/refine loops, with only the mutated residue and any clashing residues selected. With hindsight, re-modelling a 3-residue window around the mutation may have been more appropriate.
5. Check for steric clashes once more in the resulting model; if there are any, select the new clashing residues and model/refine loops again. If there are major clashes that cannot be resolved this way, it suggests this mutation is not stable on its own and must be accompanied by one or more co-mutations.

The mutant models were then compared to the original structures, paying particular attention to changes in hydrogen bonding, chemical environments of key residues, and the volume and electrostatics of the binding pocket. The normalised Discrete Optimised Protein Energy (zDOPE) scores were compared for the modelled mutant and the wild-type.

2.11 Caveats and Limitations of the Methodology

The analysis of SDPs is only as good as the information that goes into it. There are potential issues to do with incompleteness and impurity; SDPs are more significant when derived from diverse and distinct specificity groups.

The current sequence databases are incomplete representations of the sequence diversity found in nature, so many true SDPs may not be identified. This problem will shrink as the sequence databases continue to grow, especially with metagenomic sequences. Version 4.1 of CATH-Gene3D was used in this project, but the FunFams in v4.2 are significantly larger and more diverse.

The functional purity of the CATH classification at the level of FunFams and structural clusters of FunFams must also be considered. While FunFams are currently among the most functionally coherent classifications in use (Jiang *et al.*, 2016), the intrinsic complexity of the sequence to function relationship in proteins, as well as phenomena like moonlighting and promiscuity, means that most FunFams are not EC-pure. Note that 3.40.640.10 is the only catalytic domain in TAms, so there is no issue of EC annotations coming from other domains. The FunFam classification is better for some superfamilies than others. This is because the equation for the Functional Coherence Index (more specifically, the R_{sdp} value) in the FunFHMMer algorithm (Das, Lee, *et al.*, 2015) is optimised based on the 11 families that make up the core structure function

linkage database (SFLD). TAms are not among the core SFLD, which may mean that the FunFam classification of TAms is slightly too fine or too coarse, depending on if the average ratio of SDPs to conserved positions required for functional diversity between two groups of TAms is more than or less than 1:4 (which is the threshold used by FunFHMMer). An example is FunFam 63148, which may be diverse enough to warrant splitting into around 3 smaller groups.

Many structural clusters of FunFams are even more heterogenous – at the EC3 level. While this is not the case for the SC studied here, it is a general consideration for more heterogenous SCs; a more careful analysis of SDPs in the subset of FunFams with desirable functions versus the other FunFams may be required to understand the variations that lead to changes at the EC4 level as opposed to more drastic EC3 changes.

The modelling of single point mutations based on a single crystal structure does not take into account large-scale conformational changes that may occur at different stages of the enzyme reaction; therefore the mutation modelling is limited to the specific structural context in which it is carried out. Also, the zDOPE scores only indicate if a mutation is stable in the ground state of the enzyme; it gives little indication as to the effect on the stabilisation of transition states, which may be crucial for the catalytic rate.

A limitation of studying isolated mutations in a few residues is that it can never offer a complete picture of all the subtle contributions towards differing substrate specificities. For example, the rigidity of the protein scaffold, which was posited as an explanation for the difference between omega-TAms from *Chromobacterium violaceum* and *Pseudomonas aeruginosa* (Sayer *et al.*, 2013), or the electrostatic surface of the entrance to the binding pocket, which influences the k_{on} parameter by attracting substrates into the binding site at a faster rate than random diffusion. These effects are usually achieved by many residues together – each one contributing only a small part. For this reason, they are unlikely to show up in the SDP analysis.

It is important to point out that the novel mutations in SDPs shouldn't be linked too strongly to one particular FunFam – the mutant sequences are only distantly related to the existing FunFams, so the findings are best viewed at the level of the SC.

3 RESULTS

3.1 Metagenome TAm Statistics

Salvaging full domain sequences from gene fragments in the metagenome contigs more than doubled the number of domains retrieved for analysis (see table 3.1). For the weak hits, this allowed for the identification of many more novel mutations in SDPs, especially for the maras3 contigs, which were of lower quality.

Metagenome statistics			
Sample Name →	**Pilluana**	**Maras3**	**Total**
Contigs	857,886	1,437,380	2,295,266
Total ORFs	1,460,829	2,197,977	3,658,806
Full-length ORFs (% of total)	301,436 (20%)	290,016 (13%)	591,452 (16%)
3.40.640.10 strong hits (E-value < inclusion threshold)			
Domains in full-length genes	854	572	1426
Domains in fragments	725	816	1541
Totals	1579	1388	2967
3.40.640.10 weak hits (0.001 > E-value > inclusion threshold)			
Domains in full-length genes	551	397	948
Domains in fragments	444	553	997
Totals	995	950	1945

Table 3.1 *The number of domain hits for 3.40.640.10 FunFams in each metagenome sample.*

It should be noted that only 48 of the 134 FunFams in 3.40.640.10 have a high alignment diversity (DOPS > 70). The number of weak hits may be artificially inflated by the stringent inclusion thresholds of the low diversity funfams. A weak hit is more significant for a high DOPS FunFam, since most of the natural diversity of that FunFam is already known, so not meeting the more flexible inclusion threshold is more likely to indicate true sequence and function novelty. Also, the hits in table 3.1 are not purely TAms – they include other enzymes in the functionally heterogenous 3.40.640.10 superfamily. To focus in on the TAms in high diversity FunFams, the hits were broken down by FunFam and filtered by DOPS and EC (see tables A1 and A2).

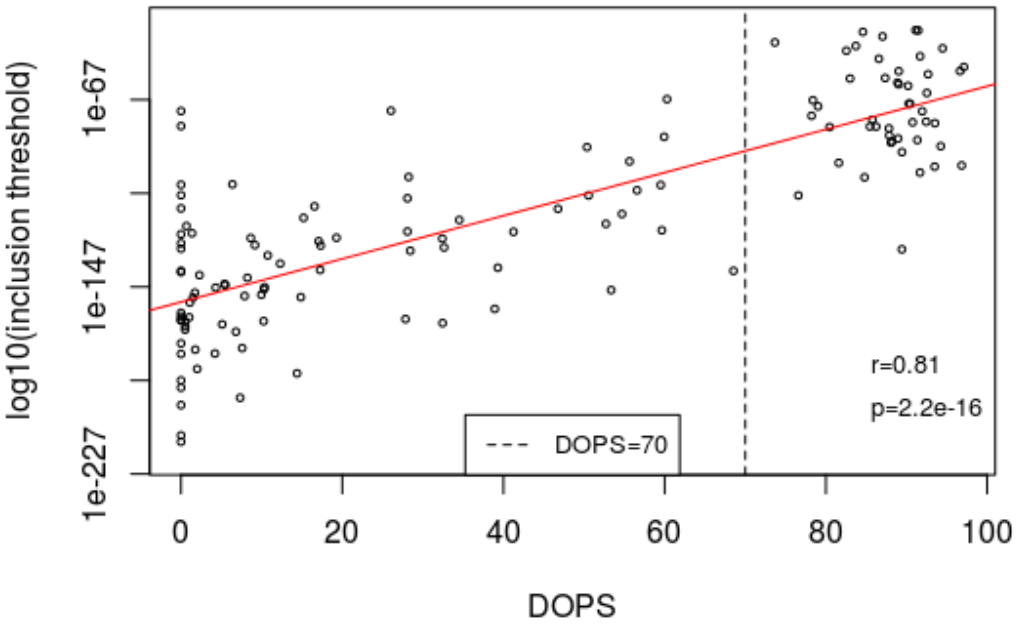

Figure 3.1a *All 134 v4.1 FunFams in 3.40.640.10 are plotted as DOPS vs E-value inclusion threshold. The r- and p-values displayed refer to Pearson's product-moment correlation.*

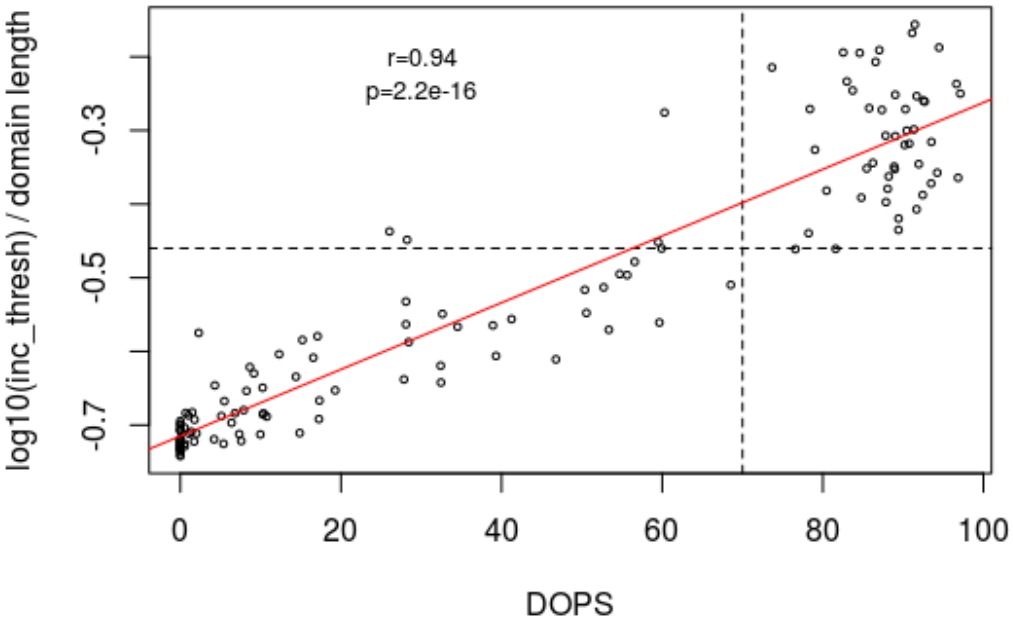

Figure 3.1b *When the domain length of each funfam is taken into account, the correlation between DOPS and inclusion threshold becomes stronger. Scores on the horizontal dashed line would be considered weak hits (potentially novel FunFams) to all but 4 of the low DOPS FunFams, but would meet the inclusion threshold of all of the high DOPS FunFams. The r- and p-values displayed refer to Pearson's product-moment correlation.*

Figure 3.1 explains the rationale behind leaving low diversiy (DOPS < 70) FunFams out of the analysis of weak hits (distant homologues). The average DOPS for v4.1 FunFams in 3.40.640.10 is 42.32, while it has increased to 54.77 for v4.2. This means many of the weak hits in low DOPS v4.1

FunFams would in fact be accepted into the updated FunFams with higher DOPS. These sequences could be useful for fleshing out the existing low DOPS FunFams but are far less likely to represent novel FunFams, which is why they have been left out of the analysis in this project. Some examples of low DOPS FunFams are 30569, 62445 and 63291, which together have 96 weak hits and 0 strong hits in the two metagenomes. It is much more likely that this statistic is due to the low information content and stringent inclusion thresholds, rather than these FunFams having had a very large selection pressure to functionally diverge in this environment.

Table A1 gives some insight into the different TAm compositions of the two samples, pilluana and maras3, while table A2 shows the potential of each sample environment for identifying putative novel FunFams. For instance, the pilluana metagenome contains almost double the weak hits to aromatic amino acid aminotransferase (FF63364). Interestingly, comparing the two tables shows that the number of strong hits is not a good predictor of the weak hits related to each FunFam. For instance, FF63364 has very few strong hits but the second highest number of distantly related hits in potentially novel FunFams. This means that users of the genomescan tool need to be aware of the FunFam-specific inclusion thresholds in order to calculate the number of truly novel sequences, since it will not be apparent just from looking at the total number of reported hits.

3.2 Weak Hits Grouped by Structural Cluster

Version 4.1 of the CATH database only has structural cluster information at 9Å resolution. The FunFams are too structurally heterogenous at this level (all FunFams in 3.40.640.10 fit into 2 structural clusters). Therefore, the 5Å structural clusters for v4.2 FunFams were used instead. This separates the 3.40.640.10 FunFams into 9 distinct structural clusters. At this resolution, we can be confident that a multiple sequence alignment of structurally similar FunFams will align the structurally equivalent residues correctly. The v4.2 structural clusters are applicable to v4.1 sequence information since the sequence profiles have remained broadly the same.

Table 3.2 shows the combined weak hits for both metagenomes. Structural cluster 4 was chosen for further analysis (not including 62814, which has no weak metagenome hits). In these four FunFams, 126 of the weak domain hits came from full length genes while 169 weak hits came from gene fragments. It also breaks down as 158 from the pilluana sample and 137 from maras3.

Structural clusters 1 and 3 would also be candidates for additional studies of TAms within this superfamily.

FF	Modal EC (SwissProt annotations)	DOPS	FunFam Size	Structural Representative	5Å Structural Cluster	FunFam hits	SC hits (of which in a TAm FF)
63355	2.6.1.1	91.129	656	1u08B02	1	58	220 (217)
63364	2.6.1.39	92.682	178	4gebB02		110	
63298	2.6.1.9	84.591	243	1geyA01		5	
63324	2.6.1.1	89.043	60	5bj4B02		17	
62993	2.6.1.9	79.023	20	3cq5A02		7	
63270	2.6.1.5	91.953	93	3dydB02		16	
63395	4.4.1.14	87.359	236	3ihjA03		3	
63401	2.6.1.83	87.036	62	3ei5B02		0	
63411	2.6.1.9	91.689	85	3eucA02		4	
63379	2.6.1.1	91.465	505	3tatE02	2	5	5 (5)
62618	2.6.1.37	88.078	29	1m32E02	3	8	212 (49)
62953	2.6.1.44	94.245	92	3r9aA02		31	
63117	3.7.1.3	88.956	60	3e9kA02		8	
63436	2.6.1.52	94.475	161	3e77C01		10	
63415	2.8.1.7	82.542	558	1p3wA02		155	
62814	2.6.1.19	91.342	85	1ohwD02	4	0	295 (295)
63097	2.6.1.19	91.677	88	1sffC02		72	
63154	2.6.1.11	90.450	273	4addD02		51	
63148	2.6.1.62	96.847	86	4ba5B02		117	
62757	2.6.1.62	84.795	78	3du4A02		55	
63242	2.5.1.73	85.465	21	3wkrA02	5	0	270 (45)
63248	4.1.1.28	85.773	328	3rchB02		171	
63300	2.1.2.1	89.025	478	1dfoD01		14	
63333	2.6.1.50	86.577	391	1b9iA01		45	
63070	2.3.1.37	92.444	104	2bwoB02		1	
63285	4.1.1.15	78.394	268	4q6rB03		17	
63207	2.3.1.47	73.671	499	1fc4B01		19	
63292	2.3.1.47	80.463	22	3wy7B02		3	
62877	1.4.4.2	87.824	188	1wyuB03	6	2	126 (0)
63076	1.4.4.2	92.527	189	1wyuG02		6	
63127	4.1.99.1	90.200	67	2vlhB02		1	
63385	5.4.3.8	96.643	176	2hp1A02		42	

62952	4.1.2.5	86.233	150	3wlxA01		75	
63132	2.6.1.99	81.600	30	3bwoD02	7	1	32 (1)
63352	2.5.1.48	82.995	834	2gqnB01		31	
62900	4.1.1.18	89.427	70	2vycD02	8	26	26 (0)
62879	2.9.1.1	90.763	67	3w1hA02	10	11	11 (0)

Table 3.2 *Weak hits grouped by 5Å structural clusters. FunFams in green are probable TAms (based on the most common SwissProt EC annotations). FunFams with DOPS < 70 are not included. Hits include domains in gene fragments as well as full-length genes.*

3.3 SDPs in SC4 FunFams

GroupSim Score	Residue positions in PDB structures					
	FunFam 62757		FunFam 63097	FunFam 63148		FunFam 63154
	3du4	4cxq	1sff	4ba5	4e3q	4add
0.79093	254	257	242	262	259	226
0.69785	149	160	141	156	153	141
0.64389	226	229	214	234	231	198
0.64334	223	226	211	231	228	195
0.64172	316	317	296	323	321	280
0.63539	119	130	117	126	123	111

Table 3.3 *The top 6 GroupSim scores for the 90% non-redundant multiple sequence alignment of structural cluster 4 FunFams. Residue positions are shown for structural representatives of each FunFam and other structures referred to in this study.*

Table 3.3 shows the top scoring residues from the SDP-prediction tool GroupSim. Figure 3.2 shows a sequence logo – a graphical representation of the conserved residues and motifs within each of the SC4 FunFams. The FunFams 63097 and 63154 appear to be similar to each other based on several SDPs – R149, E223 and Q254. FunFam 63148 is the most diverse – several of the SDPs have a lower conservation than the other FunFams. For instance, the top-scoring SDP, residue 254, is split between isoleucine, valine and methionine. This diversity is mirrored in the EC annotations for FunFam 63148, which are evenly split between reactions 2.6.1.62, 2.6.1.18 and 2.6.1.77 – each represent about 25% of the total EC annotations in v4.2 of the FunFam.

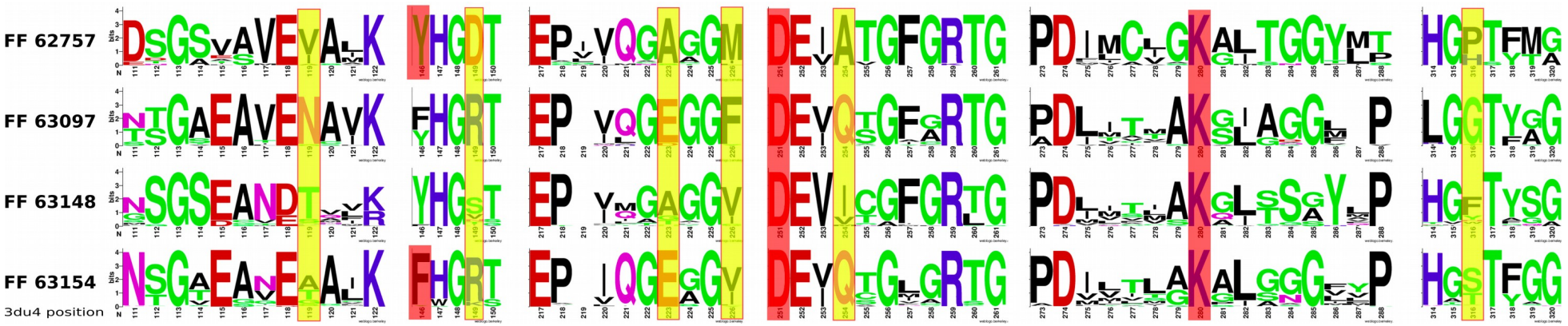

Figure 3.2 *Sequence logo of selected regions of the multiple sequence alignment between 90% non-redundant members of 4 of the FunFams in structural cluster 4. The positions are based on the 3du4 sequence (FunFam 62757 structural representative). Known catalytic residues (based on M-CSA annotations) are highlighted in red, while the top 6 predicted SDPs are highlighted in yellow. The height of each stack corresponds to the sequence conservation at that postion, while the height of the letters is proportional to the relative frequency of each amino acid at that position. Sequence logos were generated using WebLogo (Crooks et al., 2004).*

The locations of the SDPs in 3D space is shown in figure 3.3. Five of the six top-scoring SDPs are relatively close to the catalytic residues or cofactor in the active site.

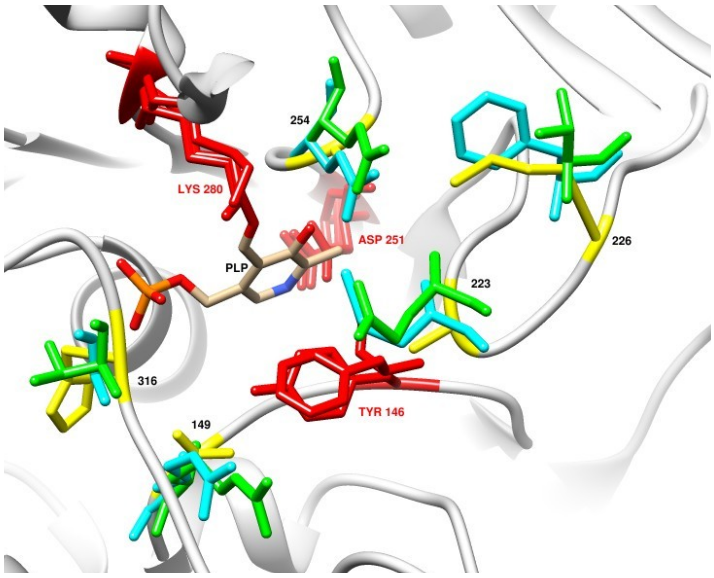

Fig 3.3 *Specificity determining positions in SC4 FunFam representatives. The view is looking into the binding pocket from outside. Catalytic residues are shown in red and SDPs are shown in yellow (FunFam 62757), cyan (FunFam 63097) and green (FunFam 63154). Residues are numbered according to the 3du4 sequence.*

3.4 Enrichment of SDPs in the Secondary Shell

The enrichment of SDPs in secondary shell residues was calculated, where SDP residues are defined as the top 10% of GroupSim scores and secondary shell residues are defined as being within 5Å of one of the 3 main catalytic residues or the PLP cofactor. The calculation, adapted from Dessailly et al. (2013), is the proportion of SDPs that are in the secondary shell (P_{sdp}) minus the proportion of all residues in the secondary shell (P_a), so that a positive score indicates some level of enrichment.

- $P_{sdp} = 11/24 = 0.458$
- $P_a = 40/270 = 0.148$
- Enrichment score = 0.458 – 0.148 = 0.310

P_{sdp} changes depending on the threshold chosen to define an SDP. If only residues with GroupSim scores greater than 0.6 are chosen, $P_{sdp} = 4/6 = 0.67$, giving a higher enrichment score. Either way, the enrichment scores suggest that the SDPs are good proxies for functionally important sites, based on the assumption that most functionally important sites are within a certain radius of the active

site. Figure 3.4 gives a visual representation of how the SC4 SDPs are distributed throughout the domain, relative to the active site and the other domains in the native structure. The proximity of residue 316 to the active site of the other monomer in the homodimer shows the importance of dimerisation for the function of these TAms.

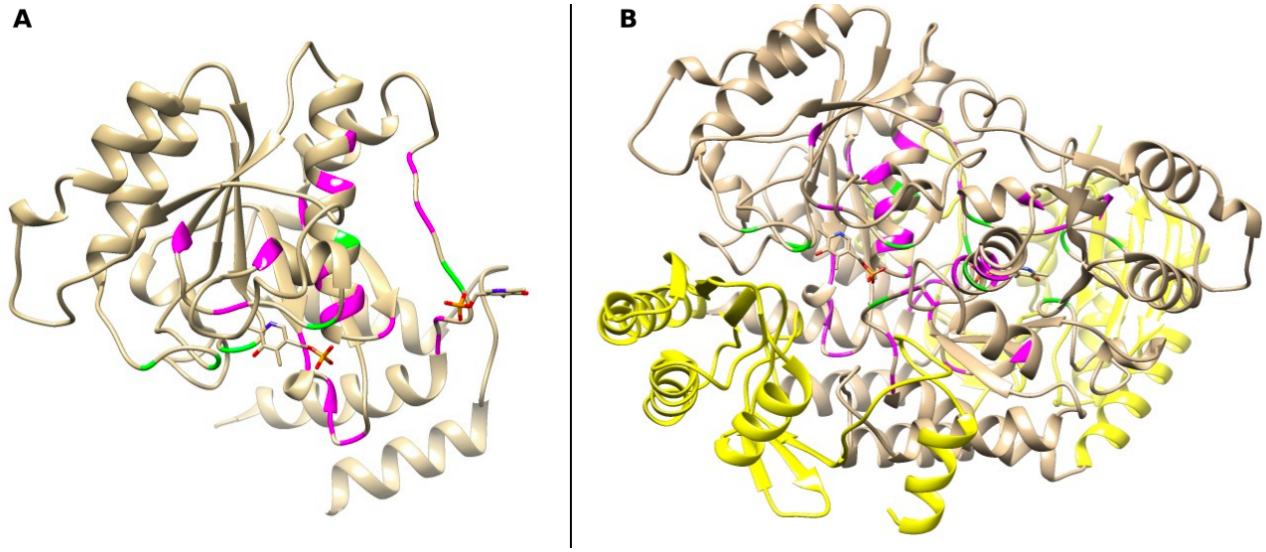

Figure 3.4 *SDP enrichment in the secondary shell. (a) Major domain of 1dty, a member of FunFam 62757. The PLP cofactors for both active sites of the homodimer are shown in stick representation, but only one domain is shown. The positions with GroupSim scores > 0.6 within structural cluster 4 are coloured green and the rest of the top 10% of GroupSim scores are coloured magenta. The SDPs are significantly enriched in secondary shell residues, especially the loop regions around the active site, as compared to the background proportion of all residues in the secondary shell. The long loop on the right of the structure is in proximity to the PLP from the other active site. (b) The complete 1dty homodimer, with small domains coloured yellow.*

3.5 Novel Mutants in SDPs

The putative novel FunFam sequences selected from the metagenomes are close enough in sequence (often around 40-50% identity to the representative structures of the closest matching FunFam) to be tentatively considered part of the same structural cluster. Therefore, a sequence alignment of these weak hits with the existing SC4 FunFam members could be used to identify novel mutations in SDPs. Table 3.4 shows all of these mutations observed in the metagenome sequences.

	Position in 3du4					
	119	**149**	**223**	**226**	**254**	**316**
Known amino acids in v4.1 FunFams	ACGHIKLMNSTV	-CDIKNQRSTV	ADEGST	ACFILMVY	AIMQV	-AFGHMNPSTVWY
Novel mutations	DFQQ	AAAAAAAAAAAAAEEEEEEEEEEGGGGGGGGGGGGHHLLLLLLLLLLLLLMMMMMMMWW	IIIIIIKKLLLLLNNNPPRRRRRRRRRRRRRRRRVVVVVVVVVVW	DGNNNNNPQSSSSSTX	FFFFFFFFFFFFFKKLLLLLLLLLLLRSWYYY	DDDDDIIIIIIIIILLLLLLLLLQQQQQ

Table 3.4 *Novel mutations in key SDPs. The first row indicates all types of amino acids found in that position in all FunFams within the structural cluster, while the second row shows each instance of a novel mutation in that position, to give an idea of the relative frequency of different mutations.*

The analysis of the novel mutants gives some insight into the preference for certain mutations in the SDPs. This is useful for narrowing the candidates that can be taken forward for further investigation *in silico* and *in vitro*.

3.6 Modelling Observed SDP Mutations in Representative Structures

The observed novel mutations are highly likely to be stable and functional (except in rare cases of recently formed pseudogenes) and so there must be some way that the general structure of SC4 TAms can accommodate them without requiring too many concurrent mutations. With this assumption, the most frequent mutations in SDPs were modelled *in silico*. The results of this modelling are presented for the highest-scoring SDP – residue 254 in 3du4. This position can be called the ‘DEVx’ motif, where x is the variant residue. DEVx occurs in a loop region at the end of a beta-strand in the central beta-sheet. The aspartate at the start of the motif is a 100% conserved catalytic residue that through a hydrogen bond helps stabilise the PLP cofactor and activate it as an electron sink by increasing its basicity. The next residue, a glutamate, faces away from the active site but plays a structural role and is also 100% conserved in all four FunFams. An isoleucine or valine in the third position stabilises the PLP cofactor with Van der Waals interactions with its aromatic ring. The fourth position, the SDP, sits at the back of the small binding pocket (often called ‘S pocket’), which is responsible for size exclusion near the amino donor/acceptor end of the substrates (Malik, Park and Shin, 2012). This residue is a conserved glutamine in FunFams 63097

and 63154, and an aliphatic residue in the other two (conserved alanine in FunFam 62757 and valine, isoleucine or methionine in FunFam 63148).

3.6.1 FunFam 62757 – Adenosylmethionine—8-amino-7-oxononanoate Transaminase

The DEVx concensus in FunFam 62757 is ‘A’, but it has been observed to mutate into ‘F’ (4 occurences) and ‘L’ (1 occurence) in distantly-related metagenomic sequences. The A257F mutation was modelled from the PDB structure 4cxq.

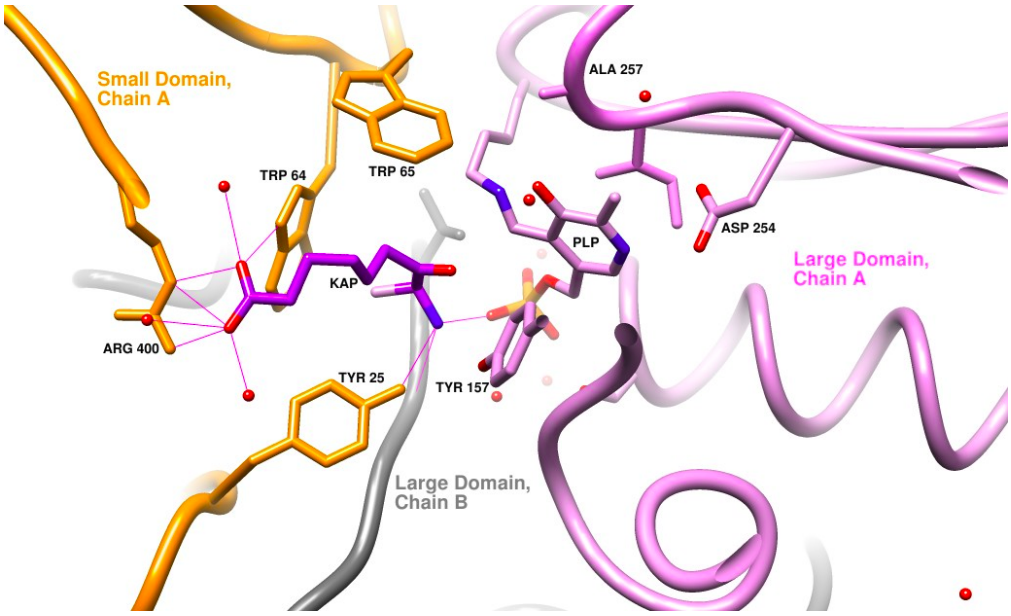

Figure 3.5 *The active site structure of 4cxq. The amino acceptor, 7-keto-8-aminopelargonate (KAP), is shown in dark purple. KAP enters the binding pocket from the left in this diagram.*

In the wild-type structure (PDB: 4cxq, see fig. 3.5), the large domain (FunFam 62757) holds the PLP cofactor in place and provides 3 of the 4 catalytic residues. However, the binding site for KAP is at the interface between the two domains, and KAP forms most of its hydrogen bonds with residues from the smaller domain (CATH superfamily 3.90.1150.10): R400 and Y25. Interestingly for such a strongly predicted SDP, A257 is not in the direct proximity (< 5Å) of the substrate. However, this may change over the course of a reaction, where the substrate forms an external aldimine with the PLP cofactor, bringing it closer to the small pocket of the binding site formed by W65 and A257. This is an unstable intermediate, which normally cannot be captured in a crystal structure without the use of an inhibitor.

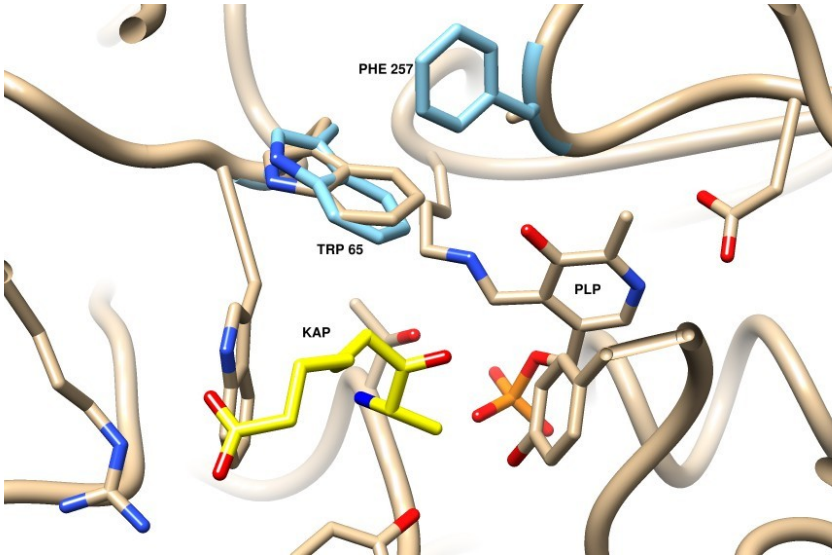

Figure 3.6 *A257F mutation of 4cxq. The mutated F257 can form a T-shaped pi-stacking interaction with W65, in which the W65 sidechain is tilted roughly 1Å closer towards the KAP substrate (in yellow), but doesn't result in a steric clash or a significant change in binding pocket electrostatics.*

The A257F mutant model is shown in figure 3.6. The A257F mutant model has a zDOPE (normalised Discrete Optimised Protein Energy) score of -1.31, as compared to the wild-type score of -1.27. This indicates that the mutation increases the stability of the enzyme. The same mutation also increased the stability of 3lv2, a sinefugin (SAM analogue) bound structure. In both structures, the A257F mutation has no major effect on the size or electrostatics of the active site pocket, so it is unclear what effect the mutation has on substrate specificity. W65 may in fact be covariant with A257, which together with the A257F mutation could cause a larger change around the active site.

3.6.2 FunFam 63097 – GABA Transaminase

Unlike in FunFam 62757, the DEVx SDP is not adjacent to a bulky tryptophan residue from the small domain. In 1sff, the representative structure for FunFam 63097, the equivalent small domain residue is A51, which may mean a larger binding pocket. As noted by Liu et al. (2004), the amide nitrogen of Q242 forms a hydrogen bond with the 3' phenolic oxygen of PLP. It seems likely to be more directly involved in the reaction mechanism than the equivalent aliphatic residues in FunFams 62757 and 63148. However, no fewer than 7 different novel mutations have been observed at this position: 'L' (5 occurences), 'F' (5), 'K' (2), 'Y' (2), 'R' (1), 'S' (1), and 'W' (1). The Q242L mutation was modelled on the 1sff structure. The mutant showed an improved zDOPE score (-1.52) compared to the wild-type (-1.50).

3.6.3 FunFam 63148 – Aminotransferase Class-III

The DEVx variant residue is ‘I’, ‘V’ or ‘M’ in FunFam 63148, and was observed to mutate into ‘L’ (6 occurences) and ‘K’ (1 occurrence) in the metagenome sequences. The I262L mutation was modelled on chain B of 4ba5 (see fig. 3.7), an amine:pyruvate TAm from *Chromobacterium violaceum* (Sayer *et al.*, 2013). In this structure, the inhibitor gabaculine is covalently bound to PLP, mimicking the external aldimine intermediate of a transamination reaction. The I262L mutant has the same zDOPE score (-1.89) as the wild-type structure, suggesting that the mutation can be accomodated without disrupting the stability of the enzyme. The mutation causes the side chain of the neighbouring W60 residue to flip almost 180°, creating more room in the small pocket and potentially allowing larger substrates into the binding site.

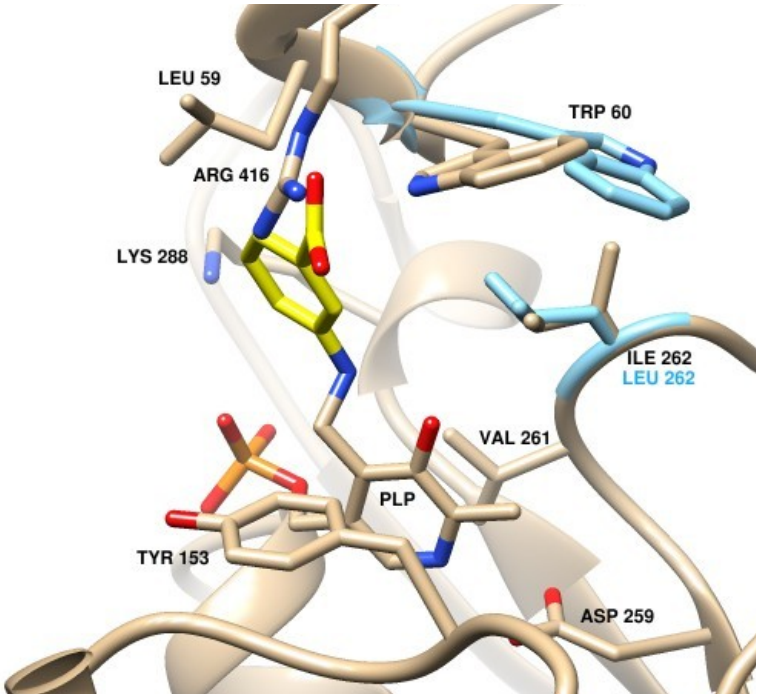

Fig 3.7 *I262L mutation in 4ba5, chain B. The gabaculine inhibitor is coloured yellow, and selected residues in the I262L mutant model are coloured light blue. The small pocket of the binding site is partly formed by I262 and W60. R416 is the so-called ‘flipping argnine’, which is responsible for the dual substrate specificity in many TAms.*

4 DISCUSSION

4.1 DEVx SDP in the Literature

To the best of my knowledge, there is no mention of the significance of this residue in the literature on DAPA TAms (FunFam 62757). Likewise for GABA TAms (FunFam 63097), except for one mention of Q242, stating that it shares a hydrogen bond with PLP, in the paper connected to the 1sff sructure (Liu *et al.*, 2004). This demonstrates the strength of the SDP analysis for uncovering important residues that are not obviously significant from studying the crystal structure.

For FunFam 63148, which contains by far the most studied enzymes of the FunFams analysed here, the story is different. Rausch et al. (2013) identify I259 as a small pocket residue and one of the top 7 scoring residues in an SDP analysis based on a phylogenetic grouping of class III TAms into 6 clades. It is therefore included as part of the signature of omega-TAms. Midelfort et al. (2013) used a bioinformatic approach to identify variant residues in homologues of a *Vibrio fluvialis* TAm. The most beneficial variants turned out to be identified from more remote homologues using the ProteinGPS tool; this supports the rationale behind analysing remote homologues in metagenome sequences in this study. The I259V and W57F mutations provided substantially more room for the ethyl moiety of the imagabalin precursor in the small pocket of the active site. This corroborates the findings here, where a the equivalent I262L mutation alone may be enough to create the necessary expansion of the small pocket of omega-TAms. Contente et al. (2016) comes to a similar conclusion about the DEVx residue. In the *Halomonas elongata* aminotransferase (HEWT), the I258A mutation was predicted to allow the side chain of the neighbouring W56 greater rotational flexibility, thus removing its interference with a bulky para-substituted aromatic substrate. Enzyme assays of the I258A mutant confirmed an improved activity over the wild-type. The mutation modelling in the present study suggests an I258L mutation may even be more effective than I258A in stabilising the flipped position of W56. Dourado et al. (2016) present a comprehensive study of the rational design of an amine TAm for activity towards a bulky substrate: 2-acetylbiphenyl. The triple mutant W57G/R415A/I259M was designed to introduce a S/pi interaction between the suphur atom of M259 and the aromatic phenyl group of the ketone substrate. The W57G mutation enlarges the small pocket, allowing the aromatic ring of the substrate to occupy the space vacated by W57. Han et al. (2015) carried out alanine scanning

mutagenesis. I261A showed reduced activity compared to WT; however this is only for substrates S-alpha-MBA and 2-oxopentanoate.

In summary, the identification of the DEVx residue as an important SDP and potential target for mutagenesis is corroborated by several other studies. This study goes further in suggesting novel mutations at this position, rather than the standard alanine substitution approach, in order to alleviate the steric constraint.

4.2 Biocatalytic Applications

The use of the CATH FunFam classification combined with mining metagenomic sequences has proved successful in identifying key SDPs in class III TAms and may also be applied to other enzymes, particularly less well characterised ones. This information can easily be plugged into existing protocols for rational engineering or directed evolution of enzymes, where it can improve the quality and/or reduce the size of mutant libraries to be tested. The added value of the methods in this study will need to be determined through practical experimentation by testing the mutant activities against target substrates. It is important to realise that the identification of potentially useful mutations only serves as a starting point for the rational design of new enzyme functions. As Sirin et al. (2014) notes, most single-point mutations do not lead to significant improvements in activity. Engineering substrate specificity towards more sterically demanding substrates may require additional mutations in both the small and large pockets of the active site.

The selection pressures that are have driven the observed mutations in the context of bacterial metabolism are not necessarily relevant to target molecules in the pharmaceutical and chemical industries. Most TAms act on relatively small aliphatic substrates, whereas most drug molecules are larger and contain bulky ring structures. Currently no TAms that accept sterically demanding ketones (with bulky substituents on either side) have been discovered in nature. However, the broad substrate scope that is achieved by TAms sharing the same structure is testament to their potential to be bioengineered for unnatural biocatalysis. The residues that are observed to correlate with functional divergence in nature are likely to be the same residues that will enable more drastic changes in substrate specificity. Stereoselectivity must also be considered. D-amino acids are extremely rare in nature and so the vast majority of amino acid TAms act on L-amino acids. There is no natural selective pressure for changing the enantiopreference of most

TAms, so this needs to be achieved through rational mutagenesis. Luckily, it can be done with as little as one mutation (Svedendahl *et al.*, 2010).

4.2.1 FunFam 63148

FunFam 63148, essentially synonymous with omega-TAms, is the most studied of the SC4 TAms, especially in recent years. Wilding et al. (2016) have showed that an omega-TAm isolated from *Pseudomonas sp.* Strain AAC has similar activities with substrates ranging from beta-alanine all the way up to 12-aminododecanoic acid. This extreme level of promiscuity makes this and similar TAms good candidates for industrial applications. The I262L mutation in 4ba5 potentially increases the size of the S pocket, which would allow a larger R-group on both sides of the amine.

4.3 Shortcomings of the Project

While it is useful to focus on the evolution of single domains, the domain context must also be considered for a more full picture. I realised too late that the small domain (CATH superfamily 3.90.1150.10 for most TAms), while not providing the key catalytic residues, is just as important as the large domain in binding the substrate and thus determining substrate specificity. The binding cleft is at the interface between the two domains, and the substrates in several of the structures discussed here are stabilised by more interactions with the small domain than with the large domain. Rausch et al. (2013) identify the consecutive small domain residues G55, L56 and W57 as three of the top seven scoring residues in their SDP analysis. It would be interesting to know how strong the correlation is between small domain and large domain FunFams i.e. are there mostly one-to-one or one-to-many relationships, or is there a lot of domain promiscuity (many-to-many relationship), which may be a source of functional diversity.

Another factor that was not yet addressed was the correlation of mutations. While the single point mutations discussed in this study were all stabilising, it would have been useful to see which, if any, of the SDPs are mutually dependent with regards to the acquisition of mutations. Many of the large domain SDPs are adjacent to small domain residues in 3D space, and so there may be inter-domain as well as intra-domain covariation of several SDPs.

4.4 Future Directions

There a few hypotheses generated here that can be directly tested. For instance, an omega-TAm with the I→L mutation in the DEVx residue could be screened against a range of substrates, with the expectation of an improved substrate profile towards bulkier compounds. Another hypothesis is that FunFam 63148 would be more functionally coherent if split into around 3 smaller groups. The correlation between single SDPs or combinations of SDPs and EC annotations could be tested within this FunFam.

In terms of further bioinformatic analysis, there is much more that can be done. The focus of this project was the top scoring SDP in the large domain, but the other top 6 SDPs should be looked at as well. For instance, the second highest scoring SDP (residue ~ 150) is on the other side of the binding pocket, so an additional mutation here could allow expansion of the binding pocket in two directions. As touched upon above, some mutations may be correlated, so figuring this out would lead to a better understanding of the evolution of new specificities as well as enabling a better rational design of mutant libraries. Molecular dynamics and ligand docking approaches, combined with a greater range of template structures, would allow more comprehensive predictions before any lab assays are required. On a broader scale, the addition of the small domain into the sequence to function equation may clear up much of the functional impurity observed when just considering the large domain. Another useful piece of analysis would be to examine the context of the metagenome hits; many of the TAm genes are likely to be in operons, so the predicted functions of the surrounding genes woud give a strong indication of the metabolic pathway, which in turn could be used to predict the natural substrate of the TAm in question. This information could be used to present a stronger case for some SDPs being implicated in divergence of substrate specificity. The analysis of operons would be somewhat limited by the quality of the contig assembly: a minority of the TAm genes are in contigs large enough to identify the adjacent genes. The quanitity of the input information could also be improved by sampling different environments or using sequences already deposited in metagenome databases.

The scope of the analysis of TAms could be widened by including sructural cluster 1 (see table 3.2), which maps onto Pfam families PF00155 (class I) and PF00266 (class V). There are no M-CSA entries that correspond to an SC1 FunFam, so there is an opportunity to contribute to the knowledge of a less well characterised group of enzymes, where the SDP analysis will be arguably more valuable than for the well characterised class III TAms.

5 CONCLUSIONS

This project has analysed SDPs within structurally similar groups of homologous class III TAms in order to provide insights into the evolutionary mechanisms driving functional divergence. The CATH FunFam classification was used as the basis for a set of multiple sequences alignments representing different specificity groups. FunFams were carefully selected based on the information content of the multiple sequence alignment and the prevalent EC annotations. Metagenomic sequences from two different Peruvian salt evaporation pond samples were used to cast a wide net in the search for distant homologues with novel mutations in SDPs. The potential impact of mutations in the DEVx motif was modelled in various TAms. The importance of this residue corroborates findings in previous structure-based and mutagenesis studies of omega-TAms (FunFam 63148), while also uncovering its importance in the less well-studied DAPA-TAm and GABA-TAm. The findings will need to be validated in the lab through mutagenesis and enzyme assays. This will determine if the methods used in this project can add value to bioengineering approaches, which can be costly and time consuming. Most of the efforts around engineering new TAms involve increasing the size of the active site pocket to abolish the steric exclusion of bulky substrates; this project has shown one way in which this could be achieved.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) ‘Basic local alignment search tool’, *Journal of Molecular Biology*. doi: 10.1016/S0022-2836(05)80360-2.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) ‘Gapped BLAST and PSI-BLAST: A new generation of protein database search programs’, *Nucleic Acids Research*. doi: 10.1093/nar/25.17.3389.

Baud, D., Jeffries, J. W. E., Moody, T. S., Ward, J. M. and Hailes, H. C. (2017) ‘A metagenomics approach for new biocatalyst discovery: Application to transaminases and the synthesis of allylic amines’, *Green Chemistry*. doi: 10.1039/c6gc02769e.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: A flexible trimmer for Illumina sequence data’, *Bioinformatics*. doi: 10.1093/bioinformatics/btu170.

Capra, J. A. and Singh, M. (2008) ‘Characterization and prediction of residues determining protein functional specificity’, *Bioinformatics*. doi: 10.1093/bioinformatics/btn214.

Clay, D., Koszelewski, D., Grischek, B., Gross, J., Lavandera, I. and Kroutil, W. (2010) ‘Testing of microorganisms for ω-transaminase activity’, *Tetrahedron Asymmetry*. doi: 10.1016/j.tetasy.2010.07.009.

Contente, M. L., Planchestainer, M., Molinari, F. and Paradisi, F. (2016) ‘Stereoelectronic effects in the reaction of aromatic substrates catalysed by: Halomonas elongata transaminase and its mutants’, *Organic and Biomolecular Chemistry*. doi: 10.1039/c6ob01629d.

Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004) ‘WebLogo: A sequence logo generator’, *Genome Research*. doi: 10.1101/gr.849004.

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G. and Orengo, C. A. (2015) ‘Functional classification of CATH superfamilies: A domain-based approach for protein function annotation’, *Bioinformatics*. doi: 10.1093/bioinformatics/btv398.

Das, S., Sillitoe, I., Lee, D., Lees, J. G., Dawson, N. L., Ward, J. and Orengo, C. A. (2015) ‘CATH FunFHMMer web server: Protein functional annotations using functional family assignments’, *Nucleic Acids Research*. doi: 10.1093/nar/gkv488.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A. and Sillitoe, I. (2017) ‘CATH: An expanded resource to predict protein function through structure and sequence’, *Nucleic Acids Research*. doi: 10.1093/nar/gkw1098.

Dessailly, B. H., Dawson, N. L., Mizuguchi, K. and Orengo, C. A. (2013) ‘Functional site plasticity in domain superfamilies’, *Biochimica et Biophysica Acta - Proteins and Proteomics*. doi: 10.1016/j.bbapap.2013.02.042.

Dourado, D. F. A. R., Pohle, S., Carvalho, A. T. P., Dheeman, D. S., Caswell, J. M., Skvortsov, T., Miskelly, I., Brown, R. T., Quinn, D. J., Allen, C. C. R., Kulakov, L., Huang, M. and Moody, T. S. (2016) ‘Rational design of a (S)-selective-transaminase for asymmetric synthesis of (1S)-1-(1,1’-biphenyl-2-yl) ethanamine’, *ACS Catalysis*. doi: 10.1021/acscatal.6b02380.

Eddy, S. R. (1998) ‘Profile hidden Markov models’, *Bioinformatics*. doi: 10.1093/bioinformatics/14.9.755.

EDDY, S. R. (2009) ‘A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE’, in *Genome Informatics 2009*. doi: 10.1142/9781848165632_0019.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. and Bateman, A. (2016) ‘The Pfam protein families database: Towards a more sustainable future’, *Nucleic Acids Research*. doi: 10.1093/nar/gkv1344.

Fuchs, M., Farnberger, J. E. and Kroutil, W. (2015) ‘The Industrial Age of Biocatalytic Transamination’, *European Journal of Organic Chemistry*. doi: 10.1002/ejoc.201500852.

Genz, M., Vickers, C., van den Bergh, T., Joosten, H. J., Dörr, M., Höhne, M. and Bornscheuer, U. T. (2015) ‘Alteration of the donor/acceptor spectrum of the (S)-amine transaminase from vibrio fluvialis’, *International Journal of Molecular Sciences*. doi: 10.3390/ijms161126007.

Gordon, A. and Hannon, G. J. (2010) ‘Fastx-toolkit. FASTQ/A short-reads pre-processing tools’, *http://hannonlab. cshl. edu/fastx_ toolkit*.

Han, S.-W., Park, E.-S., Dong, J.-Y. and Shin, J.-S. (2015) ‘Active-Site Engineering of ω-Transaminase for Production of Unnatural Amino Acids Carrying a Side Chain Bulkier than an Ethyl Substituent’, *Applied and Environmental Microbiology*. doi: 10.1128/aem.01533-15.

Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., Endy, D. and Salit, M. (2017) ‘Measurements of translation initiation from all 64 codons in E. coli’, *Nucleic Acids Research*. doi: 10.1093/nar/gkx070.

Hirotsu, K., Goto, M., Okamoto, A. and Miyahara, I. (2005) ‘Dual substrate recognition of aminotransferases’, *Chemical Record*. doi: 10.1002/tcr.20042.

Höhne, M., Schätzle, S., Jochens, H., Robins, K. and Bornscheuer, U. T. (2010) ‘Rational assignment of key motifs for function guides in silico enzyme identification’, *Nature Chemical Biology*. doi: 10.1038/nchembio.447.

Hwang, B. Y., Cho, B. K., Yun, H., Koteshwar, K. and Kim, B. G. (2005) ‘Revisit of aminotransferase in the genomic era and its application to biocatalysis’, *Journal of Molecular Catalysis B: Enzymatic*. doi: 10.1016/j.molcatb.2005.09.004.

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D’Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., et al. (2016) ‘An expanded evaluation of protein function prediction methods shows an improvement in accuracy’, *Genome Biology*. doi: 10.1186/s13059-016-1037-6.

Katoh, K. and Standley, D. M. (2013) ‘MAFFT multiple sequence alignment software version 7: Improvements in performance and usability’, *Molecular Biology and Evolution*. doi: 10.1093/molbev/mst010.

Lee, D., Das, S., Dawson, N. L., Dobrijevic, D., Ward, J. and Orengo, C. (2016) ‘Novel Computational Protocols for Functionally Classifying and Characterising Serine Beta-Lactamases’, *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1004926.

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., Orengo, C. and Lees, J. (2018) ‘Gene3D: Extensive prediction of globular domains in proteins’, *Nucleic Acids Research*. doi: 10.1093/nar/gkx1069.

Li, D., Luo, R., Liu, C. M., Leung, C. M., Ting, H. F., Sadakane, K., Yamashita, H. and Lam, T. W. (2016) ‘MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices’, *Methods*. doi: 10.1016/j.ymeth.2016.02.020.

Liu, W., Peterson, P. E., Carter, R. J., Zhou, X., Langston, J. A., Fisher, A. J. and Toney, M. D. (2004) ‘Crystal structures of unbound and aminooxyacetate-bound Escherichia coli γ-aminobutyrate aminotransferase’, *Biochemistry*. doi: 10.1021/bi049218e.

Madera, M. (2008) ‘Profile Comparer: A program for scoring and aligning profile hidden Markov models’, *Bioinformatics*. doi: 10.1093/bioinformatics/btn504.

Malik, M. S., Park, E. S. and Shin, J. S. (2012) ‘Features and technical applications of ω-transaminases’, *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-012-4103-3.

Mathew, S. and Yun, H. (2012) ‘ω-Transaminases for the production of optically pure amines and unnatural amino acids’, *ACS Catalysis*. doi: 10.1021/cs300116n.

MEHTA, P. K., HALE, T. I. and CHRISTEN, P. (1993) ‘Aminotransferases: demonstration of homology and division into evolutionary subgroups’, *European Journal of Biochemistry*. doi: 10.1111/j.1432-1033.1993.tb17953.x.

Midelfort, K. S., Kumar, R., Han, S., Karmilowicz, M. J., McConnell, K., Gehlhaar, D. K., Mistry, A., Chang, J. S., Anderson, M., Villalobos, A., Minshull, J., Govindarajan, S. and Wong, J. W. (2013) ‘Redesigning and characterizing the substrate specificity and activity of Vibrio fluvialis aminotransferase for the synthesis of imagabalin’, *Protein Engineering, Design and Selection*. doi: 10.1093/protein/gzs065.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. (2004) ‘UCSF Chimera - A visualization system for exploratory research and analysis’, *Journal of Computational Chemistry*. doi: 10.1002/jcc.20084.

Rausch, C., Lerchner, A., Schiefner, A. and Skerra, A. (2013) ‘Crystal structure of the ω-aminotransferase from Paracoccus denitrificans and its phylogenetic relationship with other class III amino- transferases that have biotechnological potential’, *Proteins: Structure, Function and Bioinformatics*. doi: 10.1002/prot.24233.

Rho, M., Tang, H. and Ye, Y. (2010) ‘FragGeneScan: Predicting genes in short and error-prone reads’, *Nucleic Acids Research*. doi: 10.1093/nar/gkq747.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C. and Jaffe, D. B. (2013) ‘Characterizing and measuring bias in sequence data’, *Genome Biology*. doi: 10.1186/gb-2013-14-5-r51.

Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W. and Hughes, G. J. (2010) ‘Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture’, *Science*. doi: 10.1126/science.1188934.

Sayer, C., Isupov, M. N., Westlake, A. and Littlechild, J. A. (2013) ‘Structural studies of Pseudomonas and Chromobacterium ω- aminotransferases provide insights into their differing

substrate specificity', *Acta Crystallographica Section D: Biological Crystallography*. doi: 10.1107/S0907444912051670.

Sirin, S., Kumar, R., Martinez, C., Karmilowicz, M. J., Ghosh, P., Abramov, Y. A., Martin, V. and Sherman, W. (2014) 'A computational approach to enzyme design: Predicting W-Aminotransferase catalytic activity using docking and MM-GBSA scoring', *Journal of Chemical Information and Modeling*. doi: 10.1021/ci5002185.

Svedendahl, M., Branneby, C., Lindberg, L. and Berglund, P. (2010) 'Reversed Enantiopreference of an ω-Transaminase by a Single-Point Mutation', *ChemCatChem*. doi: 10.1002/cctc.201000107.

Temperton, B. and Giovannoni, S. J. (2012) 'Metagenomics: Microbial diversity through a scratched lens', *Current Opinion in Microbiology*. doi: 10.1016/j.mib.2012.07.001.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. and Barton, G. J. (2009) 'Jalview Version 2-A multiple sequence alignment editor and analysis workbench', *Bioinformatics*. doi: 10.1093/bioinformatics/btp033.

Webb, B. and Sali, A. (2016) 'Comparative protein structure modeling using MODELLER', *Current Protocols in Bioinformatics*. doi: 10.1002/cpbi.3.

Weber, F. and Sedelmeier, G. (2014) 'Top 200 Pharmawirkstoffe', *Nachrichten aus der Chemie*. doi: 10.1002/nadc.201490364.

Wilding, M., Peat, T. S., Newman, J. and Scott, C. (2016) 'A β-Alanine Catabolism Pathway Containing a Highly Promiscuous ω-Transaminase in the 12-Aminododecanate-Degrading Pseudomonas sp. Strain AAC', *Applied and Environmental Microbiology*. doi: 10.1128/aem.00665-16.

Appendices

Scripts developed for this project can be found at *https://github.com/greglv93/novel-transaminases-in-metagenomes*.

FunFam	Name	Modal EC	STRONG Metagenome hits		
			pilluana	maras3	pil-mar
63333	DegT/DnrJ/EryC1/StrS aminotransferase family enzyme	2.6.1.87	229	132	97
63355	Probable aspartate aminotransferase 1	2.6.1.7	141	116	25
63385	Glutamate-1-semialdehyde 2,1-aminomutase 1	2.6.1.93	93	77	16
63298	Histidinol-phosphate aminotransferase 1, chloroplastic	2.6.1.9	41	30	11

63324	Putative aspartate aminotransferase	2.6.1.1	25	16	9
63097	4-aminobutyrate aminotransferase GabT	2.6.1.19	15	7	8
63148	Aminotransferase class-III family protein	2.6.1.96	26	20	6
62618	Bifunctional phosphonoacetaldehyde hydrolase/aminoethylphosphonate transaminase	2.6.1.37	6	1	5
63401	LL-diaminopimelate aminotransferase, chloroplastic	2.6.1.83	7	2	5
63379	Aspartate aminotransferase P2, mitochondrial	2.6.1.1	20	16	4
62777	Diaminobutyrate--2-oxoglutarate transaminase (Diaminobutyrate--2-oxoglutarate aminotransferase)	2.6.1.76	20	20	0
62953	Serine--pyruvate aminotransferase, mitochondrial	2.6.1.51	14	15	-1
63364	Aromatic amino acid aminotransferase	2.6.1.39	4	7	-3
63074	Alanine--glyoxylate aminotransferase 2, mitochondrial	2.6.1.44	9	12	-3
63436	Phosphoserine aminotransferase 1, chloroplastic	2.6.1.52	29	38	-9
63154	Probable acetylornithine aminotransferase, mitochondrial	2.6.1.11	78	109	-31

Table A1 *Strong domain hits to TAm FunFams, ordered by the difference between the two samples. The numbers of hits includes both full genes and domains from gene fragments. FunFams with fewer than 5 hits in at least one of the samples are not shown. The modal EC is the most common EC annotation for SwissProt (manually reviewed) entries that fall into the latest version (v4.2) of that FunFam. Low DOPS (<70) and non-transaminase (EC ≠ 2.6.1.-) FunFams are not shown. The class III FunFams of interest are highlighted in yellow.*

Closest FunFam	**Name**	**Modal EC**	**WEAK Metagenome hits**		
			pilluana	**maras3**	**pil-mar**
63364	Aromatic amino acid aminotransferase	2.6.1.39	71	39	32
63148	Aminotransferase class-III family protein	2.6.1.96	71	46	25
62757	Related to Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	2.6.1.62	34	21	13
62953	Serine--pyruvate aminotransferase, mitochondrial	2.6.1.51	21	10	11
63385	Glutamate-1-semialdehyde 2,1-aminomutase 1	2.6.1.94	26	16	10
63436	Phosphoserine aminotransferase 1, chloroplastic	2.6.1.52	8	2	6
63324	Putative aspartate aminotransferase	2.6.1.1	11	6	5
63074	Alanine--glyoxylate aminotransferase 2, mitochondrial	2.6.1.44	16	12	4
62993	Histidinol-phosphate aminotransferase	2.6.1.9	5	2	3
63270	Putative nicotianamine aminotransferase B	2.6.1.5	9	7	2
62777	Diaminobutyrate--2-oxoglutarate transaminase	2.6.1.76	7	6	1

	(Diaminobutyrate--2-oxoglutarate aminotransferase)				
63333	DegT/DnrJ/EryC1/StrS aminotransferase family enzyme	2.6.1.87	23	22	1
63438	Histidinol-phosphate aminotransferase	2.6.1.9	7	9	-2
63391	Uncharacterized aminotransferase C27F1.05c	2.6.1.11	6	10	-4
63097	4-aminobutyrate aminotransferase GabT	2.6.1.19	32	40	-8
63154	Probable acetylornithine aminotransferase, mitochondrial	2.6.1.11	21	30	-9
63355	Probable aspartate aminotransferase 1	2.6.1.7	22	36	-14

Table A2 *The same as table A1, except that these are sample hits that do not meet any FunFam inclusion thresholds, grouped by the closest matching FunFam. The class III FunFams of interest are highlighted in yellow.*