STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF ENZYMES IN COG3964 OF THE AMIDOHYDROLASE SUPERFAMILY: FROM SEQUENCE TO STRUCTURE TO FUNCTION

A Dissertation

by

ARGENTINA ORNELAS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Frank M. Raushel
Committee Members,	Wenshe Liu
	Donald W. Pettigrew
	Coran M.H. Watanabe
Head of Department,	David H. Russell

December 2012

Major Subject: Chemistry

ABSTRACT

The Amidohydrolase Superfamily (AHS) of enzymes is one of the most structurally and functionally studied groups of biological catalysts, exquisitely designed to carry out an extensive number of reactions defined by a similar reaction mechanism. There are approximately 11,000 genes coding for AHS proteins from about 2,100 sequenced organisms. Sequence information for these genes has been catalogued in databases, the most instrumental being the National Center for Biotechnology Information (NCBI). Despite the accessible information organized in genomic databases, there is still an extensive problem of reliability in the functional annotation of gene products assigned to the AHS.

Proteins in COG3964 of the AHS have been functionally identified as dihydroorotases and adenine deaminases. Eight proteins within three group families of COG3964 have been purified and fail to demonstrate the functionally annotated activity. A library of compounds developed by functional-group modifications was compiled and tested with these enzymes. A group of enzymes within COG3964 demonstrates the ability to hydrolyze stereospecific acetylated α-hydroxyl carboxylates. Substrate profiles were constructed for enzymes belonging to group 6 of COG3964. Atu3266, Oant2987 and RHE_PE00295 hydrolyze the *R*-isomers of a library of α-acetyl carboxylates of which acetyl-*R*-mandelate is the best substrate with catalytic efficiencies of 10^5 $M^{-1}s^{-1}$. This compound was identified after a series of modifications from a low-activity compound (V/K = 4 $M^{-1}s^{-1}$). Methylphosphonate analogs of acetyl-*R*-mandelate and *N*-

acetyl-D-phenyl glycine are inhibitors of enzymes in group 6. The structure of Atu3266 was used in docking experiments to assess the selectivity of *R*- enantiomers over their *S*-counterparts. An additional group of orthologues share less than 40% sequence similarity to enzymes from group 6. EF0837, STM4445 and BCE_5003 from group 2 show significantly lower rates for the hydrolysis of α-acetyl carboxylates, including acetyl-*R*-mandelate, hydrolyzed at values of $k_{cat}/K_m = 10^3\ M^{-1}s^{-1}$. This is also the only active compound for EF0837. Xaut_0650 and blr3349 from group 7 of COG3964 demonstrate less than 30% identity to enzymes in groups 2 and 6. These enzymes fail to hydrolyze any compound from an extended library of compounds.

An annotated selenocysteine synthase gene (SelA) from COG1921 has been identified as a gene neighbor to almost every amidohydrolase from COG3964. Atu3263, Oant2990 and EF0838 are pyridoxal-5'-phosphate dependent enzymes that were purified and assayed with D- and L- amino acids. Initial thermal-shift fluorescence assays determined that in the presence of D-cysteine, the proteins were denatured at lower temperatures.

DEDICATION

A mi familia, por su apoyo y amor – Los quiero mucho.

To my family, for their love and support.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Frank M. Raushel, for his guidance and support throughout the course of this research, and my committee members: Dr. Donald Pettigrew, Dr. Wenshe Liu, and Dr. Coran Watanabe. I would also like to extend my appreciation to Dr. Tamari Narindoshvili, Dr. Magdalena Korczynska, and to past and present members of the Raushel Lab, the best 'superfamily' that I could ask for in grad school.

Thanks to my family, a big crazy bunch of people who supported me and had confidence in me, and to Michael Dearing, for his support and understanding through this phase of my career.

Finally, thanks to Arthur Kornberg's *For the Love of Enzymes*, a book that provided me with inspiration during the tough times of graduate school.

NOMENCLATURE

NCBI	National Center for Biotechnology Information
SCOP	Structural Classification of Proteins
PDB	Protein Data Bank
GenBank	Gene Bank database
UniProtKB	Universal Protein resource Knowledge Base
TIM	TrIose phosphate isoMerase
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
AHS	AmidoHydrolase Superfamily
M_α, M_β	Metal at α-position and β-position of active site
COG	Cluster of Orthologous Groups
EFI	Enzyme Function Initiative
HEI	High Energy Intermediate
KEGG	Kyoto Encyclopedia for Genes and Genomes
ATCC	American Type Culture Collection
NYSGXRC	New York Structural GenomiX Research Consortium
SDS-PAGE	Sodium Dodecyl Sulfate-PolyAcrylamide Gel Electrophoresis
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
SelA	Selenocysteine synthase
PLP	Pyridoxal-5'-Phosphate

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

The conundrum to functional determination of gene products from newly sequenced organisms has become one of the most important issues to solve in chemical biology, especially after the exponential accumulation of sequenced genomes continues to escalate. The technology that has allowed the rapid increase in sequencing of organisms and biological samples has not transferred to the improvement of methods geared towards the complete understanding of functional annotation and metabolic roles of gene products. Understanding the biological organization and networks of organisms is essential towards the eventual improvements in medical and industrial applications, drug development, synthetic biology and the understanding of diversity in biological systems.

Over the coming years, the information from sequencing databases will continue to expand, unless there is an improvement in the comprehensive analysis of this data, sequence databases will continue to grow with a wealth of information that conceals the necessary details to the understanding of cellular metabolism, the organization to larger complexes and the evolution of biological systems. The genetic code has long been a focus of study, analysis and understanding. It is after all, the most basic narration of every organism. Determining the correct functional annotation of genes in sequenced organisms has become a critical and multi-disciplinary goal of the biochemical scientific community.

To those scientists focused on the functional understanding of genomes, proteomes and metabolomes, the analysis of one specific gene becomes fundamental to appreciate the larger formations and interactions. Over the years, the National Center for Biotechnology Information (NCBI) has become the hub for the compilation of the genomic information on a variety of organisms from all kingdoms (*1*). Other databases have become more specialized in the type of information catalogued; the more fundamental in bioinformatics is the development of sequence databases (*2*). As of 2012, the sequence database UniProtKB/TrEMBL (*3*), contained well over 16 million entries of protein sequence information (*4*). Unfortunately, the achievement of the abundance in gene product sequencing information is diminished by the large number of protein function misannotations. It is estimated that about 40% of the functional annotations in sequence databases are represented with unknown, uncertain or incorrect functional assignments (*5-8*). This problem extends as functional misannotations propagate across databases.

Functional designations of open reading frames are based on sequence similarity threshold values to other homologous sequences in databases (*5, 9-12*). This initially was helpful, as early characterization of gene products was focused on genes operating in central metabolic pathways likely to be conserved in a variety of organisms. However, as sequencing technology improved, the number of sequenced organisms increased, and the assault on the characterization of individual players in central pathways were determined; the annotation of gene products became redundant, this in part due to functional annotations based on sequence homology. These functional annotation efforts

are based on comparison of sequences from many organisms using computational tools such as PSI-BLAST to retrieve related sequences from databases (*10*). If a protein has been found to have about 40% sequence identity to another protein whose biochemical function has been experimentally confirmed, and if the functionally important residues, in essence those found in the active site of the protein, are conserved between the two species, then a reasonable assumption can be made that the two proteins have a common biochemical function. **Figure 1.1** is an adaptation of Mark Gerstein proposal to defining a functional relationship to the sequence similarity of proteins (*13-14*).

A comparison of proteins sequences continues to be an area of active research mainly because it is the easiest technique to implement when assigning functional roles to genes in a newly sequenced organism. Many of these annotations however are significantly biased and incorrect when gene product annotation is based on evolutionarily distant organisms. What this now implies is that there are a large number of chemical reactions in biological pathways of organisms that have not yet been defined.

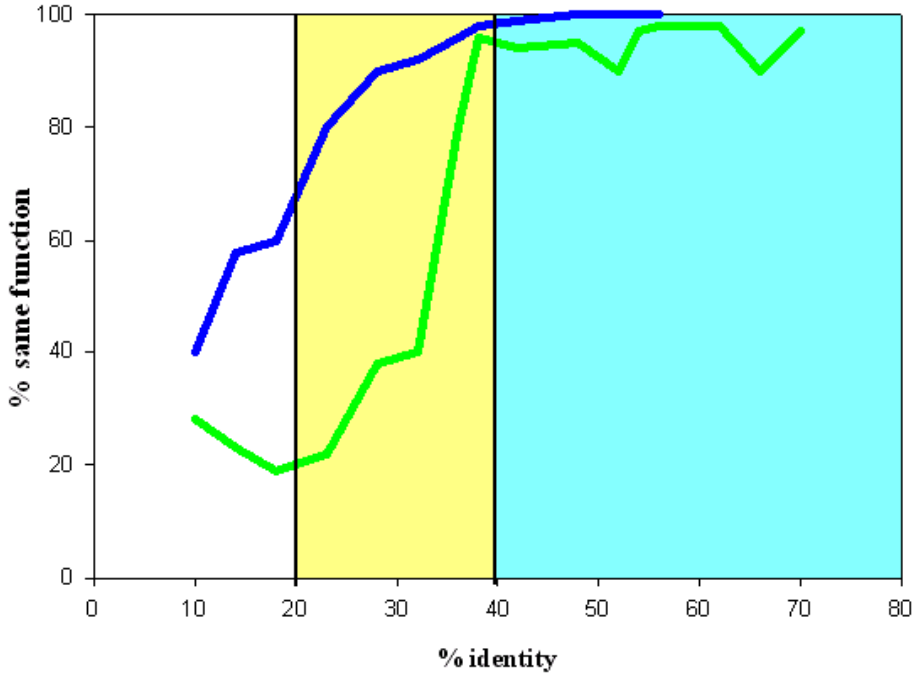


Figure 1.1: Graph relating sequence to function similarities according to Gerstein (*12-13*). The percentage of proteins with the same biochemical function is plotted against the sequence similarity or identity (enzymes, blue curve; non-enzymes, green curve). White area represents proteins in which neither fold nor function, can reliably be predicted from sequence identity. Yellow area represents proteins whose fold can reliably be predicted from sequence similarity, but whose function cannot be predicted. Blue area represents proteins whose fold and function can be reliably predicted from sequence comparisons. Image was adapted and modified from Protein Structure and Function (*15*) and from Mark Gerstein (http://bioinfo.mbb.yale.edu/lectures/spring2002/show/index_2).

Certainly it would pose a great advantage to be able to categorize and functionally characterize gene products based on comparative sequence analysis (*9-11*). It has been a long term goal to be able to detect structural properties of folded, active forms of proteins from the primary sequence forming the polypeptide (*16-17*). The biological roles of proteins from genome sequencing projects will require knowledge of the structure in addition to their function. Although there has been great success in the experimental development of methods that provide high-resolution structural information from a variety of proteins, computational structure prediction methods can provide valuable information for the large fractions of proteins whose structures cannot be determined (*18-21*). *De novo* and *ab initio* methods can predict the structure of proteins from sequence alone. However these methods require the use of other homologous proteins with a pre-determined X-ray crystallographic structure. Once again, there are certain threshold values in sequence similarity that must be shared in order to create a structural homology model for a particular protein sequence. In practice, a sequence with a greater than 40% similarity with a structurally characterized homologue can usually produce a predicted structure equivalent to that of a medium-resolution experimentally solved structure (*22*). In cases when a protein sequence does not share minimum sequence identity to any of the available from the catalog of sequence databases, a technique known as profile-based threading facilitates the development of a structure based on sequence information alone. In this method, fold assignment and alignment are assessed by a computer program that forces the particular

sequence to conform to a variety of structural folds, which are then assessed by quantitive measures of energy of the folded protein (*23, 24*).

Determining a structural model from sequence alone, whether or not the inquired sequence has closely related homologues; or whether these homologues have experimentally determined X-ray crystal structures that can be used for homology models, can only add to the improvements in the approaches to structurally and functionally characterizing gene products. Notably, the number of protein folds an amino acid sequence can conform to, is estimated to be significantly less than the number of proteins encoded by a genome (*25-27*). This has been observed as the number of conserved structural folds is more readily identified than the conservation of sequence identity, and proteins having statistically insignificant sequence similarity sometimes adopt the same fold. For example, the enzymes benzoyl formate decarboxylase (*28*) and pyruvate decarboxylase (*29*) share essentially only about 21% sequence identity, but have essentially identical folds. This places a limitation on the number of possible structural architectures. Consequently, these folds have been reused by divergent evolution or independently formulated by convergent evolution to accommodate the number of reactions represented by proteins in living organisms.

Based on the available information detailing sequence, structure and function of homologous gene products, these can be associated to explain three distinct strategies nature has used to develop the divergent evolution of enzyme function. Each of these strategies is instigated by the duplication of a progenitor gene that consequently evolves so that the original enzyme is used in cellular metabolism. The first strategy for

divergent evolution involves the enzymes that catalyze reactions in biosynthetic pathways (*30-31*). Under this scenario, when the substrate for an enzyme in a pathway is depleted, a new enzyme evolves to supply that substrate using an available precursor template of an enzyme that uses the substrate. In accordance, both the evolved enzyme and the precursor enzyme will share the ability to bind the same molecule as both, substrate and product. The mechanism of the reactions carried out by the precursor and evolved enzymes will not be related so these would be categorically characterized as members of functionally distinct suprafamilies. Enzymes in a group assigned as a suprafamily, can essentially have homologous sequences, however they catalyze distinct reactions in a metabolic pathway and do not share a common mechanistic attribute, these sequences can also conserve active site residues, yet these perform different functions in each enzyme (*32*).

In an additional hypothesis of divergent evolution that also defines the progenitor and the evolved enzyme as members of functionally distinct suprafamilies, the progenitor enzyme is selected based on the functional groups present at the active site. The active site is able to support alternate reactions, with little or no change to the identities of the amino acid residues involved in catalysis. For this approach, the active site is capable of utilizing the shared functional groups in distinct mechanistic and metabolic contexts (*33*).

In the third strategy for divergent evolution the hypothesis is that nature selects the protein for divergent evolution from a pool of enzymes based on its ability to stabilize the intermediates or transition states required for a desired transformation (*34*,

35). The mechanistic characteristics, substrate specificity and proficiency of the new enzyme are then enhanced by evolution. The simplest explanation to this event is that the progenitor enzyme and the evolved enzyme catalyze the same reaction but with different specificities. It can also be argued that nature, perhaps selects a progenitor that carries out the desired reaction as a result of promiscuous activity, perhaps at the expense of the original reaction (*36*, *37*). Under this strategy of directed evolution the new and evolved enzymes would be members of a mechanistically diverse superfamily of enzymes (*32*). A group of homologous enzymes that catalyze either the same chemical reaction with distinct substrate specificities; or different overall reactions that shares a common mechanistic feature such as partial reactions, intermediates or transition states, enabled by conserved active site residues performing the same function, are known to form a superfamily (*32*). The presence of superfamilies of proteins that share related structure and biochemical functions encourages the assumption that in addition to defining the ensemble of all possible protein folds, comprehensive structural information could provide a firmer basis than sequence for functional predictions.

The most common structural fold characterized in the structural information of enzymes obtained by X-ray crystallography is the $(\beta/\alpha)_8$-TIM barrel fold (*38*). Nearly 10% of structurally characterized proteins contain at least one domain with this structural fold. Enzymes featuring this structural design have a domain with eight parallel β-sheets, flanked by eight α-helices on the outer face of the barrel. At the center of this β/α-barrel domain is a divalent metal ion active site located at the C-terminal ends of the β-strands. Coordinating this metal center, are conserved residues originating

from the flexible loops that follow the β-strands. The nature of these active site metal ligands is conserved based on specific superfamilies of enzymes that feature this structural fold. The $(\beta/\alpha)_8$-barrel fold is one of the most diverse in terms of the number of functional superfamilies conforming to this architecture. The flexibility of this fold and the variability in the active site residues, accounts for the versatility in the divergence of enzyme function. The structural fold was initially identified from the three dimensional structure of the enzyme triosephosphate isomerase from chicken muscle (*39*). The $(\beta/\alpha)_8$ motif of the triosephosphate isomerase structure is observed in **Figure 1.2** (PDBcode = 7TIM). Recent studies have suggested that this common $(\beta/\alpha)_8$-barrel fold has evolved from the duplication and fusion of identical $(\beta/\alpha)_4$ half-barrels. The fusion of two copies of a gene that encode an ancestral $(\beta/\alpha)_4$ enzyme (imidazole glycerol phosphate synthase – HisF-C), in addition to few amino acid exchanges, was able to generate a highly stable $(\beta/\alpha)_8$ barrel protein with wild-type structural properties (*40-41*).

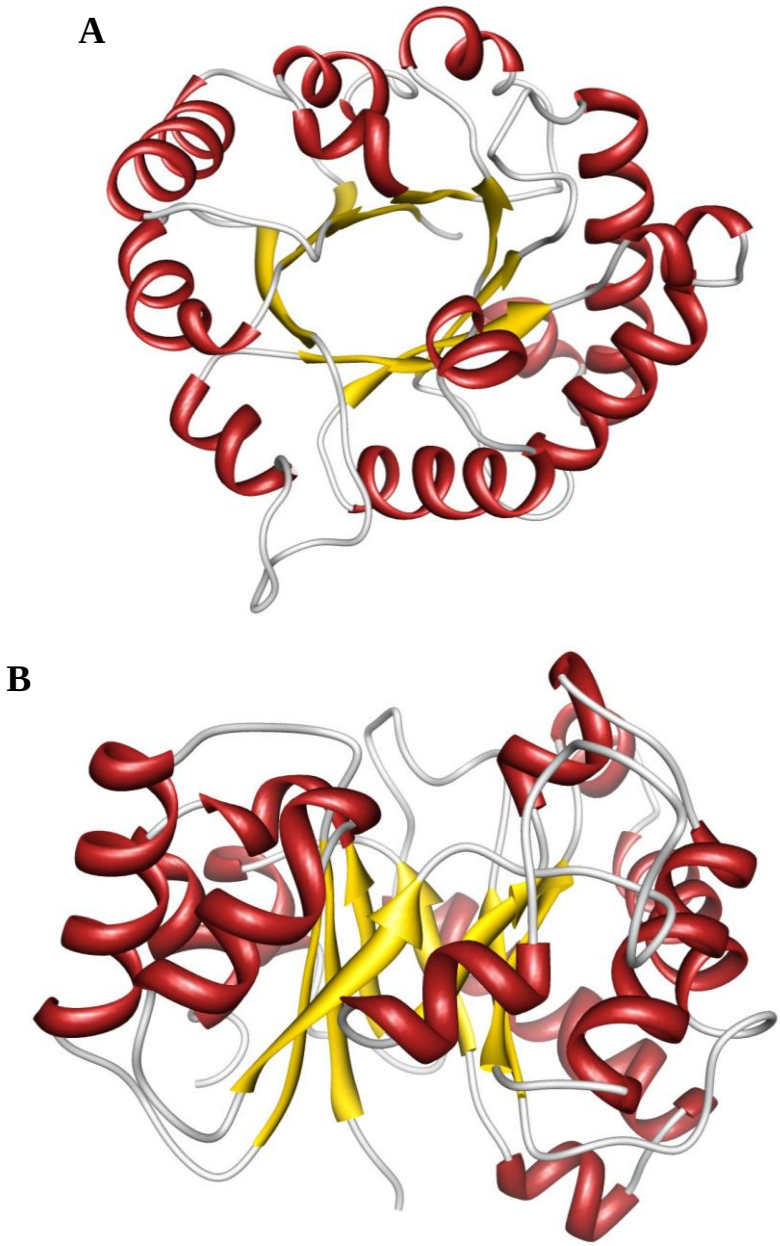


Figure 1.2: Structure of $(\beta/\alpha)_8$-barrel of triosephosphate isomerase (7TIM.pdb). **A.** the C-terminal ends of the B-strands and **B.** side view of the barrel.

According to the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop) (*16*), as of 2009 there were over 30 identified $(\beta/\alpha)_8$-barrel superfamilies, including the amidohydrolase superfamily (*32, 38 and 42*), the enolase superfamily (*32, 43-45*), the thiyl-radical superfamily (*32, 46-47*) and the crotonase superfamily (*32, 48-49*). All of these superfamilies of proteins carry out a variety of chemical reactions. Because the total number of protein folds is smaller than the total number of expressed genes in biology, if a protein can be assigned to a superfamily from sequence and structural information, at the very least the number of its possible functions can be narrowed down, and under some instances it may be possible to assign a precise function. The convergence of structural fold emphasizes the theory of divergent evolution; these enzymes can be assumed to having shared at one point, a common ancestor that has evolved its structural fold to carry out a myriad of reactions.

One of the most functionally, structurally and mechanistically documented superfamily of enzymes is the amidohydrolase superfamily (AHS). This group of enzymes was first identified based on the similarities in the three-dimensional structural fold and active site conformations of urease (URE) (*50*), phosphotriesterase (PTE) (*51*) and adenosine deaminase (ADA) (*52*). Lissa Holm and Chris Sander introduced the concept of the AHS in 1997 by unification of this metal dependent, hydrolytic and functional diverse group of enzymes (*42*). Members of this family of enzymes catalyze the cleavage not only on C-N bonds, but also C–C, C–O, C–Cl, C–S and O–P in a variety of biological molecules including amino acids, sugars, nucleic acids and organophosphate esters (*53-55*). In addition to hydrolytic reactions and deaminations, it

also carries out hydrations, decarboxylations and isomerizations (*56-59*). These enzymes all feature a $(\beta/\alpha)_8$-TIM barrel structural fold. Superimposition of the structural models revealed the common structural core consisting of an ellipsoidal TIM-barrel with a conserved metal binding site at the C-terminal end of β-strands. Until recently, it had been accepted that all amidohydrolases were metalloenzymes, however in the particular example detailing the structure and mechanism of a lignin degrading amidohydrolase, it was determined that this enzyme did not require a divalent metal for catalytic activity. (*60*). Most commonly observed in the AHS are enzymes containing a mononuclear or binuclear divalent *d*-block metal cofactor at the active site within the confines of the barrel (*42, 53*). The divalent metal ions that have been found with these enzymes are zinc, cobalt, manganese, iron and nickel. These metal ions are tethered to the protein through residues originating from loops at the C-terminal ends of specific β-strands. The various conformations of active site residues coordinating the metal ions will be discussed shortly.

Biochemical studies on the amidohydrolase superfamily reveal the essentiality that the metal cofactor confers on this group of enzymes. Binuclear metal centers have metal ions at designated M_α and M_β sites, this arrangement of metal cofactors will also require five to six protein ligands coordinating to the binuclear center. In mononuclear enzymes, the metal is located at either the M_α- or the M_β- site, and this cofactor will require four or five protein ligands. The metal center has dual catalytic roles in the function of amidohydrolases. Under both metal center conformations, the metal ions assist: (**1**) in the activation of a water molecule to enhance the nucleophilic properties of

the water-derived hydroxide, and (**2**) in the activation of the scissile bond of the substrate for bond cleavage (*53*).

As discussed earlier, the divalent mononuclear or binuclear center is coordinated by highly conserved amino acid residues originating from specific β-strands. There are currently ten subtypes of structurally characterized metal center variations in the AHS. **Table 1.1** summarizes the type of conserved residues that are found in the 10 different subtypes of metal ligands to mononuclear and binuclear metal centers in the AHS. The prototypical and most common metal center is **subtype I**. This binuclear metal center is found in PTE, URE, dihydroorotase (DHO) (*61*) and isoaspartyl dipeptidase (IAD) (*62*). A buried M_α metal is coordinated by an HxH motif at the end of β-1 and an aspartate from β-8. The solvent exposed M_β metal is coordinated by two histidine residues from β-5 and β-6. The two divalent cations are bridged by a post-translationally modified carboxylated lysine from β-4. This assembly is virtually the same in **subtype II** and **subtype X**, except that the carboxylated lysine residue is replaced with a glutamate from β-4. The PTE homology protein (PHP) from *E.coli* features the metal center conformation of **subtype II**. **Subtype X**, includes an additional conserved aspartate residue coordinating the M_β metal, this residue is found after the histidine from β-6. This subtype of coordinating ligands is observed in the enzyme adenine deaminase (ADE) from *A. tumefaciens* (*63*). The human renal dipeptidase is one of the most distantly related members of the amidohydrolase superfamily for which there is a three-dimensional crystal structure (*64*); the enzyme has coordinating residues to the M_α metal ion presented in **subtype VI**. In this enzyme the common HxH motif is replaced by an

HxD motif that coordinates to the M_α ion. The protein binds two metal equivalents of zinc, but the bridging ligand is a glutamate rather than a carboxylated lysine originating from β-strand 3. In addition the aspartate originating from β-8 does not coordinate to the M_α, it is observed that the δ oxygens from aspartate at β-1 replaces coordination from the second histidine typically found in β-1, as well as the aspartate usually forming a ligand from β-8 to M_α. The D-amino acid deacetylase from *A. faecalis* has been observed to also adopt a second metal at the M_α position; however, the enzyme is more active in the mononuclear state with a required metal at the M_β site. In **subtype V**, the coordination scheme to the M_β metal consists of a cysteine from β-2, and histidines from β-5 and β-6. In the presence of an M_α metal, this one is coordinated by an HxH motif from β-1, and an aspartate from β-8. A binuclear metal enzyme was found significantly less active (*65*). **Subtype IV** is exemplified by the *N*-acetyl-D-glucosamine-6-phosphate deacetylase (*N*-AGD). This enzyme can have a mononuclear or binuclear metal center depending on the organism. The *N*-AGD from *B. subtilis* has been structurally characterized with two metal-ions (*66*). The coordination of the M_α is carried out by the HxH motif from β-1 and an aspartate from β-8. The M_β position is coordinated by a glutamate from β-3 and two histidines from β-5 and β-6. In the mononuclear homologue from *T.maritima*, *N*-AGD only binds the metal-ion at the M_β site; however, it is postulated that the glutamate from β-3 can act as a bridging residue between the M_α and M_β ions. The mononuclear *N*-AGD still conserves those residues binding M_α (*67*). In *E.coli*, *N*-AGD homologue also presents a mononuclear metal center, however the HxH

from β-1 is not conserved, instead a glutamine and an asparagine replace the histidine motif, two residues that are not commonly found to ligate metals, suggesting that the enzyme can function with only one metal (*53, 67*). This metal coordination for a mononuclear *N*-AGD center has been identified as a different subtype and is exemplified by **subtype VII**.

Mononuclear active site **subtype III** was identified in the cytosine deaminase and adenosine deaminase enzymes from *E. coli* (*68, 69*). This mononuclear variation of the active site has an M_α coordinated by the histidines of an HxH motif from β-1, as well as a histidine from β-5. The histidine found after β-6 is conserved, but it does not bind the metal. An aspartate from β-8 forms the fourth ligand to the metal. The amino acid metal coordination for uronate isomerase from *B. halodurans* is observed in **subtype VIII** (*59*). This metal coordination maintains the HxH motif from β-1 and aspartate from β-8 to bind the M_α ion. The histidine from β-5 is conserved in all homologues, however β-6 is missing in others (*70*). **Subtype IX** is represented in the active site of three members in the AHS annotated as a tatD-deoxyribonuclease from *E. coli*. This metal binding conformation has an M_β-ion coordinated by two adjacent residues at the C-terminal of β-6, a histidine and a cysteine. The HxH motif from β-1 is not present; and a glutamate from β-8 coordinates the M_β site. A water molecule is found within hydrogen-bonding distance between an aspartatic acid following β-8 and the M_β site. This catalytic water molecule most likely also aids in the coordination to this metal. **Figure 1.3** shows

the two-dimensional representations for all the discussed metal center active sites found in AHS enzymes.

Table 1.1: Subtypes of metal ligand variations in structurally characterized metalloenzymes of the amidohydrolase superfamily.

Subtype	position	β-strand								
		1	2	3	4	5	6	7	8	Example:
I	M_α,M_β	HxH			K	H	H		D	PTE, DHO, IAD
II	M_α,M_β	HxH			E	H	H		D	PHP
III	M_α	HxH				H	h[b]		D	CDA, ADA
IV	M_α,M_β or M_β	hxh[a]		E		H	H		D[b]	N-AGD *B.subtilis* and *T.maritima*
V	(M_α), M_β	hxh[a]	C			H	H		d[b]	DAA
VI	M_α,M_β	HxD		E		H	H		d[b]	hRDP
VII	M_β	qxn[a]		E		H	H		d[b]	N-AGD *E.coli*
VIII	M_α	HxH				H			D	URI
IX	M_β						H, C		E, d[b]	tatD
X	M_α,M_β	HxH			E	H	H, E		D	ADE

[a]: These residues are in the active site but do not appear to be ligated directly to the divalent cation in the M_α site. [b]: These residues are in the active site but do not appear to be directly coordinating to the divalent cation at M_β site, however, they may be bonded to the hydrolytic water molecule.

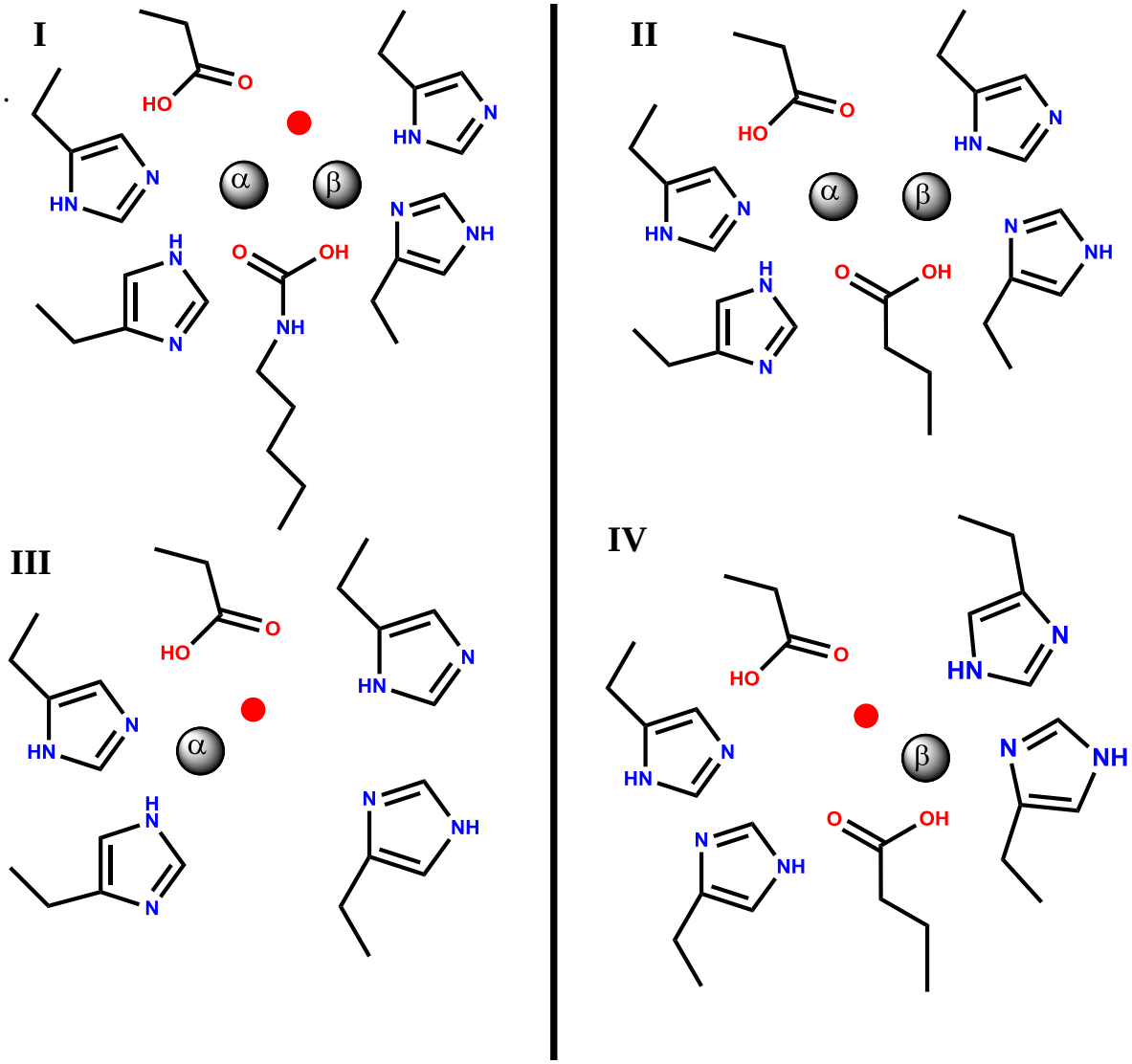


Figure 1.3: Active site subtypes in the AHS of enzymes. Two-dimensional representation of the ten different metal centers characterized in amidohydrolases. I: Dihydroorotase; II: PHP (PTE homology protein), III: cytosine deaminase; IV: N-acetyl-D-glucosamine-6-phosphate deacetylase (*N*-AGD); V: D-amino acid deacylase; VI: human renal dipeptidase; VII: *N*-AGD mononuclear; VIII: uronate isomerase; IX: tatD deoxyribonuclease; X: adenine deaminase.

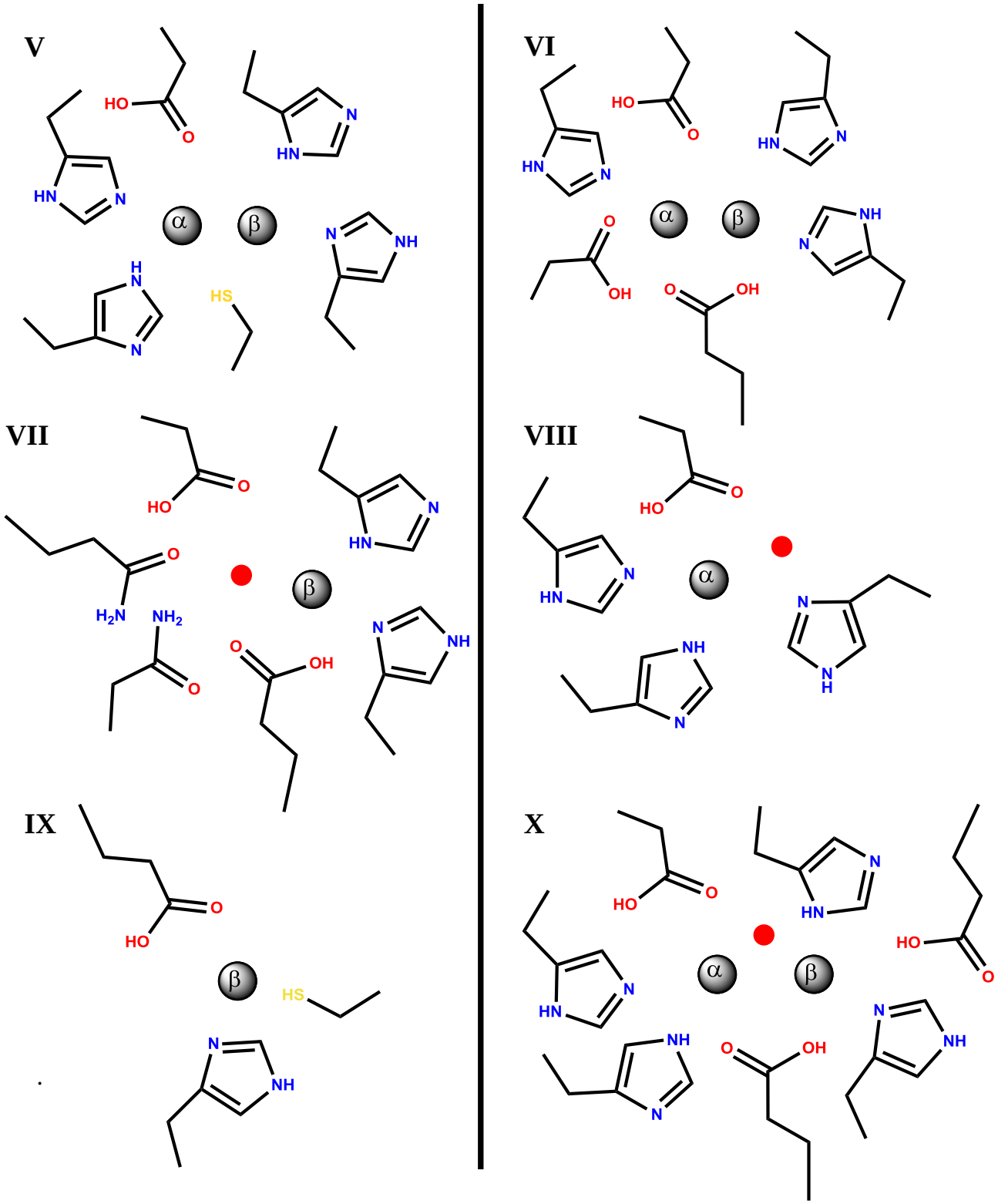


Figure 1.3 continued.

Despite the numerous strategies that have been implemented to classify the structural and functional information that is observed in the AHS, it is estimated that the extent of functional misannotations derived from protein sequence can range between 10-60%; this range is contingent to the functional database used to query a particular sequence (*71*). Databases such as GenBank NR, UniProtKB/TrEMBL, KEGG, and Swiss-Prot have been assessed to detect the level of accuracy to assign functional roles to members of six superfamilies, of which the AHS is one of them. It is estimated that the level of misannotation varies between superfamilies, but is remarkably high for members of this superfamily of enzymes. This study by Babbitt and co-workers also demonstrates that over-time the number of misannotations has increased, not only because the number of submission of genomic sequences has increased, but because of the lack of a comprehensive computational medium that can correctly assess the function of gene products from sequence alone. **Figure 1.4-A** and **B** adapted from Schnoes et. al. details the percent of misannotation found in the study in individual families within the AHS (*71*). **Figure 1.4-A** enlists the number of constituent families and number of sequences studied within the AHS, based on the annotations derived from four particular sequence databases. **Figure 1.4-B** estimates the percent misannotation of these families of the AHS.

A

Superfamily	Family	E.C. No.	Family Color	Number of Sequences Analyzed			
				Database			
				NR	TrEMBL	KEGG	Swiss-Prot
Amidohydrolase (AH)	Cytosine deaminase	3.5.4.1		82	62	75	1
	Adenosine deaminase	3.5.4.4		238	142	136	50
	N-acyl-D-amino-acid deacyclase	3.5.1.81		80	58	28	3
	L-hydantoinase	3.5.2.2		4	2	0	1
	D-hydantoinase	3.5.2.2		46	34	10	0
	Urease	3.5.1.5		267	168	112	39
	Isoaspartyl dipeptidase	—		26	25	19	1

B

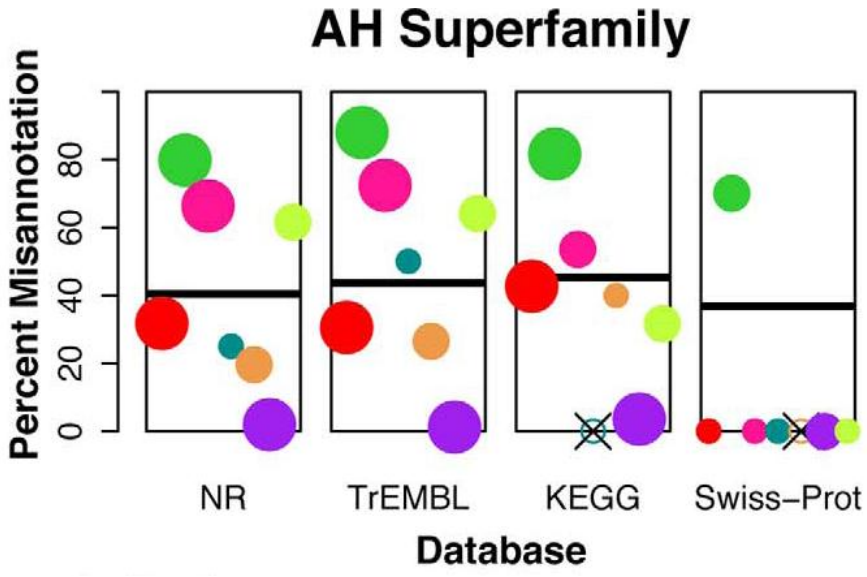


Figure 1.4: Functional missannotation in the AHS. **(A)** Sequences analyzed within constituent families with diverse functional roles, all members of the AHS. **(B)** These families are functionally annotated in four different databases; this study assessed the percent of functional misannotation in each particular database. Image adapted from Schnoes, A.M., et al. (*71*).

There are over 2,000 sequenced genomes, with the genomic information extending across various public databases. As predicted from the genomic DNA sequences of the initial 1,000 sequenced microbial genomes, it is expected that there are well over 12,000 genes coding for proteins in the AHS (*72, 73*). Knowing the extent and magnitude of this information, it has been necessary to develop a system of organization that can also serve as basis to the initial assessment of specific functional characterization of this diverse group of enzymes. It is estimated that the members in the AHS catalyze approximately over 100 different reactions, however only about 40% have been experimentally identified. (*7, 8, 55, 60, 63, 72-78*). It is also evident that there is a significant fraction of these amino acid sequences that belong to proteins or enzymes that have an unknown, uncertain or misannotated function with no empirical evidence to their catalytic activity.

The AHS has been re-organized into 24 clusters of orthologous groups/proteins (*55, 79, 80*). The clusters of orthologous groups database (COGs) has been developed as an attempt to phylogenetically classify orthologous proteins encoded in complete genomes (*79, 80*). The NCBI developed COGNITOR program organizes and categorizes newly sequenced genes to pre-existing COGs or under circumstances where there is no homology to existing sequences, the program suitably develops a new COG. Each sequence that has been assigned to the AHS belongs to one of the twenty-four COGs that have been devised based on sequence similarity to their closest homologues or orthologues. Orthologous proteins are direct evolutionary counterparts related by a common ancestor and diverged by a speciation event, rather than a duplication event, as

it is observed in paralogous proteins. Orthologous proteins show a high degree of sequence similarity and a high degree of similarity in structural fold and domain architecture, but depending on the degree of sequence similarity, they do not always catalyze the same chemical reaction. Functional studies within the AHS can be conveniently developed by focusing on specific families of proteins. **Table 1.2** lists the 24 COGs that have been identified in the AHS.

Table 1.2: Classified COGs within the amidohydrolase superfamily.

COG	Annotation as per NCBI and examples where applicable
0044	Dihydroorotase & related cyclic amidohydrolases – Examples: Dihydroorotase, allantoinase, dihydropyrimidinase, D-hydantoinase.
0084	Mg-dependent DNase – Examples: TatD family hydrolases
0402	Cytosine deaminase & related metal-dependent hydrolases – Examples: Cytosine deaminase, adenosine deaminase, S-adenosylhomocysteine deaminase, isoxanthopterin deaminase.
0418	Binuclear-metal dependent dihydroorotase.
0613	Predicted metal-dependent phosphoesterases (PHP family) – Examples: polymerase-histidinol phosphatase (PHP) family.
0804	Urea amidohydrolase (urease) α-subunit.
1001	Adenine deaminase – Examples: Binuclear adenine deaminase and N-6 methyl adenine deaminase.

Table 1.2 continued.

COG	Annotation as per NCBI and examples where applicable
1099	Predicted metal-dependent hydrolases with the TIM-barrel fold. TatD related deoxyribonuclease.
1228	Imidazolonepropionase, enamidase and related amidohydrolases
1229	Formylmethanofuran dehydrogenase subunit α
1387	Histidinol phosphatase and related hydrolases of the PHP family
1574	Predicted metal-dependent hydrolase with the TIM-barrel fold
1735	Predicted metal-dependent hydrolase with the TIM-barrel fold – Examples: Phosphotriesterase, phosphotriesterase homology protein, phospho-sugar lactone hydrolase.
1816	Adenosine deaminase – Examples: Adenosine deaminase and mononuclear adenine deaminase
1820	*N*-acetylglucosamine-6-phosphate deacetylase
1831	Predicted metal dependent hydrolases of the urease superfamily – Examples: TatD related deoxyribonuclease
1904	Glucuronate isomerase – Examples: D-glucuronate isomerase, uronate isomerase
2159	Predicted metal-dependent hydrolase of the TIM-barrel fold – Examples: γ-resorcylic acid decarboxylase, α-amino-β-carboxymuconate-ε-semialdehyde decarboxylase, isoorotate decarboxylase.
2355	Zn-dependent dipeptidase, microsomal dipeptidase homolog

Table 1.2 continued.

COG	Annotation as per NCBI and examples where applicable
3454	Metal dependent hydrolase involved in phosphonate metabolism
3618	Predicted metal dependent hydrolase of the TIM-barrel fold – Examples: sugar lactonase, 2-pyrone-4,6-dicarboxylic acid hydrolase
3653	*N*-acetyl-D-aspartate/D-glutamate deacylase
3964	**Predicted amidohydrolase dihydroorotase-like**
4464	Capsular polysaccharide biosynthesis protein

Phylogenetic associations between the different COGs in the AHS were constructed doing an all-by-all gapped BLAST sequence comparison (*10*, *81*). These generated BLAST analyses of COGs are better visualized by integration of graphical components from the software Cytoscape (*82*). Cytoscape is a bioinformatics open source platform that allows visualization of molecular network interactions and biological pathways. This program integrates networks with gene annotation and gene expression profiles. Several networks can be associated by a BLAST analysis plug-in build into the various interfaces of cytoscape. When this plug-in is adapted to associate the various sequences annotated to the amidohydrolase superfamily at specific *E*-values (expectation values), it generates a network of associated sequences based on the defined *E*-value. An *E*-value is simply a parameter describing the probability of a sequence one might expect to see by chance when searching a database of a defined size. This value takes into consideration the number of matches as well as the length of the queries (*81*). The *E*-value decreases exponentially as the number of matches of a sequence increase.

The lower the expectation value, the more significant the match becomes. For example a match at an *E*-value 10^{-70} between two proteins sequences accounts for a sequence identity of about 40%, whereas an *E*-value 10^{-100} accounts for about 60-65% sequence identity between two proteins. Sequence similarity networks for the bulk of sequences assigned to the AHS were developed. **Figure 1.5** is an image of the cytoscape generated sequence similarity network detailing the divergent evolution of members of the AHS (*55*, *72*). These networks were designed based on the notion that for every group of at least three sequences at specific BLAST stringency values, these three protein sequences will be more similar to each other than they are to any other protein sequence belonging to any other group or orthologous family (*80-82*).

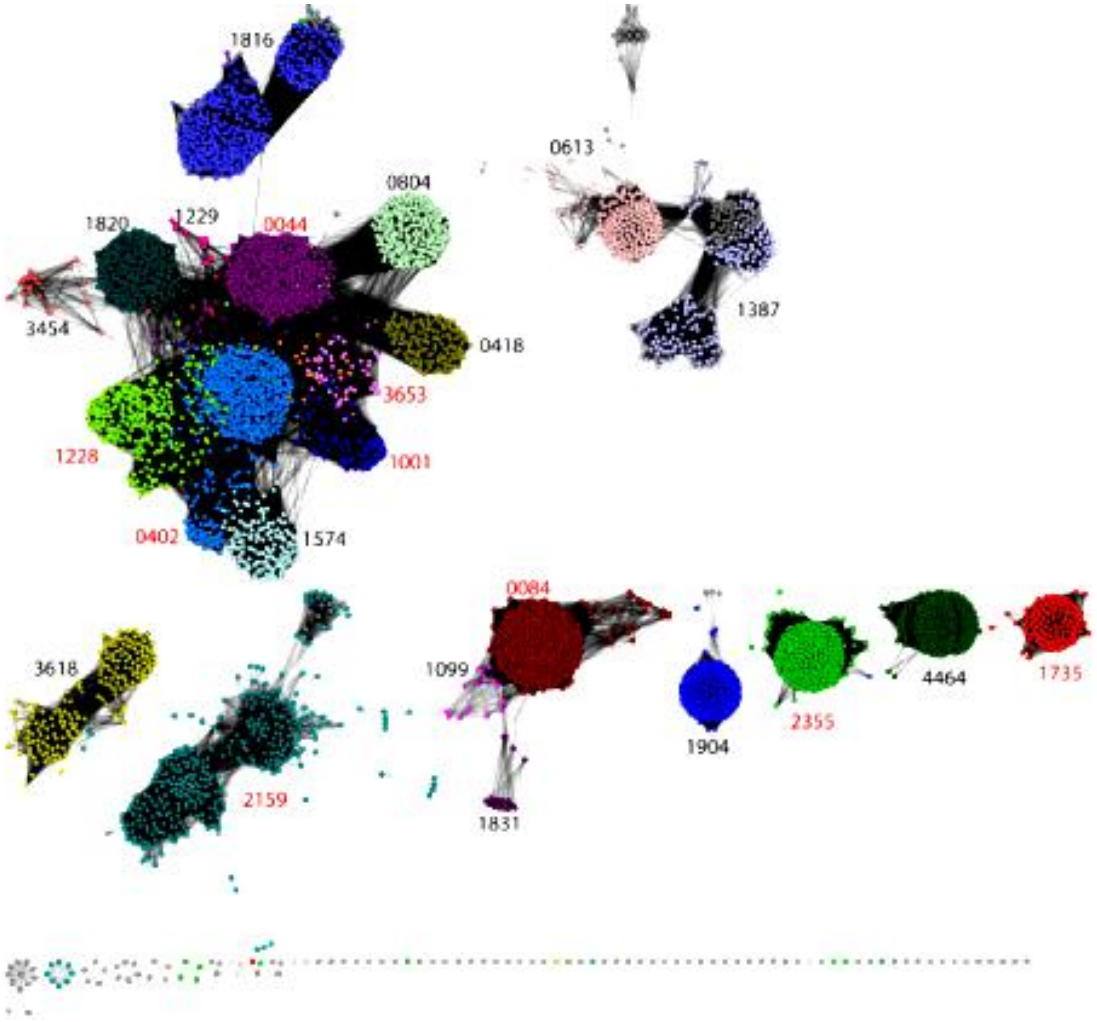


Figure 1.5: Sequence similarity network of twenty-four COGs in the AHS. Each node represents a sequence assigned to the amidohydrolase superfamily, and each edge represents a relationship between two proteins at a specified BLAST value. The network was generated at a BLAST *E*-value 10^{-10}. Image adapted from Atkinson, H.J., et al. (*81*).

Given the diverse functional roles of members in the amidohydrolase superfamily, as well as the broad, generic annotations; the focus here is directed towards the functional and structural characterization of the growing multitude of sequences assigned to this superfamily. The specific focus is to characterize those enzymes that have been sequence-based assigned to the cluster of orthologous groups (COG) 3964. This group of enzymes is functionally annotated in sequence databases as putative dihydroorotases or adenine deaminases. *In vitro* functional assessments of eight members of COG3964 have failed to demonstrate activity in the catalytic interconversion of dihydroorotate to carbamoyl-L-aspartate or in the deamination of adenine to hypoxanthine. COG3964 is listed in **Table 1.1** as a predicted amidohydrolase dihydroorotate-like. This COG however, fails to appear in the image adapted to show the overall arrangement of the sequence similarity network of members in the AHS (**Figure 1.5**). The sequence similarity network for COG3964 is presented in **Figure 1.6.** This network consists of nearly 200 non–redundant proteins all obtained from the NCBI protein cluster database and all assigned to the AHS (*83*). Each node represents a sequence in a cluster, and each edge represents the pairwise connection between two sequences with the most significant BLAST *E*-value. This cytoscape-derived sequence similarity network was developed at a PSI-BLAST *E*-value of 10^{-70}; proteins share a minimum of 40% sequence identity to other members within that group. At the stringency value of $E = 10^{-70}$ eight groups were assembled and it is predicted that enzymes within each group at this value will carry out the same catalytic function on the same substrate, but enzymes between groups will carry a different reaction on

structurally similar metabolites (*76*, *84*). The eight enzymes discussed in this study for functional characterization are members of groups 2, 6 and 7. These proteins are identified in the figure below by their locus tag.

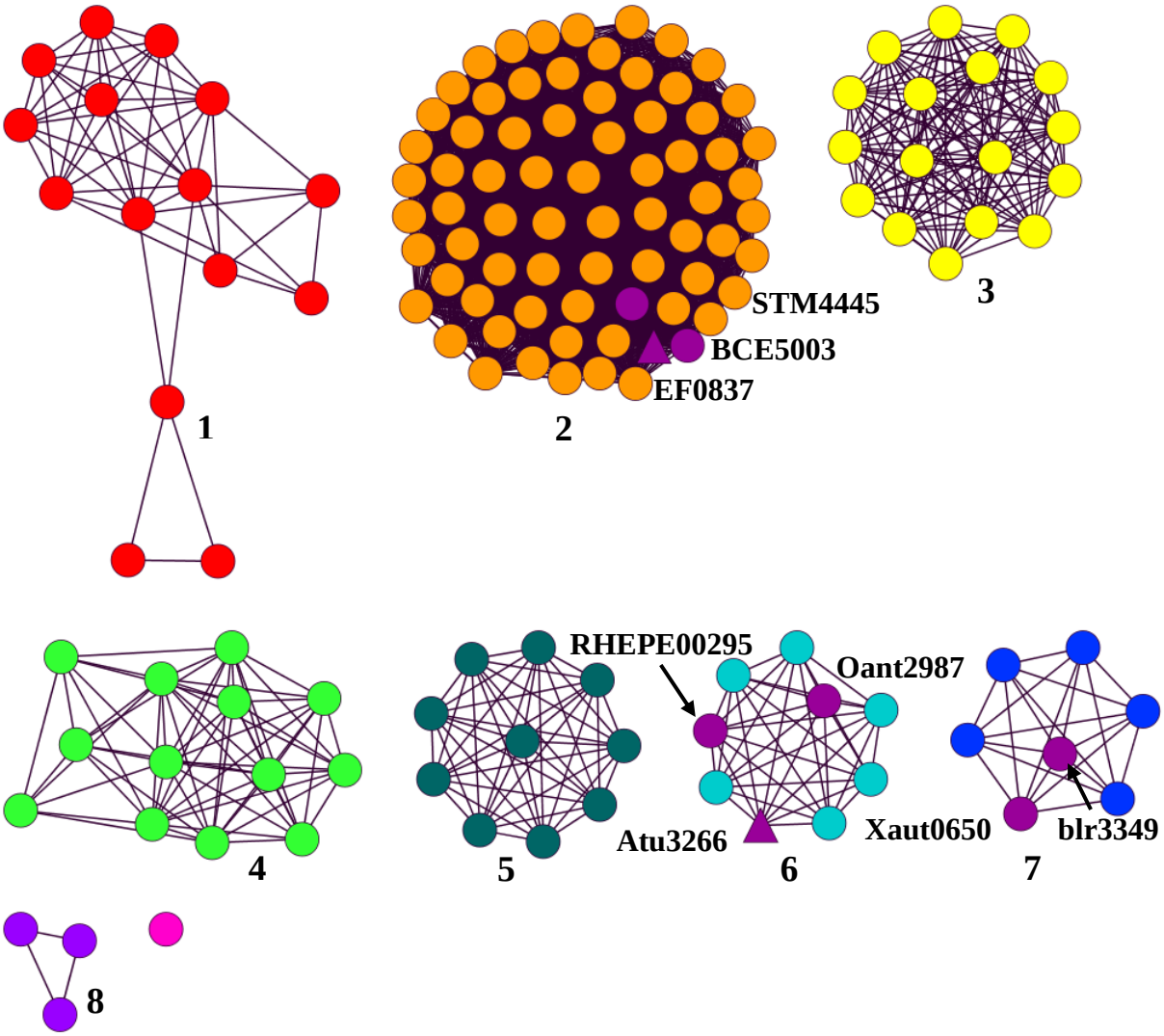


Figure 1.6: Sequence similarity network of COG3964. Those sequences for the proteins that were assessed for functional roles are identified by a purple node and a locus tag id. From group 2: EF0837, BCE_5003 and STM4445; from group 6: RHE_PE00295, Atu3266 and Oant2987; and from group 7: Xaut0650 and blr3349. Triangles in group 2 and 6 for EF0837 and Atu3266 respectively, signify an available crystal structure.

Experimentally and structurally characterized dihydroorotases are found in COG0044 and COG0418 (*61, 85, 86*), while binuclear adenine deaminases are found in COG1001 (*63*) and mononuclear adenine deaminases have been identified in COG1816 (*78*). Other deaminases are also found in COG0402. COGs 0044, 0418, 3653, 1001, and 0402 surround the area that is considered to hold sequences from COG3964 in the sequence similarity network observed in **Figure 1.5**. Enzymes in COG3964 share less than 20% sequence identity with any member from COGs 0044, 0418, 0402, 1001, and 1816. Sequence alignment comparisons also show that enzymes in COG3964 lack specific residues important in other characterized proteins for substrate selectivity and involved in the catalytic processes of dihydroorotate hydrolysis (*85*) and adenine deamination (*63, 78*). In addition sequences in group 4 of the sequence similarity network of COG3964 are also found to be assigned to COG3653, which has been characterized as *N*-acetyl-D-amino acid deacetylase cluster (DAA) (*83, 87*). Proteins from every group generated at the specified BLAST *E*-value of COG3964, were selected as targets for functional and structural characterization, however, only the identified proteins in **Figure 1.6** were able to be cloned, expressed and purified. Two enzymes had been previously structurally characterized, EF0837 from *Enterococcus faecalis* V583, found in group 2, has an X-ray crystal structure in the Protein Data Bank (PDB) with the PDB code: 2ICS. Additional targets in this group include BCE_5003 from *Bacillus cereus* ATCC 10987, and STM4445 from *Salmonella enterica subsp. Enterica serovar typhimurium* str. LT2. The other protein structure is found in group 6. Atu3266 from *Agrobacterium tumefaciens* C58 is identified in the protein data bank with the PDB

code: 2OGJ. Additional proteins studied from group 6 include: Oant2987 from *Ochrobactrum anthropi* ATCC 49188 and RHE_PE00295 from *Rhizobium etli* CFN 42. In group 7 two proteins were acquired for functional identification in this family of proteins: Xaut_0650 from *Xanthobacter autotrophicus* Py2, and blr3349 from *Bradyrhizobium japonicum* USDA 110. This group currently does not contain a crystallized protein structure. Herein, the focus is to determine the functional roles of these enzymes as members of a misannotated COG by identifying the substrate variability between these groups of proteins, and assembling a substrate profile for each group.

Functional correction of COG3964 focuses on specific groups of enzymes using a multidisciplinary sequence/structure-based strategy. The goal of an integrative approach for functional characterization has been adopted for a variety of enzyme superfamilies (*4*). This comprehensive approach employs strategies such as bioinformatics, structural genomics, computational docking, genomic operon interrogation, substrate screening and substrate modification. Sequence alignment comparisons show that enzymes in COG3964 lack specific residues observed in other enzymes that have been characterized as adenine deaminases and dihydroorotases (*63, 78, 85*). Initial interrogation of these functions was carried out with Atu3266 from group 6 and EF0837 from group 2. Although these proteins conserve a similar active site metal center observed in the crystal structure of *E.coli* dihydroorotase (DHO), they lack key substrate coordinating residues in the active site (*85*). Similar observations were found in the structural comparisons between the crystal structure of a characterized binuclear

adenine deaminase from *Agrobacterium tumefaciens* and COG3964 proteins, EF0837 and Atu3266 (*63*). Screening experiments with other analogues of dihydroorotate and adenine confirmed that these enzymes were clearly misannotated.

The availability of crystal structures for two members of COG3964 facilitated functional interrogation of this group of enzymes utilizing docking experiments. A library of ground-state and high-energy intermediates (HEI) of potential AHS substrates included molecules with electrophilic centers of restricted size based on the pocket size of the active site of the crystallized proteins, the molecules were extracted from the KEGG database of metabolites. This approach has been previously used in the AHS to determine substrate specificity for enzymes of unknown function, as well as to develop substrate inhibitors (*75, 88, 89*). Initial docking results suggested the possible role of Atu3266 and EF0837 in the hydrolysis of modified amino acids. A library of *N*-acetyl, *N*-succinyl, *N*-carbamoyl and *N*-formyl D- and L- amino acids was tested against the purified proteins. Initial screens determined the ability of Atu3266 to hydrolyze *N*-acetyl-D-amino acids, in particular *N*-acetyl-D-serine and *N*-acetyl-D-threonine. Determination of these activities resulted in the modification at various positions of the amino acid, including the C-1 carboxylate moiety, the C-2 chiral carbon, the acetyl portion representing the modification at the *N*-terminal, as well as the side chain position. Continual substrate modification led to the discovery of a substrate with increased activity. Acetyl-*R*-mandelate was found to be the best substrate in groups 2 and 6 of COG3964, and was found to be hydrolyzed at rates of 10^4 $M^{-1}s^{-1}$ by an enzyme previously characterized as a D-amino acid deacetylase from group 4 (GOX1177) (*87*).

There is a wealth of information that can be derived from insights and inquires of the operon neighbors of a misannotated gene (*91-95*). This computational assisted approach focuses on the adjacency and placement of genes as means to directly imply functional interactions between genes in the same operon. These insights can provide potential function of a selected gene, subject to the annotation of the adjacent neighbors. Genomic operon analyses of organisms with enzymes in COG3964 have identified a common phylogenetic profile pattern found in about 70% of the operons for organisms that have an amidohydrolase in this COG. A gene functionally annotated as a selenocysteine synthase or SelA belonging to COG1921, is found adjacent to the COG3964 annotated gene. **Figure 1.7** depicts the presence of a SelA gene (COG1921) in the vicinity of amidohydrolases in group 2, EF0837 and BCE_5003; and in group 6, Atu3266 and Oant2987. 100% of the organisms in group 2 of COG 3964, as well as 6 of the 9 organisms in group 6 have an annotated SelA gene adjacent to the AH encoding gene.

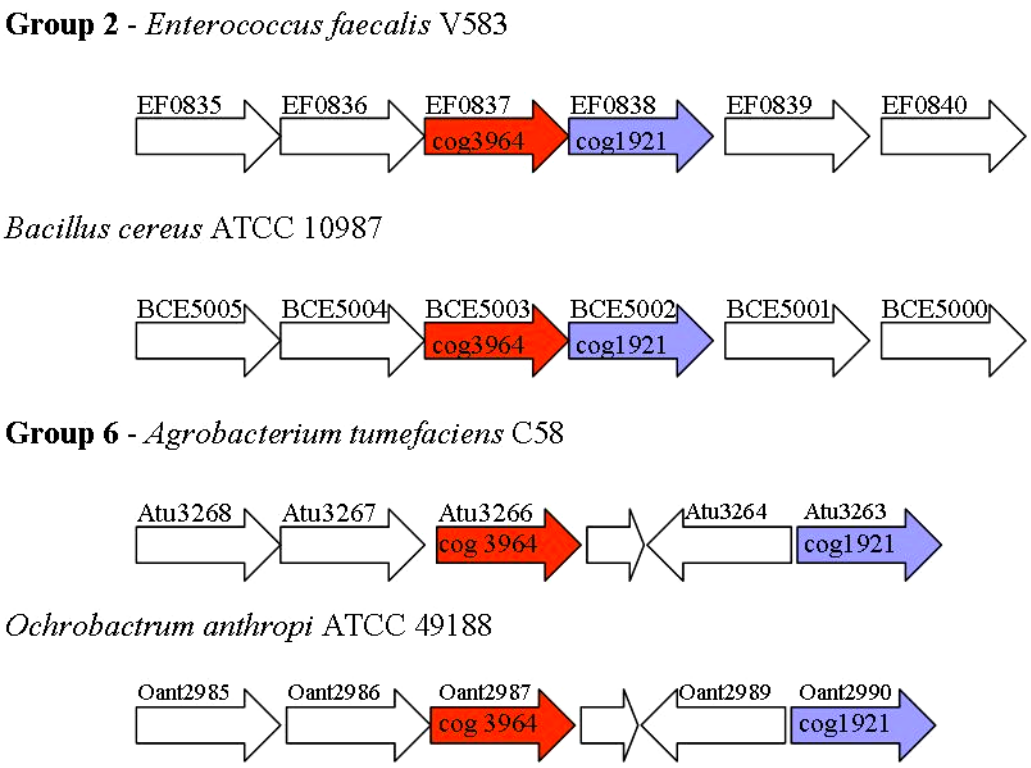


Figure 1.7: Phylogenetic profiles of organisms with COG3964 enzymes. Schematic illustrates the genomic operon for *E. faecalis* and *B. cereus* containing a COG3964 gene encoding an amidohydrolase in group 2 of the sequence similarity network, and an adjacent COG1921 selenocysteine synthase gene. In group 6 of the similarity network, the same manifestation of the two genes is observed in *A. tumefaciens* and *O. anthropi.*

Enzymes commonly annotated as SelA also belong to COG1921; collectively these enzymes are dependent on a pyridoxal-5'-phosphate cofactor and have been given the functional annotation of catalyzing the synthesis of selenocysteine, the 21st amino acid. Assessment of the operon encoding true selenocysteine synthases reveals that these enzymes require the presence of additional adjacent factors in the synthesis of selenocysteine charged tRNA; however, the presence of an amidohydrolase is not one of those requirements (*95, 96, 97*). Analysis of the genomic operon has been a powerful tool in previous studies for the annotation of functionally unknown amidohydrolases (*59, 98*). Here, the consistent presence of the SelA gene adjacent to a misannotated and functionally uncertain COG3964 gene inspired the study of COG1921 as means to determine a possible connection and establish the functional roles of COG3964. Three genes in COG1921 adjacent to genes encoding amidohydrolase enzymes were cloned, expressed and purifed. EF0838, Atu3263 and Oant2990 are identified in **Figure 1.8.** This image illustrates the sequence similarity network for COG1921 detailing the presence of the proteins that were targeted, cloned, expressed and purified as possible functional support to COG3964 AHs. True bacterial selenocysteine synthases have been identified in the sequence similarity network for COG1921 in group 1. The majority of COG1921 genes found as neighbors to AH enzymes are found in groups 2, 3 and 7.

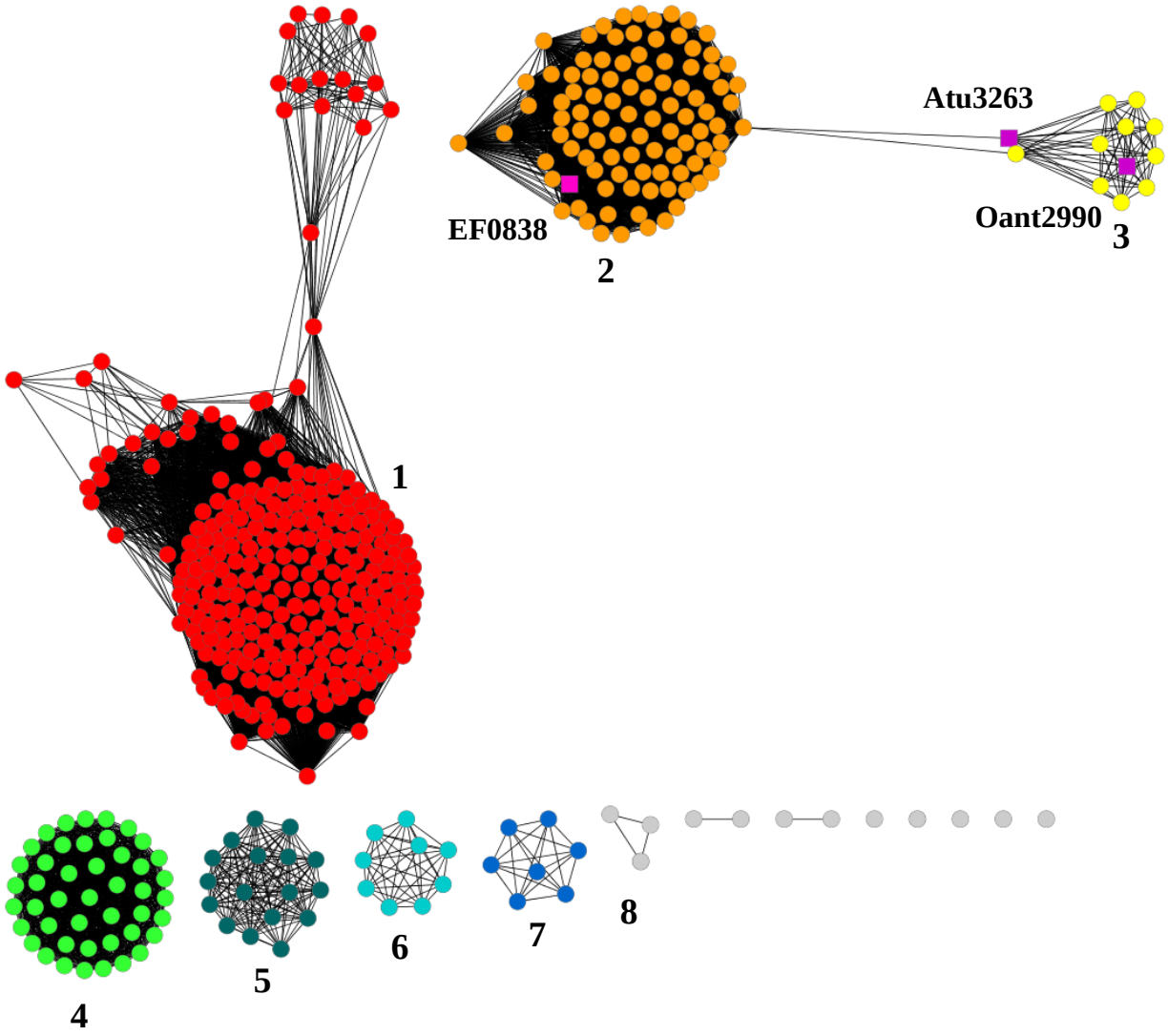
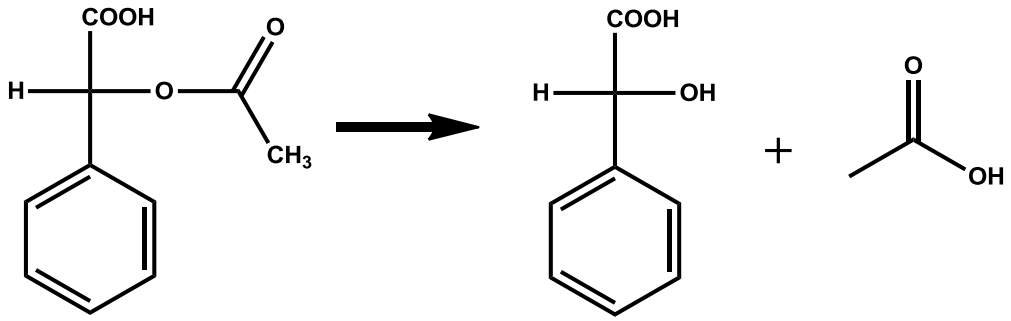


Figure 1.8: Sequence similarity network of COG1921. This network was prepared at an *E*-value 10^{-70}. Enzymes in COG1921 are annotated as selenocysteine synthases. Based on literature, the only characterized bacterial SelA proteins have been found in group 1. The majority of SelA annotated enzymes in groups 2 and 3 are found as neighbors to COG3964 assigned AHs.

As mentioned previously, the best found substrate for various AH enzymes from COG3964 has been acetyl-*R*-mandelate. Determination of this substrate was based on substrate development. Initial screens of a library of *N*- modified D- and L- amino acids and dipeptides determined that *N*-acetyl-D-serine was the first compound showing hydrolytic activity in the presence of enzymes from COG3964. Upon modifications to this substrate at various positions, it was observed that the ester bond was hydrolyzed significantly faster than the amide bond. The best determined substrate was the *R*-enantiomer of the α-acetyl carboxylate with a phenyl group attached to the central chiral carbon, or acetyl-*R*-mandelate (**Scheme 1.1**).



Scheme 1.1: Reaction catalyzed by enzymes: Atu3266, Oant2987 and RHE_PE00295, all found in group 6 of COG3964 sequence similarity network. Hydrolysis was observed at lower rates in EF0837, STM4445 and BCE_5003 enzymes found in group 2.

Mandelate has been observed to be a metabolite used by all *Pseudomonas sp.* (*99*), as well as various strains of *Arthrobacter*, *Asotobacter*, *Bacillus*, *Nocardia*, *Rhizobium* and *Rhodopseudomonas* (*99-102*). None of the organisms encoding a COG3964 amidohydrolase have been found to use *R*- or *S*- mandelate as a metabolite. Degradation of this compound can be carried out in at least as many different pathways or variations as glucose (*99, 103*), and although it is not a common compound, a large number of organisms can sustain growth on either one or both enantiomers of mandelate as sources of carbon and energy for growth (*99*). The first step in the degradation of mandelate is carried out by a stereospecific mandelate dehydrogenase, however in various organisms, even before this step takes place, a mandelate racemase often provides with the correct isomer that can be metabolized (*103-105*). Acetyl-*R*-mandelate is not a metabolite that has yet been identified in organisms degrading mandelate. Recently, it was found that the whole cells of the strain *Pseudomonas* sp. ECU1011 was able to produce chiral mandelate by enantioselective deacetylation of a mixture of acetyl *R,S*-mandelate (*106*). The cells were specific for the hydrolysis of the *S*-enantiomer and typically all strains of *Pseudomonas aeruginosa* can grow on the S-(+), but not the R-(-) isomer of mandelate (*99, 103, 105*).

Esterase activity of acetyl-*S*-mandelate has also been detected by carboxypeptidase A., a pancreatic, zinc containing metalloenzyme that mediates the hydrolysis of peptides having a free carboxyl function at the terminal α-amino acid moiety and which must be of the L- or *S*- configuration (*107, 108*). Here however, no substrate with the *S*- configuration was active. At this stage, it is not known if acetyl-*R*-

mandelate is a bona fide substrate, or if the enzymes that were able to carry out the deacetylation of the compound are merely exhibiting signs of substrate promiscuity.

Enzymes in the AHS can carry out a multitude of reactions on structurally similar compounds; in COG3964, because there is no defined functional role, the activity that has been determined cannot be assessed as being the activity that the enzyme was designed to catalyze.

Here, functional annotation of COG3964 is studied and assessed to provide a more definitive metabolic function to the various proteins that have been misannotated to this cluster of orthologous groups. The functional and structural characterization of enzymes in group 6 found to catalyze, at enzymatic competent rates, the deacetylation of α-acetyl carboxylates is discussed in Chapter 2. Chapter 3 focuses on the structure and functional diversity that is found in group 2 of COG 3964. The enzymes in this group were hydrolytically less active than those found in group 6 with the library of compounds developed for enzymes in group 6. In Chapter 4, the focus is on enzymes from group 7; these enzymes have a completely different substrate and do not carry out the hydrolysis of α-acetyl carboxylates. The number of compounds tested in group 7 was extended to include a variety of cyclic compounds including: diketopiperazines, hydantoins, nucleosides, nucleotides and lactone sugars. Enzymes in groups 2, 6 and 7 share less than 35% sequence identity.

In addition, functional investigation is extended to the determination of the functional roles of operon proteins. In Chapter 5, enzymes from COG1921 or SelA, annotated as selenocysteine synthases, are discussed. These enzymes encoded in the

neighboring genes of organisms containing a COG3964 amidohydrolase are cloned, expressed, purified and studied in the presence of various amino acid substrates to determine their possible connection towards the correct functional identification of amidohydrolases in COG3964.

CHAPTER II

FUNCTIONAL ANNOTATION AND THREE-DIMENSIONAL STRUCTURE OF INCORRECTLY ANNOTATED DIHYDROOROTASES FROM COG3964 IN THE AMIDOHYDROLASE SUPERFAMILY

The development of a comprehensive strategy for the functional annotation of proteins and enzymes whose sequences have been deposited in public databases has proven to be a difficult and persistent challenge. Utilization of homology-based sequence comparisons for functional annotation of newly sequenced genes can often lead to the annotation of the wrong function when unreasonable threshold values are used (*109-111*). The end result is often a misrepresentation of the potential metabolic transformations contained within a given organism. In addition to the misannotation of enzyme function there is a significant fraction of the total gene inventory that is simply not annotated (*112-114*). This observation suggests that a substantial segment of the metabolic landscape remains to be discovered.

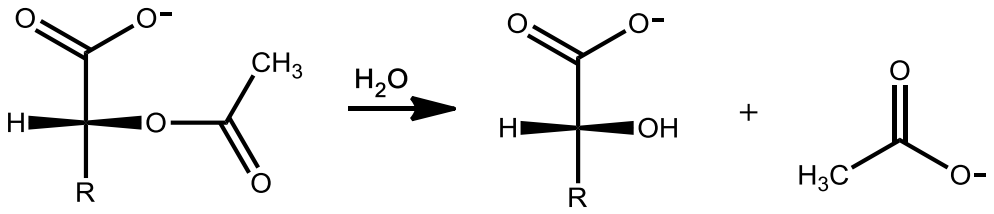
In many cases, functional annotation of newly sequenced genes can be strategically approached by the integration of structural genomics (*91,115*), computational docking (*88, 116*), high-throughput screening (*117*), and genome context analysis (*92, 118*). We have focused our efforts toward the development of a simple and integrated strategy for functional annotation that is based on an assault on the amidohydrolase superfamily (AHS) of enzymes. This superfamily was first identified and recognized from the three-dimensional structural similarities between urease,

adenosine deaminase and phosphotriesterase (*42*). The amidohydrolase superfamily is characterized structurally by an active site that contains a mono- or binuclear metal center embedded at the C-terminal end of a distorted $(\beta/\alpha)_8$-barrel protein fold. Typically identified by an HxH motif after the C-terminal end of β-strand **1**, the enzymes from the AHS also contain other coordinating ligands to the metal center at the ends of β-strands **4**, **5**, **6** and **8** (*53*). The metal center serves in the activation of a hydrolytic water molecule for nucleophilic attack and to stabilize the transition state. Most of the reactions catalyzed by this diverse superfamily involve the hydrolysis of C-O, C-N or P-O bonds. However, some members of this superfamily also catalyze decarboxylation, hydration or isomerization reactions (*56-58*). Enzymes from the AHS catalyze reactions in the metabolism of carbohydrates, amino acids, nucleic acids and the degradation of organophosphate esters.

The AHS has been organized into 24 Clusters of Orthologous Groups (COG) (*79, 80*). One of these clusters, COG3964, represents one of the smaller homologous groups of amidohydrolases, with approximately 200 sequences identified to date. A sequence similarity network for this COG is presented in **Figure 2.1** at an *E*-value cutoff of 10^{-70}. Some members of COG3964 have been annotated as dihydroorotases, which catalyze the interconversion of dihydroorotate and *N*-carbamoyl aspartate (*61, 85*), while other members have been annotated as catalyzing the deamination of adenine (*63*). Experimentally annotated dihydroorotases are found in COG0044 and COG0418. Structurally and experimentally characterized adenine deaminases are found in COG1001 (*63*) and in COG1816 (*78*). A straightforward comparison of the amino acid

sequences found within the proteins from COG3964 with the experimentally annotated adenine deaminases and dihydroorotases clearly indicates that the residues required for substrate recognition of adenine and/or dihydroorotate are not conserved. This observation suggests that either these annotations are clearly wrong for COG3964, or that a previously unrecognized novel constellation of amino acids has evolved for the deamination of adenine or the hydrolysis of dihydroorotate.

In this chapter we report the three-dimensional crystal structure and substrate profile for three enzymes from COG3964; Atu3266 from *Agrobacterium tumefaciens* C58, Oant2987 from *Ochrobactrum anthropi* ATCC 49188, and RHE_PE00295 from *Rhizobium etli* CFN 42. All of these enzymes are part of group 6 in COG3964. Neither of these enzymes is able to catalyze the hydrolysis of dihydroorotate or the deamination of adenine. We have utilized a focused compound library and computational docking to discover that these three enzymes from COG3964 actually catalyze the hydrolysis of acetylated α-hydroxyl carboxylates as shown in **Scheme 2.1**. The best substrate identified to date is acetyl-*R*-mandelate. This compound is enzymatically hydrolyzed with a k_{cat}/K_m value of 2.8 x 10^5 M^{-1} s^{-1}.



Scheme 2.1: General architecture of compounds found to be hydrolysable by enzymes from group 6 of COG3964. A generic α-acetyl carboxylate is hydrolyzed to yield an acetate molecule and a α-hydroxyl acid.

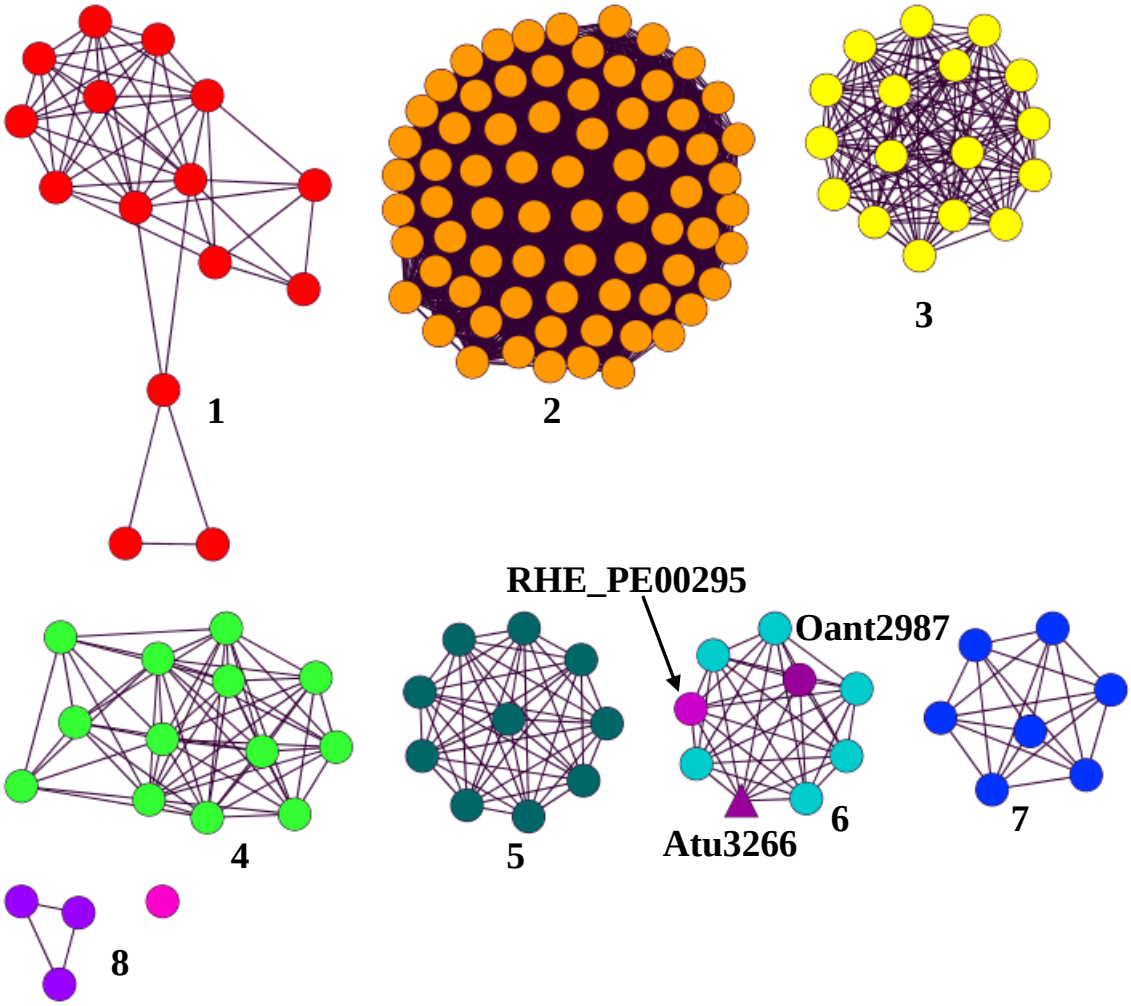
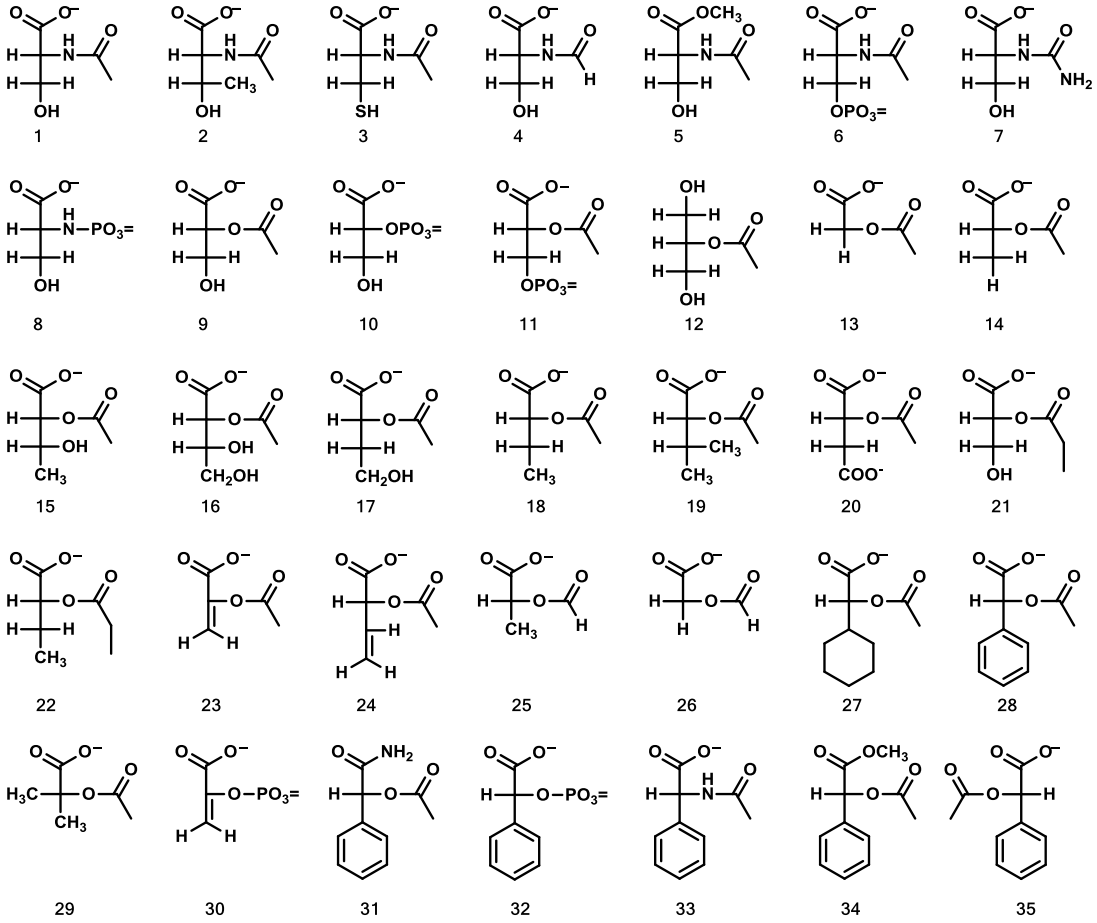


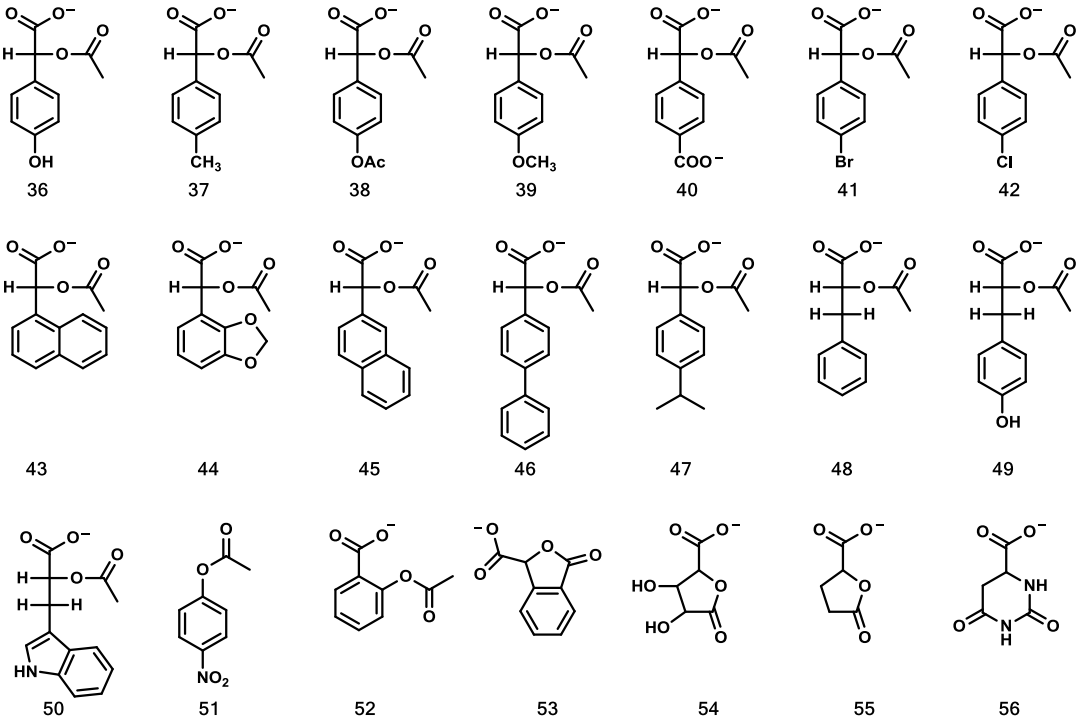
Figure 2.1: Sequence similarity network of COG3964 with group 6 enzymes. Each node in the diagram represents a sequence that has been annotated to this cluster of orthologous groups, and each edge represents the pairwise connection between two sequences at a BLAST *E*-value of 10^{-70}. Group 6 contains the amidohydrolases discussed in this chapter and these are shown in purple (Atu3266, Oant2987 and RHE_PE00295). Atu3266 has been structurally characterized and its crystal structure is found in the protein database (PDB: 2OGJ). This protein is identified by a triangle.

MATERIALS AND METHODS

Materials: The genomic DNA from *Ochrobactrum anthropi* strain ATCC-49188 was purchased from the American Type Culture Collection (ATCC). DNA oligonucleotide synthesis and sequencing reactions of plasmid constructs were conducted at the Gene Technology Laboratory at Texas A&M University. The chemicals utilized in the expression, purification, and screening of Atu3266, Oant2987 and RHE_PE00295 were obtained from Sigma-Aldrich, unless stated otherwise. The purified protein RHE_PE00293 (gi|86360569) was obtained from the Enzyme Function Initiative protein core (EFI:505180). The preparation of the acylated compounds tested for substrate activity was conducted by Dr. Tamari Narindoshvili at Texas A&M University. The structures for many of these compounds are provided in **Scheme 2.2**. The pET-20b(+) expression vector and Rosetta-gami™ B(DE3) pLysS competent cells were obtained from Novagen. The acetate and formate detection kits were purchased from Megazyme. *N*-acetyl-DL-phenyl glycine was purchased from Chem-Impex International. 3-Bromomandelic acid and 4-bromomandelic acid were purchased from Oakwood Products, Inc. Mutants of Atu3266 were prepared using the QuikChange™ site directed mutagenesis kit from Stratagene. Molecular biology materials were acquired from Invitrogen including all polymerase enzymes and T4-DNA ligase. Restriction endonucleases were acquired from New England biolabs. The His-tag purification Ni^{2+}-binding NTA resin was obtained from Thermo Scientific.



Scheme 2.2: Diagram of compounds (1-35 and 36-56) tested for activity with Atu3266, Oant2987 and Rhe_PE00295, compounds.



Scheme 2.2 continued.

Gene Cloning: The gene for Oant2987 (gi|153010310) from *Ochrobactrum anthropi* was amplified by PCR using the primer pairs 5'-ACAGGAGCCATATGATTTCCGGTGAACAGGCGAAGCCG-3' containing an *Nde*I restriction site and 5'-ACGCGAATTCCCAGCGCCACGAATAGCCATGGCTATGGC-3' having an *Eco*RI restriction site. Oant2987 was inserted into a pET-20b(+) vector that had been previously digested with *Nde*I and *Eco*RI. The gene for Atu3266 was initially cloned by the New York SGX Research Consortium for Structural Genomics from *Agrobacterium tumefaciens* (gi|159185666). The poor overexpression and solubility of this clone in BL21(DE3) cells led to the removal of the gene from its original TOPO-isomerase vector, pSGX3(BC), using the restriction sites CCGC$^{\downarrow}$GG and CA$^{\downarrow}$TATG of the construct and removing the gene by digestion with *Sac*II and *Nde*I restriction enzymes. The gene was transferred to a pET-20b(+) vector previously digested with the same restriction enzymes. Both cloned constructs contained a $(His)_6$ C-terminal purification tag. The plasmids were transformed into XLI-Blue competent cells. Colonies were selected from LB plates containing 100 μg/mL ampicillin. The plasmids were then purified using the QIAprep spin miniprep kit. The fidelity of both inserts (Oant2987 and Atu3266) was verified by DNA sequencing.

Expression and Purification of Oant2987 and Atu3266: The plasmid containing the gene for Oant2987 was transformed into BL21(DE3) and plated on LB agar-ampicillin plates. A single colony was inoculated into overnight cultures containing 5 mL of LB medium supplemented with 100 μg/mL of ampicillin. These cultures were utilized to make 1 L cultures containing the same concentration of ampicillin and grown

at 30 ^{o}C in a shaker-incubator. After an OD_{600} of 0.3 was obtained, the cells were supplemented with 150 μM of 2,2'-bipyridyl to coordinate excess ferric iron from the medium. At an OD_{600} of 0.6, the cells were induced with 200 μM IPTG and the addition of 1.0 mM $Zn(OAc)_2$. The temperature upon induction was reduced to 20 ^{o}C and the cultures were allowed to continue to grow for an additional 12 hours. The cells were harvested by centrifugation and frozen at -80 ^{o}C. The frozen cell pellet was thawed and re-suspended in 60 mL of 50 mM HEPES buffer, pH 7.6 (buffer A). The cells were supplemented with 10 μg/mL of phenylmethanesulfonyl fluoride (PMSF) and lysed by sonication at 0 ^{o}C. The supernatant solution was treated with 2% w/v protamine sulfate and centrifuged. The protein was precipitated with ammonium sulfate between 40% and 80% saturation. The protein pellet was resuspended with a minimum amount of buffer A and loaded onto a Superdex 200 gel filtration column. The fractions of interest were collected and subsequently loaded onto a Resource Q column. The protein was eluted from the column with a gradient of 20 mM HEPES buffer, pH 7.6, containing 1 M NaCl (buffer B). The purity of Oant2987 was greater than 95% based on SDS-PAGE.

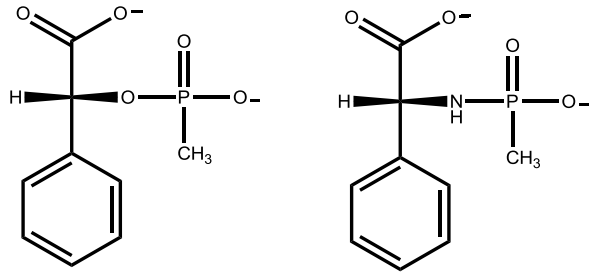
The plasmid containing the gene for Atu3266 was transformed into Rosetta gami B (DE3) pLysS electrocompetent cells. A single colony containing the gene of interest was inoculated into an overnight culture containing 5 mL LB medium, 100 μg/mL of ampicillin and 20 μg/mL of chloramphenicol. The 5 mL starter cultures were used to inoculate 1 L of LB medium. The cells were incubated at 30 ^{o}C in a shaker-incubator. When the OD_{600} reached 0.3, the cells were treated with 150 μM 2,2'-bipyridyl. The cells continued to grow to an OD_{600} of about 0.6, at which point 200 μM of IPTG and 1.0 mM

$ZnCl_2$ were added. The temperature was reduced to 20 ^oC and the cells were grown overnight for 16 hours. The cells where harvested by centrifugation at 8000g for 10 minutes and then frozen at -80 ^oC. The cell pellet was resuspended in binding buffer containing 20 mM HEPES, 500 mM NaCl and 5.0 mM imidazole at pH 7.6. The cells were supplemented with 10 μg/mL of PMSF and lysed by sonication at 0 ^oC. The cell debris was removed by centrifugation and then Atu3266 was purified using a His-tag affinity column. The cell lysate supernatant was filtered though a 0.2 μm cellulose acetate sterile filter and loaded onto Ni^{2+}-NTA column equilibrated with binding buffer. The column was washed with binding buffer until the A_{280} remained constant and below an absorbance of 0.1. The protein of interest was eluted using a gradient of 0-500 mM imidazole in a buffer solution containing 10 mM HEPES, and 250 mM NaCl at pH 7.6. After dialysis, the protein was judged to be >95% pure by SDS-PAGE.

Preparation of Atu3266 Mutants: Site directed mutagenesis of Atu3266 was implemented to probe the catalytic role of residues found in the active site of Atu3266. Arginine-177 is a conserved amino acid found in all the sequences of group 6. This variable residue is found two residues away from the carboxylated lysine that bridges the α- and β- metal. The residue has also been observed in the sequences from group 1 and 2. In groups 3 and 5 this arginine is replaced with a lysine, and in group 7 the residue is replaced with a histidine. The mutants designed to analyze the role of arginine 177 include R177A and R177H. In addition the aspartate from β-strand 8 was mutated. The mutant is referred to as D291N. Kinetic assays in the presence of acetyl-*R*-mandelate were carried out to characterize each of these three mutants.

Protein Concentration and Metal Analysis: Protein concentration was determined spectrophotometrically using a SPECTRAmax-384 PLUS UV-vis spectrophotometer. The concentration was obtained from the absorbance at 280 nm using the extinction coefficients determined from the amino acid sequence (web.expasy.org/protparam/). The extinction coefficients for Rhe_PE00295, Oant2987 and Atu3266 are 32,680 $M^{-1}s^{-1}$, 43,605 M^{-1} cm^{-1} and 43,969 M^{-1} cm^{-1}, respectively (*119*). The metal content of the purified proteins was determined by inductively coupled plasma emission – mass spectrometry (ICP-MS) using a Perkin-Elmer Analyst 700 atomic absorption spectrometer. The samples were prepared by heating 1.0 μM enzyme with 1% (v/v) HNO_3 for 30 minutes.

Synthesis of Dipeptide Libraries, N-Acyl Amino Acids and Methyl Phosphonates: Syntheses of dipeptide libraries (L-Xaa-L-Xaa, D-Xaa-L-Xaa, and L-Xaa-D-Xaa), and *N*-acetyl, *N*-formyl, and *N*-succinyl derivatives of D- and L-amino acids were prepared as described previously (*7, 87*). The preparation of the *O*- and *N*- methyl phosphonate derivatives of R-mandelate (**57**) and D-phenyl glycine (**58**) respectively, were prepared as previously described (*7*). The structures of these compounds are presented in **Scheme 2.3** and their composition was verified by mass spectrometry and NMR.



57 58

Scheme 2.3: Illustration of the two compounds synthesized and used for inhibition of Atu3266 and Oant2987. On the left O-phosphonate of acetyl-*R*-mandelate (**57**) and on the right is *N*-phosphonate of *N*-acetyl-D-phenyl glycine (**58**).

Screening of N-Acyl and Dipeptide Libraries: The preliminary substrate screening of Atu3266 and Oant2987 was initiated by mixing each protein with the *N*-acyl libraries of the D- and L- amino acids (*N*-formyl and *N*-succinyl were screened similarly), and three dipeptide libraries (L-Xaa-L-Xaa, D-Xaa-L-Xaa and L-Xaa-D-Xaa). The assays were conducted as previously described using a Cd-ninhydrin assay for the detection of free amines (*87*). Each assay was buffered with 20 mM HEPES, pH 7.6 and each library contained 17-19 modified amino acids (L- and D- cysteine were not included). Each assay contained ~100 μM of each component and 0-1000 nM of Atu3266 or Oant2987 was added to initiate the reaction. Rhe_PE00295 was not used in the screening phase of substrate search due to limited quantities of the enzyme.

A negative control was prepared without the addition of either Atu3266 or Oant2987. The initial screening reactions were conducted at 30 ^{o}C for 15 hours. The formation of free amino acids was detected using a modified Cd-ninhydrin assay (*120*).

Each 70 μL reaction mixture was quenched with 280 μL of ninhydrin reagent. The entire mixture was heated at 85 ^{o}C for 15 minutes and then cooled. A 250 μL aliquot was transferred to a 96-well UV-visible micro plate and the extent of total free amino acids was measured at 507 nm.

Screening Methods for Deacetylase Activity: All compounds having a hydrolysable acetyl moiety were screened using the acetic acid kit, KACETAF, from Megazyme®. The catalytic activities of Atu3266, Oant2987 and Rhe_PE00295 were monitored by the formation of acetate and the subsequent reduction of NAD^+ by the coupled activity of acetyl-coenzyme A synthetase, citrate synthase and L-malate dehydrogenase. The reaction was monitored spectrophotometrically at 340 nm. The 250 μL reaction mixture contained 1.0 mM of the test compound, 20 mM HEPES, pH 7.6 and 75 μL of the coupling system containing 128 mM TEA buffer pH 8.4, 5.0 mM NAD^+, 3.1 mM ATP, 3.2 mM $MgCl_2$, 0.15 mM CoA, 4 U of L-malate dehydrogenase, 0.6 U of citrate synthase, and 0.3 U of acetyl CoA synthetase. Each compound was tested for ester hydrolysis at 30 ^{o}C and the reaction was initiated by the addition of 1.0 μM of Atu3266, Oant2987 or RHE_PE00295.

Screening of O-Propionyl Compounds: The hydrolysis of *R*-2-(propionyloxy)-butanoate (**22**) and *R*-3-hydroxy-2-(propionyloxy) propanoate (**21**) was assessed using a pH-sensitive colorimetric assay (*121*). The hydrolysis of the ester bond releases a proton that was detected using a pH indicator dye, cresol purple. The screening reactions were carried out in 2.5 mM Bicine buffer, pH 8.3, containing 0.2 M NaCl and 1.0 mM of each compound with up to 1.0 μM enzyme. Each reaction contained 0.1 mM cresol purple in

1% DMSO and the change in absorbance was monitored at 577 nm. The extinction coefficient under these reaction conditions was determined to be 1.51 x 10^3 M^{-1} cm^{-1} using acetic acid as a titrant.

Data Analysis: The kinetic constants k_{cat}, K_m, and k_{cat}/K_m were determined for Atu3266, Oant2987 and RHE_PE00295 for selected substrates by fitting the initial velocity data to equation 1. Competitive inhibition constants were determined using equation 2 for a competitive inhibitor. In these equations v is the initial velocity, E_t is the enzyme concentration, k_{cat} is the turnover number, A is the substrate concentration, K_m is the Michaelis-Menten constant, I is the inhibitor concentration and K_{is} is the slope inhibition constant.

$$v / E_t = k_{cat} [A] / (K_m + [A]) \quad (1)$$

$$v / E_t = k_{cat} [A] / K_m (1 + I / K_{is}) + A \quad (2)$$

Crystallization and Structure Determination of Atu3266: Selenomethionine (SeMet) substituted Atu3266 from *Agrobacterium tumefaciens* with the 6x-His tag intact was crystallized by mixing 1.5 μL of protein (5 mg/mL in a buffer of 10 mM HEPES pH 7.5, 150 mM NaCl, 10 mM methionine, 10% glycerol and 5mM DTT) with 1.5 μL of reservoir solution containing 0.1 M HEPES pH 7.5, 25% PEG3350, 0.2 M ammonium sulfate and equilibrating against the same reservoir solution by the hanging drop vapor diffusion method. Crystals appeared after three days and were flashed-cooled in liquid nitrogen for data collection.

A native dataset and a complete MAD dataset from single crystals were collected at 100 K on beamline 31-ID at APS using a Mar CCD 225 detector. Crystal diffracted to 2.6 Å, and belongs to the orthorhombic space group $P2_12_12_1$ with 6 molecules in the asymmetric unit. Data were indexed, integrated and scaled using the program HKL2000 (*122*). Selenium sub-structure was determined using SHELXD (*123*). The phase refinement and density modifications were carried out using SHAARP and SOLOMON (*124, 125*). Model building was done using ARP/wARP (*126*). Further model building and refinement of the structure was carried out in iterative cycles using 'O' and CNS 1.1 (*127, 128*). Extra residual density was observed near the NZ atom of Lys175 and the lysine residue was modeled as a carboxylated lysine. In addition, two zinc ions and one imidazole were located per molecule and included in the later stages of refinement. The atomic coordinates and structure factors for the Atu3266 structure has been deposited in the Protein Data Bank under accession code 2OGJ. Crystal, diffraction data and refinement details are given in **Table 2.1.**

Table 2.1: Data collection and refinement statistics for crystallized Atu3266

beamline	NSLS beamline X12C		
	Data Set - Unit Cell Parameters		
a(Å)			90.65
b(Å)			139.2
c(Å)			206.13
$\alpha = \beta = \gamma = 90^o$			
space group	$P2_12_12_1$		
	Data Collection Statistics		
	Peak	Inflection	Remote
wavelength (Å)	0.9793	0.9798	0.94
resolution range (Å)	50-2.6	50-2.8	50-2.8
redundancy	5.8	5.5	5.2
Mean $I/\sigma(I)$	14.5	13.2	12.8
R_{merge}[a]	0.121	0.125	0.118
no. of unique reflections	139906	138603	138965
	Phasing Statistics		
Phasing power[b] (ano)			0.21/0.28
FOM[c]: (centric/acentric)			0.85
	Refinement Statistics		
resolution range (Å)			2.62-39.7
no. of reflections (work)			139800
no. of reflections (test)			4173
no. of protein atoms			16450
no. of heterogen atoms			22
no. of solvent atoms			113
R_{factor}[d]/R_{free}			0.256/0.302
rmsd bond length (Å)			0.016
rmsd bond angles (°)			1.6
<B-values> (Å^2)			58.3
PDB entry code			2OGJ

$R_{merge} = \Sigma_j \mid I_h - \langle I_h \rangle \mid / \Sigma I_h$ where $\langle I_h \rangle$ is the average intensity over symmetry equivalents. [b]As defined in SHARP. [c]FOM = figure of merit as defined by SHARP. [d]$R_{factor} = \sum \mid F_{obs} - F_{calc} \mid / \sum \mid F_{obs} \mid$; R_{free} same as R_{factor} but for test set.

Computational Docking: A non-redundant virtual library of compounds from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (*129*) was filtered based on possible amidohydrolase reactions and subsequently prepared as high-energy intermediates (HEI). The HEI compounds are generated by the addition of an activated water to the metabolite molecules from KEGG, which generates approximations of the transition state with variable substrate protonation states (*75, 88*). This library contains ~4,200 metabolites that generated ~22,500 HEI molecules which adopted ~177,700 conformations and ~54,000 configurations. This virtual HEI library of compounds, which compliments the AHS active site better than the ground state molecules, was docked against Atu3266.

The Atu3266 structure was prepared for docking as previously described (*130*). All six chains from the asymmetric unit of Atu3266 (PDB: 2OGJ) were aligned against chain A, and the best rotamer positions of the amino acids were chosen based on the $2f_of_c$ density maps. For example, the position of residues 177, 268 and 267 near the active site were chosen based on chain D positions. Histidine protonation states were manually defined in the vicinity of the two zinc ions, directing polar hydrogens away from the metal center. All other polar hydrogens were automatically generated. The charges on the α- and β-zinc ions were assigned as 1.4 and 1.3 respectively, and the remaining charge difference was distributed to the metal ligating residues (charge: His-77, His-79, His-175, and His-231 (0.25 each); Asp-291 (-0.9); and Kcx-175 (-0.8)) to give each metal an apparent charge of 2.0.

Molecules were docked into the active site of Atu3266 with DOCK3.6 internal release 73. The docking was performed using receptor and ligand bin sizes of 0.4 Å, an overlap of 0.1-0.2 Å, a distance tolerance of 1.5 Å and color matching turned off. The docked molecules were subjected to 250 cycles of rigid-body minimization and were scored based on van der Waals, electrostatic and solvation terms. Each pose of the scored HEI metabolites was filtered based on a maximum distance of 4 Å between the reactive center of the HEI small molecule to the metal center formed by the two zinc atoms of Atu3266. Finally, the top 500 scored molecules were manually inspected to identify potential substrates. Some of the docked substrates were further subjected to minimization with SZYBKI (*131*); this was used to confirm substrate fit and validate hydrogen bond network identification.

Since many molecules were synthesized *de novo* for this project, they were not present in the KEGG library. To investigate the stereospecificity of the compounds hydrolyzed by Atu3266 and Oant2987, a dedicated library containing all of the experimentally tested compounds (**Scheme 2.2;** compounds **1-56**) was generated as previously described (*75, 88*). To further identify new ligands the HEI dedicated library was expanded to contain molecules with the acetylated α-hydroxyl carboxylate motif [O=C(O)C(R_2)OC(=O)C], and also encompassed longer backbones [O=C(O)X_nC(R_2)X_nOC(=O)X_n]. These new compounds were docked and scored in the same manner as the HEI KEGG library molecules.

Network Analysis: All the protein sequences ascribed to COG3964 were obtained from the protein cluster database in NCBI (*83*). The annotation for these sequences is not consistent throughout the whole COG. Although the majority of the proteins seem to be annotated as dihydroorotases, there is significant amount of individual proteins throughout that have been annotated as adenine deaminases. There is one specific group that has been ascribed to two COGs. Group 4 as observed in **Figure 2.1**, has been annotated to COG3964 as well as COG3653. Many of the proteins in COG3653 have been annotated as D-amino acyl deacetylases and D-glutamate deacylases. A D-aminoacylase from group 4 has been previously characterized (*87*). Gox1177 from *Gluconobacter oxydans* (gi|58039630) has been shown to hydrolyze a variety of *N*-acetyl-D-amino acids, with the best rates observed in the deacetylation of *N*-acetyl-D-tryptophan, *N*-acetyl-D-leucine, and *N*-acetyl-D-phenylalanine. Gox1177 was also screened for activity with *N*-acetyl-D/L-phenyl glycine and acetyl-*R*/*S*-mandelate.

A sequence similarity network was constructed using the program Cytoscape (*82*). The connections between the sequences shown in the network were based on permissive and specified BLASTP *E*-values (*10*). COG3964 has been functionally studied at a threshold BLASTP *E*-value of 10^{-70}.

Sequence Alignment: An alignment between the sequences of Atu3266 and Oant2987 from group 6, EF0837 (gi|29375425) from group 2 (which will be discussed in further detail in the succeeding chapter) and those from the characterized binuclear adenine deaminase from *E.coli* (b3665) (gi|16131535) and dihydroorotase also from *E.coli* (b1062) (gi|16129025), was designed using Jalview alignment editor (*132*). These alignments were then manipulated in a text editor in order to align the β-strands that arrange the formation of the TIM-barrel, as well as the residues from these strands that are coordinating to the α- and β- metal ions.

RESULTS

Purification of Atu3266 and Oant2987: Atu3266 was purified to homogeneity and the identity of this protein was confirmed by sequencing the first 8 amino acids from the N-terminus, done by the Protein Chemistry Lab at Texas A&M University. Oant2987 was purified to homogeneity using gel filtration and anion exchange chromatography. Each protein was >95% pure based on SDS-PAGE analysis. ICP-MS confirmed the presence of 2.0 ± 0.2 equivalents of Zn per subunit of Atu3266 and 1.2 ± 0.1 of Zn per subunit of Oant2987 with less than 0.1 equivalents of iron and manganese. RHE_PE00295 was never purified but was still analyzed for metal content. This enzyme proved to have 0.7 ± 0.1 equivalents of Fe per subunit. **Table 2.2** summarizes the metal content of these enzymes. The addition of supplemental zinc to Oant2987 or Rhe_PE00295 did not improve the catalytic hydrolysis of acetyl-*R*-mandelate. The table

also contains metal analysis for the prepared Atu3266 mutants: D291N, R177H and R177A.

Table 2.2: Metal content of enzymes from group 6. Analysis was conducted by inductively coupled plasma emission mass spectroscopy for purified members of group 6 in COG3964. Each quantity represents equivalents of metal per monomer of protein

Enzyme	Zn^{2+}	Fe^{2+}	Mn^{2+}	Ni^{2+}	Cu^{2+}	Total
Atu3266	2.0	<0.1	<0.01	0.4	n/a	2.4
Oant2987	1.2	0.1	0.2	0.2	0.1	1.7
RHE_PE00295	0.2	0.7	0.3	0.2	0.1	1.5
Atu3266-D291N	1.8	0.4	0.1	0.1	n/a	2.3
Atu3266-R177H	1.5	0.15	0.1	<0.1	n/a	1.7
Atu3266-R177A	0.7	0.1	0.1	<0.1	n/a	~1.0

n/a = values were well below the detectable limits

Three-Dimensional Structure of Atu3266: The three-dimensional structure of Atu3266 was determined to a resolution of 2.6 Å (PDB ID: 2OGJ). The six protomers in the asymmetric unit form a dimer of trimers as illustrated in **Figure 2.2**. The hexamer appears as two discs (made of three protomers) stacked against each other. Each protomer consists of two domains; a TIM barrel domain and a second domain consisting of two β-sheets formed by both N- and C-terminal residues as shown in **Figure 2.3**. The structure reveals a distorted $(\beta/\alpha)_8$-TIM barrel fold that is similar to other structurally characterized enzymes from the AHS (*53*). The active site is dominated by a binuclear divalent metal center that is reminiscent of the metal centers found in phosphotriesterase (*51*) and dihydroorotase (*61*) as illustrated in **Figure 2.4**. The structure of Atu3266 was determined in the presence of imidazole in the active site that is coordinated to the β-metal at a distance of 2.2 Å (not shown). This imidazole molecule is only observed in two monomers from the six making the hexameric structure. The α-metal is coordinated by His-77 and His-79 at a distance of 2.2 Å and 2.4 Å, respectively. These residues are positioned at the C-terminal end of β-strand 1. This metal ion is also coordinated to Asp-291 from β-strand **8** and a carboxylated Lys-175 from β-strand **4**. The β-metal is coordinated to the carboxylated lysine from β-strand **4**, and to His-208 and His-231 from β-strands **5** and **6**. A water molecule is at 3.1 Å from the β-metal ion. The two metal ions are 3.3 Å apart.

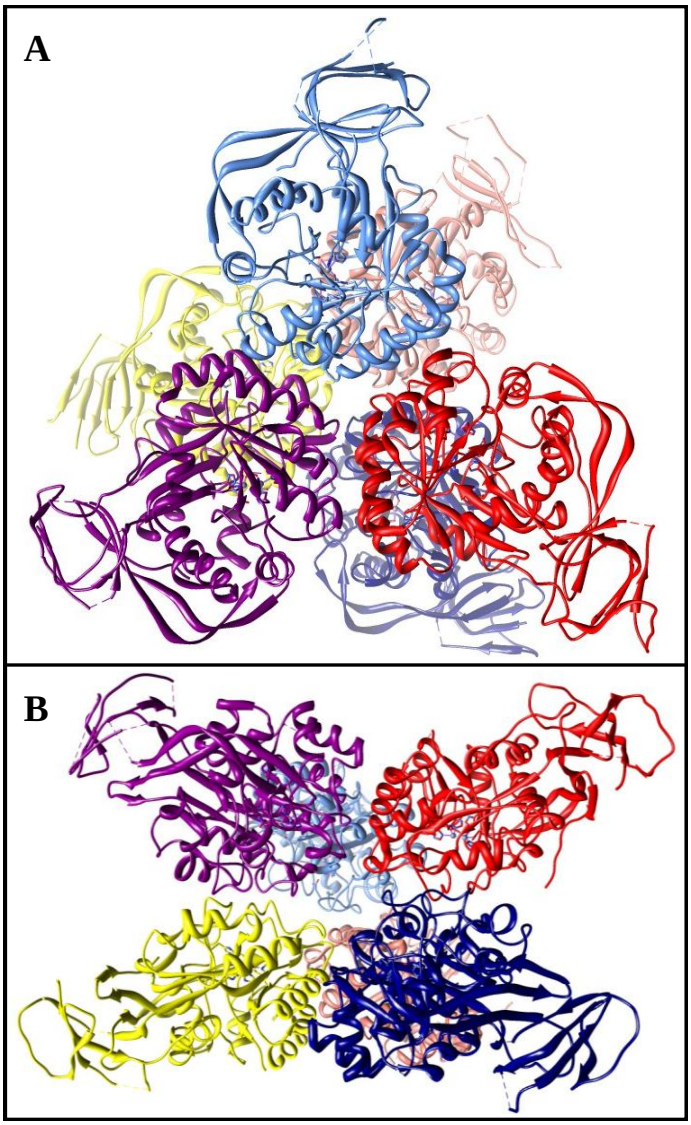


Figure 2.2: Ribbon representation of the hexameric structure of Atu3266. Three dimensional structure shows arrangement of the subunits in face-on (**A**) and edge-on (**B**) views of the hexamer.

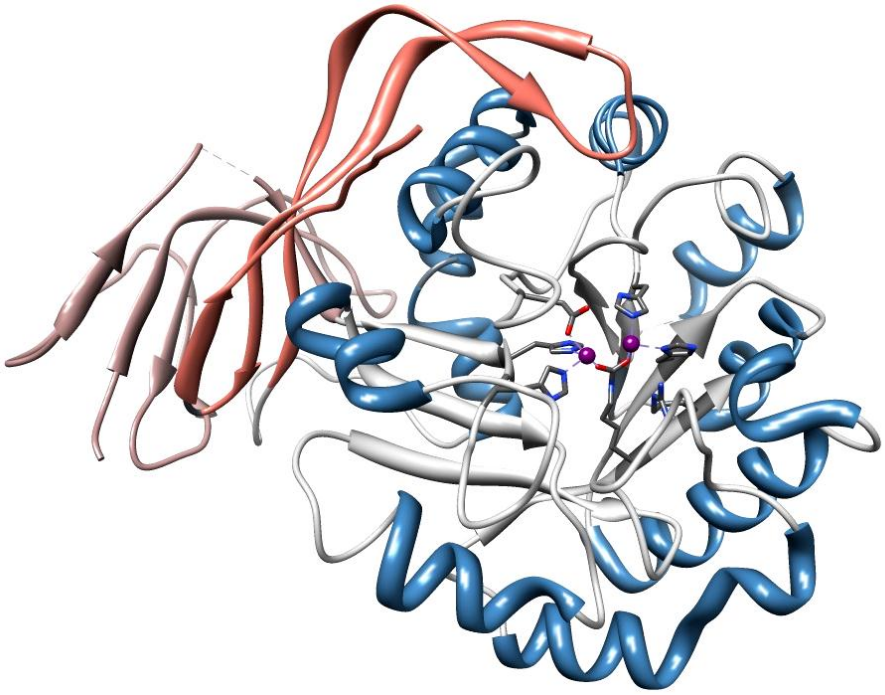


Figure 2.3: Ribbon representation of the monomeric structure of Atu3266. The metal center is depicted in purple; the metal coordinated ligands are in dark gray. The central β-barrel is shown in light gray; the helices are shown in blue. In rose and tan color are the N- and C- terminal domains of the enzyme, respectively.

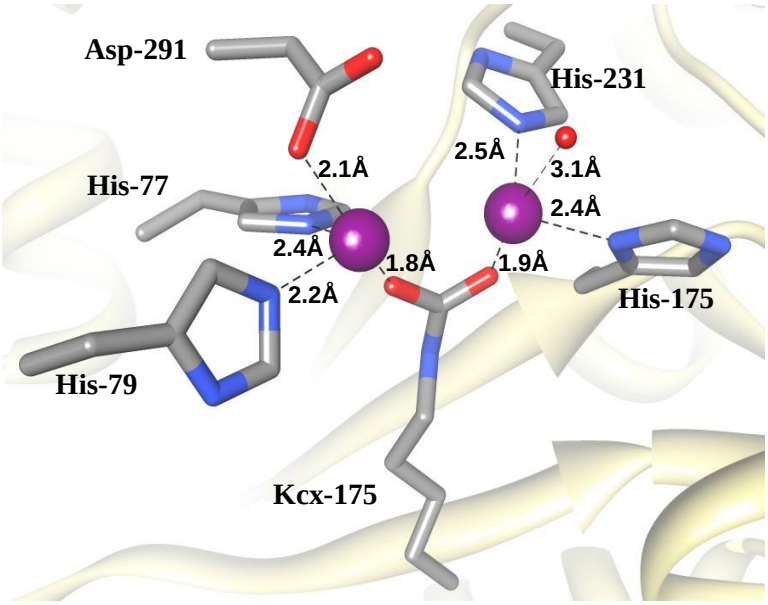


Figure 2.4: Active site of Atu3266. Model depicts the residues that coordinate the binuclear metal center. The enzyme ligands are shown in gray, the binuclear metal center is shown in magenta, the coordinating water molecule to the β-metal is shown in red.

Substrate Specificity of Atu3266 and Oant2987: D- and L-Dihydroorotate (**56**) were the first two compounds to be tested as substrates for Atu3266, since this enzyme has been annotated in various databases as a dihydroorotase. However, no hydrolysis of these compounds could be observed and thus this annotation is not correct. This experiment was followed by the utilization of a focused library of *N*-acetyl, *N*-succinyl and *N*-formyl derivatives of the common D- and L- amino acids and multiple libraries of L-Xaa-L-Xaa, L-Xaa-D-Xaa, and D-Xaa-L-Xaa dipeptides. From the more than 1200 compounds tested in the initial screening only two compounds were found with very weak rates of hydrolysis. The values of k_{cat}/K_m for *N*-acetyl-D-serine (**1**) and *N*-acetyl-D-threonine (**2**) are 4.0 M^{-1} s^{-1} and 2.0 M^{-1} s^{-1} respectively. This finding prompted the separate synthesis of *N*-acetyl-D-cysteine (**3**) and *N*-formyl-D-serine (**4**), but these compounds were not detectable substrates for Atu3266. Methylating the α-carboxylate (compound **5**) or phosphorylating the side chain hydroxyl (compound **6**) of the weak substrate, *N*-acetyl-D-serine, abolished the catalytic activity. The *N*-carbamoyl (compound **7**) and *N*-phosphoryl (compound **8**) derivatives of D-serine were not catalytically active. However, the substitution of the α-amino group with an α-hydroxyl group, as in 2-acetyl-*R*-glycerate (**9**) resulted in a two order of magnitude improvement in k_{cat}/K_m, relative to *N*-acetyl-D-serine.

The dramatic increase in the rate of hydrolysis of 2-acetyl-*R*-glycerate (**9**) prompted the exploration of modifications to this sub-structure. However, 2-phospho-D-glycerate (**10**) and 3-phospho-2-acetyl-D-glycerate (**11**) were not hydrolyzed and no activity could be observed with 2-acetyl-glycerol (**12**). Surprisingly, excision of the

hydroxymethyl substituent to form acetyl glycolate (**13**) increased k_{cat}/K_m by a factor of 27, relative to compound **9**. Further substitutions to the sub-structure of acetyl glycolate (**13**) did not result in any further improvements (compounds **14** to **28**) in catalytic activity except for the addition of a phenyl group. Acetyl-*R*-mandelate (**28**) is hydrolyzed with a k_{cat}/K_m of 2.8 x 10^5 M^{-1} s^{-1}. This is a factor of 20 better than acetyl glycolate (**13**) and nearly 5 orders of magnitude better than the initial hit, *N*-acetyl-D-serine (**1**). No activity was observed with phospho(enol)pyruvate (**30**), the amide of acetyl-*R*-mandelate (**31**), 2-phospho-*R*-mandelate (**32**), acetyl-D-phenyl glycine (**33**), the methyl ester of acetyl-*R*-mandelate (**34**), acetyl-*S*-mandelate (**35**), and 2-acetoxy-isobutyric acid (**29**). Further modifications to acetyl-*R*-mandelate did not improve the rate of hydrolysis (see compounds **36** to **50**), relative to acetyl-*R*-mandelate (**28**). Other modifications, including *p*-nitrophenyl acetate (**51**), aspirin (**52**) and various lactones (**53-55**) were not active. The kinetic constants for the catalytically active substrates are provided in **Table 2.3**.

Table 2.3: Kinetic parameters for Atu3266 and Oant2987. Constants relate to compounds in **scheme 2.2**

	Atu3266			Oant2987		
Substrate	k_{cat} (s^{-1})	K_m (mM)	k_{cat} / K_m ($M^{-1}s^{-1}$)	k_{cat} (s^{-1})	K_m (mM)	k_{cat} / K_m ($M^{-1}s^{-1}$)
1	n.a.	n.a.	4	n.c.	n.c.	n.c.
2	n.a.	n.a.	2	n.c.	n.c.	n.c.
3	<0.001	n.a.	n.a.	<0.001	n.a.	n.a.
9	12 ± 1	25 ± 4	(4.8 ± 0.1) x 10^2	0.6 ± 0.1	2.1 ± 0.5	(3 ± 1) x 10^2
13	40 ± 3	3.1 ± 0.5	(1.3 ± 0.2) x 10^4	4.2 ± 0.2	1.1 ± 0.2	(4.2 ± 0.8) x 10^3
14	0.8 ± 0.1	0.5 ± 0.1	(1.6 ± 0.2) x 10^3	0.4 ± 0.1	0.6 ± 0.1	(7 ± 1) x 10^2
15	0.5 ± 0.1	10 ± 2	50 ± 10	n.a.	n.a	40
16	0.6 ± 0.1	9.5 ± 0.7	58 ± 6	n.a	n.a.	60
17	1.5 ± 0.1	7 ± 1	(2.1 ± 0.3) x 10^2	n.a.	n.a.	200
18	3.1 ± 0.3	6 ± 1	(5.0 ± 0.8) x 10^2	0.81 ± 0.08	1.2 ± 0.1	(8.1 ± 0.8) x 10^3
19	1.1 ± 0.2	9.5 ± 3.0	(1.2 ± 0.4) x 10^2	0.23 ± 0.02	1.4 ± 0.1	(2.2 ± 0.2) x 10^2
20	0.41 ± 0.04	2.1± 0.5	(2.0 ± 0.5) x 10^2	0.30 ± 0.02	0.71 ± 0.02	(4.3 ± 0.4) x 10^2
21	n.a.	n.a.	(1.1 ± 0.6) x 10^2	n.a.	n.a.	30
22	n.a.	n.a.	(1.8 ± 0.6) x 10^2	n.a.	n.a.	30
23	n.a.	n.a.	50 ± 10	n.c.	n.c.	n.c.
24	n.a.	n.a.	(2.4 ± 0.5) x 10^2	2.0 ± 0.1	2.0 ± 0.5	(1.0 ± 0.2) x 10^3
25	0.10 ± 0.03	0.4 ± 0.1	(2.5 ± 0.2) x 10^2	<0.05	n.a.	n.a.
26	0.21 ± 0.01	0.3 ± 0.1	(6.6 ± 1.0) x 10^2	<0.02	n.a.	n.a.
27	2.3 ± 0.2	30 ± 10	67 ± 22	n.a.	n.a.	40
28	280 ± 5	1.0 ±0.1	(2.8 ± 0.2) x 10^5	130 ± 1	0.7 ± 0.1	(1.8 ± 0.2) x 10^5
36	34 ± 2	1.1 ± 0.1	(3.4 ± 0.3) x 10^4	6.0 ± 0.2	0.21 ± 0.02	(3.0 ± 0.3) x 10^4
37	70 ± 7	7 ± 1	(1.0 ± 0.1) x 10^4	60 ± 4	3.4 ± 0.5	(1.7 ± 0.2) x 10^4
38	43 ± 5	4 ± 1	(1.1 ± 0.2) x 10^4	4.1 ± 0.3	0.5 ± 0.1	(8.0 ± 0.8) x 10^3
39	30 ± 9	10 ± 3	(3.0 ± 0.9) x 10^3	60 ± 6	18 ± 2	(3.3 ± 0.3) x 10^3
40	n.a	n.a	(3.0 ± 0.3) x 10^3	50 ± 5	5 ± 1	(1.0 ± 0.2) x 10^4
41	n.a	n.a	(5.0 ± 0.4) x 10^3	36 ± 5	8 ± 1	(4.5 ± 0.5) x 10^3
42	190 ± 20	12 ± 3	(1.6 ± 0.1) x 10^4	150 ± 10	3.0 ± 0.5	(5 ± 1) x 10^4
43	n.a	n.a	(2.2 ± 0.2) x 10^3	30 ± 2	4.0 ± 0.5	(7.5 ± 0.9) x 10^3
44	n.a	n.a	(1.0 ± 0.1) x 10^3	50 ± 1	5 ± 1	(1.0 ± 0.2) x 10^4
45	10 ± 1	2.0 ± 0.2	(5.0 ± 0.5) x 10^3	40 ± 5	11 ± 3	(3.6 ± 0.9) x 10^3
46	7.0 ± 0.2	0.6 ± 0.1	(1.0 ± 0.1) x 10^4	10 ± 1	0.8 ± 0.2	(1.2 ± 0.3) x 10^4
47	10 ± 1	1.4 ± 0.3	(7.0 ± 0.2) x 10^3	20 ± 2	2.0 ± 0.5	(1.0 ± 0.2) x 10^4
48	50 ± 4	10 ± 1	(5.1 ± 0.1) x 10^3	60 ± 6	9 ± 1	(7.0 ± 0.7) x 10^3
49	90 ± 9	7 ± 1	(1.3 ± 0.1) x 10^4	80 ± 6	3 ± 0.2	(2.7 ± 0.2) x 10^4
50	17 ± 3	5 ± 1	(3.5 ± 0.1) x 10^3	28 ± 10	n.a.	(1 ± 0.3) x 10^3

n.c. = denotes parameter that was not calculated, mainly because there was no detection of acetic acid. n.a. = denotes parameter that was not available, this was under the event that the substrate did not saturate the enzyme and the rates obtained were fit to a linear curve to obtain a *V*/*K* for the observed activity.

Activity by RHE_PE00295: The EFI provided enzyme RHE_PE00295 was assayed with only four of the multitude of *N*-acetyl-D-amino acids and α-acetyl carboxylates. These included: *N*-acetyl-D-serine, 2-acetyl-*R*-glycerate, acetyl glycolate and acetyl-*R*-mandelate. The rates are comparable to those observed with Atu3266 and Oant2987 with *N*-acetyl-D-serine, 2-acetyl-*R*-glycerate and acetyl glycolate. The rate was an order of magnitude lower for acetyl-*R*-mandelate. RHE_PE00295 shares 85% sequence identity to Atu3266. Additional enzymes in group 6 of this COG are expected to have the same rate constants in the hydrolysis of compounds that have shown activity with the purified enzymes discussed here. **Figure 2.5** shows the logarithmic graph for activities of Atu3266, Oant2987 and RHE00295 with *N*-acetyl-D-serine, 2-acetyl-*R*-glycerate, acetyl-glycolate and acetyl-*R*-mandelate.

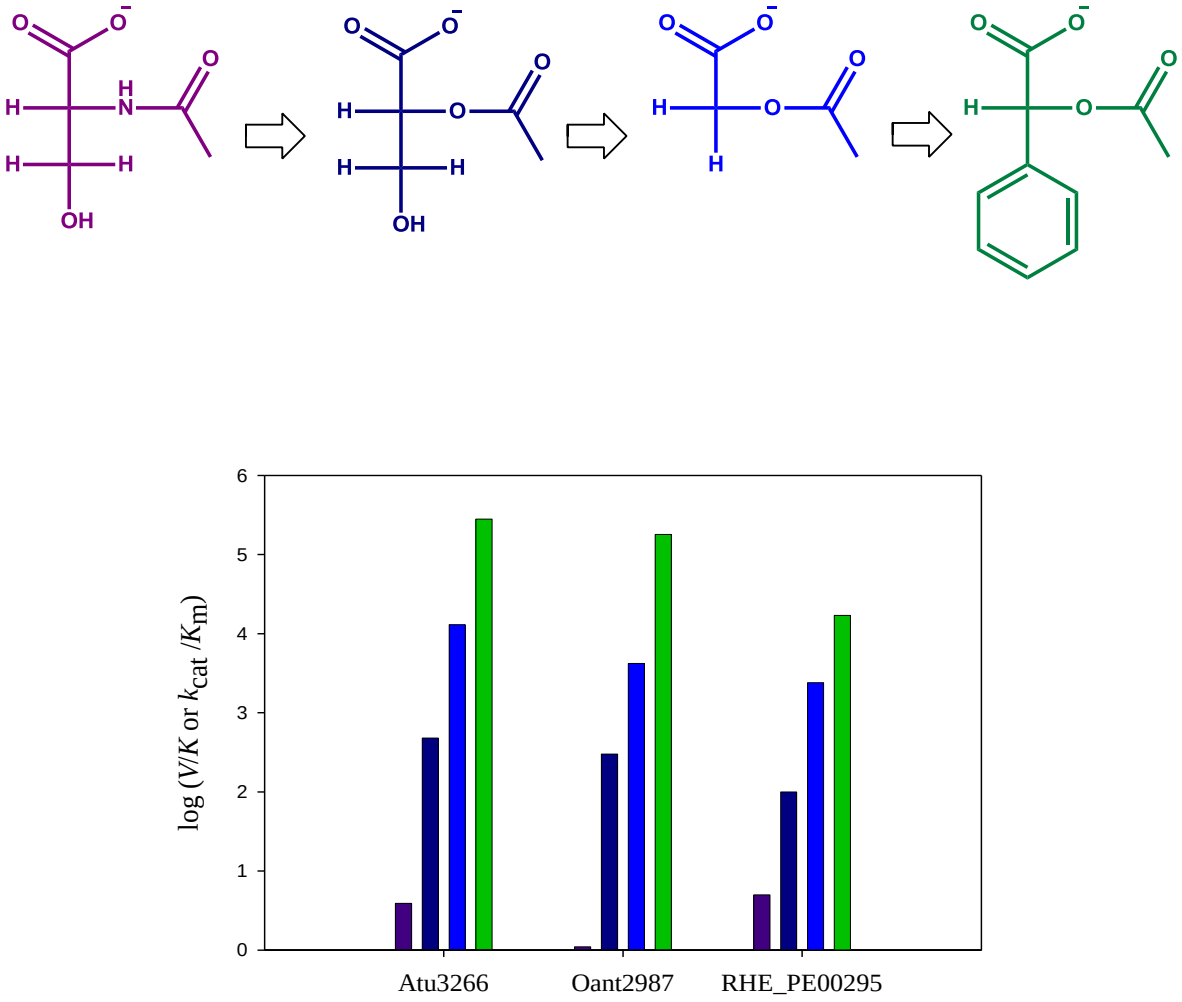


Figure 2.5: Comparison of rates in hydrolysis of selected compounds. Graph details the rate increase in each enzyme (Atu3266, Oant2987 and RHE_PE00295) as substrates were modified beginning with *N*-acetyl-D-serine, then 2-acetyl-*R*-glycerate, acetyl-glycolate and finally, the best characterized substrate acetyl-*R*-mandelate.

Inhibition by N-methylphosphonate-D-phenyl glycine: The *N*-methylphosphonate and *O*-methylphosphonate derivatives of D-phenyl glycine (**57**) and *R*-mandelate (**56**), respectively, were synthesized and tested as inhibitors for the hydrolysis of acetyl-*R*-mandelate (**28**) by Atu3266 and Oant2987. The *O*-methylphosphonate inhibitor of *R*-mandelate was unstable. However, the *N*-methylphosphonate analog of D-phenyl glycine (**57**) was found to be a competitive inhibitor in the deacetylation of acetyl-*R*-mandelate. A fit of the data to equation 2 gave an inhibition constant of 35 ± 2 μM for Atu3266 and 40 ± 2 μM for Oant2987. **Figure 2.6** shows the inhibition curves for both enzymes.

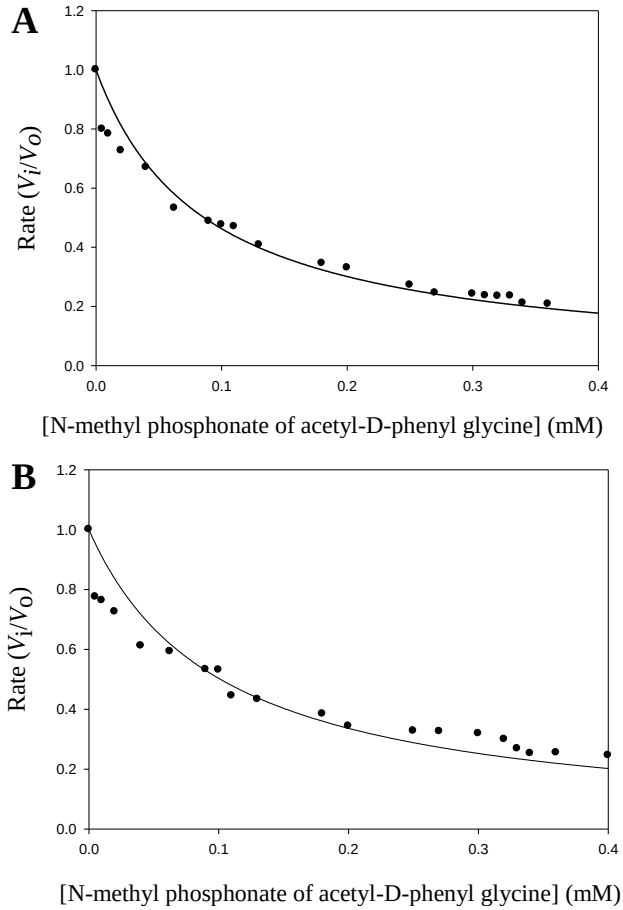


Figure 2.6: Inhibition curves for activity of Atu3266 **(A)** and Oant2987 **(B).** Assays were carried out in the presence of *N*-methyl phosphonate of acetyl-D-phenyl glycine.

Mutational Analysis of Atu3266: Only three mutants were constructed to probe particular amino acids in the vicinity of the active site. In-depth site-directed mutagenesis studies were not carried out because of the lack of a bona fide substrate. Mutation of D291N abolished hydrolytic activity for Atu3266 with acetyl-*R*-mandelate. The side chain of this residue is implicated in the shuttling of the proton from the bridging hydroxide molecule to the leaving group (*53, 63, 85*), in this scenario it would be the α-hydroxylate from the product molecule which would be mandelic acid. **Table 2.4** shows the catalytic constants for the mutation of D291N as well as R177H, and R177A. The mutation of arginine-177 to histidine showed a decreased activity in the hydrolysis of acetyl-*R*-mandelate compared to the wild-type that was close to four-orders of magnitude lower, while the activity of the R177A mutant is nearly the rate observed of the hydrolysis of the WT enzyme for *N*-acetyl-D-serine. The thoroughness of site-directed mutagenesis studies was not developed because of the ambiguity in substrate specificity and determining what constitutes a true substrate with these enzymes, but arg-177 is observed to have some role in the hydrolysis of the best found substrate for this group of enzymes.

Table 2.4: Catalytic rate constants for selected Atu3266 variants.

Enzyme	k_{cat} (s^{-1})	K_m (mM)	k_{cat} / K_m ($M^{-1}s^{-1}$)	Difference
Atu3266	280	1.0	2.8×10^5	0
D291N	<0.001	n.d.	n.d.	n.d
R177H	-	-	20	-1.4×10^4
R177A	-	-	3.2	-8.8×10^4

n.d. = not determined.

Computational Docking to the Active Site of Atu3266: Computational docking was initiated using the three-dimensional crystal structure of Atu3266. The scored molecules obtained from docking of the full HEI KEGG library were filtered, based on distance constraints to the catalytic metal center. This provided a top-500 list from approximately 22,500 unique docked HEI molecules, which resulted in 133,986 scored molecules that were then filtered, based on distance constraints that kept 43.2% of molecules that had energy scores from -183.18 to -126.02. Molecules that had high-quality poses in this range included carbamylated and acetylated amino acids and thus docking was enriched for molecules such as *N*-acetyl-L-lysine (ranked #40), *N*-acetyl-D-methionine (ranked #171), *N*-acetyl-D-cysteine (**3**) (ranked #218), *N*-acetyl-L-leucine (ranked #313), and *N*-acetyl-D-phenylalanine (**48**) (ranked #440). These docking results are shown in **Figure 2.7**.

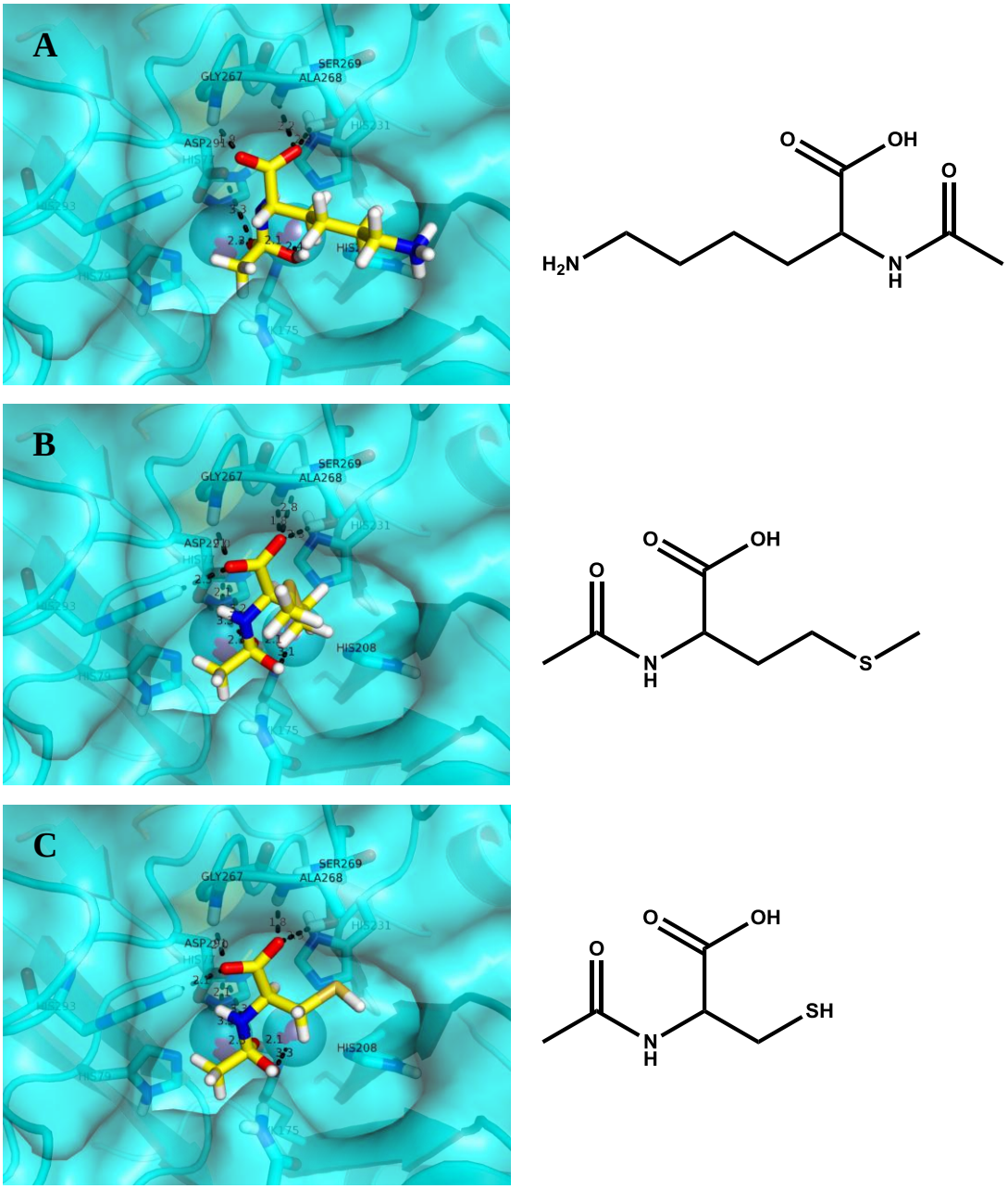


Figure 2.7: Docking results and models of *N*-acetyl-D-/L-amino acids. Five well-posed acylated amino acids were found in the top 500 compounds of the docking list, from the 10 modified amino acids present in the KEGG HEI library: (**A**) L-lysine is ranked 40th, (**B**) D-methionine is ranked 171st, (**C**) D-cysteine is ranked 218th, (**D**) L-leucine is ranked 313th and (**E**) D-phenylalanine is ranked 440th. These five acetylated amino acids are placed in the 98th percentile of the all the KEGG metabolites that can undergo an amidohydrolase reaction.

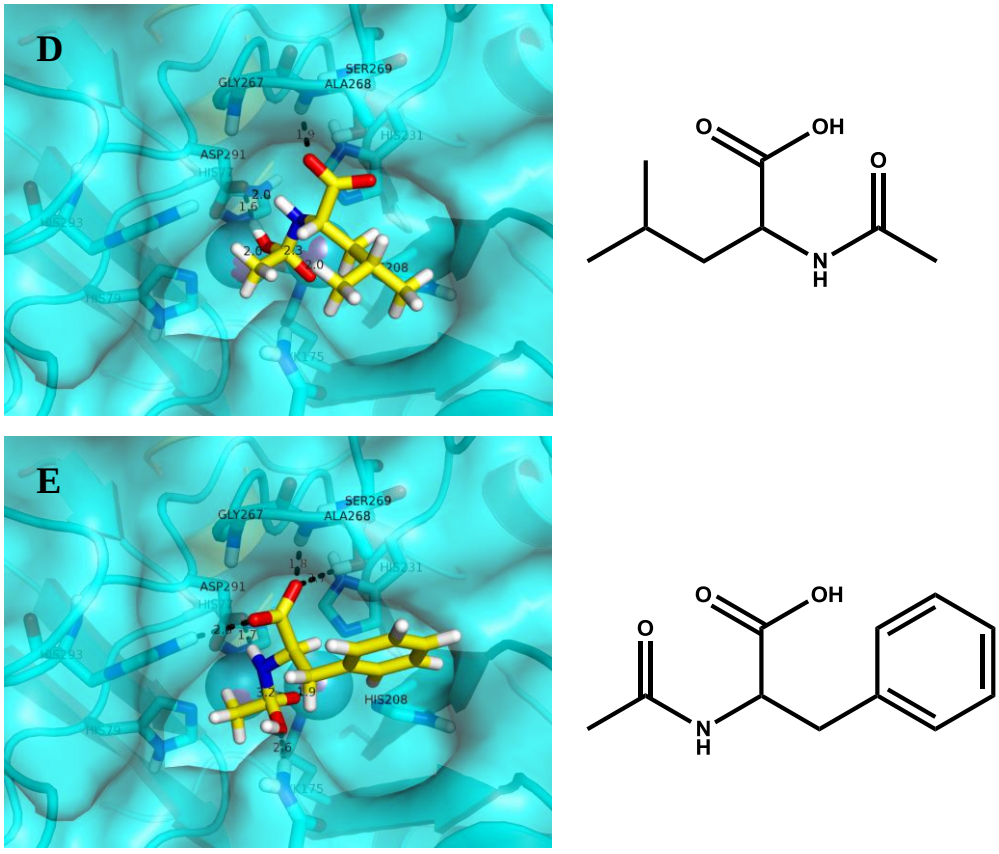


Figure 2.7 continued.

The dedicated HEI library docking experiments provided insights into a common binding pose for the compounds that contain the HEI acetylated α-hydroxyl carboxylates. The apparent competent pose for these molecules clearly illustrates the transition that could occur between the ground state molecule (**Figure 2.8-A)** and the transition state mimic for 4-methyl acetyl-*R*-mandelate **(Figure 2.8-B**). A detailed examination revealed that the competent pose of the acetylated α-hydroxyl carboxylates directs the ester group towards the metal cluster, where the carbonyl oxygen of the acetyl portion to the molecule is positioned over the α- and β-zinc atoms. In this orientation, the methyl group of the acetyl moiety is surrounded by the hydrophobic pocket lined by Ile-87, Leu-140, Cys-142, and the imidazole side chain of His-79 and His-293. This hydrophobic pocket is large enough that it can accommodate an ethyl group. This pose allows the bridging oxygen of the ester moiety to orient towards Asp-291, thus facilitating the enzymatic hydrolysis of the acetylated α-hydroxyl carboxylates.

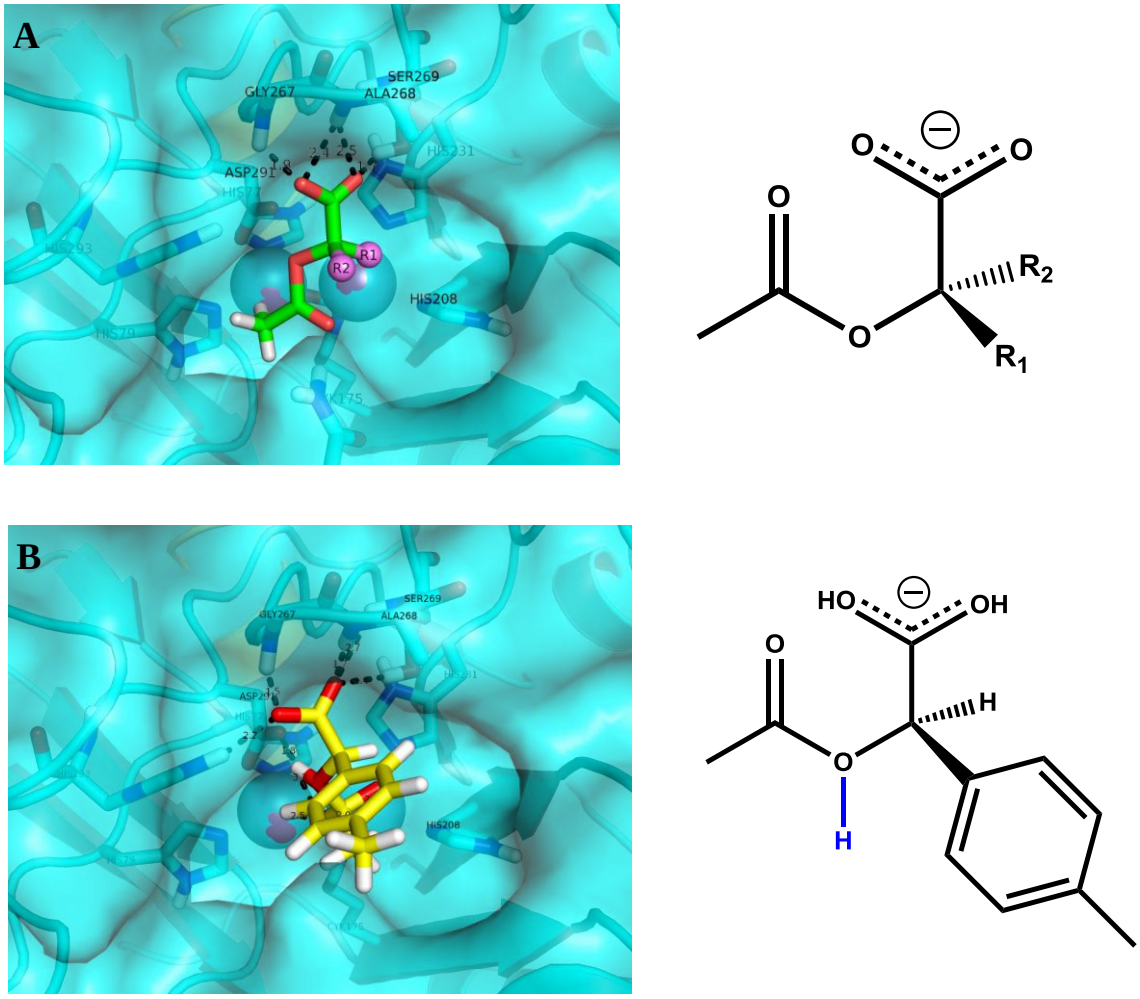


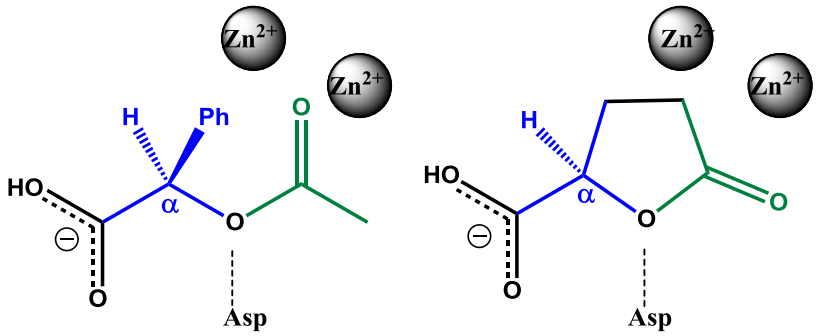
Figure 2.8: Docking models of interactions of ground state and HEI molecules. (**A**) Interactions of a ground state generic α-acetyl carboxylate, where the carboxyl moiety interacts with β-loop 7 of Atu3266. (**B**) Interaction of a transition state mimic of 4-methyl acetyl-*R*-mandelate.

The docking calculations indicate that the two remaining fragments of the molecule, the carboxylate and the mostly nonpolar substituent, are necessary for the correct binding and orientation of the substrates in the active site. Critical to the binding is the carboxylate, which forms a hydrogen bond network with the loop formed between β-strand **7** and α-helix **7** of the AHS domain. These hydrogen bond interactions include the amide nitrogens of Gly-267 and Ala-268, and the hydroxyl side chain of Ser-269. In some instances, a hydrogen bond is additionally formed between the carboxylate of the substrate and the Nε hydrogen of His-293.

The second and more variable portion of the substrates for Oant2987, Atu3266 and RHE_PE00295, contain mostly nonpolar substitutions at the chiral α-carbon of the acetylated α-hydroxyl carboxylates. All molecules identified as substrates possess the *R*-stereochemistry at C2 of the acetylated α-hydroxyl carboxylates, including the best substrate acetyl-*R*-mandelate (**28**). For this substrate, the top-placed conformer identified by docking has an energy score of -145.79. Therefore, compared to the HEI KEGG library, this molecule would have ranked #143 placing it in the 99.3^{th} percentile of the molecules downloaded from KEGG. In this optimal pose, the aromatic phenyl ring moiety of acetyl-*R*-mandelate (**28**) provides *pi*-stacking interactions with the side chain nitrogen atom of Asn-144. The distances between these two components vary depending on the substrates docked; however, in the catalytically competent poses these molecules adopt conformations where the center of the ring is 3.1-3.5 Å away from the amino acid side chain.

The docking experiments of the dedicated HEI library tested eliminated 5-member ring lactones as substrates; however, the docking does not discriminate between 5-membered cyclic acid sugar lactones. These molecules are found in the top-500 list of the docked HEI molecules but are not turned over by the enzyme and include molecules **54** and **55**. Even though these small acid sugars fit into the active site, they adopt many different catalytically non-competent poses. The two major orientations observed include a pose where the cyclic oxygen is pointing straight down into the bottom of the pocket, which would require a 90^{o} rotation of the ring between the ground state binding mode and the HEI molecule docked state, or where the cyclic oxygen is too far away from Asp-291 to facilitate catalysis. Nonetheless, even in a pose that most resembles the catalytically competent substrate chemotype of the acetylated α-hydroxyl carboxylate (**Scheme 2.4**), the cyclic acid sugar lactones have non-superimposable rotamer orientations to the docked pose of the best substrate acetyl-*R*-mandelate (**28**). This indicates that it is unlikely that both compounds would be turned over by the same enzyme. The first difference is observed in the orientation at the α-carbon (**Scheme 2.4**) of the acetylated α-hydroxyl carboxylate substrates, as these molecules adopt the *R*-configuration while the sugar acid lactones are in the *S*-configuration. The second discrepancy in the positions of the compounds is at the acetyl portion of the acetylated α-hydroxyl carboxylate backbone (**Scheme 2.4**), which is flipped between the two compound classes. As such, the acid sugar lactones adopt an orientation where the carbon ring of the molecule is placed close to the metal ions and the carbonyl oxygen points towards the hydrophobic pocket created by Ile-87, Leu-140, Cys-142, His-79 and

His-293. Consequently, these modifications in the orientation of the functional groups create an environment that would not promote or accelerate the hydrolysis of the lactone ring.



Scheme 2.4: Schematics for the difference in rotamer flexibility between α-acylated carboxylates that are substrates to group 6 enzymes and a rotamer with *R*-configuration found in lactones used to test for activity.

DISCUSSIONS

Three-Dimensional Structure of Atu3266: The three-dimensional structure of Atu3266 shows that this enzyme possesses a binuclear metal center, reminiscent of other enzymes in the amidohydrolase superfamily such as dihydroorotase, phosphotriesterase and urease (*50-53, 61*). The buried α-metal is coordinated by His-77, His-79, Asp-291 and the carboxylated Lys-175. The more solvent exposed β-metal is coordinated to His-208, His-231, the carboxylated Lys-175 and a water molecule. The individual subunits associate to form a hexameric oligomer.

COG3964: There are 24 clusters of orthologous groups in the AHS and COG3964 is one of the smallest, with about 200 sequences (**Figure 2.1**). The most common annotation in NCBI for this cluster of enzymes is dihydroorotase, and in some cases, adenine deaminase. The basis for this annotation is not clear, but the experiments reported here demonstrate that this annotation is incorrect. When the *E*-value cutoff is set to 10^{-70}, the sequence similarity network for COG3964 segregates into 8 groups with 3 or more sequences per group and Atu3266, Oant2987 and RHE_00295 belong to Group **6** of COG3964. An amino acid sequence comparison between those enzymes in Group **6** of COG3964 and the structurally characterized dihydroorotases and adenine deaminases **(Figure 2.9)** demonstrates that the key residues that have been implicated in substrate binding and recognition in these enzymes are not present in Atu3266, Oant2987 or other enzymes in their respective group. This is supportive that the functional annotation of COG3964 needs to be revised, as this new assemblage of proteins does not have the necessary elements observed previously to carry out the function they have been assigned.

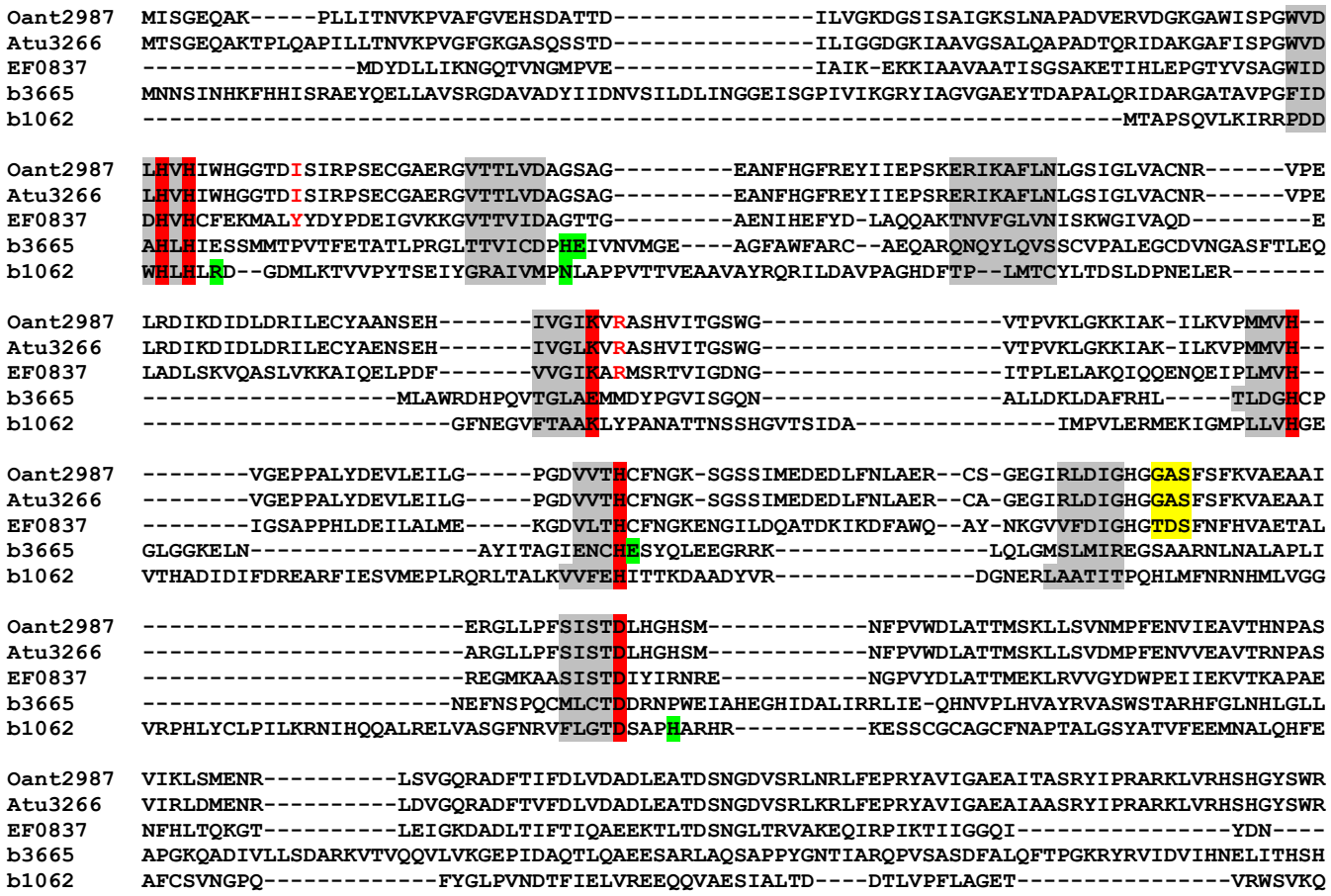


Figure 2.9: Sequence alignment of selected group 6 enzymes (Atu3266 and Oant2987) from COG3964. Alignment includes sequence from group 2 (EF0837). The sequences from characterized adenine deaminase-ADE (b3665) and dihydroorotase-DHO (b1062) are also added. Highlighted in gray are residues incorporated in the β-strands forming the $(\beta/\alpha)_8$-TIM barrel. In red are the metal coordinating ligands. Highlighted in green are substrate recognizing residues in *E. coli* ADE and DHO. In yellow are the residues in Oant2987 and Atu3266 that coordinate to the α-carboxylate of acetyl-*R*-mandelate, and in red font are residues within active site distance of COG3964 proteins: In Atu3266, R-177 conserved in sequences in groups 2 and 6 and I-87 only conserved in group 6 sequences.

Determination of Substrate Profile: The first indication of catalytic activity by Atu3266 was observed in a library of *N*-acetyl-D-amino acids. After each component of the library was tested independently, it was observed that *N*-acetyl-D-serine (4 M^{-1} s^{-1}) and *N*-acetyl-D-threonine (2 $M^{-1}s^{-1}$) were the only compounds in the *N*-acetyl-D-Xaa library to be hydrolyzed. Screening for hydrolytic activity was also conducted using a comprehensive library of dipeptides. However, none of the dipeptide libraries showed detectable catalytic activity. Subsequently we focused on the activity obtained from *N*-acetyl-D-serine, and utilized various modifications including changes in stereochemistry, side chain group functionality and carboxylate substituents, but none of these changes generated a better substrate. However, when the amide linkage was changed to an ester, a 100-fold increase in activity was observed. Atu3266 hydrolyzes the ester linkage of acetyl-*R*-glycerate with a rate constant of 400 $M^{-1}s^{-1}$. When the hydroxymethyl group was removed and replaced with hydrogen (acetyl glycolate) the rate of hydrolysis increased by an additional two orders of magnitude (1.3 x 10^4 M^{-1} s^{-1}). The highest activity was achieved in the deacetylation of acetyl-*R*-mandelate with a rate constant of 2.8 x 10^5 M^{-1} s^{-1} by Atu3266, 1.8 x 10^5 $M^{-1}s^{-1}$ by Oant2987 and 2.1 x 10^4 $M^{-1}s^{-1}$ by RHE_PE00295.

Additional modifications to the acetyl-*R*-mandelate substructure failed to improve the catalytic activity. Substitutions and modifications to the C1 carboxylate eliminated deacetylation activity. These alterations included methyl ester and amide formation, phosphorylation, and reduction to a hydroxy methyl group. At C2 the most beneficial modification was the change of the amide nitrogen to oxygen to form an ester.

Modification of the acetyl group to a formyl, succinyl or carbamoyl did not prove to be a better substituent than the acetyl. The stereochemistry at C2 must be in the *R*-configuration. The acetyl group could be extended to a propionyl group but the rate of hydrolysis is reduced. The presence of the phenyl group at the side chain position proved to be the best substitution, but the addition of functionality to the aromatic ring decreased the rate of hydrolysis.

Comparisons to Other N-acetyl D-Amino Acid Deacetylases: Other enzymes within the amidohydrolase superfamily have been shown to catalyze the hydrolysis of *N*-acetyl-D-amino acid derivatives. For example, enzymes within COG3653 have been shown to be *N*-acetyl-D-amino acid deacetylases (*87*). In COG3653 the binuclear metal center is not bridged by a carboxylated lysine residue and the α-carboxylate group of the substrate is ion paired and hydrogen bonded to conserved arginine and tyrosine residues (*87*). These motifs are not conserved in the enzymes of COG3964, where instead, the carboxylate group of the substrate is recognized by the backbone residues following β-strand **7**. Gox1177 from *Gluconobacter oxidans* carries out the hydrolysis of acetyl *R*-mandelate and *N*-acetyl *D*-phenyl glycine a rate of $1.5 \times 10^4\ M^{-1}s^{-1}$ and $1.4 \times 10^4\ M^{-1}s^{-1}$ respectively. This COG3653 enzyme is also part of COG3964. **Figure 2.10** details the location of the enzyme in respect to the enzymes characterized here. Gox1177 best hydrolyzes *N*-acetyl-D-tryptophan and *N*-acetyl-D-leucine at the rates of $4.1 \times 10^4\ M^{-1}s^{-1}$ and $3.2 \times 10^4\ M^{-1}s^{-1}$ respectively (*87*).

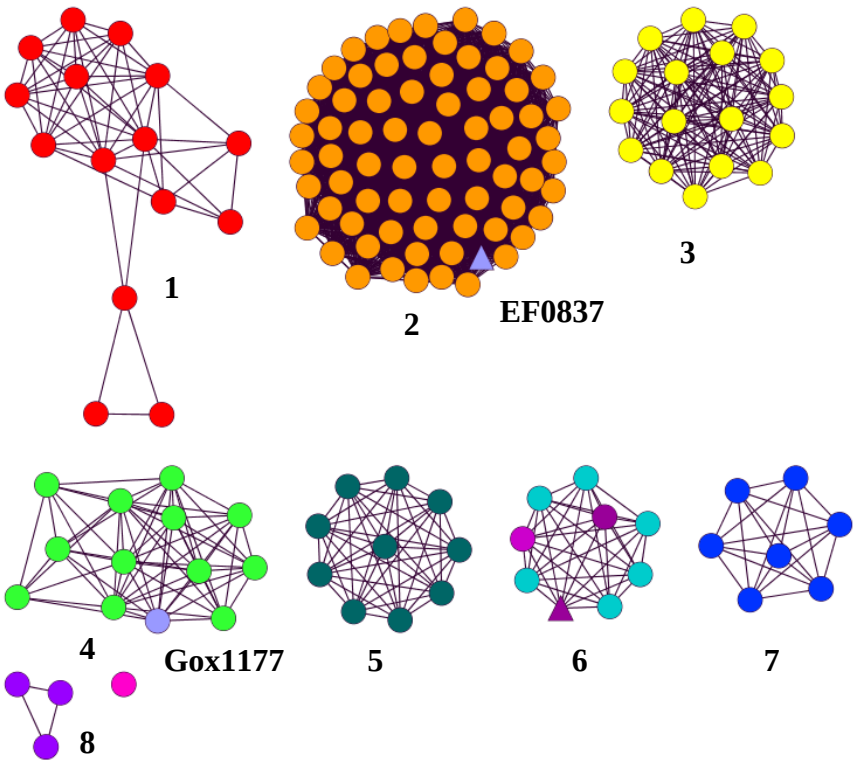


Figure 2.10: Sequence similarity network highlighting Gox1177 (group 4) and EF0837 (group 2). The nodes indicating these sequences are shown in light purple.

At this point it is not clear as to whether the enzymes identified in this investigation from COG3964 are generic esterases, or whether there is a more specific substrate. Acetyl-*R*-mandelate has apparently not been identified as a physiological metabolite in bacteria. In those organisms that have been found to use mandelate as a carbon source, there is no indication of a requirement of a deacetylase for the hydrolysis of acetyl-*R*-mandelate (*99-101*).

Computational Docking for Identification of Substrate-Enzyme Interactions: Docking of the HEI KEGG virtual metabolite library to the Atu3266 structure gave a top-500 list that satisfied the distance constraints from the catalytic center. This list indicated that one of the highest scoring and well-posed molecules is *N*-acetyl-D-cysteine that is not turned over by the enzyme at the limit of detection. Upon further analysis, a chemotype emerged from docking, as the list encompasses well-posed carbamylated or acetylated amino acids. These included derivatives of hydrophobic amino acids such as *N*-acetyl-D-methionine, *N*-acetyl-L-leucine and *N*-acetyl-D-phenylalanine and other metabolites such as *N*-acetyl-L-lysine, *N*-carbamoyl-L-aspartate, *N*-acetyl-D-glucosaminate and oxalureate. A closer examination of the KEGG database revealed that the database only included 10 acetylated amino acids (Met, Cys, Phe, Leu, Lys, His, Trp, Asp, Glu, Gln). This indicated a high degree of enrichment of the docking top-500 list for the acetamido(carboxy)methyl backbone from the overall pool of HEI KEGG compounds. Thus, the initial docking experiments did not include the two catalytically competent acetylated amino acids, *N*-acetyl-D-serine and *N*-acetyl-D-threonine. However, they were subsequently docked in the dedicated HEI library and obtained scores that would respectively rank them #321 and #265 in the top-500 list of the HEI KEGG library. The poses of the two amino acids are suboptimal as they either leave the side chain hydroxyl unsatisfied or allow it to partially interact with the carboxylate binding loop. The pattern of containing the *N*-acetylated amino acid backbone is carried over to the dedicated HEI library list where the chemotype is found

to accommodate many different functional groups that pose well in the active site, as in *N*-acetyl-D-phenyl glycine (**33**) shown in **Figure 2.11**.

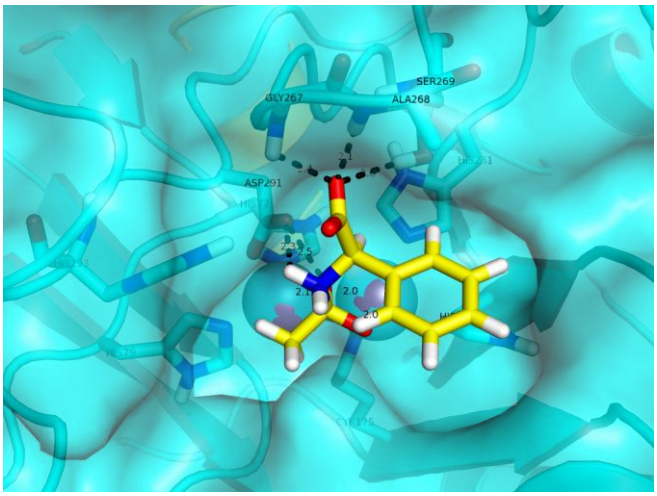


Figure 2.11: Docking model of *N*-acetyl-D-phenyl glycine in active site of Atu3266.

Based on analysis of the focused HEI library docking results, molecules that are turned over by the enzyme contain an acetylated α-hydroxyl carboxylate backbone that is hydrolyzed by the enzyme to release an acetyl group. This acetyl moiety is positioned over the zinc atoms, thus directing the methyl group into a small hydrophobic pocket that is formed by residues Ile-87, Leu-140, Cys-142, His-79 and His-293. This pocket is large enough to accommodate ethyl groups as seen by the substrate propionyl-*R*-mandelate, **(Figure 2.12-A)** and 2-(propionyloxy)butanoate (**22**) (**Figure 2.12-B**).

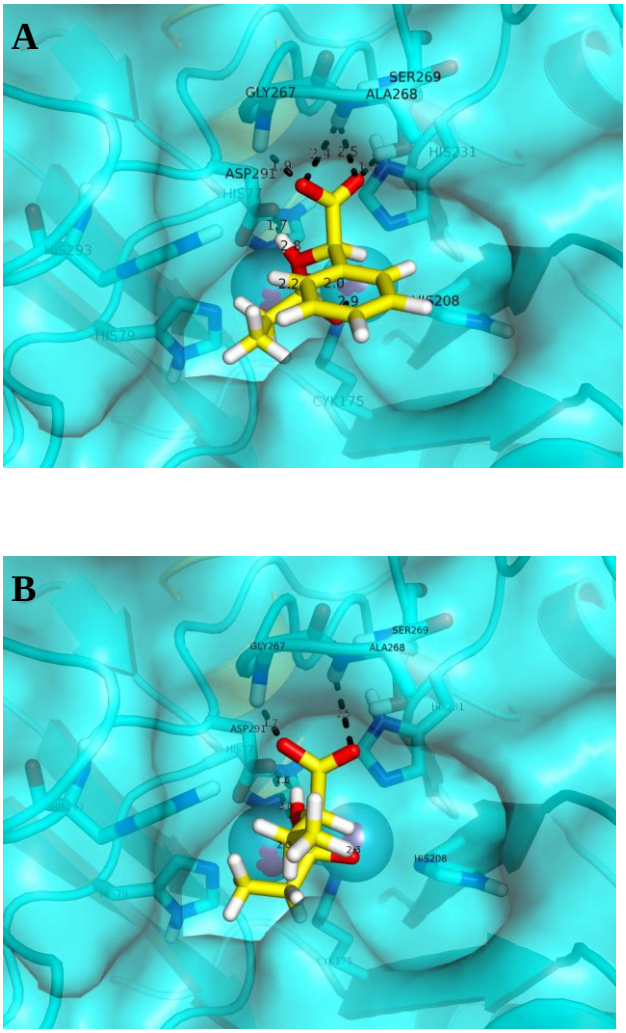


Figure 2.12: Docking models of Atu3266 in the presence of propionyl-oxy compounds: (**A**) propionyl-*R*-mandelate and (**B**) 2-(propionyloxy)butanoate.

In addition, this apparent catalytically competent pose orients the carboxylate of the substrate towards the loop formed between β-strand **7** and α-helix **7** of the AHS domain. This loop consists of Gly-267, Ala-268 and Ser-269. The high scoring poses indicate that any substitutions or bulky additions on the substrate carboxylate would change the electrostatics or cause steric clashes and likely render the molecules inactive by preventing binding. This phenomenon is observed in compounds where substitutions to the acetyl-*R*-mandelate (**28**) are made on the carboxylate moiety. More specifically molecules with an addition of a methyl group to form an ester (**34**) or molecules with a substitution of the carboxylic acid moiety for an amide group (**31**), demonstrate no enzymatic activity.

Further docking analysis indicates a preference for the *R*-enantiomer of the acetylated α-hydroxyl carboxylates, such as acetyl-*R*-mandelate. The *R*-enantiomers direct the hydrogen of the chiral carbon towards the floor of the active site cavity and provide ample space for the substituent (**Figure 2.12-A**), while still orienting the bridging oxygen with ideal geometry and distance to Asp-291 thus facilitating catalysis. Based on the well-positioned docking poses of the substrates, we can infer from the HEI docked molecules that the ground state *R*-enantiomers could easily enter the active site in the extended conformation (**Figure 2.8-A**). In this extended conformation, the molecules could easily be subjected to nucleophilic attack by an activated water molecule. Substrates that contain the *R*-enantiomer can easily accommodate bulky substituents, such as the phenyl ring of acetyl mandelate, which fit into the active site between the cavity walls. In this well-posed orientation, the compounds with aromatic

rings engage in perpendicular *pi* stacking interaction with the side chain nitrogen atom of Asn-144. This interaction is not present in the small branched or cyclohexane containing compounds (**13-14, 18-19, 22, 24-27, 48-50**), which may explain the difference in affinity parameters. This is illustrated by the different binding modes and interactions of *R*-2-acetoxybutanoate (**18**) and *R*-2-acetoxy-2-cyclohexylacetate acid (**27**) (**Figure 2.13 A and B**).

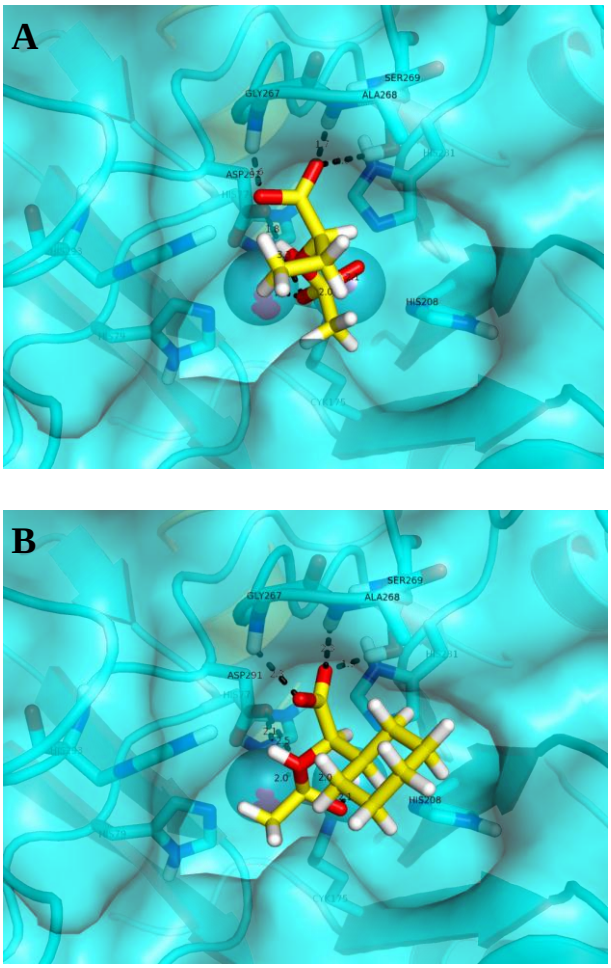


Figure 2.13: Docking models of non-aromatic compounds. Illustrations show binding mode of R-2-acetoxybutanoate (**A**) and R-2-acetoxy-2-cyclohexylacetic acid (**B**) in the active site of Atu3266. Results illustrate the lack of *Pi* stacking interactions.

Only a few ionic interactions are observed for charged substituents on the acetylated α-hydroxyl carboxylates, these charged interactions are made with Asn-144 (**Figure 2.14 A and B**) and Lys-236 (**Figure 2.14 C and D**) for some of the substrates.

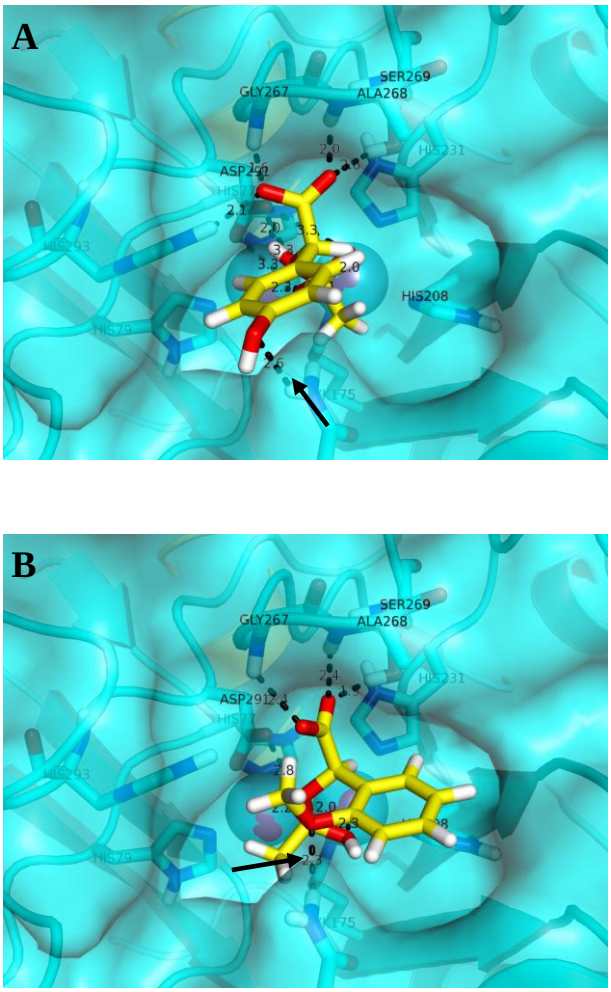


Figure 2.14: Docking models of compounds with phenyl ring substituents. Ionic interactions are shown for various compounds with additional substituents to acetyl-*R*-mandelate molecule. Interactions of molecules with Asn-144 of the protein are observed with: **(A)** 4-hydroxy acetyl-*R*-mandelate and **(B)** (R)-2-acetoxy-2-(2-methoxyphenyl) acetic acid. Interactions with Lys-236 of Atu3266 are observed with: **(C)** (R)-2-acetoxy-2-(3,4-diacetoxyphenyl)acetic acid and **(D)** (R)-2-acetoxy-2-(4-acetoxyphenyl) acetic acid.

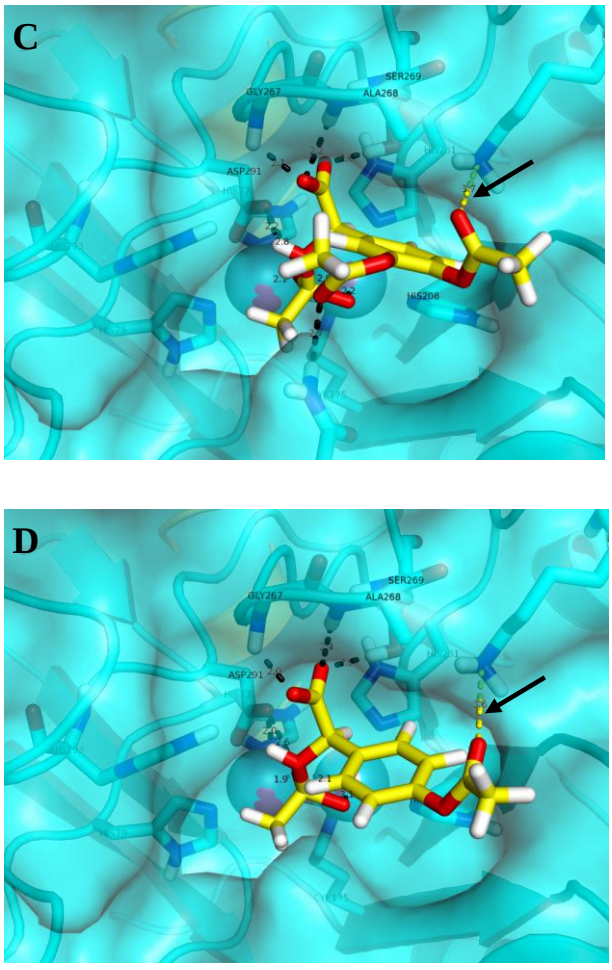


Figure 2.14 continued.

Attempts were made to dock the *S*-enantiomer of acetyl mandelate; however those experiments mostly resulted in non-competent orientations wherein the bridging oxygen was too far from the catalytic Asp-291. The highest scoring distance restrained pose of acetyl *S*-mandelate (**35**) that had an encouraging fit in the active site, orientated the carboxylate towards the loop formed between β-strand **7** and α-helix **7** of the AHS domain and placed the ester group on top of the metal ions, as previously seen with the *R*-enantiomer. However, even though this pose falls within the hydrogen bond distance restraints from the metal ions, it has a nonproductive conformation. **Figure 2.15** illustrates how this HEI compound adopted a geometry where the carbonyl oxygen points away from the nucleophile. The *S*-enantiomer of acetyl mandelate (**35**) also positions the phenyl ring at the bottom of the active site, allowing the binding pocket to surround the top of the ring. This enantiomer would thus prevent turnover of many active substrates (**36-42, 45-47**) that have substitutions on the phenyl ring, most notably in the para position.

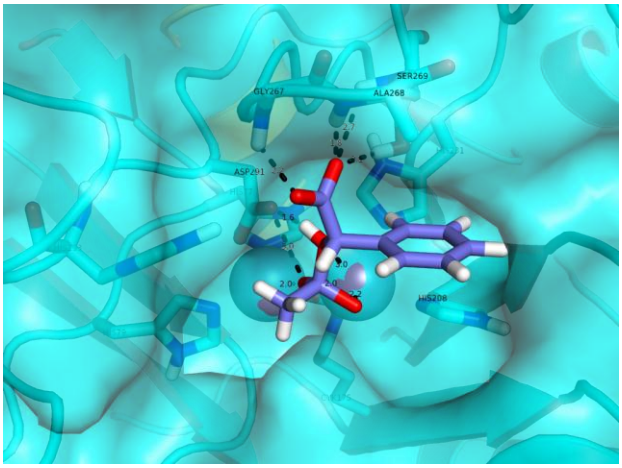


Figure 2.15: Docking model of acetyl-*S*-mandelate. HEI of compound was docked into the active site of Atu3266 to observe overall conformation.

Finally, the HEI form of acetyl-*S*-mandelate that is docked in the active site does not have an extended confirmation of the acetylated α-hydroxyl carboxylate backbone. Instead, it is bent on itself and the docking results indicate that the hydrogen on the chiral α-carbon comes into close proximity of the methyl hydrogen on the acetyl moiety (1.58 Å). This orientation suggests that significant geometric rearrangement of the HEI substrate would be required to adopt a catalytically favorable pose before the compound could undergo hydrolysis of the ester bond.

From the dedicated virtual libraries, we have been able to screen compounds with slightly longer $O=C(O)X_nC(R)X_nOC(=O)X_n$ backbones. Docking results suggest that larger molecules can be accommodated into the active site, such as 2-phenyl-β-acetoxypropanoic acid, 3-carbamoyloxy-2-phenylpropionic acid, *O*-acetyl-*S*-carnitine and *S*-3-acetoxy-4-(dimethylamino)butanoic acid (**Figure 2.16-A**). Of these, *S*-3-acetoxy-4-(dimethylamino)butanoic acid (**Figure 2.16-B)** and racemic 2-phenyl-β-acetoxypropanoic acid (**Figure 2.16-C)** were tested as possible substrates for Atu3266, however only 2-phenyl-β-acetoxypropanoic acid is active, albeit with a very poor *V/K* of 200 $M^{-1}s^{-1}$. This indicates that a small extension between the chiral carbon and the ester can be tolerated but that these substrates have a non-optimal confirmation compared to the smaller molecules that contain the simple acetylated α-hydroxyl carboxylate chemotype.

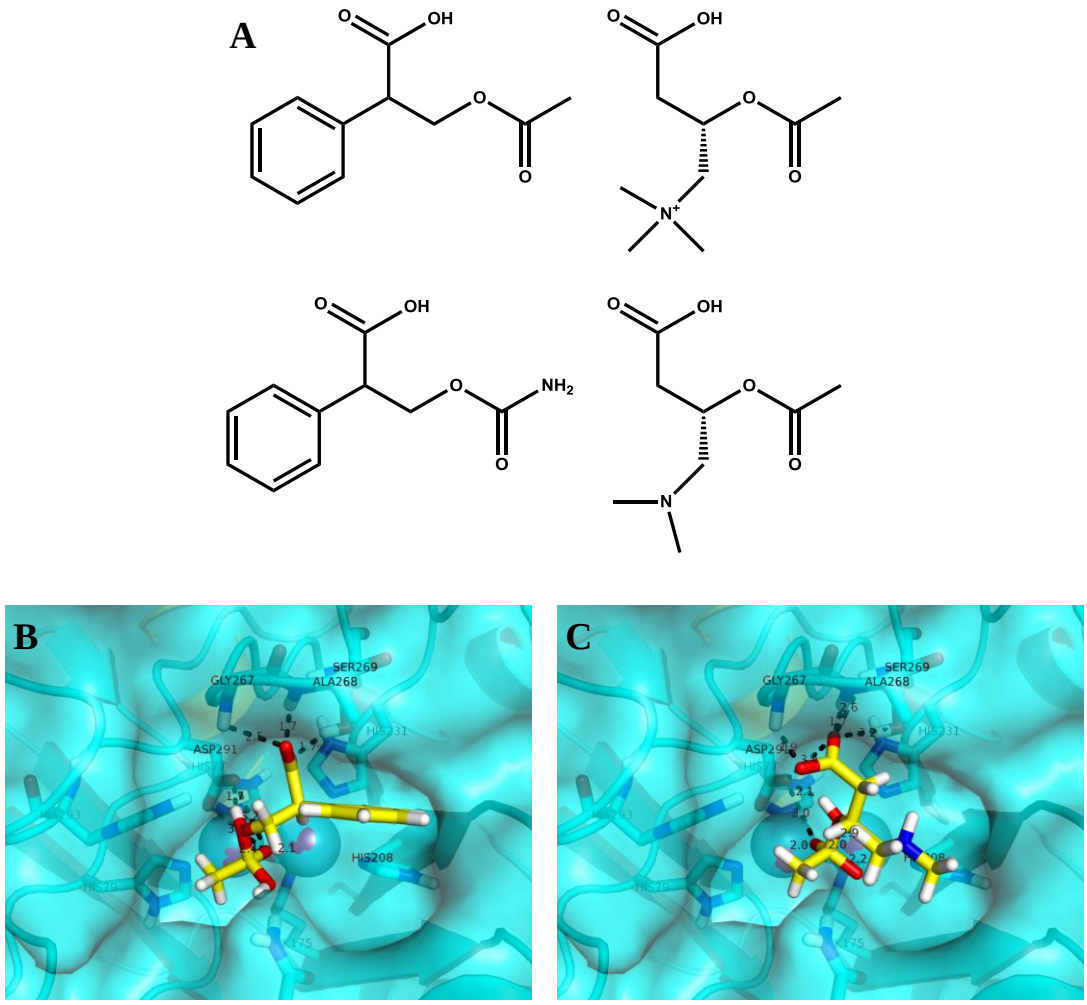


Figure 2.16: Docking models of β-acylated carboxylates. (**A**) Molecules with one carbon insertions around the chiral carbon of the acetylated α-hydroxyl carboxylate backbone that dock into the active site of Atu3266. Starting from top left counterclockwise: 2-phenyl-β-acetoxypropanoic acid, 3-carbamoyloxy-2-phenylpropionic acid, S-3-acetoxy-4-(dimethylamino) butanoic acid, and *O*-acetyl-S-carnitine. Two compounds were tested to see if substitutions on the acetyl or carboxyl side are tolerated. (**B**) Active compound 2-phenyl-β-acetoxypropanoic acid places the extra carbon to the top of the active site and displays low turnover with Atu3266. (**C**) Non active (S)-3-acetoxy-4-(dimethylamino) butanoic acid.

Strategy for Functional Annotation: The functional annotation of Atu3266 as an enzyme that was able to catalyze the hydrolysis of an acetyl group from substituted α-hydroxyl acids with *R*-stereochemistry began with the identification of the weak substrate, *N*-acetyl D-serine. The combination of focused chemical libraries and computational docking to the three-dimensional crystal structure resulted in the identification of acetyl-*R*-mandelate as a substrate that was 5 orders of magnitude better than the initial hit. In addition to Oant2987 and RHE_PE00295, we predict that 6 other enzymes from Group 6 of COG3964 will have the same substrate profile as Atu3266. These proteins are listed in **Table 2.5**. The identification of function for Group 6 of COG3964 can facilitate the functional annotation for many of the other groups of proteins in this COG. In preliminary experiments, a protein from Group **2** of COG3964 **(Figure 2.10)** has been purified and crystallized (EF0837 from *Enterococcus faecalis*, gi|29375425, PDB id: 2ICS). This protein can catalyze the hydrolysis of acetyl-*R*-mandelate with a value of k_{cat}/K_m that is 3 orders of magnitude lower than for Atu3266. These results suggest that the substrate profile for this enzyme will be different than for Atu3266. Efforts to identify the substrate for the enzymes of group 2 are discussed in the following chapter.

Table 2.5: Additional enzymes found in group 6. Locus tag and respective organism encoding other proteins in group 6 of COG3964.

Locus tag	Organism
Atu3266	*Agrobacterium tumefaciens* C58
Oant2987	*Ochrobactrum anthrop*i ATCC 49188
RHE_PE00295	*Rhizobium etli* CFN 42
pRL110419	*Rhizobium leguminosarum* bv. Viciae 3841
RHECIAT_PA0000241	*Rhizobium etli* CIAT 658
Rleg2_5803	*Rhizobium leguminosarum* bv. Trifolii WSM2304
Arad_7942	*Agrobacterium radiobacter* K84
Avi_5090	*Agrobacterium vitis* S4
Rleg_6703	*Rhizobium leguminosarum* bv. Trifolii WSM 1325

CHAPTER III

STRUCTURAL STUDIES, SUBSTRATE DIVERSITY AND FUNCTIONAL ANNOTATION OF ORTHOLOGUES IN COG3964 ENZYMES: INSIGHTS FROM EF0837, BCE_5003 AND STM4445

In the hierarchy of metabolic pathways, those that are amphibolic in nature comprised of anabolic and catabolic functions have been at the core of scientific research. There are a myriad of peripheral pathways that can utilize a large number of organic compounds as substrates. In chemical biology, these pathways are critical in the understanding of biological systems. However, due to their very complexity and consequent labor investment, it's probably a disincentive in applying the effort needed to unravel them. Yet, the very understanding of peripheral pathways may well provide evolutionary lessons that could not be learned from more central pathways. Besides their complexity, one of the challenges involved in the study of peripheral pathways is their selected availability within specific organisms; this however prompts the curiosity of evolving pathways between species and across kingdoms. In chemical biology the study of these pathways may be hindered due to the lack of validity of functional designations of genes from sequencing projects, where it is estimated that approximately 40% of newly sequenced genes have an unknown, uncertain, or incorrect function (*6, 7*).

In this chapter the extent of functional misannotations is further observed as a new group of enzymes belonging to COG3964 demonstrates substrate differentiation from their assigned annotation in sequence databases, as well as substrate variation from

the function discovered in a distinct group of COG3964, the previously discussed group 6. This in part, has been implied based on lack of activity with compounds that are substrates to the characterized enzymes that have the same annotation, as well as lower kinetic constants with compounds identified as substrates in group 6.

Based on the organization of sequence networks using similarity thresholds, it is observed that at an *E*-value of 10^{-70}, there are 8 major groups formed in COG3964. At an E-value of 10^{-70}, there is a minimum 40% sequence identity between two connected sequences belonging to a group. EF0837 from group 2 and Atu3266 from group 6 share less than 35% sequence identity. It is postulated that two sequences sharing a $\geq$ 40% sequence identity, belonging to the same family will most likely carry out the same function (*13*). **Figure 3.1** shows the organization of COG3964 at an *E*-value 10^{-70}, while **Figure 3.2** focuses on the progression of group 2 in COG3964 as similarity thresholds become more stringent. Beginning with an *E*-value 10^{-70} (40% identity) it is observed that the sequences corresponding to enzymes assigned to COG3964 are segregated into 8 distinct groups, group 2 being the largest with about 70 sequences. Increasing the level of stringency, or decreasing an *E*-value will separate group 2 into sub-groups in which the enzymes are related by a higher identity. At an *E*-value of 10^{-80} the sequences found in group 2 begin to separate into two sub-groups. At an *E*-value of 10^{-90} there are two distinct sub-groups in which the sequences of each sub-group share a minimum 55% sequence identity with at least one other connected member. At an *E*-value of 10^{-100} four sub-groups have now formed, with the smallest sub-group comprised of three sequences. In each sub-group, excluding the smallest comprised of three sequences, an enzyme has

been selected as a target for functional characterization. The minimum sequence identity shared between two connected nodes at an *E*-value 10^{-100} is about 60%.

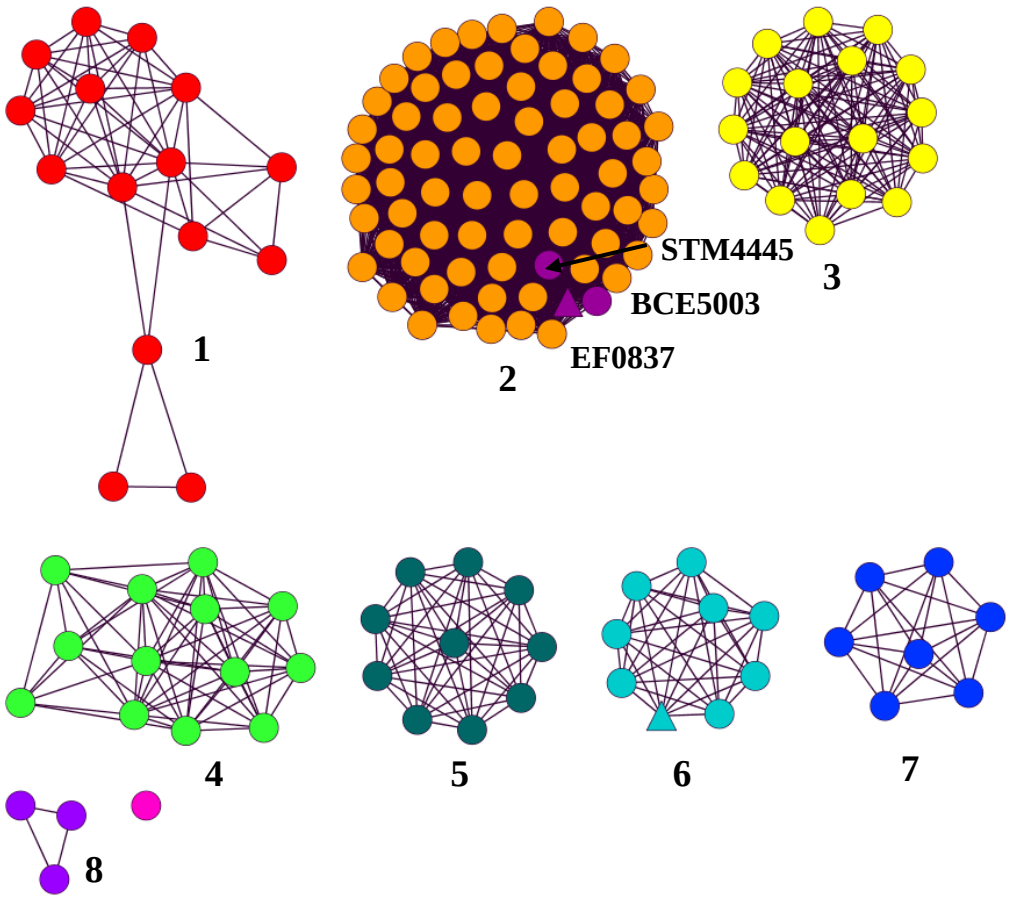
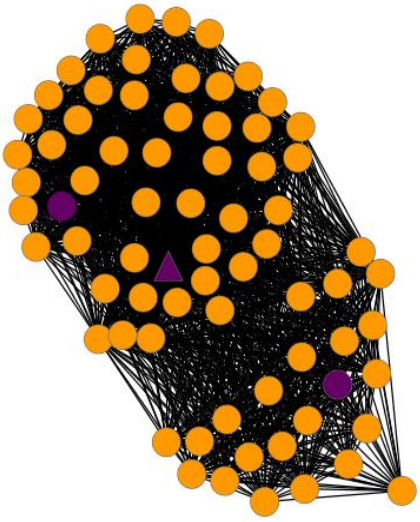
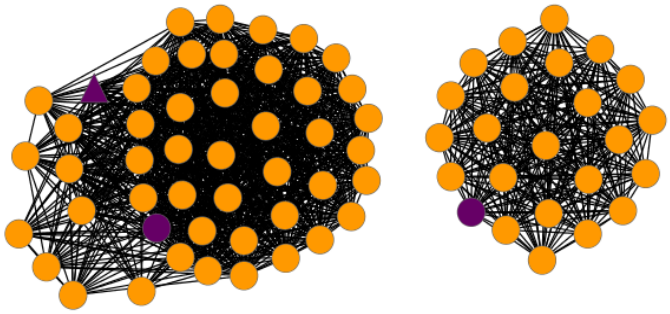


Figure 3.1: Sequence similarity network of COG3964 with group 2 enzymes. Network was prepared at a BLAST *E*-value 10^{-70}. Group 2 contains enzymes that will be functionally studied in this chapter: EF0837, STM4445 and BCE_5003. EF0837 has been structurally characterized and its three-dimensional structure is in the protein database (PDB ID: 2ICS).

Group 2 in COG3964 – BLAST *E*-value 10^{-80}



Group 2 in COG3964 – BLAST *E*-value 10^{-90}



Group 2 in COG3964 - BLAST *E*-value 10^{-100}

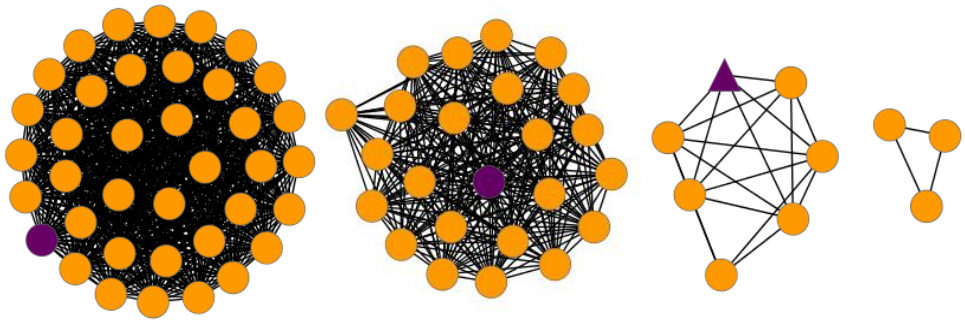


Figure 3.2: Sequence similarity network at increasing stringency values. Network is prepared for group 2 at decreasing *E*-values 10^{-80}, 10^{-90}, 10^{-100}, or increasing identity per group. As these values decrease, the groups are re-organized with each group having more similar sequences.

In group 2 of COG3964, three targets were selected for functional studies. These targets share between 42% - 45% sequence identity between each other. These enzymes are identified by their locus tag as follows: **EF0837**, gi|29375426, from *Enterococcus faecalis* V583; **BCE_5003**, gi|42784049, from *Bacillus cereus* ATCC 10987; and **STM4445**, gi|16767691, from *Salmonella enterica*, subsp. Typhimurium str. LT2. These enzymes were identified from the NCBI database as belonging to the protein cluster COG3964. They have been cloned, purified and screened for functional activity. EF0837 has been crystallized, and its structure is available through the protein structure database (PDB code: 2ICS). The crystal structure shows the prototypical $(\beta/\alpha)_8$ TIM-barrel fold of amidohydrolases. This structure also presents a bound adenine at the binuclear metal active site. Proteins of group 2 have been annotated as dihydroorotases and adenine deaminases. From the 24 COGs of amidohydrolases, there are two additional clusters annotated and characterized as dihydroorotases (COG0044 and 0418) (*43, 79, 86*), and 3 clusters annotated and characterized as deaminases (COG1001, 1816 and 0402) (*63, 76-78, 84*). EF0837, BCE_5003 and STM4445 do not exhibit turnover for the hydrolysis of dihydroorotate or the deamination of adenine. Instead, the enzymes selected from group 2 have shown turnover for the deacetylation of α-acetyl carboxylates, but at much lower rates than those found in group 6 of this same COG. EF0837 and Atu3266 from group 6 share 33% sequence identity, and their structural differences will be further discussed in this chapter. In addition we have exploited the structural differences and similarities between these two enzymes to probe secondary-residue elements in the active site of EF0837 by mutating key amino acids. This was done to reproduce the active site of

Atu3266 in the EF0837 enzyme and observing the effect in the hydrolysis of acetyl-*R*-mandelate, also the best hydrolyzed substrate among group 2 enzymes. A variety of other substrates, including modified amino acids and dipeptides, hydantoins, dihydropyrimidines, acylated sugars, and lactones were screened with no improvement in turnover or sign of catalytic activity.

In addition, we investigated the genomic context of enzymes in group 2. All of these enzymes have a neighboring selenocysteine synthase gene (SelA) adjacent to the gene coding for a COG3964 protein. Selenocysteine synthase is a pyridoxal-5'-phosphate dependent enzyme involved in the synthesis of selenocysteine (*95*-*97*). This enzyme and additional orthologs from the SelA COG (COG1921) will be further discussed in chapter V. In this chapter we present their emergence as an operon neighbor to enzymes of group 2 in COG3964, and we postulate their associations as neighbors to amidohydrolases of COG3964.

MATERIALS AND METHODS

Materials: LB medium for growth of EF0837 and BCE_5003 was purchased from TPI Research Products International Corporation. Compounds tested for activity with enzymes in group 2 of COG3964 were purchased from Sigma-Aldrich, unless stated otherwise. Various hydantoins, diketopiperazines, dihydroorotate analogs were synthesized by Dr. Tamari Narindoshvili and Dr. Chenfu Xu. Acylated sugars were purchased from Carbosynth. All libraries of modified amino acids and dipeptides were synthesized as previously stated (*7*, *87*). Mutants of clone EF0837 were prepared using

the Quik Change™ site directed mutagenesis kit from Stratagene. DNA sequencing and oligonucleotide synthesis was performed by the Gene Technologies Lab at Texas A & M University. Molecular biology materials were purchased from Invitrogen. The His-tag purification Ni^{2+}-binding NTA resin was obtained from Thermo Scientific. Determination of the metal content of all purified proteins in group 2 in COG3964 was determined using ICP-MS. Screening tests were performed in 96-well microtiter plates with a SPECTRAmax 384-plus spectrophotometer from Molecular Devices. The acetic acid assay kit was obtained from Megazyme.

Target Selection, Cloning, Expression and Purification of STM4445, EF0837 and BCE_5003 for Functional Analysis: The plasmid containing the gene for EF0837 gene product (gi|29375425) was prepared by the New York Structural Genomix Research Consortium (NYSGXRC). This target was identified for structural-functional studies within the amidohydrolase superfamily. Given the determination of its crystal structure, functional studies for this enzyme were carried out. Additional members of group 2 in COG3964 were selected based on genomic DNA availability. BCE_5003 also a NYSGXRC target, shares 42% sequence identity to EF0837, and was obtained for functional studies. STM4445 was supplied by the Enzyme Function Initiative (*4*) as a functional target. This enzyme shares 45% sequence identity to EF0837.

The 1107 base pair gene was cloned from *Enterococcus faecalis* V583, into a modified pET26b vector (Novagen) via TOPO cloning. The bacterial expression vector contained a T7-promoter, a C-terminal $(His)_6$-tag followed by a smt3 fusion protein and a kanamycin selective marker. This construct was identified by the clone name:

9295a1BCt3p1. As discussed, this target was nominated for structural-functional studies, and upon successful determination of its crystal X-ray structure, the structural information was uploaded in the PDB data bank. The obtained clone was acquired for functional studies from NYSGXRC. The plasmid was transformed by electroporation into *E. coli* BL21 (DE3) and plated onto an LB media agar plate containing 50 μg/mL kanamycin. A single colony was inoculated into overnight cultures containing 5 mL of LB medium and supplemented with 50 μg/mL of kanamycin. Each overnight culture was inoculated into 1 L LB medium with the same concentration of kanamycin. The cells were allowed to grow at 30 °C. The cells were treated with 0.15 mM 2,2'-bipyridyl as the OD_{600nm} approached 0.3. At an OD_{600nm} of 0.6, the cells were induced with 0.2 mM isopropyl-β-thiogalactoside (IPTG) for the over-expression of EF0837, at which time the temperature was dropped to 20 °C. The cells were further supplemented with 1mM $ZnCl_2$. The cells were allowed to grow for a period of an additional 10 hours, after which, the cells were harvested by centrifugation and quick frozen. The TOPO-clone of the target BCE_5003 was manipulated in similar fashion as that for EF0837. BCE_5003, was cloned into a pSGX3 (BC) vector. The SGX clone was identified as: 9299a1BCt4p1. This target did not go further utilization for structural studies, instead we received the plasmid construct to conduct additional functional studies for group 2 of COG3964.

The over-expression of BCE_5003 was carried out in similar fashion as EF0837. Purification of EF0837 and BCE5003 was carried out similarly. Upon thawing of a 10g/12g of cell pellet, the pellet was resuspended in approximately 70 mL binding buffer

containing 20 mM HEPES, 500 mM NaCl, and 5.0 mM imidazole at pH 7.6. The cells were treated with 10 μg/mL of phenylmethanesulfonyl fluoride and disrupted by sonication at 0 °C. Sonication cycles consisted of 2 second pulses for 4 minute sessions with 4 minute rest between pulsing sessions. Cell debris was removed by centrifugation. The supernatant solution was passed though a 0.2 μm cellulose acetate sterile filter, and loaded on to a Ni^{2+}-NTA column equilibrated with binding buffer. The column was washed thoroughly with binding buffer until the absorbance of the flow-through at 280 nm did not change and was ≤ 0.1. EF0837 was observed to elute with wash buffer (20 mM HEPES, 500 mM NaCl, 50 mM imidazole at pH 7.6). BCE_5003 was eluted with a gradient of a buffer containing 10 mM HEPES, 250 mM NaCl and 500 mM imidazole at pH 7.6. The protein elution collection aliquots were combined and each protein was concentrated. After dialysis the proteins were judged to be >95% pure based on SDS-PAGE. STM4445 protein was received from the Enzyme Function Initiative protein purification core for functional studies in this group. STM4445 was observed to be >95% pure based on SDS-gel electrophoresis.

Protein Concentration and Metal Analysis: Protein concentration was determined spectrophotometrically at 280 nm using a SPECTRAmax-384 PLUS UV-Vis spectrophotometer. The extinction coefficients at 280 nm used to determine the concentration of EF0837 (ε = 38,515 M^{-1} s^{-1}), BCE_5003 (ε = 30,035 $M^{-1}s^{-1}$), and STM4445 (ε = 16,180 $M^{-1}s^{-1}$) were derived from http://web.expasy.org/protparam/. The metal content was determined for each enzyme using a Perkin-Elmer Analyst 700 atomic absorption spectrometer by inductively coupled plasma emission mass spectrometry

(ICP-MS). Prior to ICP measurements, the protein samples were treated with concentrated nitric acid and refluxed at 100 °C for 15 min. The samples were then diluted with distilled water until the final concentration of nitric acid was 1%.

Crystallization and Structure of EF0837: The structure of the gene product of EF0837 was solved by collaborators at the New York SGX Research Center for Structural Genomics as part of a larger effort for structural and functional characterizations within members of the amidohydrolase superfamily. Crystals for the Zn^{2+}-Adenine bound EF0837 were grown by vapor diffusion, sitting drop at pH 7, in the presence of 1.8 M sodium malonate and 10% jeffamine. Crystals exhibited the space group $P2_12_12$, and had one molecule of EF0837 per asymmetric unit. Data for the Zn^{2+} bound EF0837 was collected to a resolution of 2.3 Å at the NSLS X29A beamline in Brookhaven National Laboratory. Adenine found in the active site was not added to forming crystals and was found fortuitously after refinement of the crystal structure. This could mean that the enzyme was potentially annotated as an adenine deaminase based on the presence of adenine at its active site. It is highly unlikely that the adenine molecule was bound to the active site as a purported substrate or inhibitor molecule based on the activity that was instead discovered for this and other enzymes from this group. EF0837 was not found to deaminate adenine or other adenine analogs; it instead carried out the deacylation of acylated α-hydroxyl carboxylates, but at rates much lower than those observed with enzymes in chapter II (group 6 of COG3964).

Sequence Analysis and Target Selection for Functional Studies: Group 2 of COG3964 consists of the largest group of orthologous enzymes within the cluster, with about 70 sequences representing this group. Sequences in this group of enzymes share a minimum of 40% identity between each other. An all-by-all sequence alignment by means of the program MUSCLE (**MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation) (*133*) was constructed using exclusively all the members of group 2. An additional alignment was prepared using all members of COG3964 retrieved from the protein clusters selection on the NCBI database (*83*). The all-by-all sequence alignment of COG3964 was analyzed for conserved residues, between different groups, within the loop regions following the β-strands forming the TIM-barrel. The sequences were also organized by cluster networks as observed in **Figure 2.1**, using Cytoscape, an open source database that allows the visualization of these network associations (*82*). COG3964 was studied at a BLAST *E*-cutoff value of 10^{-70}. At this value it is speculated that members within a group carry out the same reaction. A COG3964 alignment allows distinction of residues conserved in all the sequences belonging to this cluster, as well as those conserved in individual groups only. To further enhance the functional studies of each group in COG3964, we maximized the number of selected enzyme targets that can be cloned using ATCC available genomic DNA. Some of these targets are provided by the Enzyme Function Initiative protein core. Targets included in the functional studies of group 2 of COG3964 included: LSEI_2713 (GI|116496136) from *Lactobacillus casei* ATCC 334, t4479 (GI|29144724) from *Salmonella enterica* subsp. Enterica serovar, KPN_04638 (GI:152973110) from *Klebsiella pneumoniae* subsp. Pneumoniae, and

STM4445 (GI:16767691) from *Salmonella typhimurium* str. LT2. Only one of the previously mentioned targets (STM4445) was successfully cloned, over-expressed, purified and screened for functional studies of group 2.

Substrate Screening and Methodology: The observed activities of enzymes belonging to group 2 of COG3964 were identified from large-screening experiments with compounds found to be substrates to group 6 of COG3964. These compounds consisted of acylated α-carboxylates. Many of the compounds also screened with enzymes in group 2 consisted of *N*-acetyl, *N*-carbamoyl, *N*-formyl and *N*-succinyl D- and L- amino acids. A large library of dipeptides consisting of D-Xaa-L-Xaa, L-Xaa-D-Xaa and L-Xaa-L-Xaa were also screened for activity with all the available members of group 2. In addition, we compiled a large library of lactones that have been found to be substrates for other amidohydrolases (*75*), we theorized these could be substrates for group 2 enzymes based on the ester bond undergoing cleavage of compounds that did show activity; these lactones were also tested. In addition we screened a selected library of diketopiperazines composed of L-Ala-L-Xaa and D-Ala-L-Xaa, where Xaa is any of the 20 common amino acids. Furthermore, based on annotations found on sequence databases, protein structure databases, and homologue annotations at larger *E*-BLAST values (less % identity constraints), we complied and screened a set of nucleotides (adenine, cytosine, *N*6-acetyl adenine, and *N*6-methyl adenine) and nucleosides (adenosine, cytidine). These were screened for deamination. We also tested a variety of other nitrogenous bases (dihydroorotate, orotic acid, dihydrouracil, 6-methyl dihydrouracil, 5,6-dihydro-5-methyl uracil) for hydrolysis.

The hydrolytic activities of EF0837, BCE_5003 and STM4445 were measured using a variety of assay methods. Hydrolysis of acylated compounds were monitored using the acetic acid assay kit from Megazyme. The reaction was monitored spectrophotometrically using the formation of acetate and subsequent reduction of NAD^+ in a coupling system using acetyl-coenzyme A synthetase, citrate synthase and L-malate dehydrogenase. The reaction was examined at 340 nm. The hydrolysis of various lactones was monitored using a pH-sensitive colorimetric assay (*121*). Protons released from the carboxylate product were measured using the pH indicator cresol purple. The reactions were performed in 2.5 mM Bicine at pH 8.3. The change in absorbance at 577 nm was monitored in a 96-well UV-visible plate. Compounds having a scissile amide bond within a ring, such as hydantoins, diketopiperazines, dihydroorotate and other analogs of dihydroorotate were monitored directly at 225 nm. A full UV spectrum was obtained at a concentration 0.7 mM of the compound in the presence of 20 mM phosphate buffer (time=0). A spectrum was then re-read after addition of 1.0 μM enzyme (t=5 min), and at t = 1h, 5h and 10h to observe for spectral changes within the 220-230 nm range. Modified amino acids and dipeptides were screened for activity using the modified Cd-ninhydrin assay. Hydrolytic activity was monitored at 507 nm.

Mutant Development of EF0837: Site directed mutagenesis of EF0837 was utilized to probe the catalytic role of identified residues within short distances of the active site. These residues were observed to be characteristically conserved in sequences of group 2, but were absent in sequences from group 6. We specifically identified the residues from structural comparisons between EF0837 (PDB ID: 2ICS) and Atu3266

(PDB ID: 2OGJ). The residues selected for mutations are conserved in all sequences of group 2, but not in those of group 6. The goal was to examine if these site directed mutations resulted in an increase of activity with acetyl-*R*-mandelate at rates comparable to those in group 6 enzymes. Single, double and triple mutants were designed to represent gradual modification of EF0837 to mimic the active site of Atu3266 deacetylase enzyme. This would allow seeing the effect of these mutations progressively. The rates of wild-type EF0837 are three orders of magnitude lower than those of Atu3266 catalyzing the hydrolysis of acetyl-*R*-mandelate. This compound has also proven to be the best among enzymes in group 2. The mutations targeted are as follows: Y70I, Q125N, and Y274H. Each of these was modified individually starting from the single mutant Y70I; then the double mutant, Y70I/Q125N and lastly the triple mutant, Y70I/Q125N/Y274H.

Data Analysis: The kinetic parameters, k_{cat}, K_m and k_{cat} / K_m for the enzymes EF0837, STM4445, and BCE_5003 were determined by fitting the initial velocity data to equation 3.1, where v is the initial velocity and E_t is the total enzyme concentration, k_{cat} is the turnover number and K_m is the Michaelis constant. Enzymes did not display saturation kinetics with any of the substrates shown to be active, and under these circumstances a K_m was extrapolated from the graphical analysis, or otherwise it was not determined. Under these conditions, a V/K value is provided expressing the rate of the reaction as an apparent rate constant at high substrate concentrations (*134*).

$$v / E_t = k_{cat} [A] / (K_m + [A]) \qquad \text{(eq. 3.1)}$$

Operon Context Analysis: Phylogenetic and genomic context profiles of the proteins ascribed to COG3964 were analyzed using the microbesonline database (*135*). Analyses of these profiles demonstrate the steady presence of a protein adjacent to the gene encoding for the amidohydrolase belonging to group 2 of COG3964. All organisms that express the COG3964 amidohydrolase from group 2, also express an enzyme that has been annotated as a COG1921 protein. Determination of the correct function of COG1921 enzymes found within the vicinity of a COG3964 protein has also been a focus of study as means to correct the functional annotation of these amidohydrolases. COG1921 is a group of proteins annotated in the protein clusters database as L-seryl-tRNA(Sec) selenium transferase. This cluster is composed of pyridoxal phosphate dependent transferases. The function for these characterized enzymes and analysis of COG1921 will be further discussed in chapter V. Here we present analysis of the operon as means to relating the function of the amidohydrolases from COG3964 and the PLP-dependent enzymes form COG1921.

RESULTS

Sequence Analysis, Target Selection and Genomic Operon Context Analysis: EF0837 and STM4445 share 45% sequence identity, STM4445 and BCE_5003 share 43% identity and EF0837 and BCE_5003 share 42% identity. These enzymes essentially share the same sequence identity between each other. At this level of identity, it is expected that EF0837, STM4445 and BCE_5003 carry out the same reaction, on fairly analogous compounds. Based on the sequence alignment, it is expected that members of

group 2 have the same coordinating metal ligands as EF0837. An HxH motif from β-strand **1**, histidines from β-strands **5** and **6**, an aspartate from β-strand **8** and a carboxylated lysine originating from β-strand **4** that bridges the two metals ions. This assemblage of residues is the same that has been observed in phosphotriesterase, urease, dihydroorotase, isoaspartyl dipeptidase and other members of the amidohydrolase superfamily (*42, 50, 51, 61, 62*). This emphasizes the functional diversity of this active site as means of carrying out various different reactions. This is also the coordination of the metal center that was observed in the previous chapter for the crystal structure of homolog Atu3266.

The residues that are often essential in substrate recognition and catalysis are the secondary elements in the loops surrounding the active site. These have not been identified in group 2 enzymes, but based on sequence alignments, and structural comparisons, several residues have been suggested. **Figure 3.3** presents the structural differences between Atu3266 and EF0837. Members of group 2 have a conserved tyrosine eight residues from the HxH motif. This tyrosine faces into the active site of the enzyme. In EF0837, this tyrosine is residue number 70 and is conserved exclusively in all members of group 2 in COG3964. In group 6, this residue is replaced by an isoleucine. After β-strand 3, there is a variable glutamine-aspartate residue pair (Gln-125, Asp-126) in which the glutamine is observed to intrude into the active site of EF0837. In STM4445 the aspartate is replaced by an asparagine and in BCE_5003 the glutamine is not present, while the adjacent residue is the conserved aspartate also observed in EF0837. In the structure of Atu3266, an asparagine is found in the same

location as the glutamine of the structure of EF0837, but these residues do not coincide in the sequence alignment. An arginine of EF0837 (Arg-156) is conserved in all members of group 2 and 6. This residue is found one residue away from the carboxylated lysine. However, this residue is replaced by a lysine or a histidine in other groups of the COG. Group 7 for example, which will be discussed in the succeeding chapter, replaces this arginine at the end of β-strand 4 with a histidine. In addition, there is a variable lysine residue (Lys-216) after β-strand 6 in the crystal structure of EF0837 that reaches into the active site of the structure. This residue is also conserved in group 6 sequences but not group 7. In the structure of Atu3266, the lysine after β-strand 6 is not facing into the active site; instead it faces away from the active site and is solvent exposed. Perhaps the most important distinction of the sequences in group 2 is that they lack the conserved Gly267-Ala268-Ser269 triad that has been assigned as forming an interacting backbone of loop 7 to substrates found hydrolyzed by enzymes in group 6. The backbone of these three amino acids is observed, based on docking models, to interact with the carboxylate group of the purported substrates. In group 2 sequences, the small glycine residue is generally substituted by a threonine, serine, or asparagine.

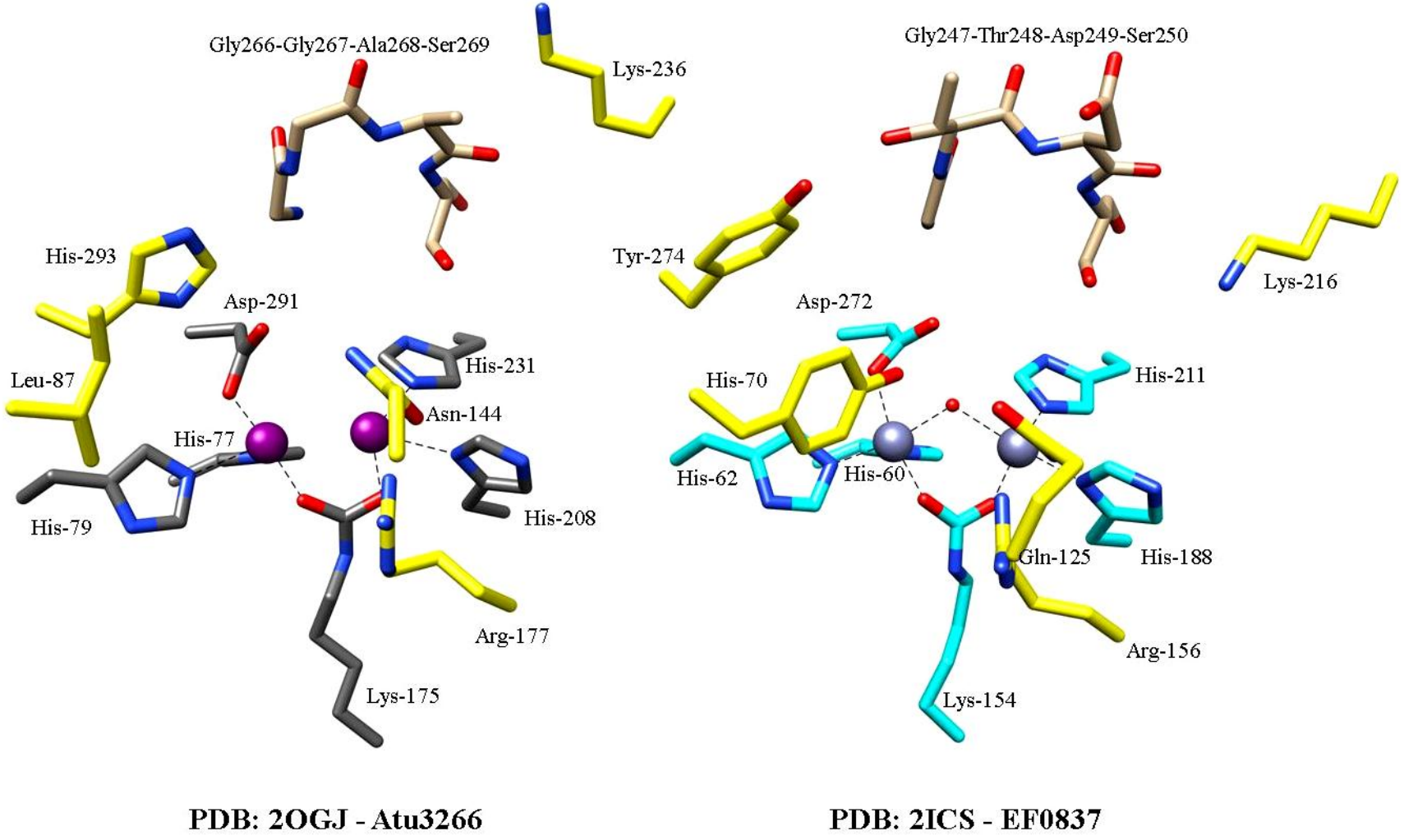


Figure 3.3: Structural comparisons of the active site of Atu3266 (from group 6) and EF0837 (from group 2 in COG3964).

The alignment presented in **Figure 3.4** highlights the specific differences within additional active site residues that may be responsible for substrate selectivity between group 2 and group 6 enzymes in this COG. In addition, the designed alignment also features the sequence of the characterized adenine deaminase from *Agrobacterium tumefaciens* str. C58 (Atu4426). The sequence of Atu4426 was manually modified to allow all corresponding residues comprising the β-strands of the TIM barrel in the sequence to overlap, as well as the metal coordinating ligands at the end of each β-strand, with those residues in the sequences of COG3964. Atu4426 has been mechanistically characterized (*63*), and its crystal structure contains similar coordinating ligands found in binuclear metal center enzymes, with the exception of the bridging residues, where a glutamate takes the place of the carboxylated lysine. Those residues that have been identified in the active site of adenine deaminases to be essential elements in catalytic activity have been highlighted in green, and these residues are not conserved in any of the COG3964 sequences. In addition, the characterized adenine deaminases assigned to COG1001 are typically larger monomeric units; these enzymes are typically 60 kDa units whereas proteins in COG3964 are 40 – 46 kDa monomers. The active site residues found to be contrasting between group 2 and group 6, as discussed earlier in this section, are highlighted in teal.

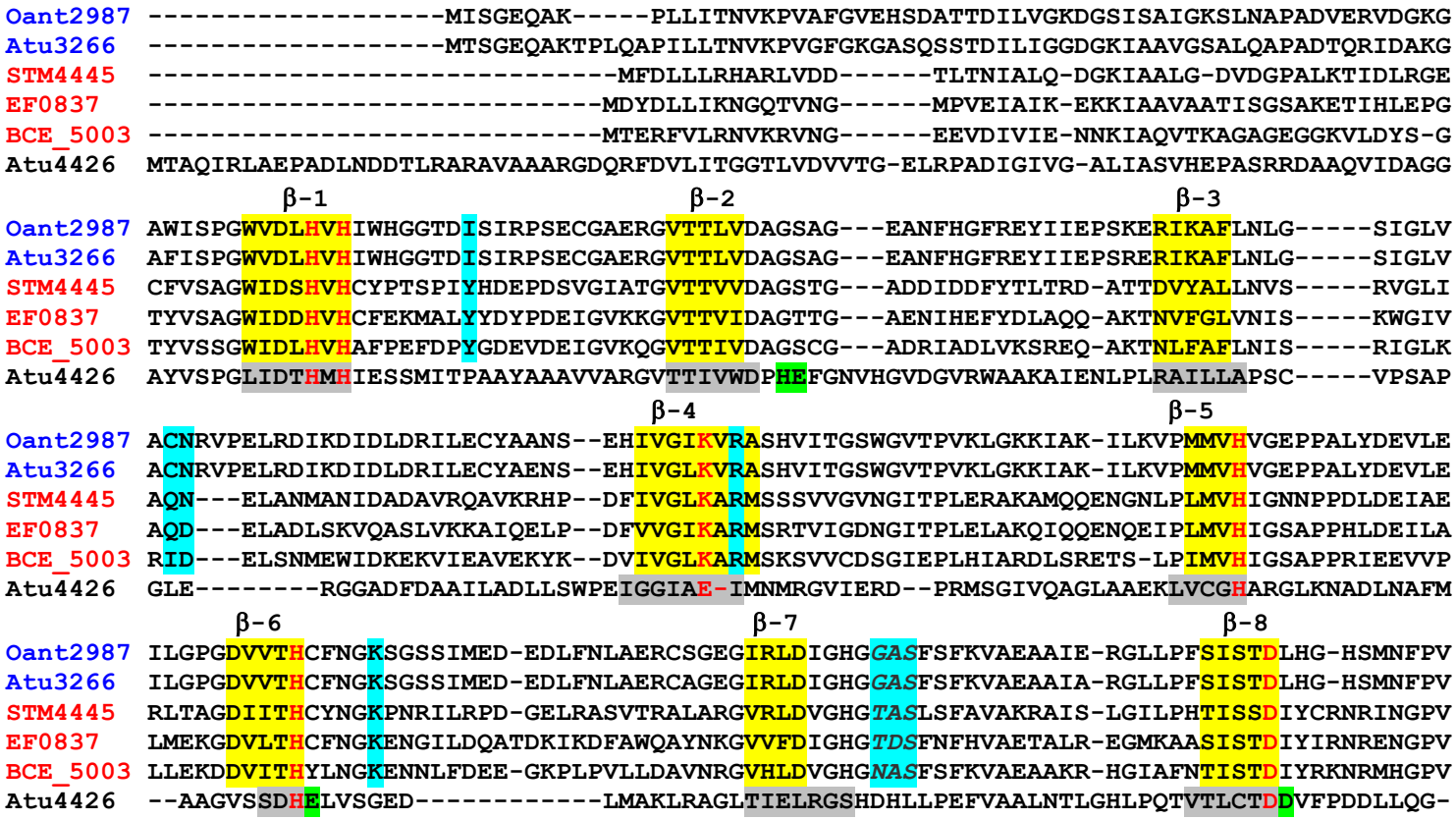


Figure 3.4: Sequence alignment of group 2 selected enzymes: **STM4445**, **EF0837** and **BCE_5003**. Alignment also includes **Oant2987** and **Atu3266** from group 6, and the characterized adenine deaminase from *Agrobacterium tumefaciens* (**Atu4426**). Highlighted in yellow and gray are the residues found in β-strands forming the β-barrel. In red font are the coordinating ligands to the binuclear metal at the center of the barrel. Highlighted in cyan are the residues that have been suggested, based on sequence comparisons, as possible determinants in substrate selectivity for enzymes in COG3964. Highlighted in green are residues experimentally identified as necessary for adenine coordination in adenine deaminases. Mutation of these residues resulted in loss of activity.

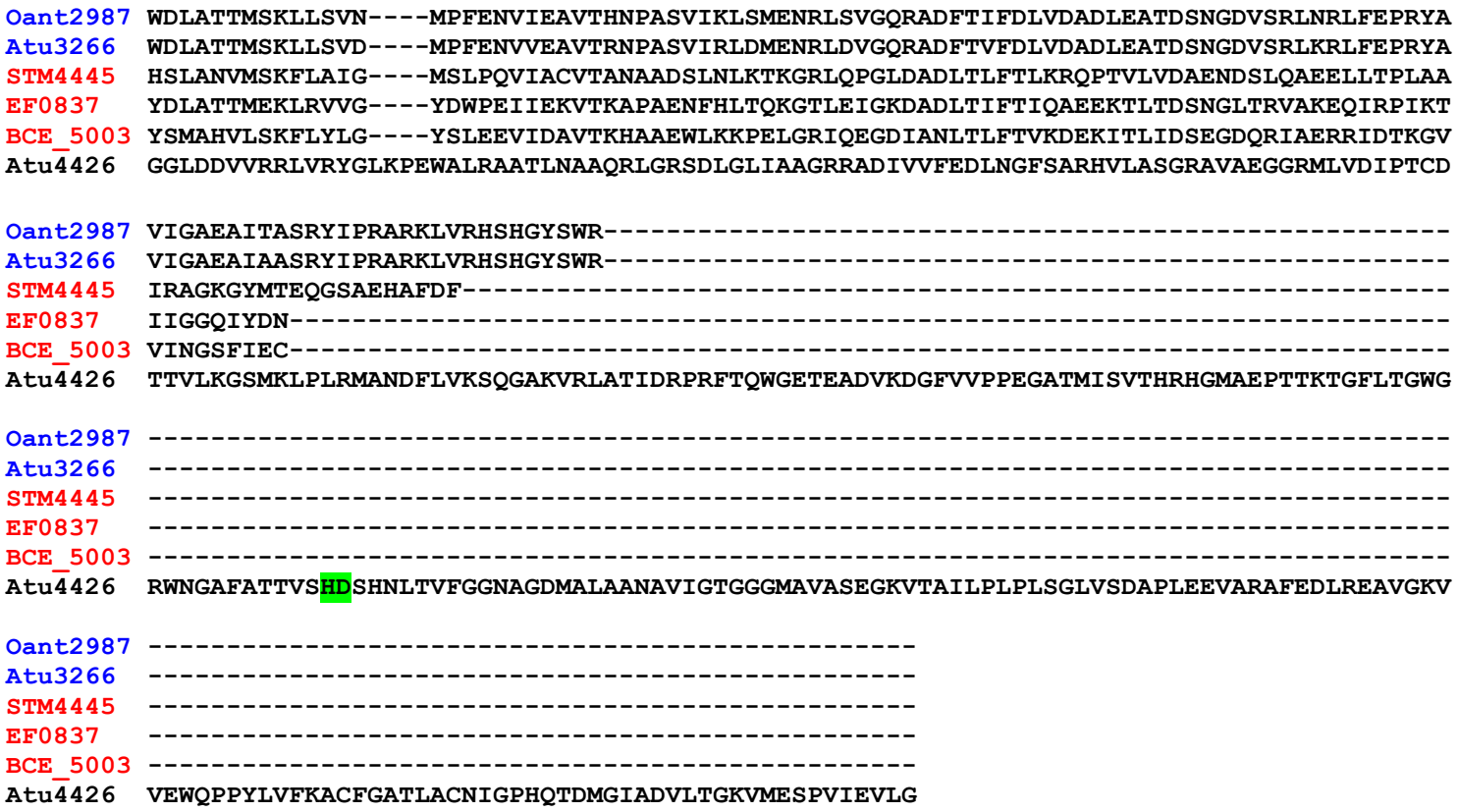


Figure 3.4 continued.

All the sequences in COG3964 encode amidohydrolases in a variety of gram-positive and gram-negative bacterial organisms. However the presence of this protein has only been identified in a limited number of organisms. For each of those organisms that encode a COG3964 amidohydrolase, the operon context of the surrounding genes adjacent to the COG3964 gene was interrogated. The operon context information is an added source for comparative and functional inquiries of enzymes organized into COG3964. The analysis of the operon within organisms encoding for various COG3964 amidohydrolases shared the presence of an additional gene adjacent to the amidohydrolases, which was further investigated in the functional characterization of amidohydrolases. There is no obvious association between the amidohydrolases of COG3964 and the pyridoxal-5'-phosphate dependent enzymes annotated as selenocysteine synthases of COG1921 that are adjacent to the amidohydrolases. **Figure 3.5** shows the general arrangement that is found in the operon of selected organisms that express a COG3964 protein in group 2. The gene encoding the amidohydrolase is identified by its locus tag within the outlined cyan arrow. Generally, there is a SelA gene that immediately follows the COG3964 gene. This one is identified as 'SelA' within an orange arrow. The function of the protein that is encoded by this gene may also be an example of the degree to which functional misannotations are represented in the entire metabolism of various organisms. In other groups belonging to COG3964, the gene for SelA may be further away. The general location of this gene in respect to COG3964 enzymes will be further discussed in chapter V.

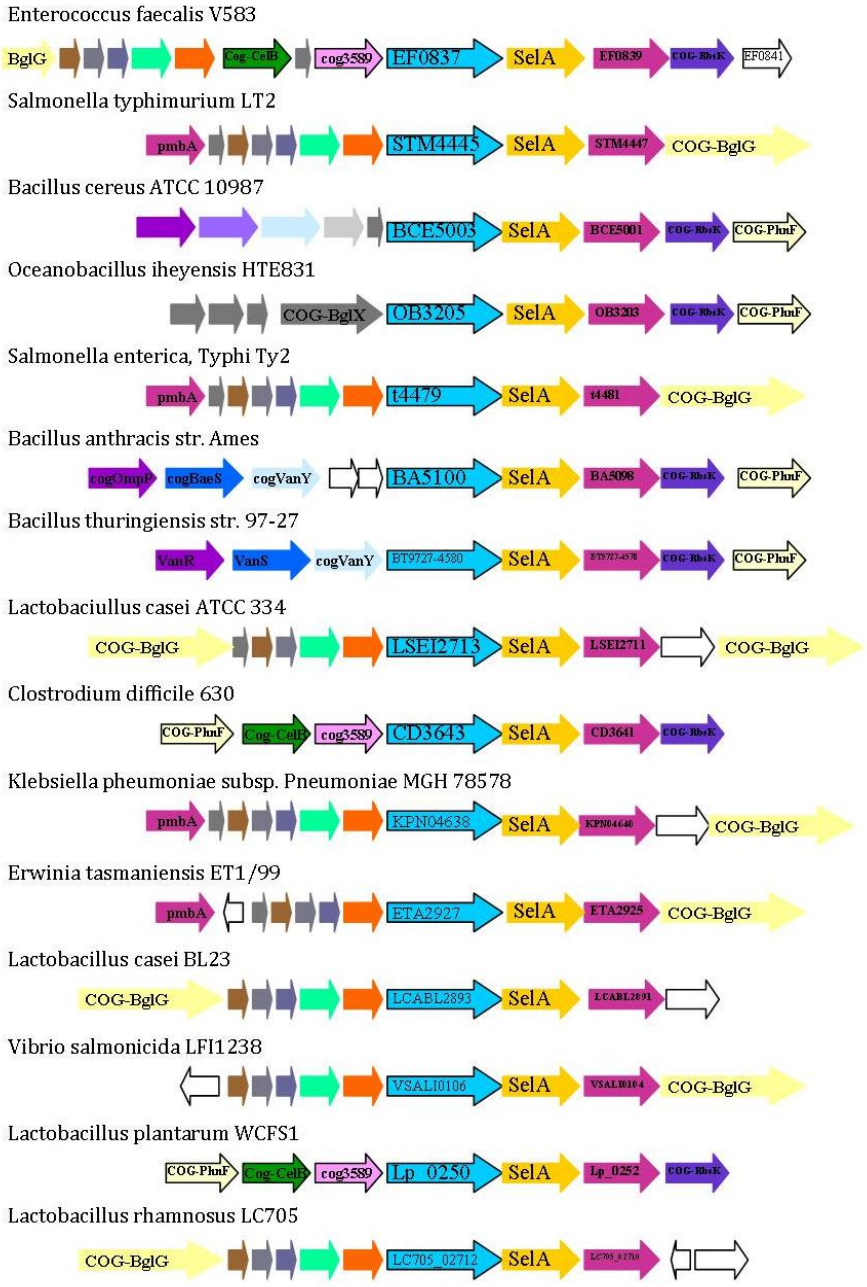


Figure 3.5: Operon context of selected group 2 organisms. Image shows the phylogenetic profiles for 15 of the 72 organisms containing an amidohydrolase belonging to group 2 of COG3964. The open reading frame for each organism shows the COG3964 amidohydrolase in the black outlined cyan arrow, followed by a selenocysteine synthase gene (SelA).

Purification Properties of EF0837, BCE_5003 and STM4445: EF0837 and BCE_5003 were soluble and purified using $(His)_6$-Trap column chromatography. The 6xHis-tag sequence was cloned at the C-terminal of the protein, and it was expected not to affect protein folding or metal binding, as proteins from group 6 also exhibited a C-terminal His-tag sequence with no major problems. 10 g of cells expressing the EF0837 gene yielded 20 mg of protein, while 12 g of cells expressing BCE_5003 yielded only 5 mg of protein. STM4445 was a protein acquired from the EFI protein core. The metal content for each of the purified proteins was determined by ICP-MS. The contents of metal for each protein are listed in **Table 3.1** below:

Table 3.1: Metal content of enzymes from group 2. Analysis of metal content for each purified enzyme was conducted by inductively coupled plasma emission mass spectrometry. Each quantity represents equivalents per monomer of protein.

Enzyme	Zn^{2+}	Fe^{2+}	Mn^{2+}	Ni^{2+}	Cu^{2+}	Total
EF0837	**1.5**	**1.1**	**0.01**	**0.01**	**n/a**	**2.6**
BCE_5003	**0.6**	**0.2**	**<0.1**	**0.2**	**<0.01**	**1.0**
STM4445	**0.5**	**0.1**	**1.1**	**<0.05**	**<0.05**	**1.7**

Each enzyme is expected to contain 2.0 equivalents of metal per monomer based on structure of EF0837 and homology sequence analysis. SD-PAGE analysis revealed the presence of a single band between 40-42 kDa for EF0837, BCE_5003 and STM4445. The molecular weights for each protein based on their sequence are 40.6 kDa, 40.5 kDa, and 40.4 kDa respectively.

Crystal Structure of EF0837: The structure of EF0837 was determined to a resolution of 2.3 Å as a monomer with a binuclear metal center at the C-terminal end of a $(\beta/\alpha)_8$ TIM-barrel (**Figure 3.6**). An adenine molecule is bound at the active site, only 2.4 Å from the β-metal. A water molecule set for activation, bridges the α- and β- metal 2.01 Å and 2.24 Å from each respective metal. The N-terminal residues 1-3 were distorted and were not included in the final model. Residues 4-55 (shown in pink) constitute the N-terminal antiparallel β-strand domain leading up to the barrel. The following chain segments are included in the eight β-strands of the barrel: β1 (residues 56-61), β2 (83-89), β3 (110-116), β4 (150-157), β5(185-189), β6 (208-210), β7(241-243), β8 (270-272). The residues constituting β-strand 8 were distorted and the strand is not shown in the structure. The C-terminal extension of the structure includes a set of three distorted α-helices (residues 274-315) sandwiched between a long loop (residues 316-330) and an additional set of antiparallel β-sheets (residues 331-371).

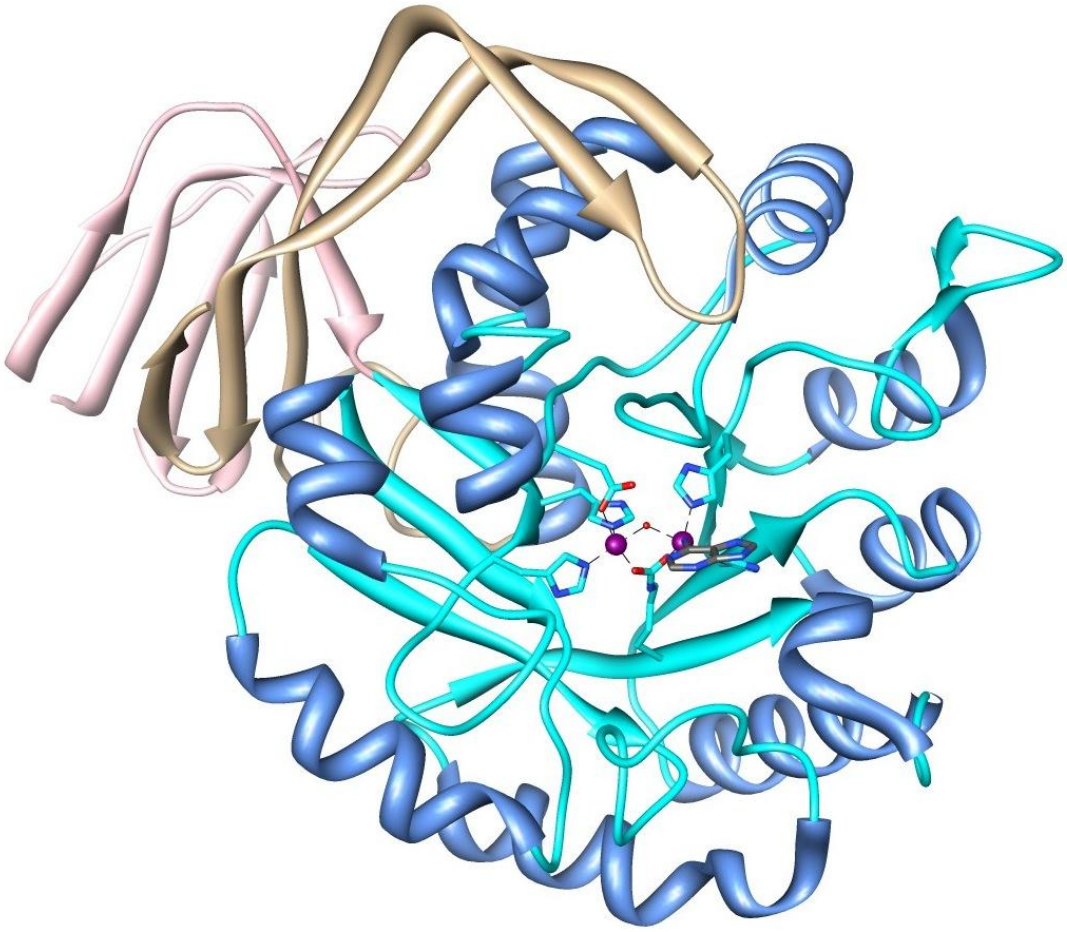


Figure 3.6: Ribbon representation of the crystal structure of EF0837. The monomeric structure has a $(\beta/\alpha)_8$ TIM-barrel colored in cyan, along with metal coordinating ligands (also in cyan). The metal center is colored in magenta. The N-terminal insertion is colored in pink. The C-terminal domain is colored in tan. The bound adenine molecule is shown in gray.

The active site of EF0837 is not much different from other characterized amidohydrolases that have a binuclear metal center and a carboxylated lysine bridging both metals. The most significant differences will be encountered in the substrate recognition elements, which at this point have not been fully identified, as an authentic substrate is not yet characterized. The active site located at the C-terminal end of the barrel is open to the solvent. The two metal ions are 3.6 Å apart. The α-Zn^{2+} is coordinated by His-60, His-62 from β1, and Asp-272 from β8. This metal is also coordinated by the carboxylated Lys-154 originating from β4. The β-metal is coordinated by the same Lys from β-4, and His-88 and His-211 from β-strand 5 and 6 respectively. The bridging water molecule serves as a 5^{th} ligand for the α-metal and it is the 4^{th} ligand for the β-metal. This water molecule is 2.0 Å and 2.2 Å from each, respectively. **Figure 3.7** shows the metal coordination for the active site with all the distances between the the α- and β-metal to the coordinating atoms in each ligand. An adenine molecule was found in the active site of EF0837. This ligand has the nitrogen bound to the C6 atom of the pyrimidine ring moiety of adenine only 2.4 Å from the β-metal. A structural and sequence comparison between the structure of EF0837 and Atu3266 show three residues that were selected to undergo mutagenesis to test out for improvements in catalytic activity. These residues were Y70, Q125, and Y274. **Figure 3.8** shows the active sites of Atu3266 and EF0837. The residues that are hypothesized to be responsible for substrate selectivity and are mutated are circled in red.

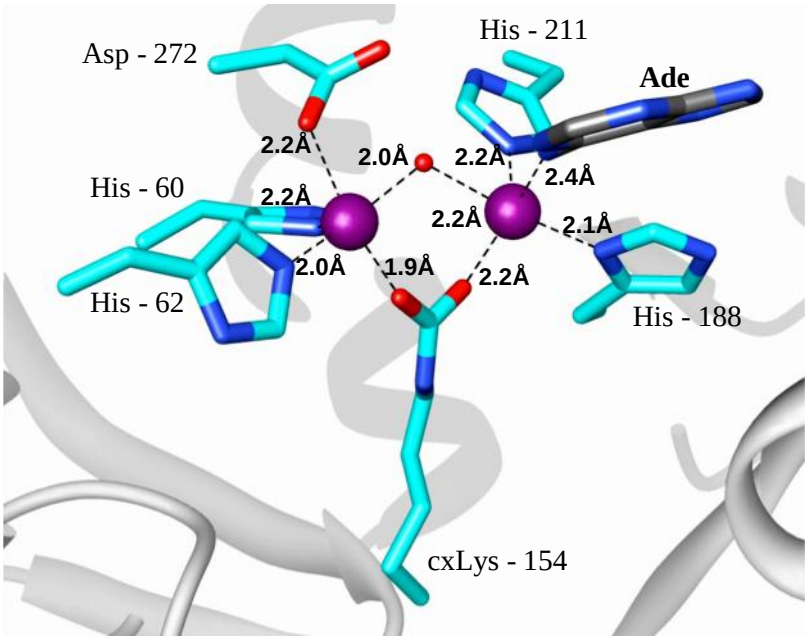


Figure 3.7: Active site of Zn/Zn-EF0837 with adenine bound. The bound molecule interacts with the β-metal. The two zinc ions are presented as magenta spheres. The adenine molecule is in gray. The bridging water molecule is in red.

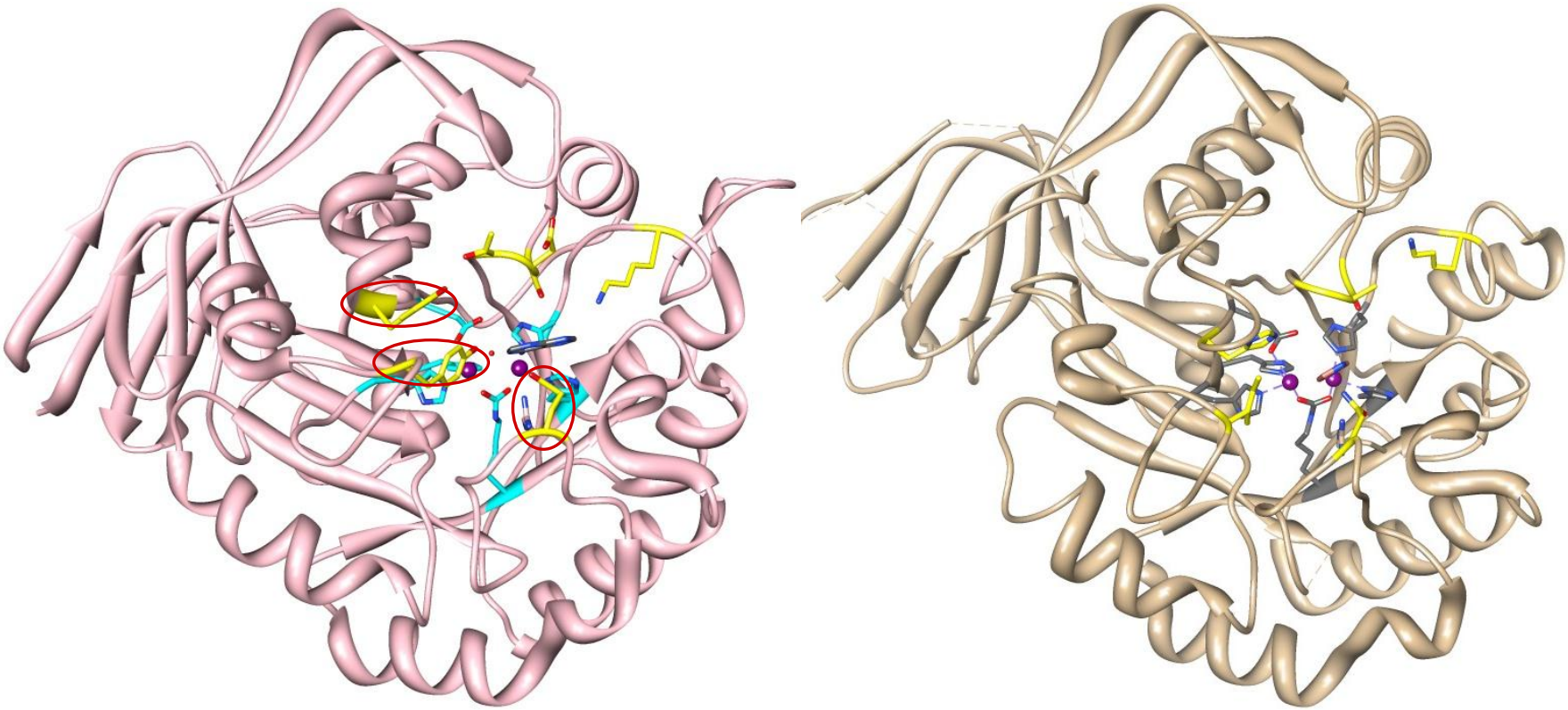


Figure 3.8: Comparison of monomeric units of Atu3266 and EF0837. The crystal structure of EF0837 (pink - left) and Atu3266 (tan - right) show their metal binding ligands (in cyan – left, and gray – right), as well as secondary residues that have been anticipated to be responsible for substrate specificity but not confirmed (in yellow). The residues circled in red in the structure of EF0837 were probed for activity with acetyl-*R*-mandelate. The mutants included single (Y70I), double (Y70I, Q125N), and triple mutant variants (Y70I, Q125N, and Y274H). None exhibited better activities for the hydrolysis of acetyl *R*-mandelate than the wild-type.

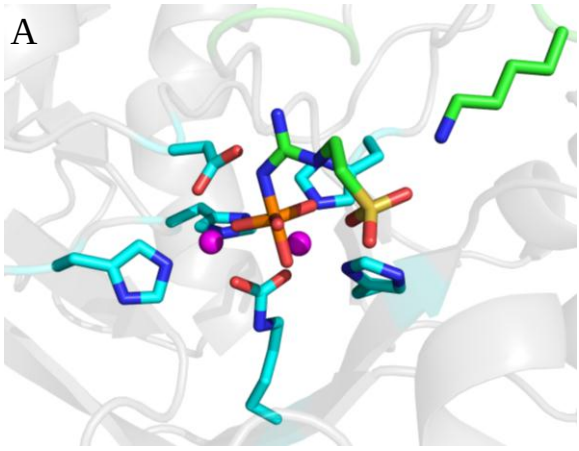
Mutagenesis of Active Site Residues and Metal Analysis of Mutants: Site directed mutagenesis was used to probe residues that are thought responsible for substrate specificity in enzymes belonging to group 2. These residues are not conserved and show variability in the sequences for enzymes in group 6, which were much more active in the hydrolysis of α-acetyl carboxylates. These residues included Y70, Q125 and Y274. The kinetic constants for the single, double and triple mutant of EF0837 are presented in **Table 3.2**. It is observed that none of the selected residues was able to provide with a more active variant for the hydrolysis of acetyl-*R*-mandelate. This suggests that although the authentic substrate may be similar to acetyl-*R*-mandelate, at least in EF0837, the selected residues that underwent mutagenesis are not directly responsible for the low activity in the deacetylation of acetyl-*R*-mandelate compared to those enzymes from group 6. In addition, the loop in Atu3266 found to interact with the carboxylate group of the substrate contains a Gly-Ala-Ser triad. This is substituted in EF0837 with a Thr-Asp-Ser triad. The loop in EF0837 has bulkier, more polar residues that were not mutated to probe for possible affinity of the carboxylate group of a substrate with the enzyme.

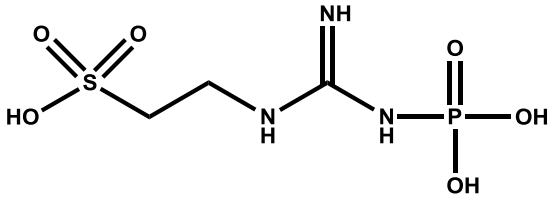
Table 3.2: Kinetic parameters of selected EF0837 variants.

Variant	V/K ($M^{-1}s^{-1}$)	Difference in rate from WT	Metal content
WT	200 ± 20	0	1.5 eq. Zn, 1.1 eq. Fe, total: 2.6 eq.
Y70I	220 ± 12	+ 1.1	2.1 eq. Zn, 0.2 eq. Fe, total: 2.3 eq.
Y70I, Q125N	80 ± 6	- 0.4	1.2 eq. Zn, 0.6 eq. Fe, total: 1.8 eq.
Y70I, Q125N, Y274H	160 ± 20	- 0.8	1.5 eq. Zn, 0.2 eq. Fe, total 1.7 eq.

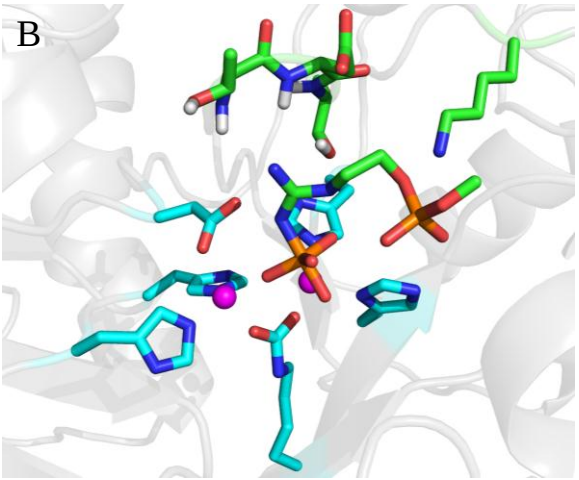
Docking of EF0837: Computational docking was integrated as an additional strategy to determine the function of group 2 enzymes. These experiments were carried out by Peter Kolb from the lab of Brian K. Shoichet at the University of California, San Francisco. The HEI KEGG library also underwent filtering methods based on the distance constraints to the metal active site and the size of the binding pocket. Many of the molecules that were suggested as substrates in the docking results to the crystal structure of EF0837 were compounds that could potentially undergo dephosphorylation, (N-phospholombricine, phosphocreatine, and phosphoarginine) and deamination, (adenosine, and adenosyl analogs). Other molecules suggested a ring opening reaction in various nucleosides and lactone containing compounds (xanthine phosphate, pterins and paederoside). Based on these suggestions a large library of compounds was compiled for screening analysis with EF0837. These compounds were obtained based on commercial availability or synthetic practicability. **Figure 3.9** shows some of the compounds obtained from docking results, and their interactions with the active site of EF0837. The top docking results contained a large number of compounds with a phosphorylated guanidine group. Based on these docking results we tested phosphoarginine, phosphocreatine, and phospholombricine.

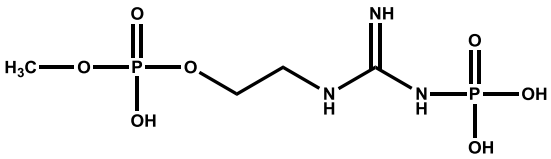
The docking results did not provide the insights that were accessible from docking experiments on Atu3266. The apparent pose of the compounds docked suggests the dephosphorylation of phospho-guanidinium moieties. However additional detailed examination does not provide with coordination of the docked molecule with any other active site residues. Activity was not observed with any of the tested docked compounds.



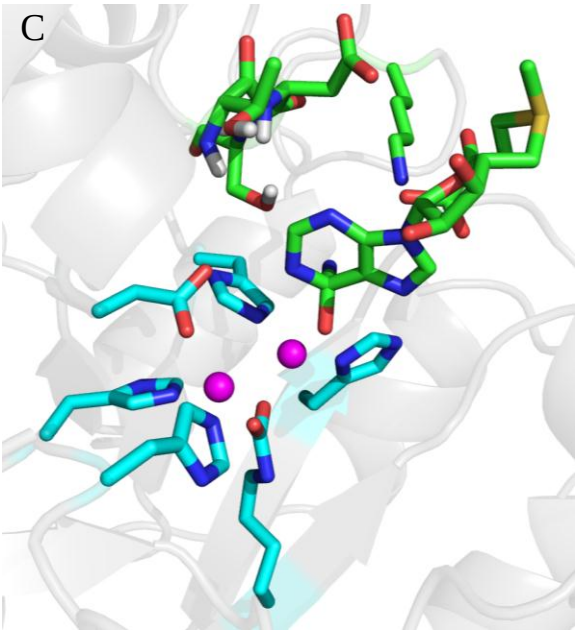


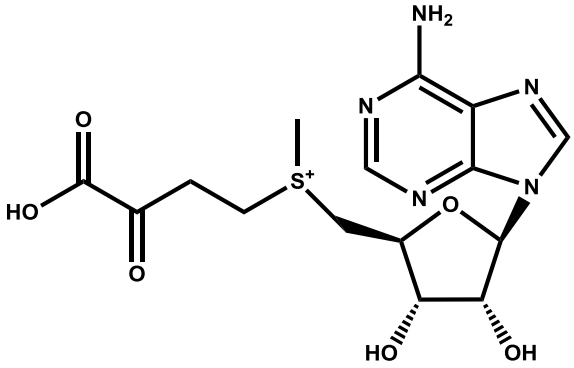
Docking rank = 2
KEGG ID: C03149
N-phosphotaurocyamine





Docking rank = 1
KEGG ID: C04503
N-phosphoguanidinoethyl
methylphosphate





Docking rank = 8
KEGG ID: C04425
S-adenosyl-4-methyltio-2-oxobutanoate

Figure 3.9: Models of initial docking results with EF0837 (PDB: 2ICS)

Activity of EF0837, STM4445, and BCE_5003: The libraries tested with Atu3266 were also used to probe for catalytic activity with enzymes in group 2. It was observed that there was no detectable turnover for EF0837, STM4445 and BCE_5003 for any of the *N*-modified amino acid or dipeptide libraries. **Figure 3.10** shows the results from screening *N*-acetyl- or *N*-succinyl- D- and L- amino acid libraries with Atu3266 and EF0837. These reactions contained the same enzyme, substrate, and buffer concentration carried out for the same period of time. These reactions indicated that EF0837 was not active in the hydrolysis of *N*-acetyl and *N*-succinyl amino acids. The reactions with *N*-formyl and *N*-carbamoyl also failed to show activity. In addition, none of the dipeptide libraries showed signs of hydrolytic activity.

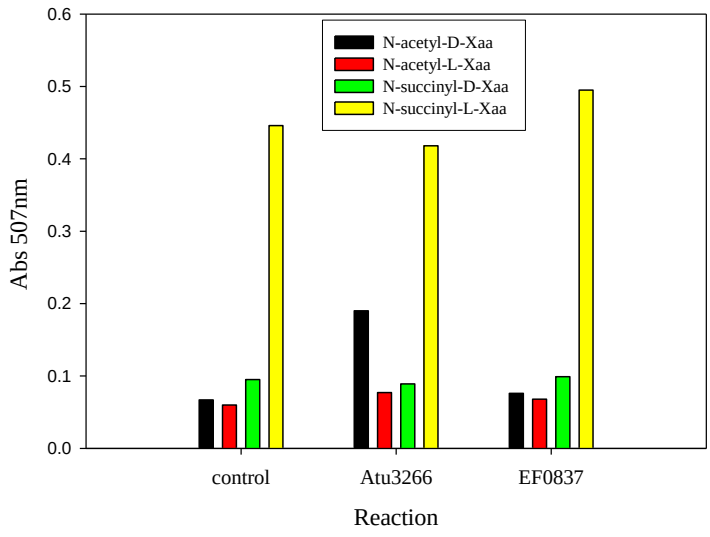
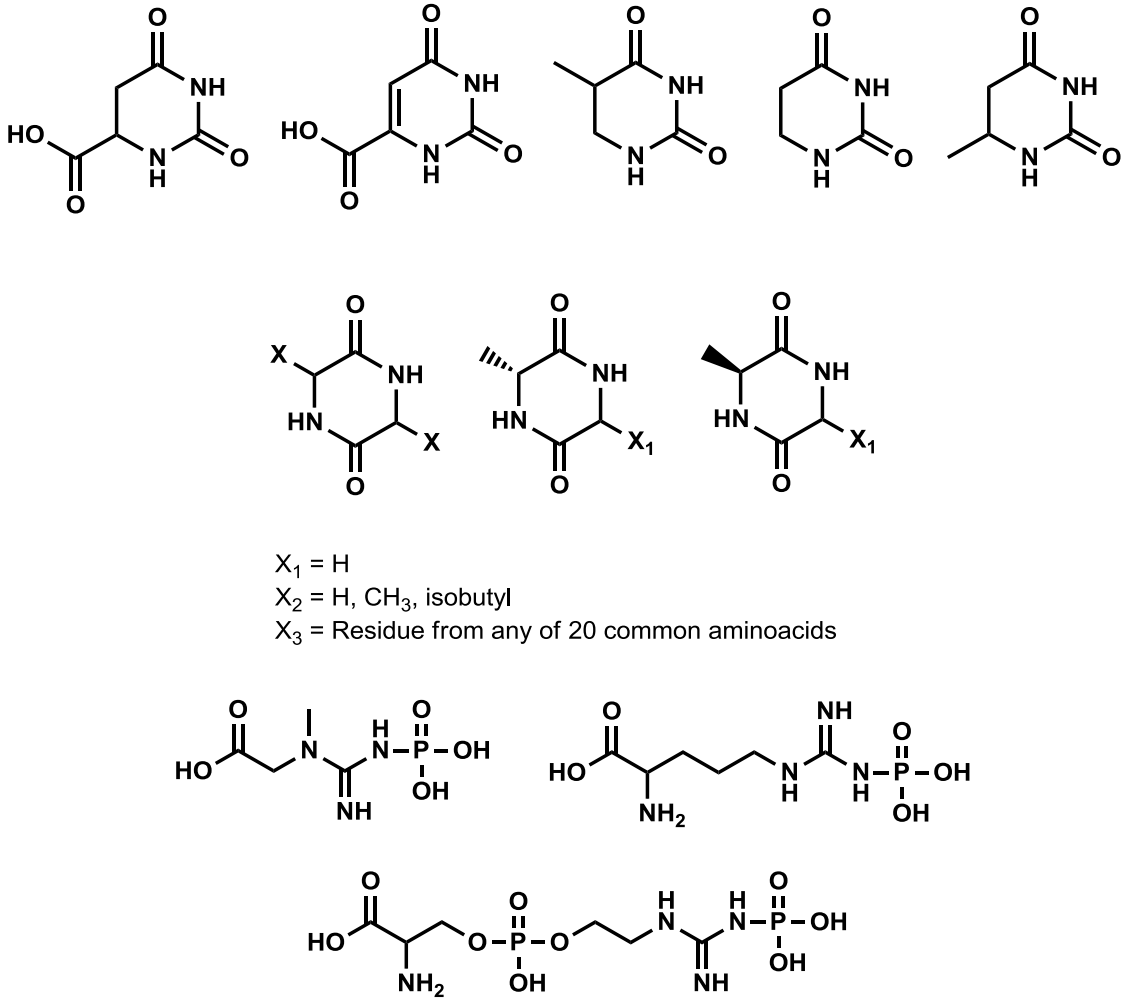
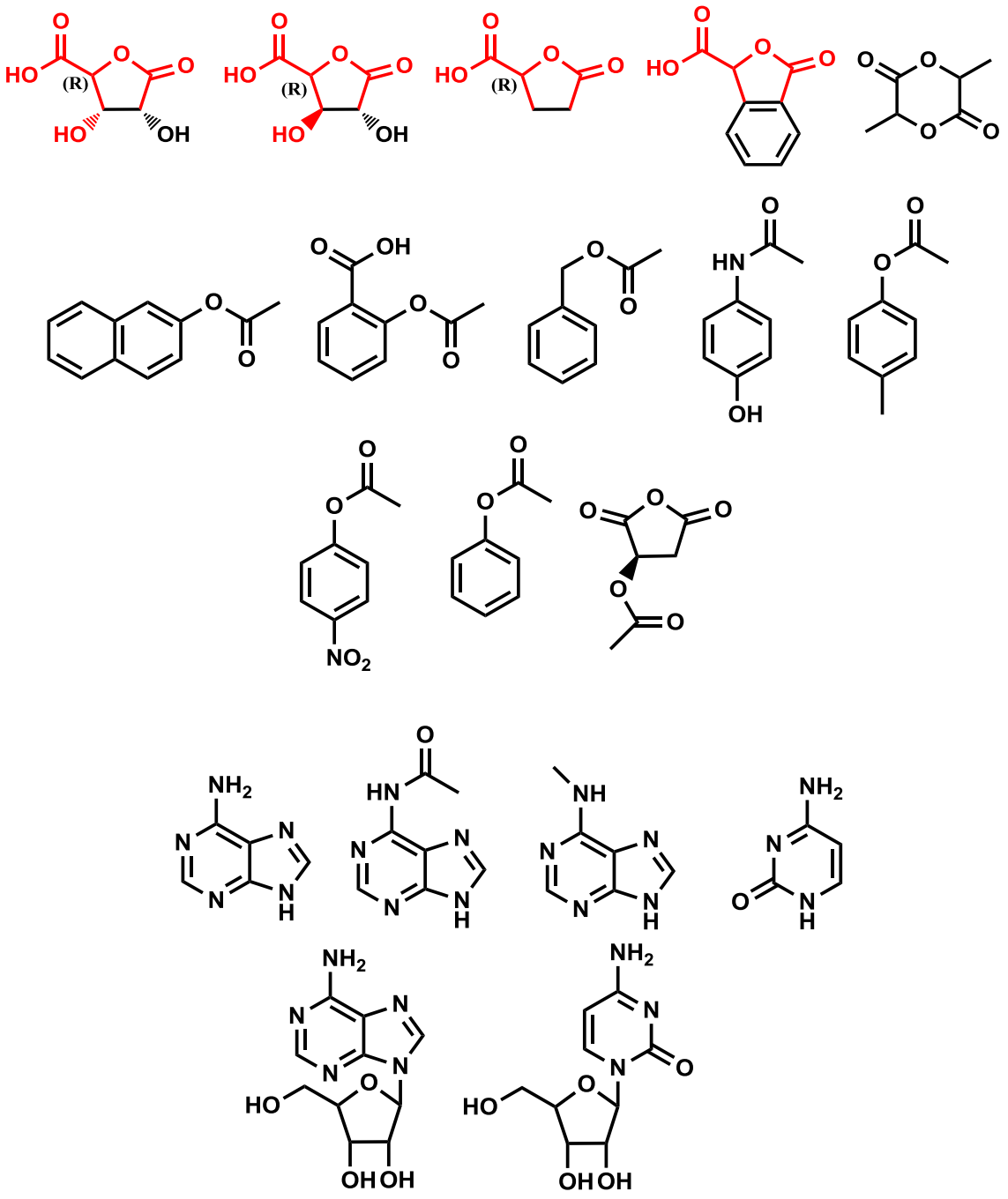


Figure 3.10: Comparisons of initial screening experiments. Reactions were set up for hydrolytic activity with *N*-acetyl-D-/L-Xaa and *N*-succinyl-D-/L-Xaa with Atu3266 and EF0837 and a control reaction. Only sign of activity was presented in Atu3266 at a 4μM enzyme concentration with the *N*-acetyl-D-Xaa library.

Additionally, a variety of lactones, nucleic bases, nucleosides and other nitrogenous bases that are observed in **Scheme 3.3** were screened; none of these compounds seem to indicate the proteins of group 2 were active in their hydrolysis. Since the large portion of the molecules obtained from docking results are not commercially available or readily synthesized, many of the compounds that were tested comprised analogs of molecules obtained from docking results.

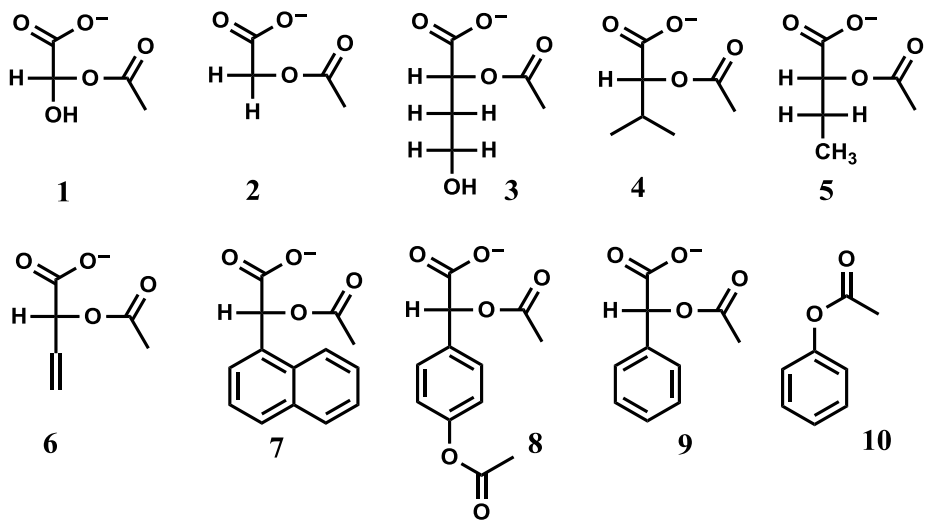


Scheme 3.3: Illustration of compounds screened for activity with EF0837, STM4445 and BCE_5003. These compounds were tested in addition to those in **Scheme 2.2**.



Scheme 3.3 continued.

The compounds in **Scheme 3.3** represent an addendum to the list of compounds that were screened for activity with enzymes in group 6 discussed in the previous chapter (**Scheme 2.2**). The compounds that were found active with group 2 enzymes were the α-acetyl carboxylates that demonstrated to be hydrolyzed by enzymes in group 6. However, not all the acylated α-hydroxyl carboxylates were shown to be active, and all the active compounds exhibited lower rates than are seen with enzymes in group 6. EF0837 did not show any activity with any compound except acetyl-*R*-mandelate, and the enzyme did not reach saturation giving a rate constant V/K of 200 $M^{-1}s^{-1}$. STM4445 and BCE_5003 were 20 times better at deacetylating acetyl-*R*-mandelate than EF0837, but are still two orders of magnitude less efficient than group 6 enzymes. Enzymes in group 2 still showed to be consistently more selective in hydrolyzing some of the acylated carboxylate compounds than are enzymes in group 6. **Scheme 3.4** and **Table 3.3** show the schematics and constants for different substrates that showed minimal activity.



Scheme 3.4: Compounds observed to undergo hydrolysis for proteins in group 2 of COG3964. Only a small number of compounds in the library of α-acylated carboxylates proved to be hydrolysable by EF0837, BCE_5003 and STM4445. The rates for the hydrolysis of these compounds are shown below.

Table 3.3: Kinetic parameters (*V*/*K*) of EF0837, STM4445 and BCE_5003. These enzymes are in group 2 of COG3964. Rates are expressed in $M^{-1}s^{-1}$

Compound	EF0837	BCE_5003	STM4445
1	n.a.	47 ± 4.1	n.a.
2	n.a.	130 ± 10	n.a.
3	n.a.	37 ± 2.2	n.a.
4	n.a.	31 ± 3.0	n.a.
5	n.a.	33 ± 3.2	n.a.
6	n.a.	61 ± 6.0	n.a.
7	n.a.	n.d.	120 ± 12
8	n.a.	n.d.	800 ± 78
9	**200 ± 2**	**(4 ± 0.3) x 10^3**	**(4 ± 0.4) x 10^3**
10	n.a	n.d.	600 ± 60

n.a. No activity – Based on the lowest detectable activity of the acetic acid coupling system, the enzyme was ruled as having no activity. n.d. Not determined – Compounds where activity was not determined due to a long lag period.

As it was observed with EF0837, there was no turnover detection for STM4445 and BCE_5003 for the hydrolysis of *N*-acetyl-D-Xaa. The activity of various α-acyl carboxylates was also tested with these three enzymes and turnovers are well below the rates found for group 6 enzymes of this same orthologous group. The best substrate determined to be hydrolyzed by enzymes in group 2 is acetyl-*R*-mandelate. EF0837 did not show hydrolytic activity with any other compound lacking a phenyl group at the C2 position chiral carbon, nor did it accept substitutions in the phenyl ring moiety.

DISCUSSIONS

Sequence and Structural Analysis of EF0837: The enzymes belonging to COG3964 have been identified as amidohydrolases based on sequence and structure similarities to other enzymes in the amidohydrolase superfamily. The three dimensional crystal structure of EF0837 was determined to a resolution of 2.3 Å. This was necessary as a first step in a structure-based approach to determining enzyme function. The enzyme exhibits a characteristic TIM β-barrel structural motif, enclosing a binuclear metal center that is coordinated by residues originating from the C-terminal ends of the β-strands forming the barrel. The crystal structure also illustrates a characteristic bridging water molecule between the two zinc metal atoms in the active site. Additionally there is an adenine molecule coordinated to the β-metal of the active site. This adenine molecule is 2.4 Å from the β-metal. The exact reason for the presence of this molecule is not well known, discussions with crystallography collaborator Dr. S. Swaminathan, lead investigator of the structural genomics initiative from the

Brookhaven National Laboratory seem to indicate that it is not the result of supplementation of adenine in the protein crystallization phase.

The structural fold and the active site arrangement of EF0837 are characteristic of various members of the amidohydrolase superfamily including dihydroorotase. EF0837 has been annotated as a dihydroorotate or as an adenine deaminase depending on the database used for sequence analysis. The crystal structure of EF0837 has been annotated as an adenine deaminase in the protein data bank, while in the NCBI database it is annotated as a dihydroorotase. The exact reasons for these annotations have not been determined. EF0837, STM4445 and BCE_5003 share less than 20% sequence identity to the characterized dihydroorotase from *E. coli* and less than 16% identity to the characterized adenine deaminase also from *E. coli.* Additionally *E. faecalis* has additional open reading frames that encode genes that are annotated as adenine deaminases and dihydroorotases.

Based on sequence and structural analysis, members of group 2 in COG3964 do not have the secondary coordinating elements that have been observed as necessary in enzymes that carry out the catalytic hydrolysis of dihydroorotate, or the deamination of adenine. In the characterized structure of dihydroorotase (*61, 85*), there is an invariant arginine residue critical for substrate recognition via electrostatic interactions with the exocyclic α-carboxylate group of dihydroorotate. Additionally there is an asparagine and a histidine that interact with the same functional group of the substrate. Mutation of any of these residues results in loss of activity (*85*). This demonstrates a high degree of selectivity in the binding of the substrate, whereas the enzymes in group 2 of COG3964

do not have these critical residues to justify the accessibility and coordination of a dihydroorotate molecule.

In binuclear adenine deaminases (*63*), a bridging glutamate replaces the carboxylated lysine found in group 2 enzymes. In addition, there are two invariant residues: a glutamate after β-strand 6 and an aspartate following β-strand 8 that if mutated, the results are significantly diminished catalytic activity by adenine deaminase. These residues are not present in group 2 - COG3964 enzymes. Whether enzymes in COG3964 have evolved to form a new assembly of residues that carry out these activities, is a possibility, however based on the lack of activity in adenine deamination, dihydroorotate hydrolysis, or activity with any other analogous compound that was tested, it is suggested that these enzymes have a completely different catalytic role.

Operon Analysis: Identification of the functional roles of enzymes in COG3964 may be supported by the gene context within the genomes of all the organisms encoding an amidohydrolase in this COG. As it was presented in **Figure 3.5**, the schematics of the open reading frame show the constant presence of an annotated selenocysteine synthase gene adjacent to the gene encoding an amidohydrolase of group 2. Selenocysteine synthase is a PLP-dependent enzyme that carries out the conversion of seryl-$tRNA^{Sec}$ to selenocysteinyl-RNA^{Sec} for selenoprotein biosynthesis. Whether this information provides a link to the functional role of EF0837, it would need to be further investigated. Based on sequence alignments between the putative SelA proteins that are neighbors to the amidohydrolases of COG3964, and sequences of characterized SelA proteins that synthesize selenocysteine, it is observed that characterized selenocysteine synthases are

larger monomeric units. Further analysis of this group of enzymes will be discussed in chapter V. Considering the substrate profiles that have been observed thus far with proteins in COG3964, it can be presumed that there is some basis to infer interactions between COG3964 amidohydrolases and SelA proteins. The first detection of activity for enzymes in COG3964 was found from screening a library of *N*-acetyl-D-amino acids, and although the rates are well-below the expectation of activity for amidohydrolase enzymes, it may be possible that we have not exhausted all the possible pool of modified amino acids and/or peptide combinations as substrates. Aside from the constant presence of the SelA annotated gene in all organisms containing a group 2 amidohydrolase, and various other members of COG3964, there is no other evidence that these two genes are dependent on one another for function.

Catalytic Function of Group 2 in COG3964: The functional role of enzymes in group 2 of COG3964 has not been fully elucidated. No other enzyme in this COG has been characterized to date, and aside from group 6 that was previously characterized in the preceding chapter, there are no validated functional roles. There is however the conclusion that the enzymes in group 2 do not carry out the hydrolysis of dihydroorotate or the deamination of adenine. Not only were these two substrates screened, but also a variety of analogous compounds that have been identified as legitimate substrates in other orthologous proteins to the true adenine deaminases (*63, 78*) and dihydroorotases (*61, 85*).

Additional cyclic nitrogenous compounds were screened for activity based on BLAST analysis to other amidohydrolases at very low stringency values, as well as the

architectural similarities to other primary target compounds. For example, various dihydroorotate analogs, such as dihydrouracil, orotic acid and a library of available diketopiperazines of L-Ala-L-Xaa and D-Ala-L-Xaa, were screened, none of which were able to be hydrolyzed by the enzymes of group 2. Docking results also suggested a wide variety of compounds that were predicted to undergo enzymatic dephosphorylation, deamination, and ring opening hydrolysis. None of these compounds were observed to turnover in the presence of EF0837, STM4445 and BCE_5003.

It is apparent that docking models for functional annotation are severely disadvantaged. The metabolites used for activity prediction are obtained from a single source of likely substrates, the KEGG database (*129*). This is problematic because metabolites that are not part of this database are not considered as likely substrates; such was the case for acetyl-*R*-mandelate. Conversely, any compound from the KEGG database that has generic properties of substrates for amidohydrolases can be a potential substrate.

Acetyl-*R*-mandelate was the only compound that was hydrolyzed by all three enzymes tested from group 2. Because the rates are significantly lower than those observed with other enzymes in group 6 of COG3964, as well as other members of the amidohydrolase superfamily, it can be inferred that this is perhaps not the correct substrate for enzymes in group 2. However it can also be assumed that the true substrate will have very similar characteristics as acetyl-*R*-mandelate, and that enzymes in group 2 carry out the deacetylation of other likely compounds. Compounds that did not feature a phenyl ring at the chiral C-2 position of the substrate did not show activity with EF0837,

and were much less active with BCE_5003, and STM4445; whereas compounds that had substitutions at the phenyl moiety of acetyl-*R*-mandelate, were much less active than the un-substituted compound. The substrate profile for group 2 enzymes of COG3964 is not fully elucidated, but clearly there are secondary elements in this group that are responsible for a different selection of substrates than those found in group 6; however, those where not identified in these studies.

CHAPTER IV

FUNCTIONAL DIVERSITY IN COG3964: SEARCHING AND ASSESSING THE FUNCTIONAL ROLES OF OTHER AMIDOHYDROLASES

Finding the functional relationships between proteins of homologous sequences at low identity values could prove to be an invaluable strategy that enhances the understanding of complex biological systems. However, given the defined gap in associating sequence to structure to function (*6*), it may be that finding sequence-structure-function relationships will require the additional genomic sequencing of a larger number of organisms, and supplementary structural studies of gene products, to discern the functional relationships between organisms with orthologous proteins. Looking at the functional relationships between orthologous proteins in organisms has proven to be in some cases an arduous task. This, in part, is due to the parameters that are selected to apply functional assignment to newly sequenced genes, and its reliability almost entirely on the degree of sequence homology to a previously annotated protein, regardless of whether or not this parental template protein has been experimentally characterized (*5*).

This case is observed in the functional designation of COG3964, belonging to the amidohydrolase superfamily. This cluster of enzymes has been identified in the GenBank (*2*), as well as other related databases as having a combination of functional annotations, neither of which is correct. Although the active site residues that coordinate to the binuclear metal center are conserved between genuine dihydroorotases and the

members of COG3964, there are important secondary structural elements that are not conserved in COG3964 sequences that must account for the difference in substrate selectivity and specificity. Other groups within this cluster have been previously discussed in the preceding chapters. Here, the search for functional annotation of COG3964 extends to a group of proteins that have not shown affinity for any of the substrates that have been previously discussed. Additional protein structure homology modeling and docking experiments have been proposed to determine if there are additional compounds within the KEGG database that can prove to be substrates to this group of enzymes. **Figure 4.1** depicts the sequence similarity network in relation to the group of enzymes discussed in this chapter.

Two enzymes in group 7 are the focus of functional characterization in this chapter. These two targets represent a group of enzymes that share greater than a 70% sequence identity between its members. Xaut_0650 (gi|154244602) from *Xanthobacter autotrophicus* Py2 and blr3349 (gi|161511104) from *Bradyrhizobium japonicum* USDA110, were identified as targets for the functional investigation of group 7 enzymes in COG3964. These enzymes were cloned, expressed and purified by the Enzyme Function Initiative (*4*). The EFI consortium has identified these proteins by EFI numbers. EFI-500436 pertains to gene blr3349, and EFI-501595 is the identification assigned to gene Xaut_0650. Here, the enzymes are identified by their locus tag. The goal for these protein targets is to develop a substrate profile much in the same manner as it has been successfully done with enzymes in group 6 of this COG. Enzymes in

group 7 share less than 26% sequence identity with Atu3266 or EF0837, two enzymes that were observed active in the hydrolysis of the ester moiety of α-acetyl carboxylates.

No enzyme in group 7 of COG3964 has been structurally characterized. Based on sequence alignment models, group 7 enzymes are expected to retain the same metal coordinating ligands observed in the active sites of EF0837 (PDB: 2ICS) and Atu3266 (PDB: 2OGJ). The conserved binuclear active site consists of an α-metal coordinated by a His-X-His motif from β-strand 1, an aspartate from β-strand 8 and a bridging carboxylated lysine from β-strand 4. This lysine also bridges to the β-metal, which is additionally coordinated by a His from β-strand 5 and a His from β-strand 6. Xaut_0650 and blr3349 have failed to demonstrate activity with any compound that has been previously tested in enzymes from group 2 and group 6 of COG3964.

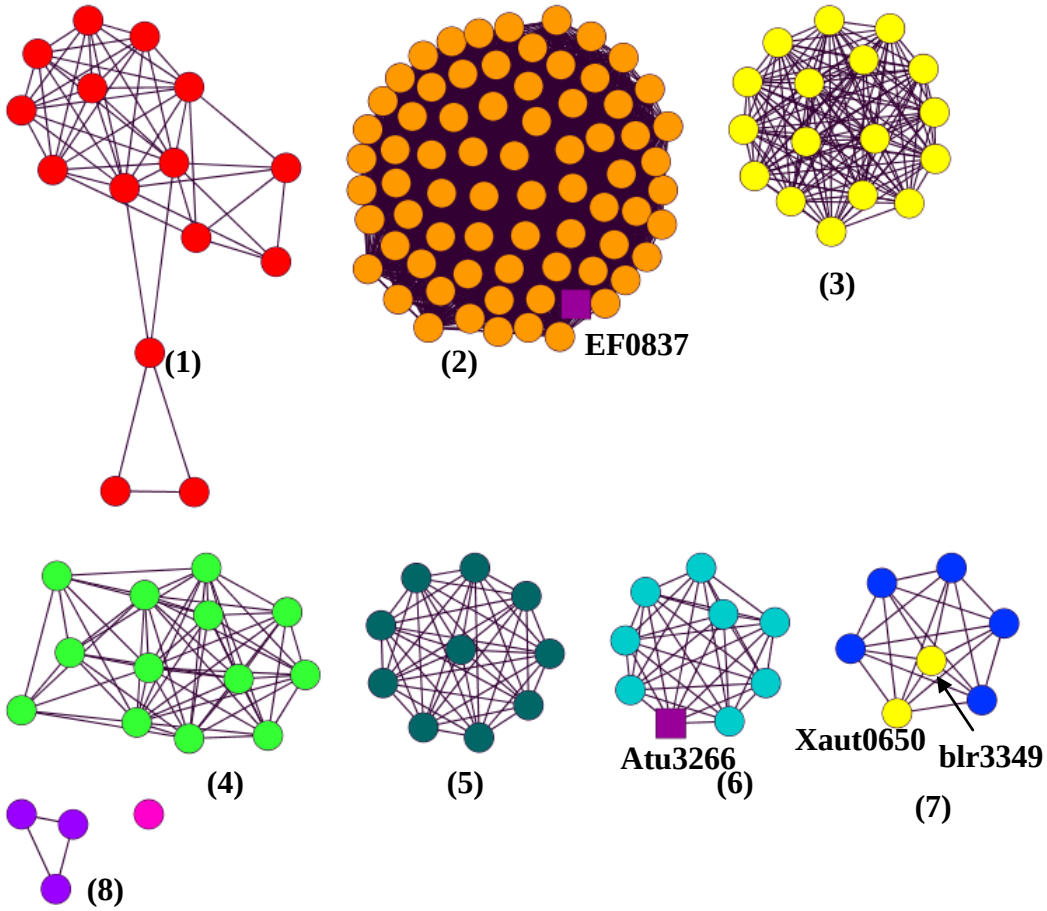


Figure 4.1: Sequence similarity network of COG3964 with group 7 enzymes. Atu3266 as shown in group 6 and EF0837 from group 2 have been previously discussed in chapters 2 and 3 respectively. In this chapter the focus is turned into the EFI-acquired proteins from group 7. Xaut_0650 and blr3349 show less than 26 % sequence identity to the amidohydrolases from group 2 and group 6.

An alignment **(Figure 4.2**) between the sequences of the target proteins from group 7 and those that belong to Atu3266, Oant2987, and EF0837 shows that there are two inserting loops in the group 7 sequences. These inserts follow β-strand 5 and β-strand 8. The sequence alignment between enzymes in group 7, 6 and 2 also present the conserved residues coordinating the active site metal center, as well as those residues that diverge from those discussed as potential substrate recognizing ligands in the sequences from group 2 and 6. Protein modeling experiments carried out for Xaut_0650 by Magdalena Korczynska from the Shoichet lab at UCSF, suggested the loop inserts are in close proximity to the active site.

Preliminary docking results carried out for the protein homology models of enzymes in group 7 showed various aminoglycosides bound in the active site. The structural homology model for the group 7 member, Xaut_0650, utilized the crystallized structures of a dihydropyrimidinase and hydantoinase. Based on these analyses, a large library of hydantoins, pyrimidines, dihydropyrimidines and diketopiperazines were screened for activity with Xaut_0650 and blr3349. Other compounds that were suggested from docking results consisted of 5'-acetylphosphoadenosine, and based on the interacting *N*-acylated sugar moiety of aminoglycosides, a small compiled library of *N*-acetyl-D-sugars was also screened for activity.

In addition to homology models and docking, the operon context is once again interrogated to investigate the neighboring genes adjacent to the gene of interest. Most enzymes in group 7 of COG3964, with one specific exception, display a different genomic operon than those observed previously; one that does not contain a SelA

(selenocysteine synthase) open reading frame in the vicinity of the amidohydrolase gene. The operon analysis for the majority of the organisms in this group contain a series of gene clusters involved in carbon monoxide utilization (*CoxMSL*) and in the co-localization of branched chain amino acids (LIV-I/LS). Only one organism displays an operon encoding a SelA gene in proximity to the amidohydrolase. *Xanthobacter autotrophicus* Py2 does have an annotated SelA (Xaut_0658) gene eight reading frames upstream from the amidohydrolase (Xaut_0650) encoding gene. This is the only case where any of the organisms in group 7 show a neighboring selenocysteine synthase encoding gene in close proximity to the annotated amidohydrolase. Interestingly enough, this annotated SelA is adjacent to selenocysteine containing proteins (formate dehydrogenase) as well as additional factors involved in selenocysteine biosynthesis (SelD factor).

The strategy for the functional annotation of amidohydrolase enzymes is further employed to determine the functional roles of proteins in group 7 that have diverged in their sequence identity from other members of COG3964 and now show different substrate specificity. In this chapter we focus on the exploration for functional determination of group 7 amidohydrolases. The search for available compounds is extended from the early library compiled of *N*-modified amino acids to α-acetyl carboxylates, compounds that were found to have initially low activity and were subsequently strategically modified to demonstrate catalytic competent rates with amidohydrolases found in COG3964.

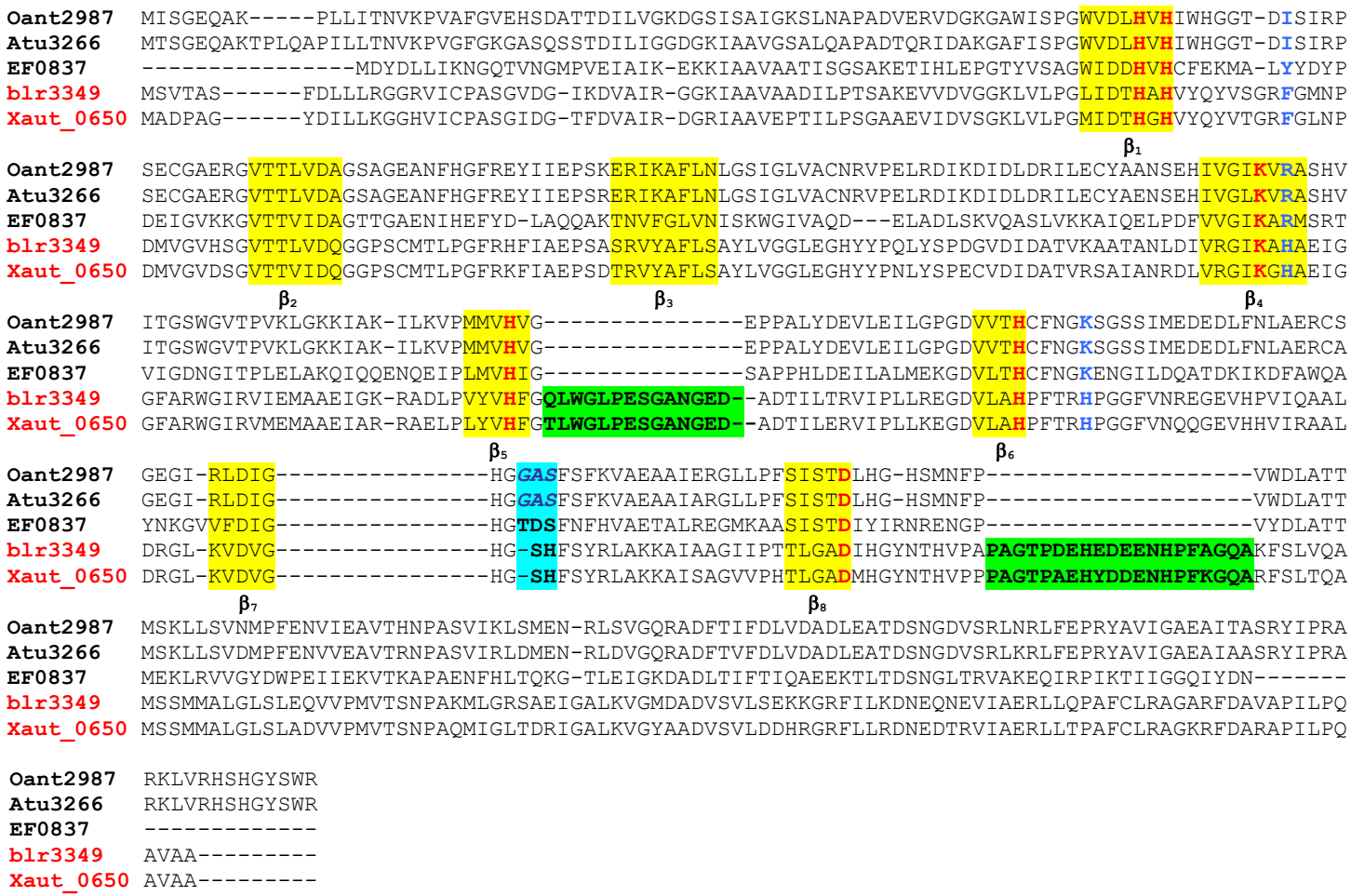


Figure 4.2: Sequence alignment with selected group 7 enzymes (in red font). Additional sequences are those from groups 2 and 6. Highlighted in yellow are the residues within the β-strands forming the TIM-barrel including those residues coordinating the binuclear metal center (red). Highlighted in green are the inserts found in sequences from group 7. Highlighted in cyan is the triad from loop 7 interacting with carboxylate of compounds active with group 6 enzymes. In blue font are residues within close distance to the active site that may be responsible for substrate diversity within COG3964.

MATERIALS AND METHODS

Materials: The compounds tested for hydrolytic activity with enzymes in group 7 were previously utilized in the functional annotations of Atu3266 and EF0837. Many of these compounds were obtained from Sigma-Aldrich unless otherwise stated. Our laboratory has complied a library of dipeptide libraries (D-Xaa-L-Xaa, L-Xaa-L-Xaa, and L-Xaa-D-Xaa), as well has an exclusive library of alanine diketopiperazines (cyclic-L-Ala-L-Xaa and cyclic-D-Ala-L-Xaa) that were used to screen enzymes in group 7 of COG3964. All *N*-acetyl-D-sugars, including 2-acetamido-3-deoxy-D-glucose were purchased from Carbosynth. The purified enzymes utilized for functional annotations were supplied by the Enzyme Function Initiative as part of the larger investigation to discern the function of various COGs within the amidohydrolase superfamily. The targets for functional annotation were selected based on genomic DNA availability for cloning the desired proteins.

Screening: Direct assays were set up to screen for hydrolysis of various hydantoin derivatives, dihydroorotate derivatives, and diketopiperazines by taking the spectra of the compound in the range of 190-300 nm. The greatest absorbance for these compounds was in the range of 220-230 nm. Previous assays monitor the activity of DHO at 230 nm for dihydroorotate and at 225 nm for dihydropyrimidinase activity (*85*). Each assay set for initial screening contained 25 mM phosphate buffer at pH 7.5, and 1 mM of the compound to be tested. The reaction was started by addition of 1-3 μM enzyme in 20 mM TRIS buffer pH 7.6. Upon addition of the enzyme, a spectral analysis was then re-taken to observe for changes from the original spectrum. Deacetylation of

5’-acetylphosphoadenosine and other *N*-acetyl-D-sugars was monitored using the KACETAF acetic acid assay kit from Megazyme™. Our libraries of modified amino acids and dipeptides were also screened by methods previously discussed in the preceding two chapters.

Sequence Alignment and Homology Models: A sequence alignment was designed using the amino acid sequences from the two enzymes purified for functional analysis belonging to group 7 of COG3964 (Xaut_0650 and blr3349), with the sequences from previously discussed enzymes, one belonging to group 2 (EF0837) and two from group 6 (Atu3266 and Oant2987). There are seven sequences comprising group 7 based on the sequence similarity network at a BLAST *E*-value 10^{-70}. The protein similarity network based on the BLAST results was created using Cytoscape (*82*). The sequence alignment was constructed using CLUSTALW alignment program (*137*) in the biology workbench database from the San Diego Supercomputer center (http://workbench.sdsc.edu/). Alignments were also performed exclusively to the sequences belonging to group 7. As previously mentioned, group 7 contains only 26% sequence identity to Atu3266 and EF0837. However, Xaut_0650 and blr3349 share 80% sequence similarity. Additional proteins in this group have a minimum 70% sequence identity to Xaut_0650 or blr3349. The protein homology model for Xaut_0650 was constructed based on the crystal structure of a dihydropyrimidinase, a hydantoinase and a putative dihydroorotase. These proteins also belong to the amidohydrolase superfamily, but not all have been functionally characterized. The models used included PDB:1GKR, a characterized L-hydantoinase form *Arthrobacter aurescen*s (*137*); PDB: 2FVK, a characterized

dihydropyrimidinase from *Saccharomyces kluyveri* (*138*); and a putative dihydroorotase from *Porphyromonas gingivalis* PDB: 2GWN. These enzymes share less than 23% sequence identity to those belonging to group 7-COG3964.

Operon Interrogation: The amino acid sequences in group 7 of COG3964 were collectively annotated as dihydroorotases. There are seven amino acid sequences in this group at the BLAST value that was selected for functional characterization of the complete COG. Analysis of the genomic context of the organisms encoding for the amidohydrolases in COG3964 was carried out using the MicrobesOnline database (*135*).

RESULTS AND DISCUSSIONS

Protein Concentration and Metal Analysis: Protein concentration was determined by measuring the absorbance with a SPECTRAmax PLUS-384 spectrophotometer. Extinction coefficients were determined at 280 nm for each of the proteins in the functional investigation of group 7. EFI target 500436 corresponds to the locus tag blr3349. The extinction coefficient for this protein is $\varepsilon = 24{,}300\ M^{-1}cm^{-1}$, while EFI target 501595, locus tag Xaut_0650 had an extinction coefficient $\varepsilon = 30{,}620\ M^{-1}cm^{-1}$. The molecular weights for each protein are 44,951 daltons and 45,521 daltons respectively. These parameters were obtained based on the sequence for each protein using the ExPASy protein parameter tool: http://web.expasy.org/protparam/. Below is table 4.1 listing the determined metal content in each enzyme obtained by inductively coupled mass-spectrometry.

Table 4.1: Metal content of enzymes from group 7. Each purified enzyme was analyzed by inductively coupled plasma emission mass spectroscopy. Each quantity represents equivalents per monomer of protein.

Enzyme	Zn^{2+}	Fe^{2+}	Mn^{2+}	Ni^{2+}	Cu^{2+}	Total
Blr3349	**1.9**	**0.3**	**<0.1**	**0.1**	**<0.1**	**2.3**
Xaut_0650	**0.5**	**0.1**	**1.1**	**<0.1**	**<0.1**	**1.6**

Follow up experiments for substrate screenings were supplemented with 2-4 equivalents of metal ($ZnCl_2$). This supplementation was not observed to improve the signs of activity for the enzymes tested.

Sequence Alignment Models with EF0837, Atu3266 and Oant2987: A sequence alignment between the two representative enzymes from group 7 and the previously discussed enzymes from groups 2 and 6 is illustrated in **Figure 4.2**. This alignment highlights all the protein metal-bound ligands and their ubiquity in the sequences across all the groups that have been assigned to COG3964. These metal coordinating residues are conserved in all the sequences belonging to COG3964. There are secondary elements that have not been probed or determined to account for the substrate diversity or specificity between all the arranged groups in this cluster. The amino acid sequences for proteins belonging to group 7 are missing a variable arginine that is found one residue away from the carboxylated lysine found at the end of β-strand 4. This residue is conserved in all sequences in groups 1, 2, 5 and 6. Group 7 replaces this variable arginine residue with a histidine. In addition, a lysine conserved in Atu3266, EF0837 and enzymes in their respective groups, found after the β-strand 6, five residues away

from the histidine that coordinates the β-metal, is also replaced in group 7 enzymes with a histidine. Most evident from sequence comparisons, is the triad of residues forming the loop after β-strand 7. In Atu3266, the backbone chain of a Gly267-Ala268-Ser269 triad has been found in docking models to interact with the carboxylate moiety from the α-acylated compounds assigned as substrates. Based on the alignments model, a glycine deletion is observed in group 7 sequences. Additionally, there are serine and histidine side chains replacing the last two residues of the motif Gly-Ala-Ser (→ Xxx-Ser-His). The residues in the loop of group 7 enzymes are much bulkier and have a polar side chains forming a loop that is predicted as the coordinating interaction and placement of the substrates hydrolyzed with other COG3964 enzymes. Sequence alignments also show two additional insertions in the sequences of enzymes from group 7. One is a long extended loop that follows β-strand 5; the other is found at the end of β-strand 8. Homology modeling for Xaut_0650 with other amidohydrolases shows that these loops extend out of the active site of the enzyme.

Preparation of Structure Homology Models of Xaut_0650 for Docking Experiments: Protein homology models were prepared by Magdalena Korczynska from the Shoichet lab at the University of California, San Francisco. Homology models were generated in the absence of an available crystal structure representing the enzymes from group 7. Sequence alignments demonstrated that there were two extended loop insertions at the end of β-strand 5 and 8. These loops are absent in the crystal structures of two enzymes in COG3964: the structure from Atu3266 from group 6, and the structure from EF0837 belonging to group 2. The two inserts were determined to be near the active site.

Figure 4.3-A represents the homology model of Xaut_0650 with the non-ordered loop inserts shown in multicolor, these loops were predicted to have multiple conformations.

These non-ordered loops were modeled after insertions that were seen in structures of other amidohydrolases: PDB codes 1GKR, 2FVK and 2GWN. 1GKR represents the crystal structure for a functionally characterized L-hydantoinase from *Arthrobacter aurescens* (*137*). 2FVK is the crystal structure of a dihydropyrimidinase from the yeast *Saccharomyces kluyveri* (*138*), while 2GWN is the crystal structure of a non-characterized putative dihydroorotase from *Porphyromonas gingivalis*. **Figure 4.3-B** illustrates the insertions present in Xaut_0650 and all enzymes in group 7 in contrast to the template proteins for modeling. Insertions in Xaut_0650 are modeled based on insertions found in the crystal structures of 1GKR, 2FVK and 2GWN. This model was prepared for initial docking experiments to increase the size of the library of compounds to screen amidohydrolases in group 7 for functional activity.

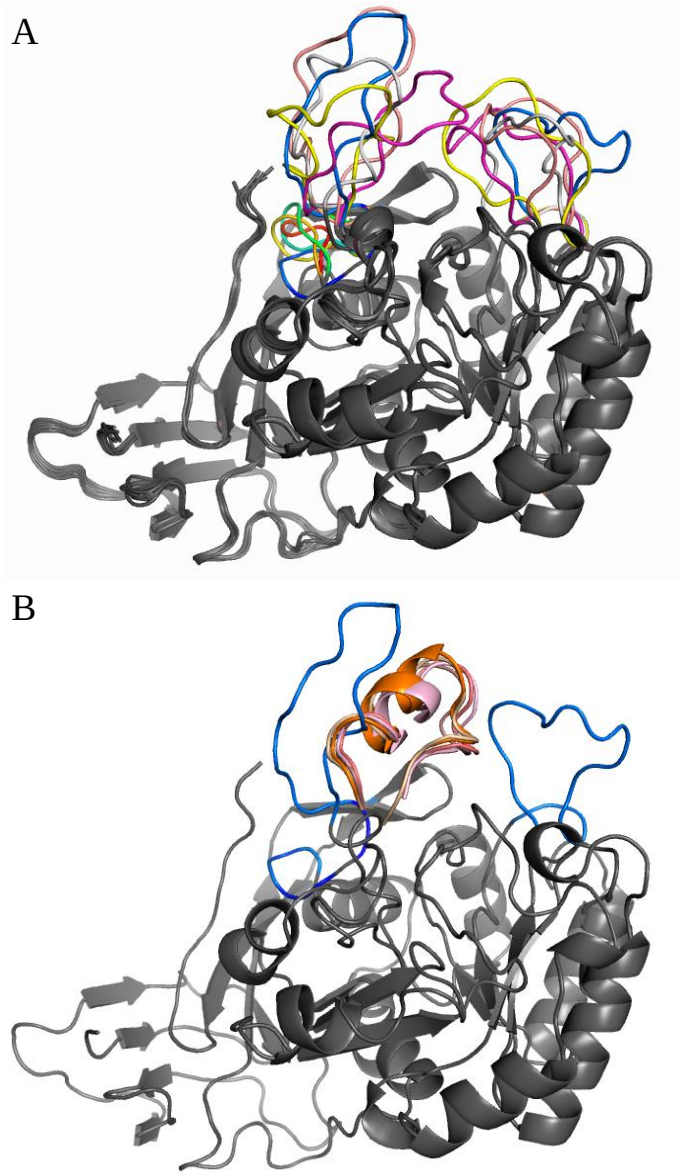


Figure 4.3: Protein homology models for Xaut_0650. (**A)** Various conformations of the homology model of protein Xaut_0650 with multitude of conformations for the inserting loops after β-strand 5 and 8. (**B)** Homology model of Xaut_0650 with the inserting loops from structures used to make model PDB: 1GKR, 2FVK and 2GWN.

Screening Analysis of Xaut_0650 and Blr3349: All the compounds previously tested for activity with enzymes in group 2 and group 6 of COG3964, were also screened for activity with the available enzymes from group 7. Xaut_0650 and blr3349 failed to show detectable activity for the hydrolysis of *N*-acetyl, formyl, succinyl and carbamoyl D- and L- amino acids. Activity was also undetectable with the library of dipeptides. All α-acetyl carboxylates (D- and L- isomers when available), hydantoins, dihydroorotate analogs, dihydropyrimidines, diketopiperazines and lactones discussed in previous chapters were screened as described, with no signs of activity. Initial docking results found substrates that were added to the growing library of molecules to test with proteins from COG3964. The molecules detected from docking results included 5'-acetylphosphoadenosine as well as sugar moieties from various aminoglycosides. **Figure 4.4** shows the results from these docking experiments.

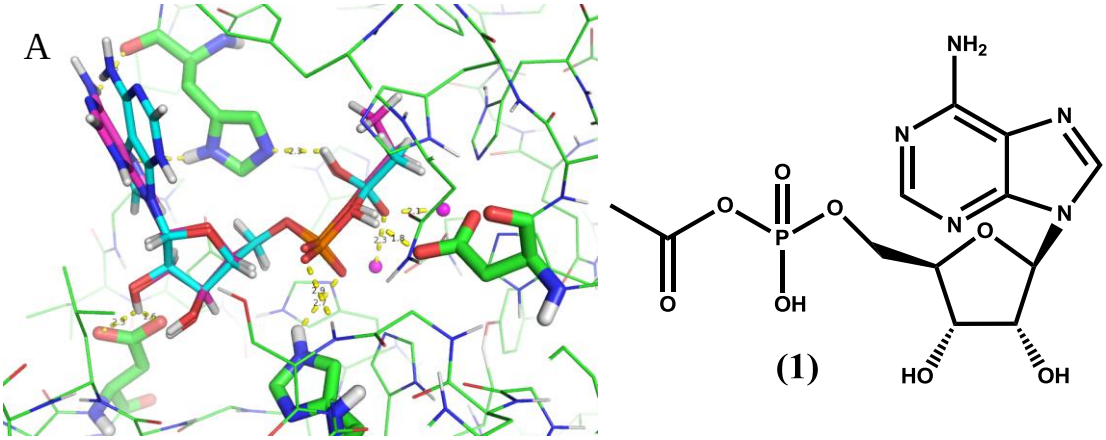
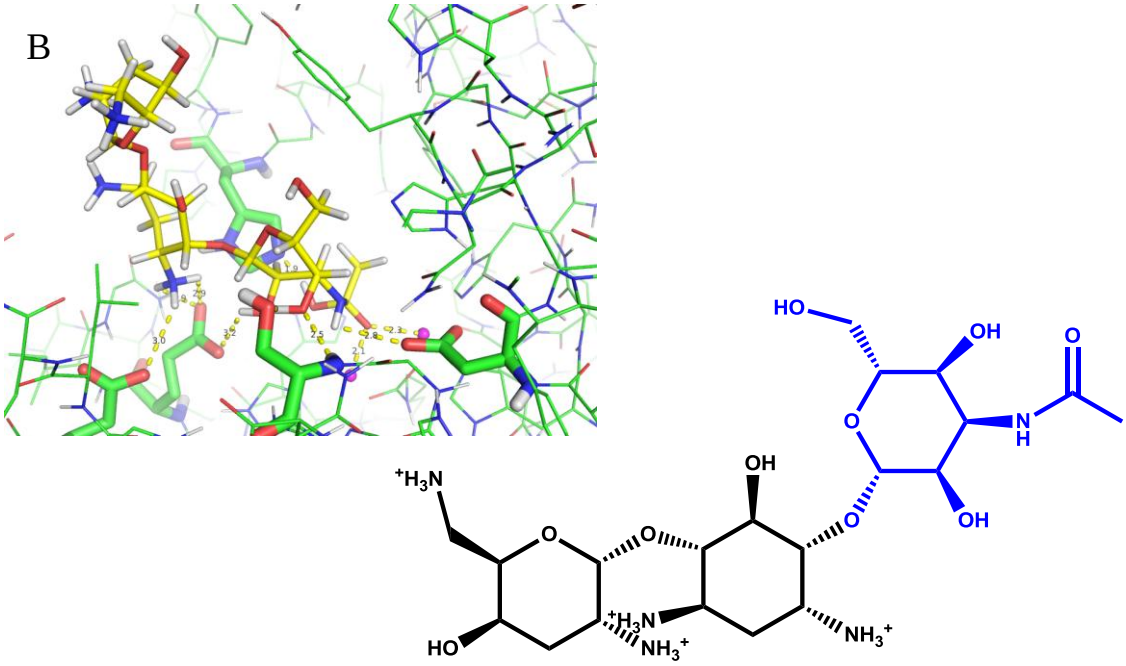


Figure 4.4: Docking results with homology model of Xaut_0650. (**A**) Docked model of 5'-acetylphosphoadenosine. (**B**) Interaction of *N*3'-acetylbramycin (KEGG ID-C03010) and (**C**) *N*3'acetylgentamycin (KEGG ID-C03009).



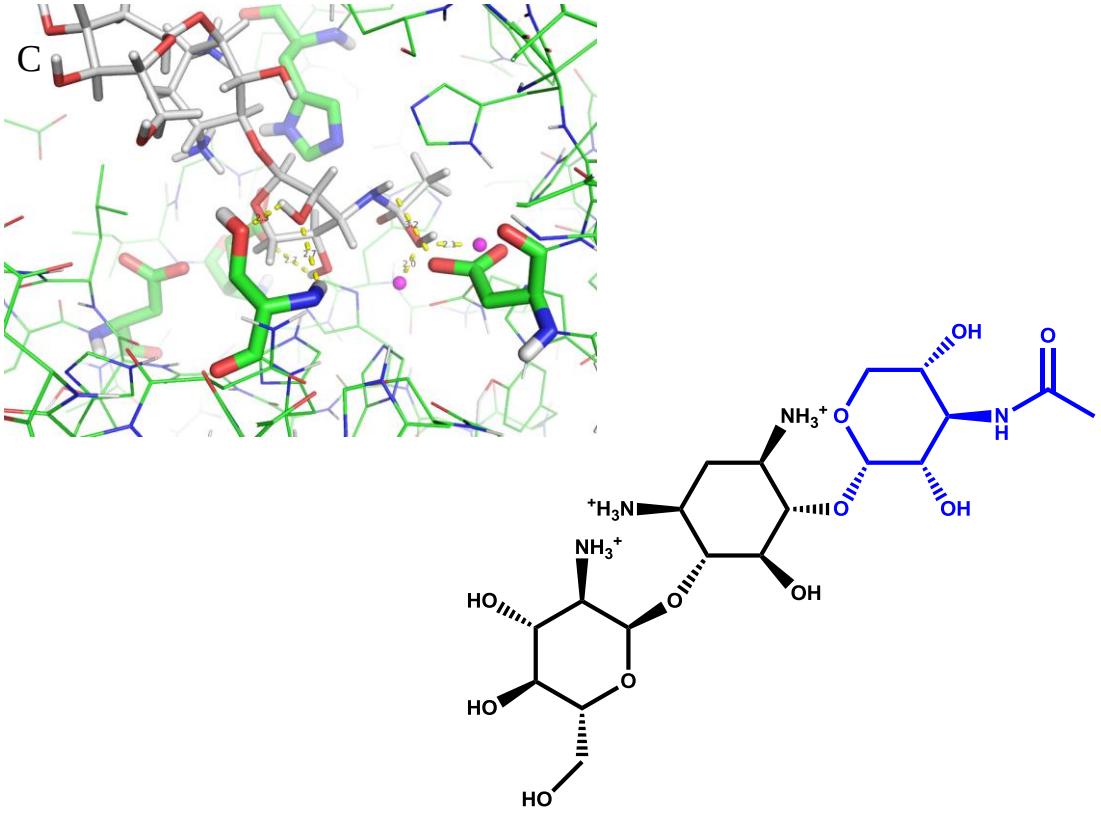
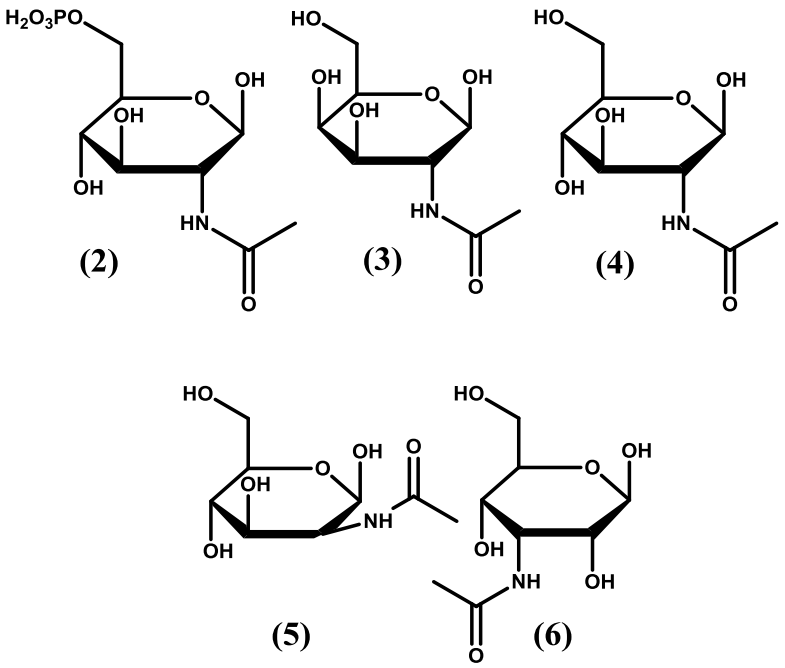
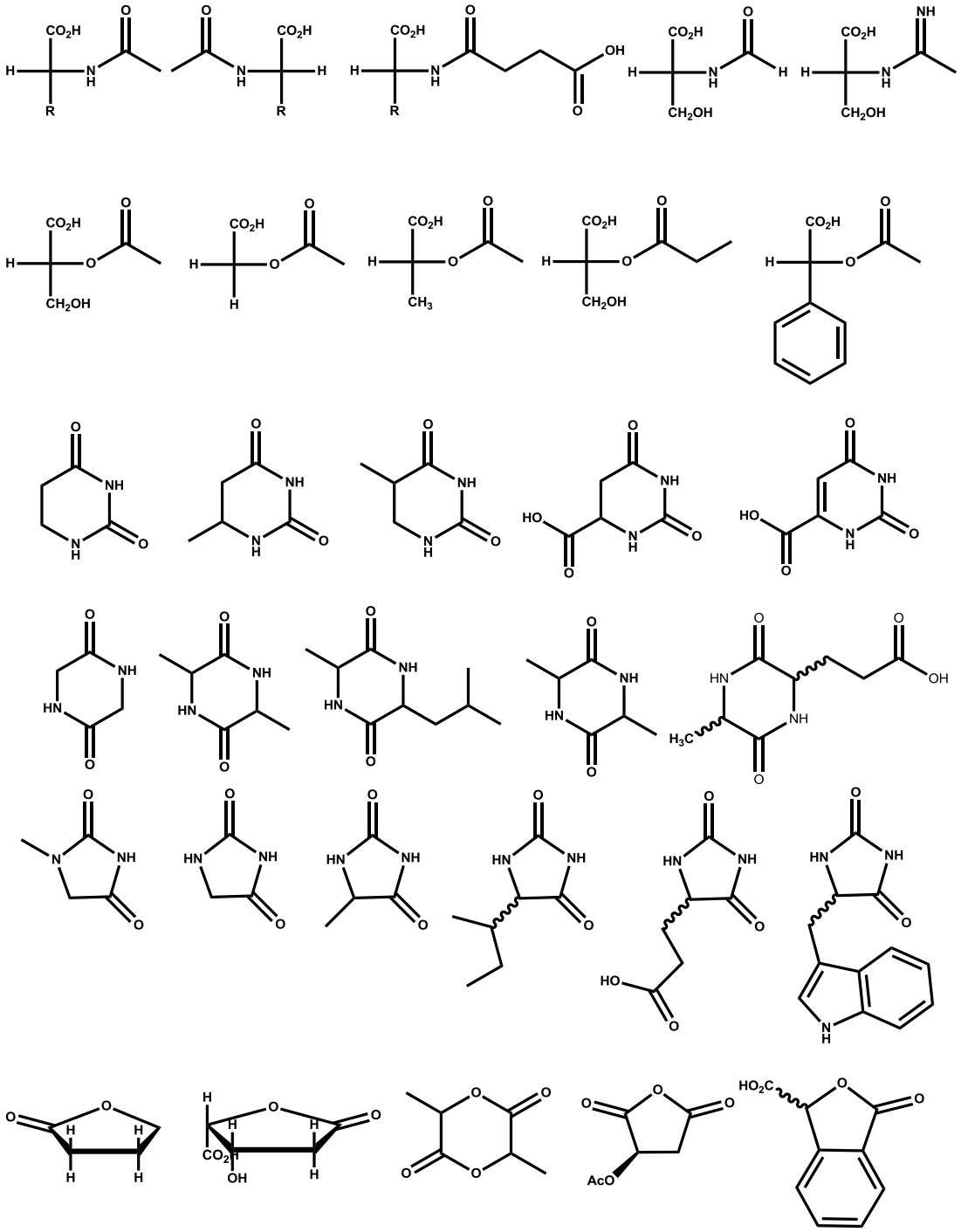


Figure 4.4 continued.

Based on these results, 5'-acetylphosphoadenosine **(1)** was screened for deacetylation showing no detection of acetate in the coupling system used. A small library of *N*-acetyl-D-sugars was compiled to test for deacetylation as well. This library included: *N*-acetyl-glucosamine-6 phosphate **(2)**, *N*-acetyl galactosamine **(3)**, *N*-acetyl glucosamine **(4)**, *N*-acetyl mannosamine **(5)** and 3-acetamido-3-deoxy-D-glucose **(6)** (**Scheme 4.1**). These compounds were selected based on the interacting sugar moiety found in the docked aminoglycosides. None of these compounds proved to be deacetylated by Xaut_0650 or blr3349. **Scheme 4.2** summarizes the type of compounds that were screened for activity with group 7 enzymes with no success.



Scheme 4.1: Compounds tested for deacetylation of *N*2' and *N*3' acyl group with enzymes Xaut_0650 and blr3349.



Scheme 4.2: Representative compounds of substrates tested for activity with group 7 enzymes.

Operon Context Analysis: Unlike previous genomic analyses of groups 2 and 6 from COG3964, the organisms encoding proteins from group 7 do not present a neighboring selenocysteine synthase gene (SelA) in the vicinity of the amidohydrolase open reading frame. Looking at the genomic operon for organisms encoding a group 7 amidohydrolase does not facilitate the determination of functional annotation of the enzymes in this group (**Figure 4.5**). Generally present in the operon encoding the COG3964 group 7 proteins, are two conserved adjacent gene clusters that carry out different reactions. One gene cluster, *coxMSL*, encodes for a molybdenum-containing iron-sulfur flavoprotein carbon monoxide dehydrogenase system (*139*). This system is composed of large, medium and small subunits of carbon monoxide dehydrogenase encoded by the structural genes *coxL*, *coxM*, and *coxS* respectively (*140*). The other cluster found in the organisms encoding the COG3964 gene is an annotated LIV-I/LS transport system. This active transport system is comprised of binding proteins (LS) and membrane components (LIV-I) operating in the uptake of amino acids. The chromosomal locus including the *liv* gene cluster consists of *livJ*, *livK*, *livH*, *livM* and *livF* genes in that order (*141*). This cluster is characterized in the HAAT family of transporters for the uptake of branched-chain amino acids. The LIV-I system in *E.coli* works in the uptake of leucine, isoleucine, valine, threonine, serine and alanine (*141*) and in other studies, the transport of phenylalanine (*142*). How these clusters facilitate the understanding of the functional roles of the amidohydrolases is not understood. Below is an illustration of the genomic operon for organisms encoding an amidohydrolase enzyme belonging to group 7.

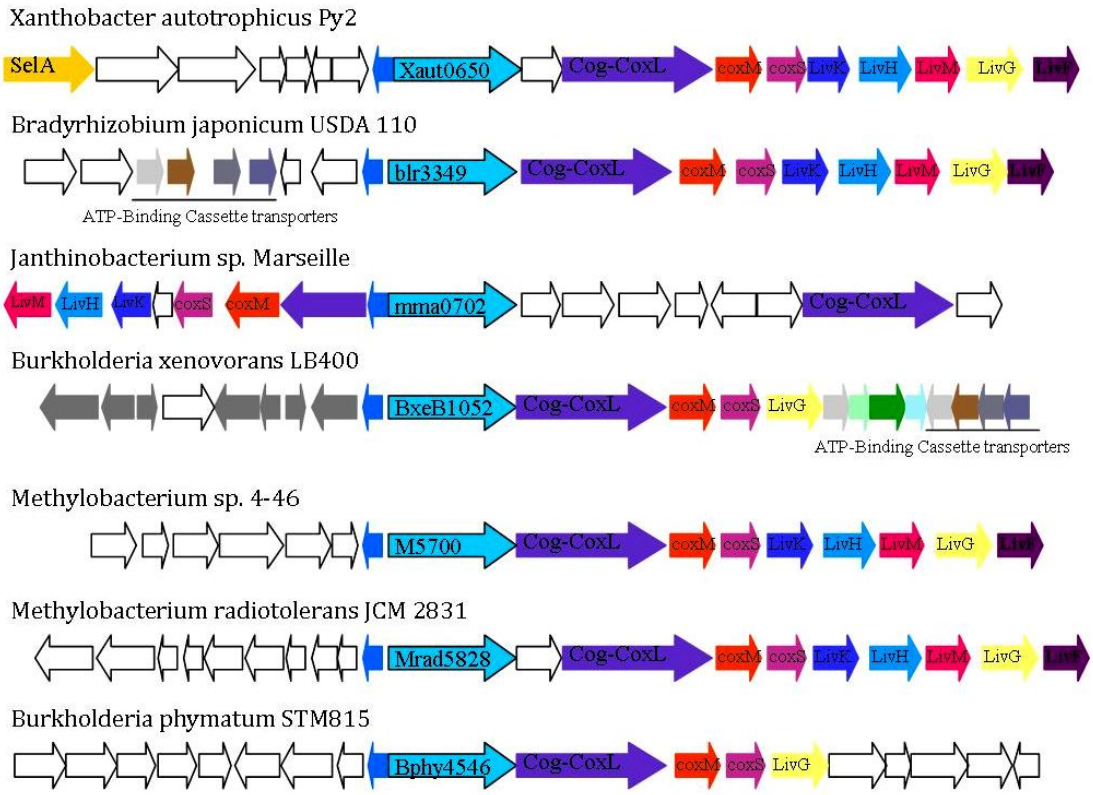


Figure 4.5: Genomic operon of organisms with group 7 enzymes. Illustration shows all the COG3964 amidohydrolases from group 7 (shown in cyan arrow with black outline). These are identified by their locus tag. In the organisms shown, only one contains an annotated SelA gene (*Xanthobacter autotrophicus* Py2). There is a sequence of genes that encode for a molybdenum containing carbon monoxide dehydrogenase (*cox*MSL) involved in carbon monoxide metabolism (*139*). Cluster Liv*KHMGF* is part of the Liv-I/LS system, a branched-chain amino acid transporter (*141*, *142*).

Only one organism encoding an amidohydrolase from COG3964 in group 7 contains a selenocysteine synthase encoding gene in proximity. *Xanthomonas autotrophicus* Py2 encodes an annotated SelA gene in the locus tag Xaut_0658. This gene is ascribed to the cluster of orthologous groups (COG) 1921. All of the SelA annotated genes previously described in chapters 2 and 3, belong to this COG. Collectively, COG1921 is annotated as PLP-dependent seryl-tRNA^{Sec} selenium transferase enzymes. Although, much in the same way as annotated amidohydrolases from COG3964, there is much speculation about the correct functional annotation about the SelA genes found as neighbors to the amidohydrolases that are part of COG1921. In *Xanthomonas autotrophicus*, the SelA reading frame is succeeded by genes that encode for various subunits of the selenocysteine incorporating protein formate dehydrogenase (Xaut_0659-Xaut_0662) (*143, 144*).The formate dehydrogenase encoding genes for various subunits of the enzyme are often observed in the open reading frame of other organisms with a genuine SelA open reading frame. These genes however, are not observed in the genomic contexts that encode for an amidohydrolase from COG3964, and selenocysteine synthase from COG1921 that is adjacent to the amidohydrolase. The SelA gene found in *Xanthomonas autotrophicus* is also found in the vicinity of an additional component for selenocysteine biosynthesis. SelD (locus tag Xaut_0666) is a selenide water kinase. This protein is responsible for the synthesis of selenophosphate (*145-147*) and activation of selenium towards its incorporation in seryl-tRNA^{Sec}. The figure below (**Figure 4.6**) illustrates the placement of all these genes from SelD (Xaut_0666) to the amidohydrolase (Xaut_0650).

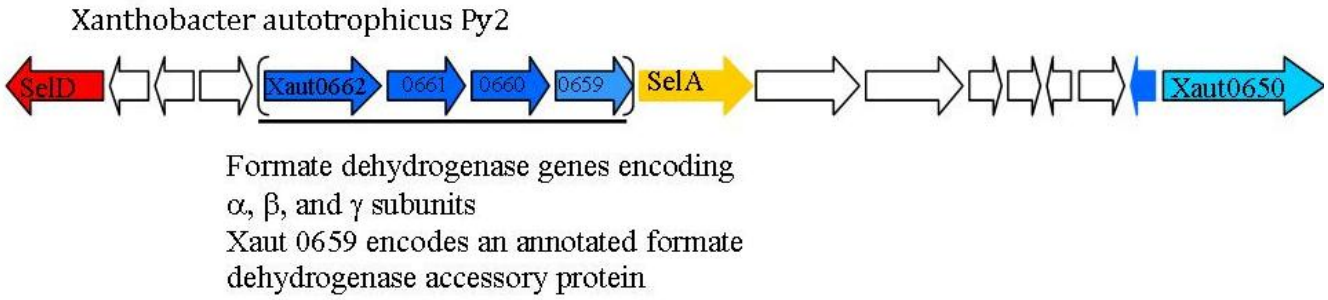


Figure 4.6: Gene operon context of Xaut_0650 from *X. autotrophicus* Py2. Operon illustrates additional proteins in the biosynthesis (SelA, SelD) and incorporation (Xaut_0659 – Xaut_0662) of selenocysteine.

Predicting the Functional Role of Xaut_0650 and Blr3349: The organization of the bulk of amidohydrolases assigned to COG3964 using similarity networks constructed by Cytoscape (*82*) has contributed to the functional analysis of this cluster of orthologous groups. It is observed that the substrate diversity is variable even between groups in each COG. COG3964 was organized based on BLAST values, and it was observed that at an *E*-value 10^{-70} specific groups are formed that can be studied as separate entities. At the specified BLAST value, it is expected that there is ~40% similarity between the sequences within each group. This has proven an interesting analysis, as not only has it been evident that there are sequence similarities that are conserved only within each group, but this translates into similar structural and eventually functional roles within individual groups in the COG.

In this chapter, group 7 of COG3964 was analyzed for functional activity. Within this group, two enzymes were adopted for functional studies. Xaut_0650 and blr3349 are the representative proteins from a group that contains 5 additional members. These seven enzymes share a minimum 70% sequence identity, and it is expected that they carry out the same metabolic reaction. Xaut_0650 and blr3349 share 80% sequence identity. These enzymes were analyzed for metal content and were further supplemented with $ZnCl_2$ to account for any possible metal deficiency. From the large library that has been developed to screen activity with other members of this COG, no compound was observed to be modified in the presence of the enzymes belonging to group 7. The list of compounds tested entailed acylated pyranose sugars, amino acids, nitrogenous compounds including nucleotides, nucleosides, diketopiperazines, pyrimidine-analogs and hydantoins, sugar lactones, and a sub-class

of α-acetyl carboxylate compounds that were synthesized. This library of compounds proved active with other proteins belonging to groups 2 and 6 of this COG. However, Xaut_0650 and blr3349 only share between 26 – 27% identity with any of the sequences belonging to group 2 or group 6, sequences for enzymes that were able to deacetylate α-acetyl carboxylates.

The most noticeable differences extrapolated from sequence alignments between sequences discussed in chapters 2 and 3 belonging to groups 6 and 2 respectively, and those assigned to group 7 are the long insertions found at the end of β-strand 5 and β-strand 8. In the absence of a crystal structure, these inserts can be assumed to be long disordered loops that can shield the metal center active site. In addition there are single residue differences that can account for substrate diversity between enzymes in group 7 and those in groups 2 and 6. Group 7 enzymes lack an arginine residue that is conserved after the bridging carboxylated lysine in enzymes belonging to group 2 and 6. This arginine is substituted with a histidine residue in all sequences from group 7. Additionally, a lysine found at the end of β-strand 6 that has been observed in the crystallized structure of Atu3266 (PDB: 2OGJ) and EF0837 (PDB: 2ICS) is also substituted in the sequences from group 7 with a histidine residue.

The three-amino acid residues found in the loop between β-strand 7 and α-helix 7 ($Gly^{266}Ala^{267}Ser^{268}$) have been observed in the docking models of Atu3266 to coordinate to the carboxylic acid moiety of various α-acetyl carboxylates. Compounds containing an α-acetyl moiety were hydrolyzed in other enzymes of COG3964, but not those that belong to group 7. In group 7 sequences; the Gly-Ala-Ser triad from group 6 sequences is replaced with Xxx-Ser-His with the first position shown as a deletion of

Gly266. This is followed by an Ala267Ser substitution and more importantly a Ser268His substitution. These substitutions place bulkier, polar residues that may have disrupted the backbone interactions with the carboxylate moiety of α-acetyl carboxylate compounds. In order to better understand the loop insertions and residue modifications a homology model for Xaut_0650 was created, but in-depth analysis have not been followed up.

Initially, the homology model for Xaut0650 was to be constructed based on the structures of Atu3266 (PDB: 2OGJ) and EF0837 (PDB: 2ICS). However, the sequences for these proteins lack the two inserting loops found in group 7 enzymes. The crystal structures of other amidohydrolases were instead used to refine the homology model. 2FVK, 1GKR and 2GWN have been annotated as a dihydropyrimidinase, L-hydantoinase and putative dihydroorotase respectively. Only the first two enzymes have been functionally characterized. The inserting loops for the homology model of Xaut_0650 were modeled after the loops extending from the structures of 2FVK, 1GKR and 2GWN. Interestingly, an alignment between the sequences used to create the homology model and the sequences from group 7 do not align the β-strands expected to form the barrel or the metal coordinating ligands. Based on the designed homology model for Xaut_0650 initial docking experiments were carried out to search for additional compounds in the KEGG database.

The docking results suggested a new line of compounds that had not been previously tested. One of the best fitting substrates was 5’-acetylphosphoadenosine. In addition there were a variety of aminoglycosides docked at the active site that showed favorable interactions with surrounding residues. 5’-acetylphosphoadenosine was tested

with no detectable activity. Because certain aminoglycosides were not purchasable, the single ring molecule of the aminoglycoside was acquired and tested for activity. The single ring molecule was the portion of the large aminoglycoside that was observed from docking models to interact with the active site. None of the *N*-acetyl sugars tested for activity were shown to be hydrolyzed.

The results from docking experiments are based on the number of compounds selected from the KEGG database to generate HEI. The best docked compounds in the homology model of Xaut0650 were not observed to be active. Whether docking is a step in the right direction to find substrates and consequentially functions for proteins in group 7, cannot yet be determined. Docking has been incorporated as an addition to an arsenal of other strategies to investigate the functional annotations of various groups in the amidohydrolase superfamily, and here for functional determination of group 7 of COG3964. One of the drawbacks with docking results is that many of the compounds observed to have the most favorable interactions at the active site of the protein structure may not be readily available for in-vitro testing. Additionally, only the compounds that have been obtained from the KEGG database are candidates for initial docking studies, and so this excludes a myriad of other compounds that can be possible substrates. A case can be made for acetyl-*R*-mandelate, a compound that was active with enzymes from group 2 and 6 of COG3964, which is not part of the KEGG database of metabolites. Acetyl-*R*-mandelate was tested after a series of modifications that began with low-level hydrolytic activity of Atu3266 with *N*-acetyl-D-serine. Whether acetyl-*R*-mandelate is the bona fide substrate Atu3266, Oant2990, and other

enzymes in COG3964 were designed to hydrolyze is not a likely scenario, but the enzymes did exhibit a high level of hydrolytic activity.

From the large selection of compounds incorporated into a library to test additional enzymes in COG3964, Xaut_0650 and blr3349 failed to show activity in the hydrolysis of any. Additional experiments to assist in the functional annotation of group 7 enzymes may require expression profiles, cloning of other neighboring genes, and the expansion of compound or substrate libraries. A systematic method has been employed to assist the functional determination of the amidohydrolases in COG3964, and it is expected that the genuine substrate for group 7 enzymes is not very different from that found to be hydrolyzed by enzymes in group 6; however, the functional role of amidohydrolases in COG3964 is still an area to be developed.

CHAPTER V

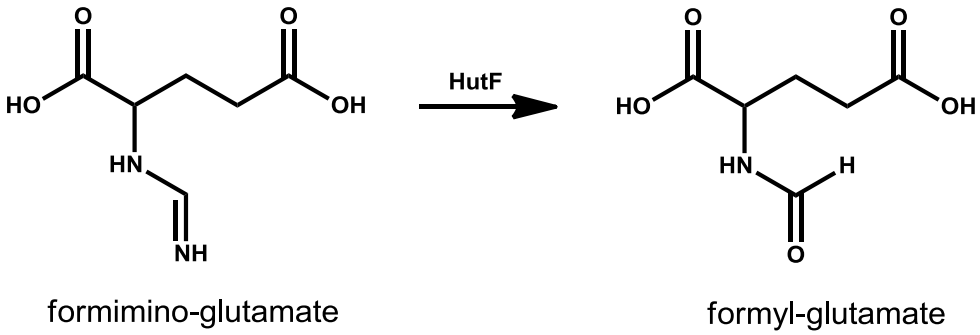
INSIGHTS INTO OPERON PROTEINS FOR FUNCTIONAL ANNOTATION OF ENZYMES IN COG3964: ASSESSING THE FUNCTIONAL RELATIONSHIP BETWEEN COG3964 AND COG1921

Function prediction and annotation of individual gene products, and their role as part of a larger and more complex machinery is fundamental in the understanding of biological systems. The adjacent functional linkages or presence of common phylogenetic profiles of open reading frames in the genomic operon of different organisms serves in various instances to help reveal the functional roles of less obvious gene products, or those that demonstrate sequence similarity below advisable threshold values for functional comparison (*90, 148-150*). This is especially useful when genes in operons associated in metabolic pathways interact with each other based on their physical assembly and location. The information to decode functional relationships can be extracted both experimentally (*151-154*) and computationally (*91*, *149-151*). Computational methods have evolved beyond the traditional methods of sequence homology, which seek correlations between amino acid sequences. One method used in rationalizing the functional linkages between proteins is searching the positions of specific genes within chromosomes (a gene neighborhood method), as well as determining the correlations for the inheritance of protein pairs across a multitude of species (phylogenetic profiles). These methods fundamentally incorporate expression patterns that can be assessed experimentally or the isolation of multi-domain complexes

that can facilitate identification of individual functional roles as various systems require additional components to function (*155, 156*).

Genomic context information can be utilized when analyzing individual reading frames to assign a functional characterization. Within the amidohydrolase superfamily, there are two examples where gene cluster analysis has proven effective as means to interrogate the functional roles of proteins.

N-formimino-L-glutamate iminohydrolase (HutF) from *Pseudomonas aeruginosa* PA01 (Pa5106, gi|15600299), catalyzes the formation of *N*-formyl-L-glutamate and ammonia from *N*-formimino-L-glutamate (*98*). This reaction is the penultimate step in the degradation of L-histidine to L-glutamate. Pa5106 was first identified as a member of the amidohydrolase superfamily based on a comprehensive amino acid sequence comparison. Initially, this enzyme was annotated as a probable chlorohydrolase; upon further examination of the genomic context, it was observed that the gene was in close proximity to genes involved in the histidine degradation pathway. HutH, HutU and HutI are involved in the conversion of *L*-histidine to *N*-formimino-L-glutamate. HutF or Pa5106 was observed to be an enzyme involved in one of the possible three pathways for the degradation of *N*-formimino-L-glutamate; its role is shown in **scheme 5.1.**



Scheme 5.1: Reaction catalyzed by AHS enzyme HutF. This enzyme is involved in one of the three possible pathways of histidine degradation.

A representation of the genomic context of the histidine degradation pathway for *P. aeruginosa* is presented in **Figure 5.1**. This analysis provided functional insights to the possible functional roles of Pa5106.

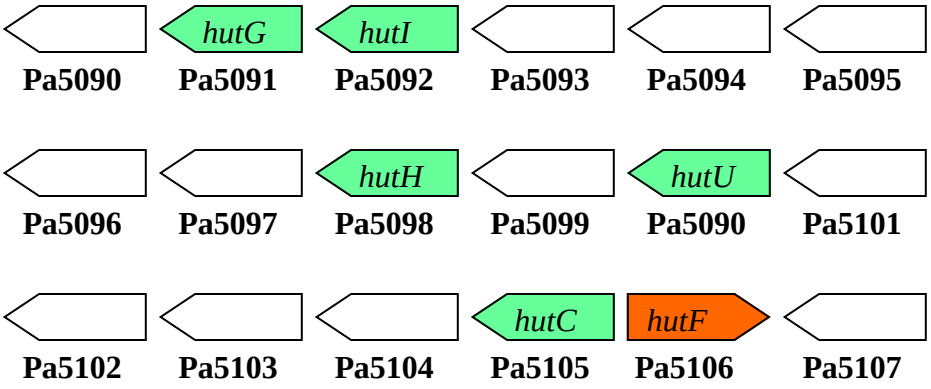


Figure 5.1: Gene operon arrangement of Pa5106. The open reading frame illustrated in the figure contains the genes of the operon involved in histidine degradation in *P. aeruginosa*. The locus tag identification is provided for each of the identified genes (HutG, HutI, HutH, HutU, HutC and HutF). The amidohydrolase coding gene HutF is distinguished in orange (*98*).

Another member of the amidohydrolase superfamily, more functionally remote and sequence divergent is the uronate isomerase (URI) identified from *Bacillus halodurans* C-125 (Bh0493). Uronate isomerase (UxaC) catalyzes the isomerization of D-glucuronate to D-fructuronate and D-galacturonate to D-tagaturonate. These isomerization reactions represent the first step in the separate metabolic pathways for the utilization of D-glucuronate and D-galacturonate. Compared to the amino acid sequences of other protein members in the amidohydrolase superfamily, Bh0493 was identified as an amidohydrolase based on weak sequence similarity (<19%) with the structurally characterized uronate isomerase from *Thermatoga maritima* (Tm0074, gi|15642839). Additional insights for the annotation of Bh0493 as an amidohydrolase enzyme were found in the active site residues originating from the ends of β-strands **1** and **8**, and the presence of a single zinc ion in the M_α active site of Bh0493. Adjacent genes expressed in the operon encoding uronate isomerase are expected to catalyze the subsequent reductions of D-fructuronate and D-tagaturonate to D-mannonate and D-altronate by UxuB and UxaB, respectively, and then dehydration by UxuA and UxaA to 2-keto-3-deoxy-D-gluconate. The function of Bh0493 was strongly supported by the gene context within the genome of *B. halodurans* (*58*, *59*).

There was an additional open reading frame in *B. halodurans* involved in the metabolism of uronic acids. Bh0705, Bh0706 and Bh0707 were identified as UxaC, UxuA, and UxuB respectively. UxuB is a D-mannonate oxidoreductase, and UxuA a D-mannonate dehydratase. These annotations led to the eventual cloning, purification, and characterization of Bh0493 and Bh0705. **Figure 5.2** is a diagram comparing the open

reading frames of the genes in the metabolism of uronic acids in *B. halodurans* and in *E.coli* K12. The comparison in the phylogenetic profiles of these two organisms contributed to the successful characterization of Bh0493 and Bh0705 as isomerases in the metabolism of D-glucuronate and D-galacturonate.

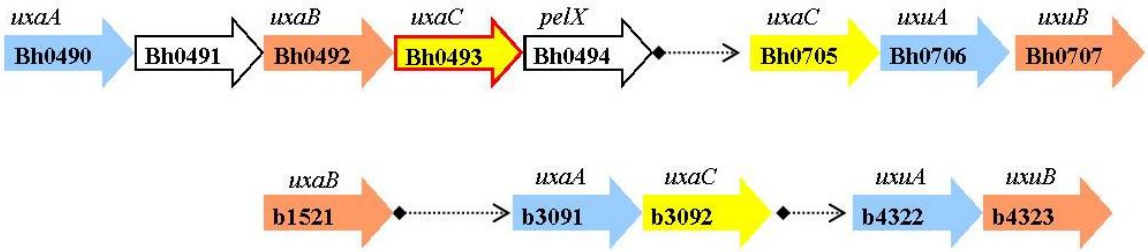


Figure 5.2: Gene operon arrangement of Bh0493. Image illustrates the chromosomal arrangement of the genes that encode for enzymes involved in the metabolism of D-glucuronate and D-galacturonate for *B.halodurans* (top) and *E.coli* K12 (bottom) (*59*).

The previous studies have served as examples in the use of gene adjacency in genomes as strategies to assign functional roles to proteins with poor sequence homology to other characterized proteins, or to proteins with functional misannotations. In the search for the correct functional roles of the amidohydrolase superfamily, we have extended functional genomics as means to understand and uncover the functional roles of COG3964. A large number of the amidohydrolases assigned to COG3964 contain a neighboring gene annotated to COG1921, also known as a SelA gene. Only groups 3 and 4, from the Cytoscape network generated at a BLAST *E*-value 10^{-70} for COG3964 sequences, fail to show an adjacent gene annotated as a SelA. The SelA gene is found at different locations in respect to the COG3964 assigned gene as seen in the phylogenetic patterns presented in **Figure 5.3**.

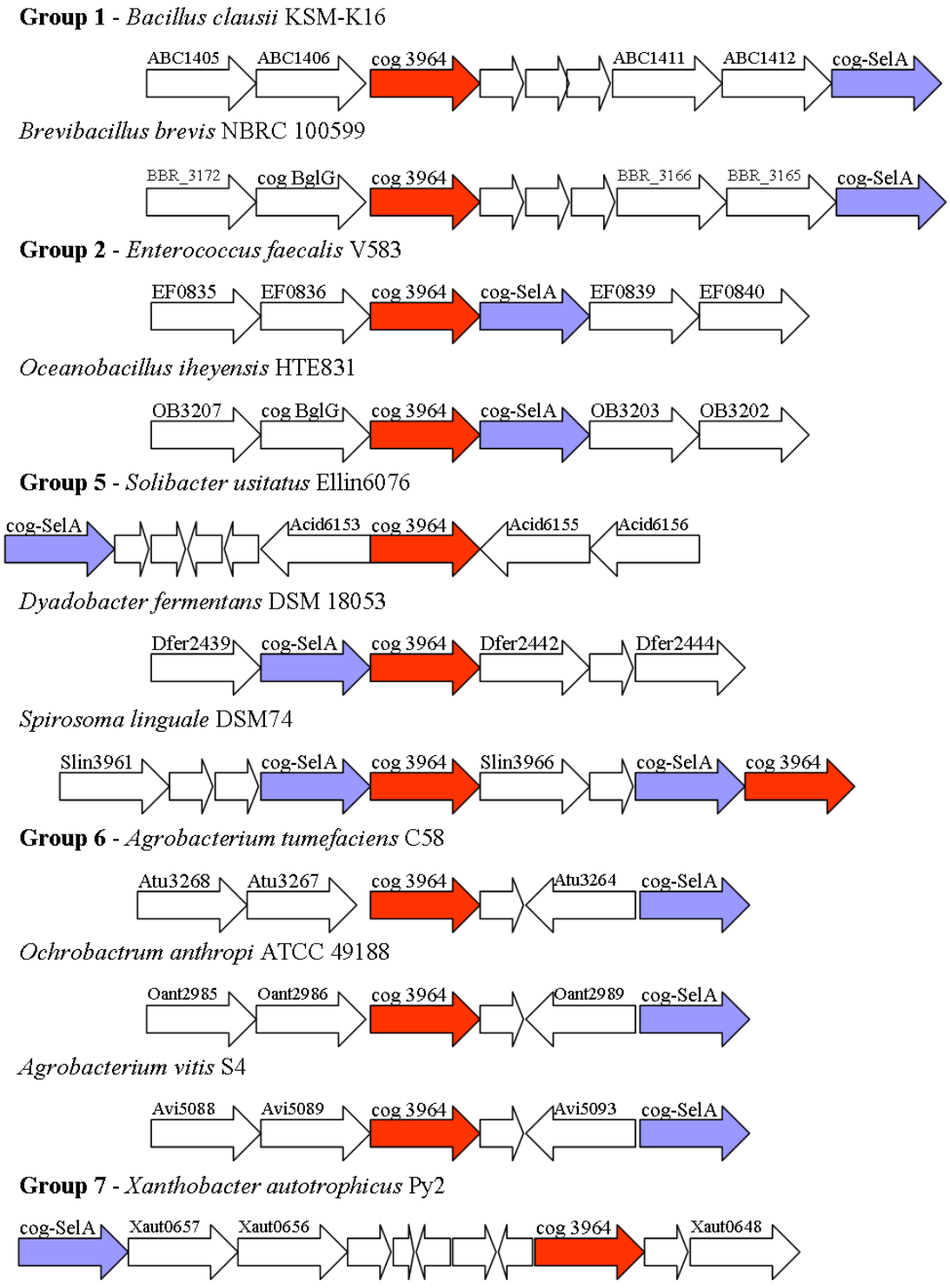


Figure 5.3: Gene operons of organisms with COG3964 (red) and COG1921 (blue) enzymes.

In the protein cluster database (*83*), COG1921 is generally annotated as a pyridoxal-5'-phosphate dependent enzyme responsible for catalyzing the conversion of a seryl-tRNA^{Sec} into selenocysteyl-tRNA^{Sec} (*96*). PLP-dependent enzymes are organized into four distinct groups based on sequence comparison and structural criteria (*157, 158*). These families include: **Fold type I -** α-/γ- families and aspartate amino-transferase; **Fold type II -** β- family and tryptophan synthase family; **Fold type IV -** D-amino acid aminotransferase subgroup, and **Fold type III -** amino acid decarboxylase and alanine racemase family. Analysis of amino acid sequences indicate that characterized SelA or selenocysteine synthase belongs to the α-/γ- superfamily of pyridoxal 5'-phosphate-dependent enzymes (*96*). There are over 550 sequences that have been assigned to COG1921 available through the protein clusters database found in the National Center for Biotechnology Information website. The sequences that encode for the PLP-dependent enzymes belong to a large group of bacteria and some archaea.

To determine the degree of relationship between amidohydrolases in COG3964 and selenocysteine synthases in COG1921 we have cloned, expressed, purified, and began analysis for the characterization of various enzymes in COG1921 that are gene neighbors to the amidohydrolases discussed in this dissertation. Atu3263 from *Agrobacterium tumefaciens* C58 (gi|159185664), Oant2990 from *Ochrobactrum anthropi* CL (gi|153010313) and EF0838 from *Enterococcus faecalis* V583 (gi|29375426) are PLP-dependent enzymes that are expressed in the operon encoding for the amidohydrolase superfamily proteins discussed in chapters 2 and 3. With the exception of glycogen phophorylase, which uses a PLP-cofactor in quite a different

manner, the majority of enzymes that use this B_6 derived cofactor act upon amino acids. The purified proteins annotated as selenocysteine synthases from COG1921 were analyzed for isothermal denaturation in the presence of various D- and L- amino acids and amino acid analogs by differential scanning fluorimetry (*117, 159, 160*).

METHODS AND MATERIALS

Materials: The selection of the genes from COG1921 to be cloned for functional studies was based on availability of a COG3964 protein for that same organism. The genomic DNA for the amplification of the various genes in the study of enzymes in COG1921 was obtained from American Type Culture Collection (ATCC). The oligonucleotide synthesis and DNA sequencing reactions were performed by the Gene Tech Laboratory of Texas A&M University. The pET20b(+) expression vector was acquired from Novagen. The T4-DNA ligase and the restriction enzymes *Nde*I, *Hind*III, and *Eco*RI, were purchased from New England Biolabs. The Herculase II Fusion DNA Polymerase was purchased from Agilent Technologies. The PCR purification and plasmid purification kits were acquired from QIAGEN. The His-Select Nickel affinity gel was purchased from Sigma. XLI-Blue *E.coli* over-expression cells were purchased from Agilent and Rosetta-gami^{TM} B(DE3)pLysS competent cells were acquired from Novagen. Amino acid compounds (20 common D-Xaa and L-Xaa), amino acid analogs (citrulline, L- and D- homoserine, L- and D- homocysteine, L-cystine, cystathionine, L- and D- phenyl glycine and *R*- and *S*- mandelate) and SYPRO® protein gel stain starter kit were acquired from Sigma unless otherwise stated.

Cloning of Gene Atu3263: The gene for Atu3263 from *Rhizobium radiobacter* (also known as *A. tumefaciens*) was amplified from the genomic DNA by standard PCR methods using the forward primer 5'-ACAGGAGCCATATGACCGAGGATATCAGAAGCAGGATC-3' and the terminal primer 5'-ACGCAAGCTTTCCCCAGTCCGGCCAGCACAGCATGC-3' with the restriction sites for *Nde*I and *Hind*III respectively. The PCR product was gel purified and digested with the respective enzymes and inserted into a pET-20b(+) vector that was previously digested with *Nde*I and *Hind*III. Upon ligation of the insert, the plasmid was transformed into XLI-Blue cells and colonies containing the plasmid were selected from LB plates containing 100 μg/mL ampicillin. The selected colony was inoculated into a 5mL culture of LB. The construct was purified using QIAprep spin miniprep kit. The fidelity of the insert was verified by DNA sequencing.

Cloning of Gene Oant2990: The gene for Oant2990 from *O.anthropi* was amplified from the genomic DNA using the forward primer 5'-ACAGGAGCCATATGACCGATGATATTCGCCGCAAGATCGG - 3' and the reverse primer 5'-ACGCAAGCTTTCCCCAATCCGGCCAACGCAGCAAACCG-3'. The amplified gene was digested with the restriction enzymes *Nde*I and *Hind*III. The gel-purified, double digested amplification product was then ligated into the expression vector pET-20b(+), previously digested by the two aforementioned enzymes, using T4 DNA ligase, and then transformed into XL1-Blue *E.coli* cells. The cells were plated into LB agar plates containing 100 μg/mL of ampicillin. A colony was selected to inoculate 5

mL of LB. The plasmid was purified and subsequently sequenced for fidelity of the gene.

Cloning of Gene EF0838: The gene for EF0838 from *E. faecalis* was amplified from the genomic DNA of *E. faecalis* ATCC®700802 using the forward primer 5'-ACAGGAGCCATATGACAATTAGTTACGAAAAATTCC-3' and the reverse primer 5'-ACGCGAATTCTAATTTCTCCTTTTTGTCCATAATTTCTTGTAATC-3'. The amplified product was doubly digested with *Nde*I and *Eco*RI. The cleaned, amplified product was inserted onto a double digested (*Nde*I and *Eco*RI) pET-20b(+) vector using T4 DNA ligase. The ligated product was transformed into XLI-Blue competent cells and plated into LB agar plates containing 100 μg/mL of ampicillin. A single colony was used to inoculate 5 mL cultures of LB. The entire coding region of the plasmid containing the gene for EF0838 was sequenced to confirm fidelity of PCR product.

Overexpression and Purification of Oant2990, Atu3263, and EF0838 (COG1921 proteins): The plasmid constructs containing the Oant2990, Atu3263 and EF0838 amplified genes, were transformed into Rosetta gami B(DE3) pLysS competent cells. A single colony was used to inoculate a 5 mL overnight culture of LB medium supplemented with 100 μg/mL of ampicillin and 20 μg/mL of chloramphenicol dissolved in ethanol. The 5 mL overnight culture was then used to inoculate 1.0 L of LB medium with 100 μg/mL of ampicillin and 50 μg/ml of kanamycin. Cells were grown at 30 ºC to reach an A_{600} ~ 0.6. Induction was initiated by the addition of 500 μM isopropyl D-thiogalactopyranoside (IPTG). The bacterial cells were allowed to grow for an additional 10 hours and were then harvested by centrifugation at 5000g for 15 minutes at

4 ºC. About 7-10 grams of cells were then re-suspended in 50-70 mL of binding buffer (20 mM HEPES pH 7.6 containing 5mM imidazole and 500 mM NaCl) and supplemented with 0.1 mg/mL of phenylmethylsulfonyl fluoride (PMSF) per gram of cell. The cells were lysed by sonication (3 second pulses for 4 minutes, with 4 minute-rest periods for a total of 6 cycles). The cell supernatant was filtered through a 0.2 μm cellulose acetate sterile membrane and loaded into a Ni^{2+}-NTA column pre-equilibrated with binding buffer. The column was washed thoroughly with 10 column volumes of binding buffer until the flow-through absorbance at 280 nm < 0.1. The proteins maintained a yellow color observed in **Figure 5.4**, this color is indicative of a bound PLP cofactor. The His_6-tagged proteins were then eluted with a gradient of elution buffer (10 mM HEPES pH 7.6, 250 mM NaCl and 500 mM imidazole). The protein obtained was bright yellow in color and showed absorbance at 420 nm. Each protein eluted at different gradient concentrations of elution buffer. The fractions containing protein were identified by a bright yellow color, as well as by SDS-PAGE. Fractions containing the protein were pooled; each protein was concentrated and dialyzed in 20 mM HEPES pH 7.6. The proteins were identified to be >95% pure based on SDS-PAGE gel electrophoresis.

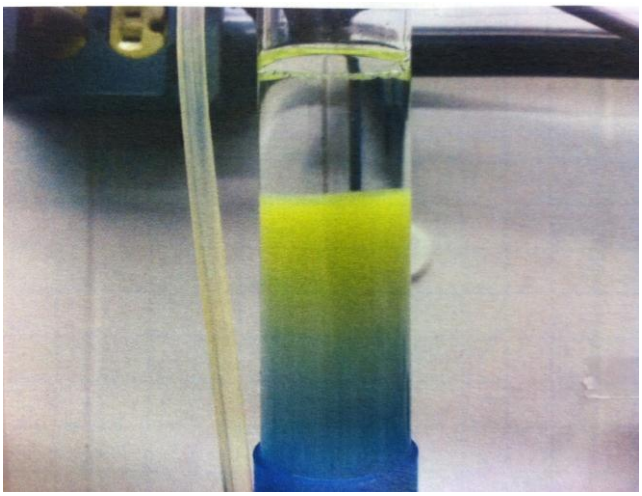


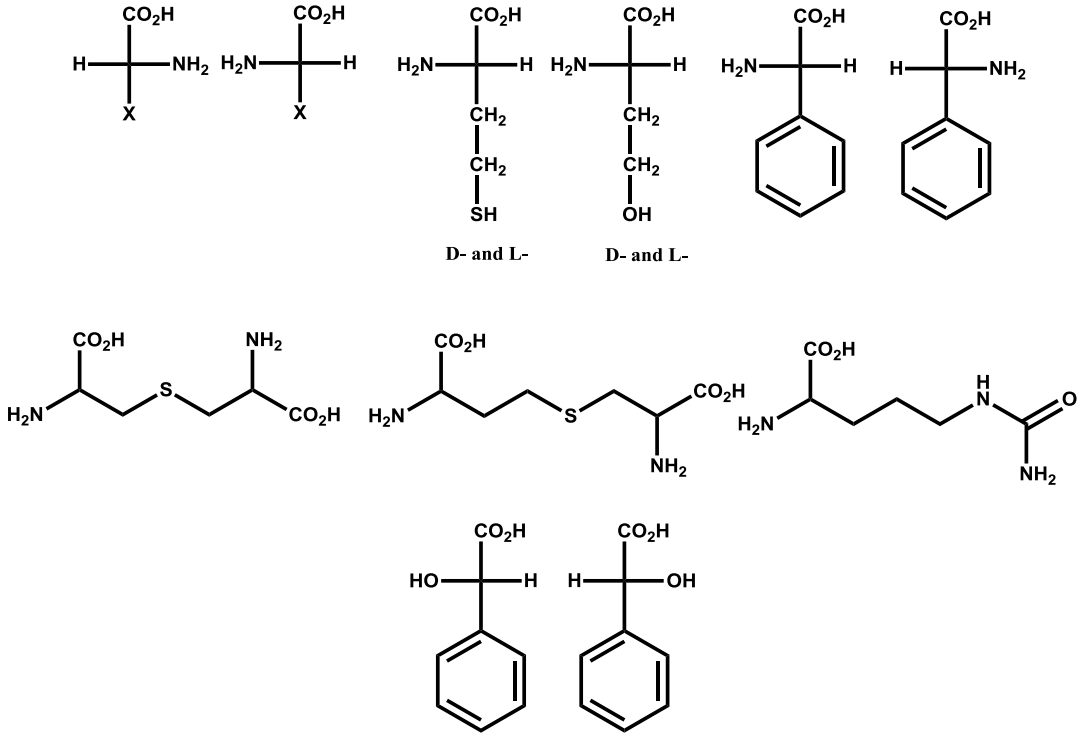
Figure 5.4: Ni^{2+} affinity column with bound Atu3263. All proteins cloned annotated in COG1921 as SelA that are adjacent to COG3964 amidohydrolase demonstrated a bright yellow color upon cell disruption.

Protein Concentration and Metal Analysis: Protein concentration was determined spectrophotometrically at 280 nm using a SPECTRAmax-384 PLUS UV-Vis spectrophotometer microplate reader (Molecular Devices Inc.). The extinction coefficients used for measurements were $\varepsilon = 37{,}025\ M^{-1}cm^{-1}$ for Atu3263; $\varepsilon = 37{,}025\ M^{-1}cm^{-1}$ for Oant2990; and $\varepsilon = 26{,}360\ M^{-1}cm^{-1}$ for EF0838. These were derived from their respective protein sequences using the ExPASy protein parameters database ((http://web.expasy.org/protparam/). Although it is expected that these enzymes are metal independent, metal analysis was determined for each enzyme by inductively coupled plasma emission mass spectrometry (ICP-MS) using a Perkin-Elmer Analyst 700 atomic absorption spectrometer. Each protein sample was treated with concentrated nitric acid and refluxed at 100 °C for 15 minutes. The samples were diluted in distilled water until the final concentration of nitric acid was 1%.

PLP Analysis and Characterization: The presence of PLP cofactor in the enzymes selected and purified from COG1921 was first observed by the bright yellow color of the crude protein extract after sonication. The spectral characteristics of purified PLP-dependent enzymes have been observed in various studies (*96, 97*). The presence of the cofactor pyridoxal-5'-phosphate was assessed by measuring the absorption spectra of each of the purified proteins within 250-500 nm.

Thermal Shift Assays: Each of the purified proteins annotated as selenocysteine synthases in this study were tested by a high-throughput screening method monitoring the temperature of protein denaturation in the presence of D- and L- amino acids. Because most PLP-dependent enzymes work on amino acid substrates (*157, 158*), a series of experiments were designed to monitor protein denaturation of the purified COG1921 proteins using the fluorescent probe SYPRO Orange. Reactions consisted of high concentrations of protein (25-60 μM) in the presence of 10 mM of a D- or L- amino acid, 10 mM HEPES buffer at pH 7.6 and 3x SYPRO Orange dye. Other compounds that were not part of the 20 common amino acids included: L-citrulline, D- and L-homoserine, cystathionine, D- and L-homocysteine, L- cystine, D- and L- phenyl glycine and *R*- and *S*- mandelate (**Scheme 5.1**). The amino acid or the corresponding compound for screening, buffer, and protein were aliquoted into one well of a polypropylene 96-well PCR plate from BIO-RAD® and allowed to incubate at 30 °C for 15 minutes. The SYPRO Orange dye was added to each sample and the plate was set on a Biorad iQ5 multicolor RT-PCR detection system instrument. The parameters for temperature variability were set to increase 1 °C per minute from a range of 55 °C – 95 °C. The

spectrum of fluorescence for each protein as a factor of temperature increase was recorded.



Scheme 5.1: Compounds assayed via thermal shift assays with purified COG1921 proteins (Atu3263, Oant2990 and EF0838).

RESULTS

Purification Properties of Atu3263, Oant2990 and EF0838: Atu3263, Oant2990 and EF0838 were soluble and purified using $(His)_6$-trap column chromatography. The $(His)_6$-tag sequence was cloned at the C-terminal of the protein. The protein was concentrated to stock concentrations of over 100 μM of protein without protein precipitation. In the table below (**Table 5.1**) are the metal analyses for Atu3263, Oant2990 and EF0838 by ICP-MS. These enzymes are not expected to bind metal ions; hence the metal concentration in each of the purified putative SelA proteins is low.

Table 5.1: Metal content for PLP-dependent enzymes from COG1921. Each quantity represents equivalents per monomer of protein.

Enzyme	Zn^{2+}	Fe^{2+}	Mn^{2+}	Ni^{2+}	Cu^{2+}	Total
Atu3263	0.06	0.03	0.01	0.01	n/a	≤ 0.1
Oant2990	0.02	0.05	<0.01	<0.02	n/a	≤ 0.1
EF0838	0.05	0.01	<0.01	n/a	n/a	≤ 0.1

n/a = quantities not applicable, below detectable limits.

Since the putatively annotated SelA purified proteins have a yellow color, an absorbance spectrum was taken for each purified protein. **Figure 5.5** illustrates that in addition to the absorbance peak contributed by the aromatic amino acids at 280 nm, there is a distinctive absorption maximum at 420 nm. This spectral property is characteristic of enzymes containing a pyridoxal-5'-phosphate prosthetic group.

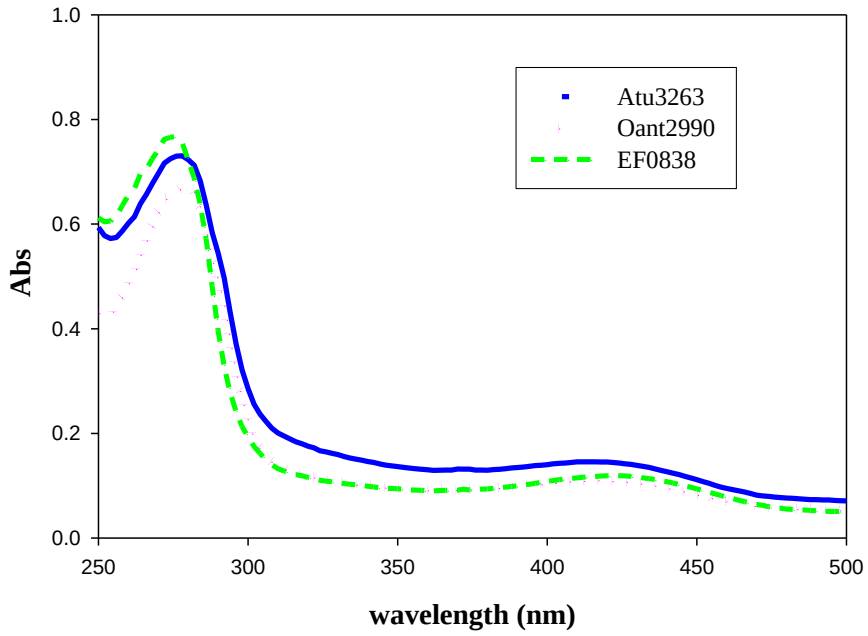


Figure 5.5: Absorbance spectrum of purified PLP-dependent enzymes. A UV-visible absorbance spectrum was obtained for each of the purified COG1921 enzymes (Atu3263, Oant2990 and EF0838) at 30 °C. The spectrum shows absorbance maxima at 280 nm and at 420 nm. Absorbance at 420 nm indicates a pyridoxal-5'-phosphate dependent enzyme.

Thermal Shift Assays: The initial experiments were performed with the three purified SelA proteins that are neighbors to the COG3964 proteins in the presence of the fluorescent dye SYPRO Orange. SYPRO Orange is a highly fluorescent molecule in non-polar environments with low dielectric constants, such as the hydrophobic residues enclosed within the folded protein (*159, 160*). When the protein begins to unfold as the temperature increases, the dye binds to the exposed hydrophobic sites of the protein and the result is a significant increase in fluorescence emission. The determination of a protein melting curve in the presence of various D- and L- amino acids and the fluorescent dye was carried out at high protein concentrations. Each of the purified proteins showed a distinctive protein melting curve based on the number of hydrophobic sites with affinity for the dye and the concentration of each enzyme in the assay. Atu3263 was assayed at a concentration of 60 μM, Oant2990 was assayed at a protein concentration of 55 μM and EF0838 was assayed at a concentration of 25 μM. These concentrations varied based on the resulting concentrations after purification and concentration of the frozen enzyme stocks. **Figure 5.6** shows the protein melting curve in the absence of amino acids. Atu3263 had a T_m of 50 °C; Oant2990 showed a T_m of about 45 °C and EF0838 had a T_m of about 43 °C. Each reaction was assayed in duplicate to confirm reproducibility of the melting point temperature of the protein.

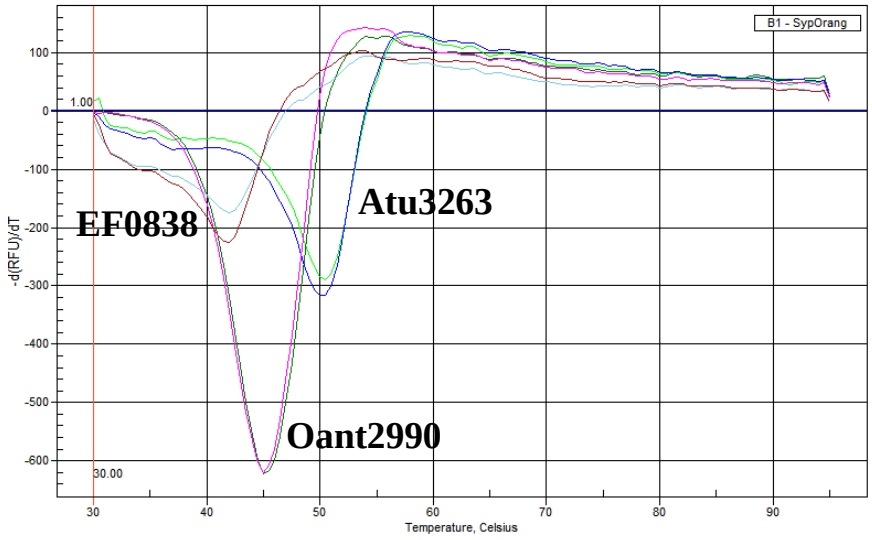


Figure 5.6: Melting curves of PLP-dependent enzymes in the absence of amino acids. Thermal shift assays of control reactions for annotated SelA proteins: Atu3263, Oant2990 and EF0838. Each reaction is shown as a duplicate.

The melting temperature values measured for the protein in the assaying buffer only, were compared to those obtained in the presence of various D- and L- amino acid compounds. Of the nearly 50+ compounds tested for melting temperature analysis of the proteins, only one compound seemed to disrupt the melting temperature of each protein. The presence of D-cysteine decreased the melting temperature for each of the proteins tested. This trend was not observed for any other L- or D- aminoacid. On some occasions L-cysteine seemed to disrupt the T_m of the protein, but this was not a reproducible trend. **Figures 5.7 A** and **B** are the melting curves in the presence of L- and D- cysteine. Other compounds that were tested but were not observed to affect the T_m of the proteins were cystine, D-/L-homocysteine, D-/L- serine, and cystathionine. Additionally, *R*- and *S*-mandelate and D- and L- phenyl glycine were used to carry out the thermal shift assays;

however none of these compounds exhibited the same effect as observed with D-cysteine.

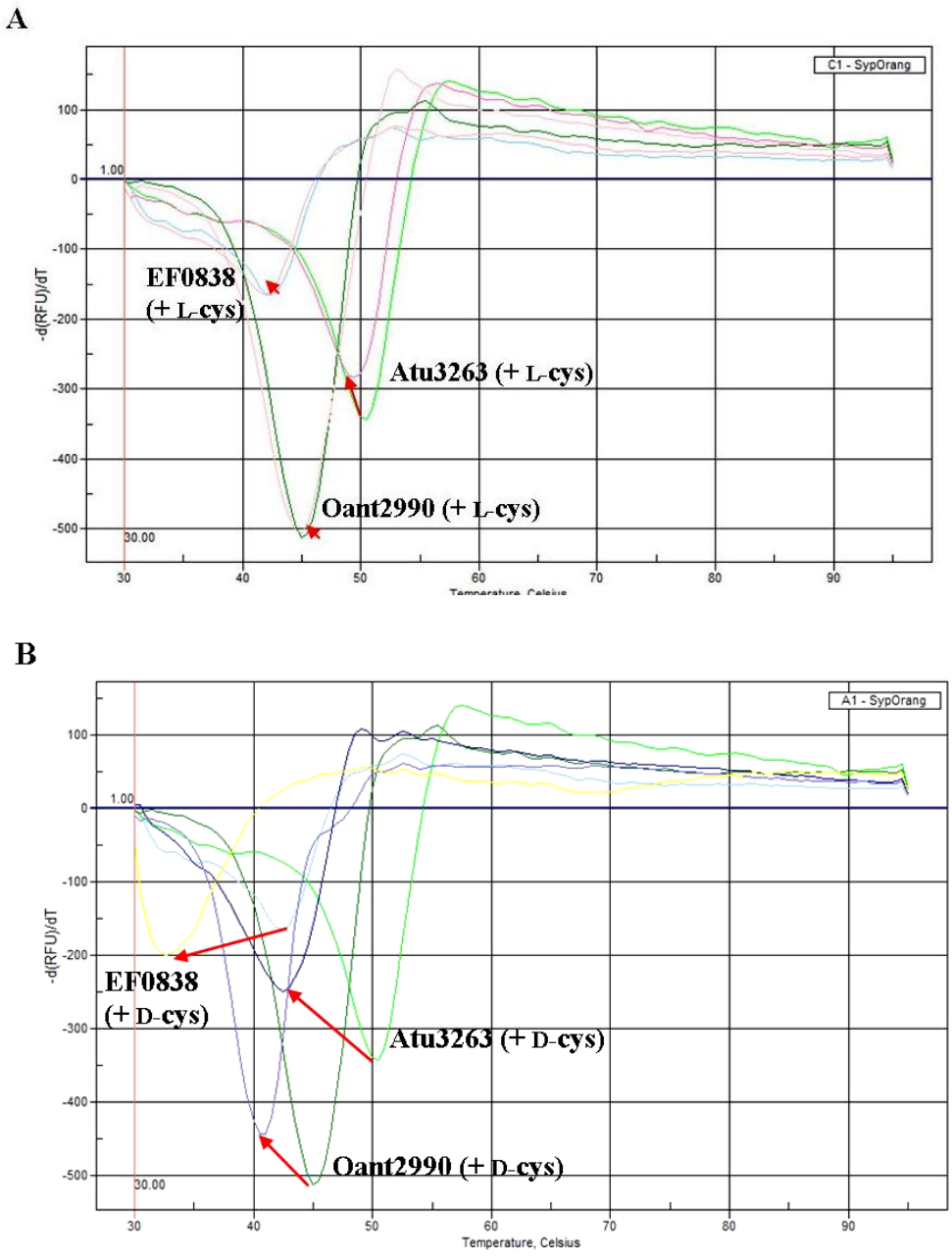


Figure 5.7: Melting curves of PLP-dependent enzymes in presence of (**A**) L-cysteine and (**B**) D-cysteine.

DISCUSSIONS

Three genes encoding a selenocysteine synthase or SelA protein were cloned, expressed and their product was purified. Atu3263, Oant2990, and EF0838 have been characterized as pyridoxal-5'-phosphate dependent enzymes. The sequences for these proteins have also been assigned to COG1921. This cluster of orthologous groups contains enzymes that have been functionally characterized as selenocysteine synthases. Atu3263, Oant2990 and EF0838 share less than 20% sequence identity to the functionally characterized selenocysteine synthase from *E.coli* (*96*). Atu3263 and Oant2990 share a 90% sequence identity, but Atu3263 and EF0838 only share a sequence identity of 32%. It can only be implied with confidence that these enzymes do maintain a PLP prosthetic group, but not that they collectively carry out the same reaction, and much less that they are involved in the synthesis of selenocysteine. The reaction catalyzed by a true selenocysteine synthase is demonstrated in **Figure 5.8**.

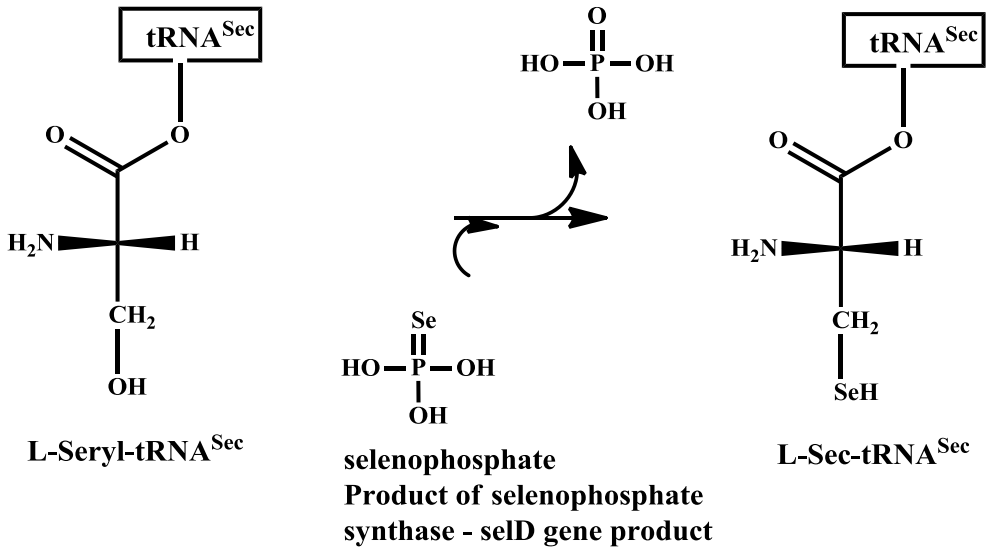


Figure 5.8: Reaction catalyzed by SelA – selenocysteine synthase. Selenocysteine synthase catalyzes the conversion of seryl-tRNA^{Sec} to selenocysteyl-tRNA^{Sec}

In the operon of the organisms that carry out the above reaction, there are several special requirements that come together for the biosynthesis of selenocysteine to occur. In addition to the presence of a PLP-dependent selenocysteine synthase (SelA) that carries out the conversion of Ser-tRNA^{Sec} to Sec-tRNA^{Sec}, other factors are employed. An mRNA secondary SElenoCysteine Insertion Sequence element (SECIS), and a translational elongation factor known as SelB that is specific for selenocysteyl-tRNA^{Sec} are necessary key elements (*95-97*). In addition, bacteria require a source of selenium; this one is provided as the product of selenophosphate synthetase or SelD. This enzyme provides the selenium-bound molecule necessary for the synthesis of selenocysteine by SelA. The discussed elements are observed in the operon of *E.coli* containing the SelA gene. These elements however are not observed in the gene operon of any of the organisms that contain a SelA adjacent to a COG3964 protein. Only one organism

encoding a COG3964 protein showed the presence of additional components for the biosynthesis of selenocysteine. *Xanthomonas autotrophicus* encodes an amidohydrolase of COG3964 in group 7. This organism also encodes a SelA (Xaut_0658) gene just eight reading frames upstream from the amidohydrolase open reading frame (Xaut_0650). A SelD (Xaut_0666) element was found eight additional reading frames upstream from the SelA gene.

There are still additional components that are missing and should be available if the SelA gene from *Xanthomonas autotrophicus* is a genuine selenocysteine synthase. Xaut_0658 is the only found gene annotated as SelA from those organisms also encoding an amidohydrolase in group 7 of COG3964. Unless those additional components for selenocysteine biosynthesis are working from different positions of the genome to carry out the desired function, it is highly unlikely that the putative SelA gene from *Xanthomonas autotrophicus* carries out the annotated function. Functional annotations of the enzymes that are adjacent to COG3964 could be a helpful strategy to unravel the function of the amidohydrolases, but given the degree of differences between characterized COG1921 and the putative COG1921 proteins presented here, this seems to be another instance of functional misannotation.

Below is the sequence similarity network for sequences in COG1921 (**Figure 5.9**). In red is group 1, this group contains characterized selenocysteine synthases from *E.coli, Moorella thermoacetica, Desulfomicrobium baculatum* and *Haemophilus influenzae* (*96*). This group also contains the sequence of the SelA from *Xanthobacter autotrophicus* (Xaut_0658) discussed in chapter IV and in the previous paragraph.

Group 2 of the similarity network for COG1921 contains the majority of the SelA annotated sequences for the genes that are found adjacent to the COG3964 genes encoding the amidohydrolases found in group 2. The SelA from *Enterococcus faecalis* (EF0838) is shown in this group. Groups 3 and 7 in the network for COG1921 contain the sequences for the selenocysteine synthases found as neighbors to genes found in groups 6 and 5 respectively of COG3964. In group 3 of the similarity network for COG1921 are the sequences for Atu3263 and Oant2990, these are also noted.

Typically the sequences found in group 1 of COG1921 encode for larger proteins. The difference in amino acid length between characterized SelA and putative SelA is about 100 residues, close to 10 kDa bigger than those that are found in groups 2, 3 and 7 of this same COG. The sequences that are found in groups 2, 3 and 7 encode for SelA annotated proteins that are neighboring genes to a COG3964 proteins. The insert of COG3964 in the larger image of COG1921 in **figure 5.9**, allows identification of those groups that have an amidohydrolase from COG3964 and a putatively annotated selenocysteine synthase from COG1921.

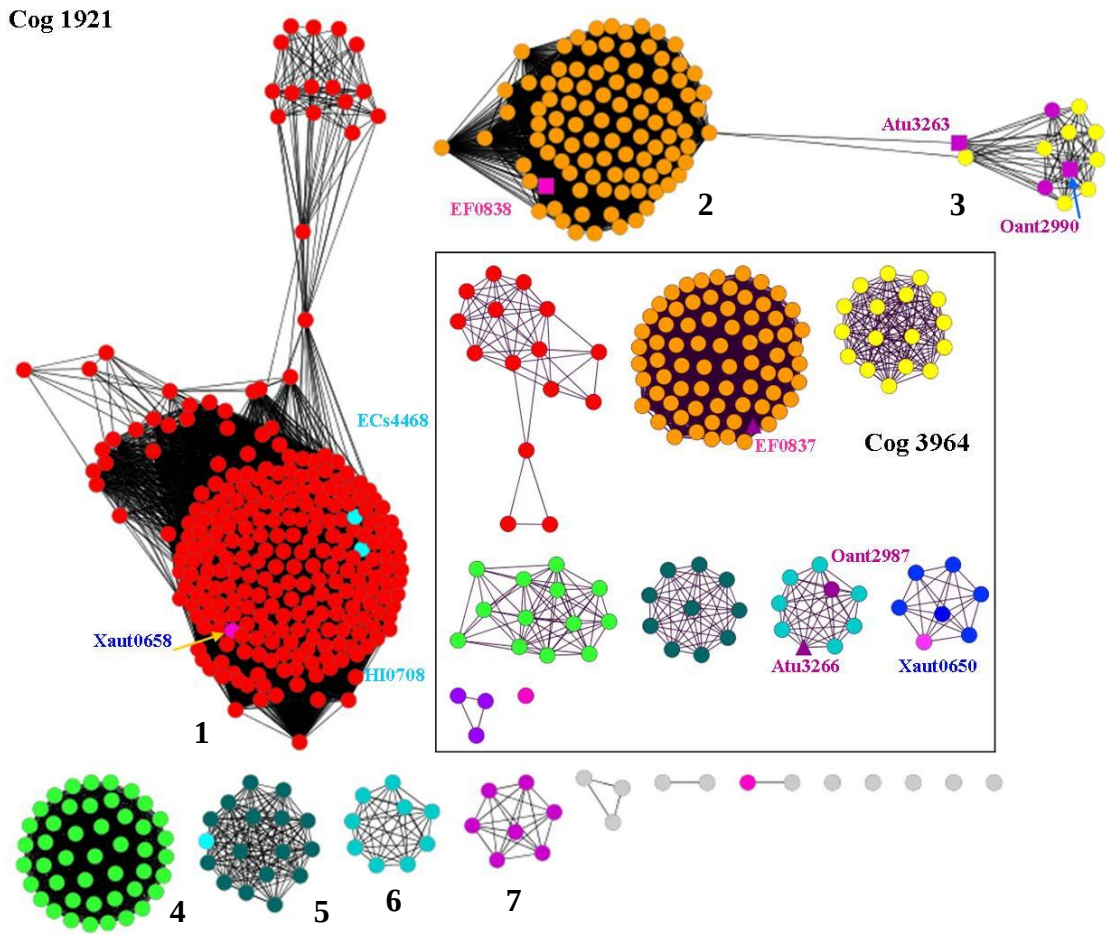


Figure 5.9: Sequence similarity networks of COG1921 and COG3964. Illustrations show relative positions of the amidohydrolases in the Cytoscape designed network of COG3964 (EF0837, Atu3266, Oant2987 and Xaut0650) and the relative grouping of the SelA annotated proteins in the network for COG1921 (EF0838, Atu3263, Oant2990 and Xaut0658). Network also shows group location of two characterized bacterial SelA within COG1921 (ECs4469 and HI0708). Most SelA enzymes that are found adjacent to a COG3964 amidohydrolases in the operon of an organism, are allocated in groups 2 (orange), 3 (yellow) and 7 (magenta).

An alignment between the sequences of the proteins purified from COG1921 that are neighbors to COG3964 proteins, and those sequences from SelA proteins that have been characterized in *E.coli* (ECs4468) and *H. influenzae* (HI0708) show the conservation of various lysine residues. This residue is essential in PLP-dependent proteins for the formation of the Schiff base between the proteins active site lysine side chain and the aldehyde group of the PLP cofactor. **Figure 5.10** shows the alignment between two characterized SelA proteins (ECs4469 and HI0708) and those cloned from organisms containing an adjacent amidohydrolase enzyme (Atu3263 Oant2990, and EF0838). The alignment shows three conserved lysine residues that are candidates for stabilizing the pyridoxal phosphate cofactor. As discussed previously, the characterized SelA proteins are found in group 1 of the sequence similarity network, while those that are neighbors to amidohydrolases are found in groups 2, 3 and 7. In *E.coli* SelA, three lysine residues were found to be conserved amongst the SelA sequences from *H.influenzae*, *M. thermoacetica*, and *D. baculatum* (*96*). These are lysines **(1)**, **(2)** and **(4)** shown in the sequence alignments. Lysine **(1)** is not conserved in the sequences of Atu3263, Oant2990 and EF0838, mutation of this lysine in *E.coli* (K224) did not result in loss of bound PLP. Mutational analysis of lysine **(2)** in *E.coli* (K295) determined this residue was essential for activity. This mutation abolished activity and resulted in the loss of pyridoxal phosphate **(***96***).** A third mutant of lysine 328 in *E.coli* SelA (lysine**(4)**) proved to destabilize the enzyme and resulted in the loss of bound PLP, but not to the extent found in the mutation of lysine **(2)**.

A fourth lysine **(3)** is also conserved in all the sequences for characterized and putative SelA. This lysine however is not implicated in any active roles in the true selenocysteine synthases.

Mutational analyses were not carried out to probe whether the corresponding lysine **(2)** of Atu3263, Oant2990 and EF0838 affected the binding of the PLP cofactor. Based on the sequence alignment there are three possible lysine residues in Atu3263, Oant2990 and EF0838 that can bind PLP, however the type of amino-containing compound that will bind has not been concretely characterized.

There is currently no X-ray three-dimensional structure for a selenocysteine synthase protein from a bacterial organism. The only current available structure for a bacterial selenocysteine synthase has been obtained from a negative-stain electron microscopy projection (*161*). The current available X-ray crystallographic structure for SelA enzymes are from archaea (*Methanococcus janaschii* MJ0158) (*97*), mouse (*162*) and human (*163*).

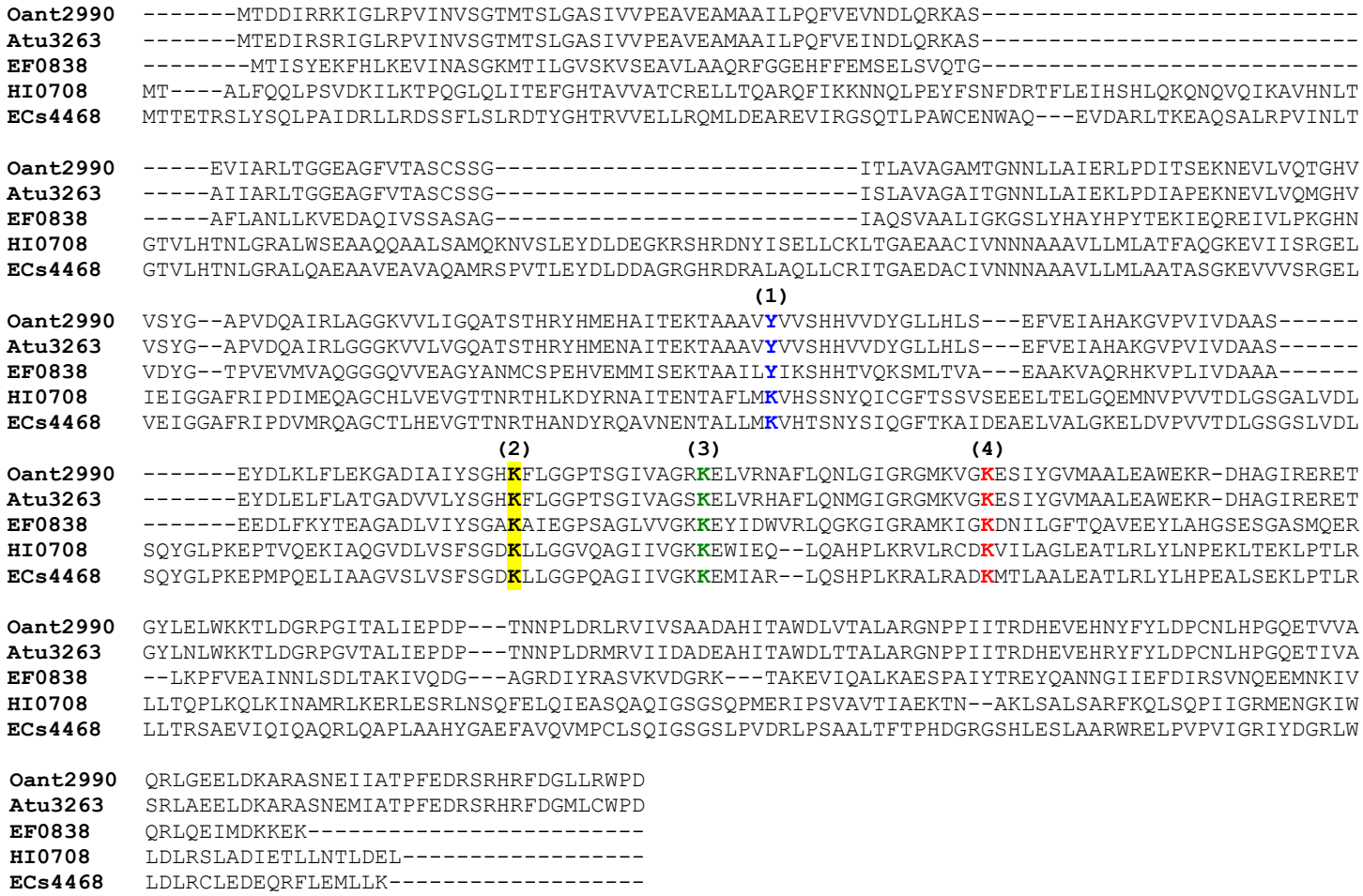


Figure 5.10: Multiple sequence alignment of selected COG1921 proteins. Oant2990, Atu3263 and EF0838 are sequences of SelA proteins adjacent to the operon of amidohydrolases. HI0708 and ECs4468 are the sequences of characterized selenocysteine synthase.

In eukarya and archaea the biosynthesis of selenocysteine is different than the pathway observed in prokaryotes. The archaeal and eukaryotic route for selenocysteine synthesis incorporates a phosphate intermediate. The seryl-tRNA^{Sec} is phosphorylated by phosphoseryl-tRNA^{Sec} kinase (PSTK) to prepare a phosphoseryl-tRNA^{Sec} intermediate. This intermediate is then used by the PLP-dependent enzyme Sep-tRNA:Sec-tRNA synthetase (SepSecS) to prepare the selenocysteinyl-tRNA^{Sec}. The sequence identity between archeal and prokaryotic selenocystine synthase is less than 20%, and the identity between eukaryotic and prokaryotic selenocysteine synthase is less than 15%. These sequence identity comparisons are strictly designated for the characterized bacterial selenocysteine synthases from *E.coli*. It is expected that because of the divergence in sequences between the human, archaeal and bacterial selenocysteine synthases, the structures will be hard to predict for a bacterial SelA. In addition, the proteins studied here (Atu3263, Oant2990 and EF0838) cannot be considered as true selenocysteine synthases, the sole basis for this is not only divergence from the sequence of characterized bacterial selenocysteine synthases, but also the absence of key elements found in the genomic operon involved in the synthesis of selenocysteine.

The thermofluor-based assays were carried out with the goal of determining connections between the COG1921 annotated genes adjacent to the COG3964 proteins. It was expected that the 2' amine analog of mandelate, or phenyl glycine, would be a good candidate to screen in the presence of the purified PLP-dependent proteins. However, this compound did not demonstrate any difference in the melting temperature point of the denaturing enzyme. Instead it was determined that D-cysteine destabilized

the protein to lower its melting point. The L-isomer of cysteine did not have this same effect. Normally, binding substrates and small molecule inhibitors have demonstrated to increase the intrinsic stability of proteins **(***160***)**. This philosophy was incorporated to detect changes of thermal unfolding of the SelA annotated proteins in the presence and absence of amino acids and amino acid-like compounds. In the screens for the PLP-dependent, selenocysteine synthase annotated proteins an inverse effect was observed. Atu3263, Oant2990 and EF0838 were thermally denatured at lower temperatures. This can be attributed to the PLP-cofactor interacting with the D-cysteine molecule rather than forming the Schiff base interaction with the active site lysine of the protein. It is interesting to observe that the destabilization effect was only caused by a D-cysteine. D- or L-selenocysteine were not tested, and additional studies would need to be carried out to further unravel the correct annotation and function of these enzymes and their possible association with the amidohydrolases of COG3964.

CHAPTER VI

SUMMARY AND CONCLUSIONS

The full breadth of capabilities in biological systems will not be entirely understood until individual biological pathways can be dissected to their individual entities and components. This means not only assigning functional annotations to new gene products based on minimum sequence similarity threshold values to components that have been previously characterized, but experimentally defining gene products and metabolites comprising individual pathways, especially those that share low-sequence and structural similarity. The understanding of biological systems is often hindered when there are misrepresentations in the assignment of individual components in pathways, either because these novel components do not share a minimal level of sequence identity to experimentally identified systems to assign a functional role, or because even when correctly characterized by sequence homology, the full range of activity and versatility has not been thoroughly investigated.

The identification of enzyme families, suprafamilies and superfamilies has developed a system that allows the bulk of the information of genome sequencing to be organized based on minimum sequence and structural similarity, substrate specificity, and mechanistic reactions. These identifications provide a starting point to initially define the various roles of gene products. Here, the roles of gene products were focused on the amidohydrolase superfamily. This superfamily is a well-characterized and documented set of enzymes that share similar structural, functional and mechanistic commonalities. Within this superfamily, there are a variety of characteristics that define

individual families. One of these is the variability of the metal-dependent active site embedded at the C-terminal domain of a $(\beta/\alpha)_8$-TIM barrel structural fold. Although identification of a new gene to this superfamily of enzymes has become an easier task, determining the exact function of new genes has become more troublesome. Given the notion that all enzymes in the amidohydrolase superfamily share a low level of sequence similarity, it is observed that unless structural information is provided, this can lead to functional misannotation. In addition, there is increased variability in the active sites of many enzymes in the amidohydrolase superfamily; ranging from the presence or absence of a metal ion center, to the nature of the residues coordinating the mononuclear or binuclear metal active site.

Various bioinformatics methods have been adopted to provide functional determination of uncertain, incorrect, or uncharacterized proteins. The sequences for enzymes found in the amidohydrolase superfamily have been organized into 24 COGs. These COGs are organized based on sequence similarity at specified BLAST *E*-values. Organization of the amidohydrolase superfamily into these COGs allows focusing on a particular set of enzymes expected to carry out a similar reaction on structurally similar compounds.

The goal here was to assess the functional annotation of enzymes in COG3964 within the amidohydrolase superfamily. In order to achieve this, the functional annotation assigned from sequence databases was interrogated for specific members of this cluster of orthologous groups. The collection of sequences assigned to COG3964 was first organized into a sequence similarity network at a BLAST *E*-value of 10^{-70}. At

this cut-off value it was observed that the approximately 200 sequences assigned to this COG arranged into eight different groups (**Figure 1.6**). Enzymes within each group share at least 40% sequence identity with another protein. These analyses suggest that each individual group can potentially carry out a reaction on different, but structurally similar metabolites, using the same defined mechanism observed in other amidohydrolases in the superfamily.

Although the collective annotation of enzymes in COG3964 is observed to be dihydroorotase or adenine deaminase, it was determined by *in-vitro* screening of enzyme targets of this COG that none of the eight enzymes screened against dihydroorotate, adenine, or additional analogues of these compounds were able to carry out the reactions of their annotations. The basis of the rationale for this functional annotation for COG3964 enzymes, even when they share less than 20% identity to characterized dihydroorotases or adenine deaminases is not known.

Functional Annotation and Three-Dimensional Structure of Incorrectly Annotated Dihydroorotases from COG3964 in the Amidohydrolase Superfamily. The enzymes from COG3964 that were able to successfully demonstrate enzyme competent rates of activity are found in group 6 of the developed sequence similarity network (**Figure 1.6**). The selected proteins from group 6 consisted of Atu3266, Oant2987 and RHE_PE00295; these were discussed in chapter 2. Initial screening analyses were carried out with the protein Atu3266 in the presence of a library of *N*-acetyl, *N*-formyl, *N*-succinyl and *N*-carbamoyl D- and L- amino acids, as well as libraries of L-Xaa-L-Xaa, D-Xaa-L-Xaa and L-Xaa-D-Xaa dipeptides. Only the *N*-acetyl-D-amino acid library was

able to show low levels of activity in the hydrolysis of the amide bond. When individual components of the library where tested, it was determined that *N*-acetyl-D-threonine and *N*-acetyl-D-serine were the only compounds in the library that were deacetylated by the enzyme, but at rates ranging between 2 $M^{-1}s^{-1}$ for *N*-acetyl-D-threonine to 4 $M^{-1}s^{-1}$ for *N*-acetyl-D-serine. Modifications to various positions of *N*-acetyl-D-serine did not improve the rates of hydrolysis of the amide bond. However, upon substitution of this amide bond to an ester bond to generate the compound acetyl-*R*-glycerate, there was a 100-fold increase in the rate of deacetylation. Eventual modification to the side chain position of the C-2 carbon, first to generate the compound acetyl glycolate, which replaced the hydroxyl methyl of acetyl-*R*-glycerate with a hydrogen, and later the substitution of a phenyl group, to generate acetyl-*R*-mandelate, increased the rate of deacetylation nearly two-orders of magnitude for acetyl glycolate and three-orders of magnitude for acetyl-*R*-mandelate from those observed with acetyl-*R*-glycerate. The best activities for Oant2987 and RHE_PE00295 were also observed with acetyl-*R*-mandelate.

Phosphonate analogs of acetyl-*R*-mandelate and *N*-acetyl-D-phenyl glycine were synthesized to determine if these acted as inhibitors, and although they did inhibit the deacetylation of acetyl-*R*-mandelate, the K_is determined were relatively high. Initially, it had been proposed to re-crystallize the enzyme Atu3266 in the presence of one of these inhibitors, but these experiments did not come to fruition. Instead, docking experiments were carried out to determine the rationale behind the ability of the enzymes in this group to hydrolyze *R*-isomeric compounds but not their *S*-isomer counterparts, as well as

to identify key components in the active site coordinating to those compounds that were hydrolyzed.

Docking experiments revealed the presence of three-key residues forming a loop between β-strand 7 and α-helix-7 that is found to coordinate the carboxylate moiety of the substrate. The backbone of the triad of residues Gly267-Ala268-Ser269 that is conserved in enzymes belonging to group 6 is found to coordinate the substrate in the active site pocket for hydrolytic attack by the activated water molecule. Docking results determined that the best conformations for substrate placement were observed in the *R*-enantiomer of acetyl-*R*-mandelate. The carboxylate group from the *S*-isomer was found within hydrogen bond distance of loop-7; however, the bridging oxygen from the ester group was too far from the catalytic Asp-291 originating from β-strand 8. In addition, it was observed that the orientation of the hydrogen at the chiral C-2 was flipped, and instead of facing the bottom of the active site pocket, and orienting the bridging ester oxygen in close proximity to Asp-291, it pushed it further away. Other molecules were synthesized that maintained the correct stereochemistry, the presence of the carboxylate group, included substituents at the phenyl moiety, or even extended the size of the acetyl group of acetyl-*R*-mandelate; however, none of these compounds exhibited better rates.

Structural Studies, Substrate Diversity and Functional Annotation of Orthologues in COG3964 Enzymes: Insights from EF0837, STM4445 and BCE5003. A variety of compounds were synthesized that contained an α-acetyl carboxylate moiety. Of all compounds screened for hydrolytic deacetylation, acetyl-*R*-mandelate was the best and one of the few hydrolyzed substrates with enzymes in group 2 as it was discussed in

chapter 3. EF0837 and Atu3266 from group 6 share a 35% sequence identity, while EF0837 and the orthologues found in group 2 (STM4445 and BCE_5003) share between 40-45% sequence identity. Acetyl-*R*-mandelate was the only compound observed to be hydrolyzed by EF0837 with a V/K = 200 $M^{-1}s^{-1}$, while STM4445 and BCE_5003 were able to hydrolyze it at slightly higher rates than EF0837, but still at much lower rates compared to those observed with enzymes in group 6 (k_{cat}/K_m = 4 x 10^3 $M^{-1}s^{-1}$ for BCE_5003 and STM4445 vs. 10^5 $M^{-1}s^{-1}$ for Atu3266 and Oant2987). None of the *N*-modified libraries of amino acids showed to be hydrolyzed by enzymes in group 2. The first sets of compounds to show activity were the α-acetyl carboxylates that had been devised previously for group 6.

Sequence alignments and structural analysis of the enzyme EF0837 reveals that the conserved triad of residues found in loop 7 of group 6 enzymes was not conserved in sequences in group 2. Docking models of Atu3266 with various α-acetyl carboxylates demonstrated that the carboxyl group was coordinated to the back bone atoms of the residues found in a loop after β-strand 7. These residues included a Gly267-Ala268-Ser269 triad. In group 2 enzymes, not all these residues are conserved. Glycine 287 is substituted by a threonine in EF0837 or a larger polar residue in other enzymes of this group, this substitution may interfere in the coordination of the N_α atom of the amino acid backbone with the carboxyl group of the substrate compounds. While in some sequences of group 2, alanine-268 and serine-269 from Atu3266 are conserved, such is the case in STM4445 and BCE_5003, in EF0837 the alanine is replaced with an aspartate, placing another polar, bulky side chain. These substitutions may account for

the difference in activity in the hydrolysis of various α-acetyl carboxylates, including acetyl-*R*-mandelate between EF0837 (V/K 10^2 $M^{-1}s^{-1}$), and STM4445 and BCE_5003 (10^3 $M^{-1}s^{-1}$).

The crystal structure of EF0837 also reveals that there are additional active site residues that may be the responsible for different substrate selectivity. A tyrosine residue that follows the HxH motif is found intruding into the active site of the enzyme. Enzymes in group 6 are missing this residue and instead feature an isoleucine residue. This tyrosine (Tyr-70) can be observed to decrease the size of the active site pocket. In addition a lysine is found to face the active site of the protein. Lysine-216 originates from the loop after β-strand 6 and faces the bottom of the active site pocket. This residue is also found in the sequences from group 6, including that for Atu3266; however the lysine faces away from the active site and instead is solvent exposed.

The crystal structure of EF0837 has an adenine molecule bound at the active site, but the enzyme does not carry out the deamination of adenine, or any other adenine-like compound. The library of compounds tested for EF0837 and the other two proteins in the group was extended to include a variety of adenine and dihydroorotate analogs, including adenosine, cytosine and dihydrouracil. None of these compounds showed activity. Why this molecule is in such close range to the M_β site (2.4 Å) is not well known. It is observed that the *N* atom bound to C6 of the aromatic ring of adenine extends toward the β-metal, but there are no other coordinations or interactions between the bound molecule and the residues surrounding the metal active site of EF0837. Crystallographers from the NYSGXRC elucidated the crystal structure of EF0837, but as

determined from their discussions, adenine was never supplemented to the enzyme during the crystallization process.

Binuclear adenine deaminases have been previously characterized in COG1001. Many of the organisms encoding a characterized adenine deaminase also contain a COG3964 amidohydrolase, including *Enterococcus faecalis*. Whether the presence of the putatively annotated adenine deaminase, EF0837, is the event of gene duplication and evolution is not well known. EF0837, STM4445 and BCE_5003 failed to show activity for the deamination of adenine or any other of the analogues used to screen for related activity (*N*-6-methyl adenine or *N*-6-acetyl adenine)

Functional Diversity in COG3964: Searching and Assessing the Functional Roles of Other Amidohydrolases. A group of proteins within COG3964 reveals the diversity of substrate specificities between the various families of enzymes as determined by the lack of activity with previously discovered substrates. This analysis was discussed in chapter 4. Enzymes in this group namely Xaut_0650 and blr3349, failed to show activity in the hydrolysis of α-acetyl carboxylates. In addition to various amino acids and dipeptides, the substrate profile for these enzymes was extended to include various cyclic compounds, including dihyroorotate and other analogs of dihydropyrimidines, including dihydrouracil and orotic acid. Additionally a large library of L- and D- hydantoins, diketopiperazines, sugar lactones and *N*-acetyl pyranose sugars was also tested. None of these compounds were observed to be hydrolyzed by the enzymes representing this group.

There is no available crystal structure representing enzymes in group 7, however a homology model based on the structure of Atu3266 (PDB: 2OGJ) and EF0837 (PDB: 2ICS) was created to develop a new library of docked compounds. It was observed from sequence analysis between group 7 enzymes and those in group 2 and 6 that there were two inserting loops, one follows β-strand 5 and the other after β-strand 8. These inserting loops form a dome above the active site cavity of the $(\beta/\alpha)_8$-barrel. Homology models were then recreated using the structure of three other members of the amidohydrolase superfamily; a characterized L-hydantoinase (PDB: 1GKR), a characterized dihydropyrimidinase (PDB: 2FVK), and a putative dihydroorotase (PDB: 2GWN). These structures served as models to determine the positions of the inserting loops found in Xaut_0650 and blr3349. There are no defined functions assigned to group 7 of COG3964, but continual search can be focused on obtaining KEGG metabolites having characteristics of substrates found to be active with enzymes in group 6, those that showed some activity in group 2 and those that are substrates in other amidohydrolases used as models for homology (1GKR, 2FVK and 2GWN).

A variety of compounds were synthesized and obtained from commercial sources. This generated a library of compounds that is in its initial stages to define the functional roles of enzymes in COG3964 of the amidohydrolase superfamily. In addition, other strategies for functional prediction and annotation have been employed as supplementary sources to define the functional roles of this COG.

Insights into Operon Proteins for Functional Annotation of Enzymes in COG3964: Assessing the Functional Relations Between COG3964 and COG1921. A detailed search of all the organisms that encode an amidohydrolase from COG3964 was carried out to determine the nature of neighboring genes in the operon. Enzymes in groups 1, 2, 5, 6 and 7 of COG3964, have an annotated SelA (COG1921) gene in the vicinity of the amidohydolase. Genuine SelA proteins carry out the conversion of seryl-$tRNA^{Sec}$ to selenocysteinyl-$tRNA^{Sec}$ in bacteria, this step is fundamental in the biosynthesis of the 21^{st} proteinogenic amino acid, selenocysteine. Only group 2 has all the genes encoding the amidohydrolase enzyme immediately adjacent to the COG1921 annotated gene. In other groups such as 1, 5, 6, and 7 the presence of a SelA gene is exclusive to only certain operons also encoding an amidohydrolase from COG3964, the reason for this is not known. SelA genes or sequences of these gene products are part of COG1921. The cloning, expression and purification of three SelA proteins was carried out as means of determining the function of these pyridoxal-5'-phosphate proteins and relating their function to those amidohydrolases from COG3964. These enzymes included Atu3263, Oant2990 and EF0838; these proteins are gene neighbors to amidohydrolases that had been previously tested for activity or characterized in chapters 2 and 3.

Although the annotated SelA enzymes were found to bind a PLP cofactor, it is very unlikely that they are involved in the biosynthesis of selenocysteine. Organisms with genuine SelA enzymes, usually encode additional factors in the neighboring genes necessary for selenocysteine biosynthesis (SECIS – mRNA structural element,

selenophoshate kinase – SelD, elongation factor – SelB). The organisms encoding a SelA adjacent to a COG3964 protein do not encode these additional components necessary in selenocysteine biosynthesis. Although it is not expected for these newly discovered SelA annotated proteins to carry out the annotated function, there is a high probability that these putative SelA enzymes do carry out reactions in amino acid substrates.

With these considerations, a library of various L- and D- common amino acids, as well as amino acid analogs were screened using thermal shift fluorescence assays. It was observed that D-cysteine was able to reduce the temperature of melting for all the PLP-dependent enzymes purified from COG1921. Other cysteine analogs were also tested including L- cystathionine and D- and L- cystine.

The significance of the thermal shift assay results have not been completely elucidated and there are a variety of factors to consider when relating the observed results with the activities found in the amidohydrolase enzymes hydrolyzing α-acetyl carboxylates. The factors that this study will need to take into consideration for the future is whether there is significant evidence that enzymes in COG3964 are functionally related to enzymes in COG1921, and if so, will the product of one of these enzymes become the substrate in another enzyme. It is interesting that the functional investigation of amidohydrolases in COG3964 was initiated by screening a variety of amino acids, which are key substrates in PLP-dependent enzymes. Subsequently, it is also remarkable that even though enzymes in COG1921 are annotated as selenocysteine synthases, the only amino acid to result in changes in thermal denaturation of the purified SelA

annotated proteins is D-cysteine. Even more fascinating is fact that the only compounds that were found to be hydrolyzed by some of the purified enzymes of COG3964 in this research were restricted to the D- or *R*- enantiomers of acylated amino acids and α-acetyl carboxylates.

The information and results obtained here serve as the basis for the functional re-annotation of COG3964. Follow up studies to determine the additional functions of enzymes not successfully determined will need to be conducted. These new investigations will still benefit from implementation of a variety of strategies to assess additional functions, these include: bioinformatics, structural genomics, genomic context analysis, docking of specific metabolites, but more importantly *in-vitro* screening experiments.

REFERENCES

1. Landsman, D., Kans, J.A., Schuler, G.D., and Ostell, J.M. (1992) Harenessing biotechnology in the 21st. Century. *Proc. Int. Biotechnol. Symp. Exp. 9,* 384-386.
2. Benson, D.A., Boguski, M.S., Lipman D.J., Ostell, J., and Ouellette B.F.F. (1998) GenBank. *Nuc. Acids Res. 26,* 1-7.
3. Bairoch, A., and Apweiler, R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nuc. Acids Res. 24,* 21-25.
4. Gerlt, J.A., Allen, K.N., Almo., S.C., Armstrong, R.N., Babbitt, P.C., Cronan, J.E., Dunaway-Mariano, D., Imaker, H.J., Jacobson, M.P., Minor, W., Poulter, C.D., Raushel, F.M., Sali, A., Shoichet, B.K., and Sweedler, J.V. (2011) The Enzyme Function Initiative. *Biochemistry 50,* 9950-9962.
5. Gerlt, J.A., and Babbitt, P.C. (2000) Can sequence determine function? *Genome Biology 5,* 1-10.
6. Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C. (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol. 5,* 1-13.
7. Xiang, D.F., Patskovsky, Y., Xu, C., Fedorov, A.A., Fedorov, E.V., Sisco, A.A., Sauder, J.M., Burley, S.K., Almo, S.C., and Raushel, F.M. (2010) Functional identification and structural determination of two novel prolidases

from COG1228 in the amidohydrolase superfamily. *Biochemistry 49,* 6791-6803.

8. Xiang, D.F., Patskovsky, Y., Xu, C., Meyer, A.J., Sauder, J.M., Burley, S.K., Almo, S.C., and Raushel, F.M. (2009) Functional identification of incorrectly annotated prolidases from the amidohydrolase superfamily of enzymes. *Biochemistry 48,* 3730-3742.

9. Bork, P., and Koonin, E.V. (1998) Predicting functions from proteins sequences – where are the bottlenecks? *Nature genetics 18,* 313-318.

10. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zheng, Z., Miller, W., and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res. 25,* 3389-3402.

11. Godzik, A., Jambon, M., and Friedberg, I. (2007) Computational protein function prediction: Are we making progress? *Cell. Mol. Life Sci. 64,* 2505-2511.

12. Marchler-Bauer, A., Lu, S. Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Greer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, G.H., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S.H. (2010) CDD: A Conserved Domain Database for functional annotation of proteins. *Nuc. Acids Res. 39,* D225-D229.

13. Gerstein, M., and Honig, B. (2001) Sequences and topology. *Curr. Opin. Struct. Biol. 11,* 327-329.

14. Hegyi, H, and Gerstein, M. (2001) Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. *Genome Res. 11,* 1632-1640.

15. Petsko, G., and Ringe, D., (2004) Deriving function from sequence. In *Protein Structure and Function*. Chapter 4-4 pp.136-137, New Science Press, London, UK.

16. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol. 247,* 536-540.

17. Koonin, E.V., Tatusov, R.L., and Galperin, M.Y. (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol. 8,* 355-363.

18. Samudrala, R., Xia, Y., Huang, E., and Levitt, M. (1999) *Ab Initio* protein structure prediction using a combined hierarchical approach. *Proteins Suppl. 3,* 194-198.

19. Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D.R., Kazmierkiwicz, R., Oldziej, S., Wedemeyer, W.J., Gibson, K.D., Arnautova, Y.A., Sunders, J., Ye, Y.-J., and Scheraga, H.A. (2009) Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA 98,* 2329- 2333.

20. Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. (1999) *Ab Initio* folding of proteins using restraints derived from evolutionary information. *Proteins Suppl. 3,* 177-185.

21. Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. (1999) *Ab Initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins Suppl. 3,* 171-176.

22. Das, R., and Baker, D. (2008) Macromolecular modeling with Rosetta. *Annu. Rev. Biochem. 77,* 363-382.

23. Miller, R.T., Jones, D.T., and Thornton, J.M. (1996) Protein fold recognition by sequence threading: Tools and assessment techniques. *FASEB J. 10,* 171-178.

24. Baker, D., and Sali, A. (2001) Protein structure prediction and structural genomics. *Science 294,* 93-96.

25. Wolf, Y.I., Grishin, N.V., and Koonin, E.V. (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol. 299,* 897-905.

26. Govindarajan, S., Recabarren, R., and Goldstein, R.A. (1999) Estimating the total number of protein folds. *Proteins 35,* 408-414.

27. Zhang, C., and DeLisi, C. (1998) Estimating the number of protein folds. *J. Mol. Biol. 284,* 1301-1305.

28. Hasson, M.S., Muscate, A., McLeish, M.J., Polovnikova, L.S., Gerlt, J.A., Kenyon, G.L., Petsko, G.A., and Ringe, D. (1998) The crystal structure of

benzoylformate decarboxylase at 1.6 Å resolution: Diversity of catalytic residues in thiamin diphosphate-dependent enzymes. *Biochemistry 37,* 9918-9930.

29. Arjunan, P., Umland, T., Dyda, F., Swaminathan, S., Furey, W., Sax, M., Farrenkopf, B., Gao, Y., Zhang, D., and Jordan, F. (1996) Crystal structure of the thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast *Saccharomyces cerevisiae* at 2.3 Å resolution. *J. Mol. Biol. 256,* 590-600.

30. Horowitz, N.H. (1945) On the evolution of biochemical syntheses. *Proc. Natl. Acad. Sci. 31,* 153-157.

31. Horowitz, N.H. (1965) In *Evolving Genes and Proteins,* ed. Bryson. V., and Vogel, H. J. New York: Academic, 15-23.

32. Gerlt, J.A., and Babbit, P.C. (2001) Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem. 70,* 209-246.

33. Traut, T.W., and Temple, B.R. (2000) The chemistry of the reaction determines the invariant amino acids during the evolution and divergence of orotidine 5'-monophosphate decarboxylase. *J. Biol. Chem. 275,* 28675-28681.

34. Babbit, P.C., and Gerlt, J.A. (1997) Understanding enzymes superfamiles: chemistry as the fundamental determinant in the evolution of new catalytic activities *J. Biol. Chem. 272,* 30591-30594.

35. Gerlt, J.A., and Babbit, P.C. (1998) Mechanistically diverse enzyme superfamilies: The importance in the evolution of catalysis. *Curr. Opin. Chem. Biol. 2,* 607-612.

36. Roodvelt, C., and Tawfik, D.S. (2005) Shared promiscuous activities and evolutionary features in various members of the amidohydrolase superfamily. *Biochemistry 44,* 12728-12736.

37. O'Brien, P.J., and Hershlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol. 6,* R91-105.

38. Gerlt, J.A., and Raushel, F.M. (2003) Evolution of function in $(\beta/\alpha)_8$-barrel enzymes. *Curr. Opin. Chem. Biol. 7,* 252-264.

39. Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D., and Waley, S.G. (1975) Structure of chicken muscle triose-phosphate isomerase determined crystallography at 2.5 Å resolution: Using amino acid sequence data. *Nature 255,* 609-614.

40. Seitz, T., Bocola, M., Claren, J., and Sterner, R. (2007) Stabilization of a $(\beta\alpha)_8$-barrel protein designed from identical half barrels. *J. Mol. Biol. 372,* 114-129.

41. Hocker, B., Lochner, A., Seitz, T., Claren, J., and Sterner R. (2009) High-resolution crystal structure of an artificial $(\beta\alpha)_8$-barrel protein designed from identical half barrels. *Biochemistry 48,* 1145-1147.

42. Holm, L., and Sander, C. (1997) An evolutionary treasure: Unification of a broad set of amidohydrolase related to urease. *Proteins 28,* 72-82.

43. Heidhart, D.J., Kenyon, G.L., Gerlt, J.A., and Petsko, G.A. (1990) Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structural homologous. *Nature 347,* 692-694.

44. Babbitt, P.C., Mrachko, G.T., Hasson, M.S., Huisman, G.W., Kolter, R., Ringe, D., Petsko, G.A., Kenyon, G.L., and Gerlt, J.A. (1995) A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids. *Science 267,* 1159-1161.

45. Babbitt, P.C., Hasson, M.S., Wedekind, J.E., Palmer, D.R., Barrett, W.C., Reed, G.H., Rayment, I., Ringe, D., Kenyon, G.L. and Gerlt, J.A. (1996) The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry 35,* 16489-16501.

46. Eklund, H., Erikkson, M., Uhlin, U., Nordlund, P., and Logan, D. (1997) Ribonucleotide reductase – structural studies of a radical enzyme. *Biol. Chem. 378,* 821-825.

47. Eklund, H., and Fontecave, M. (1999) Glycyl radical enzymes: A conservative structural basis for radicals. *Struct. Fold Des. 7,* R257-252.

48. Benning, M.M., Haller, T., Gerlt, J.A., and Holden, H.M. (2000) New reactions in the crotonase superfamily: Structure of methylmalonyl CoA decarboxylase from *Escherichia coli. Biochemistry 39, 4630-4639.*

49. Benning, M.M., Wesenberg, G., Liu, R., Taylor, K.L., Dunaway-Mariano, D., and Holden, H.M. (1998) The three-dimensional structure of 4-hydroxybenzoyl-CoA thioesterase from *Pseudomonas sp.* Strain CBS-3. *J. Biol. Chem. 273,* 33572-33579.

50. Jabri, E., Carr, M.B., Hausinger, R.P., and Karplus, P.A. (1995) The crystal structure of urease from *Klebsiella aerogenes. Science 269,* 998-1004.

51. Benning, M.M., Kuo, J.M., Raushel, F.M., and Holden, H.M. (1995) Three-dimensional structure of the binuclear center of phosphotriesterase. *Biochemistry 34,* 7973-7978.

52. Wilson, D.K., and Quiocho, F.A. (1993) A pre-transition-state mimic of an enzyme: X-ray structure of adenosine deaminase with bound 1-deazaadenosine and zinc activated water. *Biochemistry 32,* 1689-1694.

53. Seibert, C.M., and Raushel, F.M. (2005) Structural and catalytic diversity within the amidohydrolase superfamily. *Biochemistry 44,* 6383-6391.

54. Aimin, L., Tingfeng, L., and Rong, F. (2007) Amidohydrolase superfamily. *Encyclopedia of life sciences,* John Wiley & Sons, Ltd. Hoboken, NJ, 1-8.

55. Raushel, F.M. (2009) Functional annotation of orphan enzymes within the amidohydrolase superfamily. *Beilstein-Institut ESCEC proceedings,* Rüdesheim, Germany, 9-19.

56. Hara, H., Masai, E., Katayama, Y., and Fukuda, M. (2000) The 4-oxalomesaconate hydratase gene, involved in the protocatechuate 4,5-cleavage pathway, is essential to vanillate and syringate degradation in

Sphingomonas paucimolbilis SYK-6 hydration reaction. *J. Bacteriol. 182,* 6950-6957.

57. Li, T., Iwaki, H., Fu, R., Hasegawa, Y., Zhang, H., and Liu, A. (2006) α-amino-β-carboxymuconic-ε-semialdehyde decarboxylase (ACMSD) is a new member of the amidohydrolase superfamily. *Biochemistry 45,* 6628-6634.

58. Williams, L., Nguyen, T., Li, Y., Porter, T.N., and Raushel, F.M. (2006) Uronate isomerase: A nonhydrolytic member of the amidohydrolase superfamily with an ambivalent requirement for divalent metal ion. *Biochemistry 45,* 7453-7462.

59. Nguyen, T.T., Brown, S., Fedorov, A.A., Fedorov, E.V., Babbitt, P.C., Almo, S.C., and Raushel, F.M. (2008) At the periphery of the amidohydrolase superfamily: Bh0493 from *Bacillus halodurans* catalyzes the isomerization of D-galactouronate to D-tagaturonate. *Biochemistry 47,* 1194-1206.

60. Hobbs, M.E., Malashkevich, V., Williams, H.J., Xu, C., Sauder, J.M., Burley, S.K., Almo, S.C., and Raushel, F.M. (2012) Structure and catalytic mechanism of LigI: Insight into the amidohydrolase enzymes of cog3618 and lignin degradation. *Biochemistry 51,* 3497-3507.

61. Thoden, J.B., Phillips, G.N., Neal, T.M., Raushel, F.M., and Holden, H.M. (2001) Molecular structure of dihydroorotase: A paradigm for catalysis through the use of a binuclear metal center. *Biochemistry 40,* 6989-6997.

62. Thoden, J.B., Marti-Arbona, R., Raushel, F.M., and Holden, H.M. (2003) High-resolution X-ray structure of isoaspartyl dipeptidase from *Escherichia coli. Biochemistry 42*, 4874-4882.

63. Kamat, S.S., Bagaria, A., Kumaran, D., Holmes-Hampton, G.P., Fan, H., Sali, A., Sauder J.M., Burley, S.K., Lindahl, P.A., Swaminathan, S., and Raushel, F.M. (2011) Catalytic mechanism and three-dimensional structure of adenine deaminase. *Biochemistry 50,* 1917-1927.

64. Nitanai, Y., Satow, Y., Adachi, H., and Tsujimoto, M. (2002) Crystal structure of human renal dipeptidase involved in beta-lactam hydrolysis. *J. Mol. Biol. 321*, 177-184.

65. Lai, W.L., Chou, L.Y., Ting, C.Y., Kirby, R., Tsai, Y.C., Wang, A.H., and Liaw, S.H. (2004) The functional role of the binuclear metal center in D-aminoacylase: one-metal activation and second-metal attenuation. *J. Biol. Chem. 279*, 13962-13967.

66. Vincent, F., Yates, D., Garman, E., Davies, G.J., and Brannigan, J.A. (2004) The three-dimensional structure of the N-acetyl-glucosamine-6-phosphate deacetylase, NagA, from *Bacillus subtilis:* A member of the urease superfamily. *J. Biol. Chem. 279*, 2809-2816.

67. Hall, R.S., Brown, S., Fedorov, A.A., Fedorov, E.V., Xu, C., Babbitt, P.C., Almo, S.C., and Raushel, F.M. (2007) Structural diversity within the mononuclear and binuclear active sites of *N*-acetyl-D-glucosamine-6 phosphate deacetylase. *Biochemistry 46*, 7953-7962.

68. Wilson, D.K., Rudolph, F.B., and Quiocho, F.A. (1991) Atomic structure of adenosine deaminase complexed with a transition state analog: Understanding catalysis and immunodeficiency mutations. *Science 252,* 1278-1284.

69. Porter, D.J., and Austin, E.A. (1993) Cytosine deaminase: The roles of divalent metal ions in catalysis. *J. Biol. Chem. 268,* 24005-24011.

70. Benning, M.M., Kuo, J.M., Raushel, F.M., and Holden, H.M. (1994) The three-dimensional structure of phosphotriesterase: An enzyme capable of detoxifying organophosphorus nerve agents. *Biochemistry 33,* 15001-15007.

71. Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C. (2009) Annotation error in public database: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol. 5,* e1000605.

72. Pieper, U., Chiang, R., Seffernick, J.J., Brown, S.D., Glasner, M.E., Kelly, L., Eswar, N., Sauder, J.M., Bonanno, J.B., Swaminathan, S., Burley, S.K., Zheng, X., Chance, M.R., Almo, S.C., Gerlt, J.A., Raushel, F.M., Jacobson, M.P., Babbitt, P.C., and Sali, A. (2009) Target selection and annotation for the structural genomics of the amidohydrolase and enolase superfamilies. *J. Struct. Funct. Genomics 10,* 107-125.

73. Pegg, S.C.-H., Brown, S.D., Ojha, S., Seffernick, J., Meng, E.C., Morris, J.H., Chang, P.J., Huang, C.C., Ferrin, T.E., and Babbitt, P.C. (2006) Leveraging enzyme structure – function relationships for functional inference

and experimental design: The structure – function linkage database. *Biochemistry 45,* 2545-2555.

74. Xiang, D.F., Xu, C., Kumaran, D., Brown, A.C., Sauder, J.M., Burley, S.K., Swaminathan, S., and Raushel, F.M. (2009) Functional annotation of two new carboxypeptidases from the amidohydrolase superfamily of enzymes. *Biochemistry 48,* 4567-4576.

75. Xiang, D.F., Kolb, P., Fedorov, A.A., Xu, C., Fedorov, E.V., Narindoshvili, T., Williams, H.J., Shoichet, B.K., Almo, S.C., and Raushel, F.M. (2012) Structure-based function discovery of an enzyme for the hydrolysis of phosphorylated sugar lactones. *Biochemistry 51,* 1762-1773.

76. Hall, R.S., Fedorov, A.A., Marti-Arbona, R., Fedorov, E.V., Kolb, P., Sauder, J.M., Burley, S.K., Shoichet, B.K., Almo, S.C., and Raushel, F.M. (2010) The hunt for 8-oxoguanine deaminase. *J. Am. Chem. Soc. 132,* 1762-1763.

77. Hitchcock, D.S., Fedorov, A.A., Fedorov, E.V., Dangott, L.J., Almo, S.C., and Raushel, F.M. (2011) Rescue of the orphan enzymes isoguanine deaminase. *Biochemistry 50,* 5555-5557.

78. Goble, A.M., Zhang, Z., Sauder, J.M., Burley, S.K., Swaminathan, S., and Raushel, F.M. (2011) Pa0148 from *Pseudomonas aeruginosa* catalyzes the deamination of adenine. *Biochemistry 50,* 6589-6597.

79. Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000) The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nuc. Acids Res. 28,* 33-36.

80. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001) The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nuc. Acids Res. 29,* 22-28.

81. Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS Comput. Biol. 4,* e4345.

82. Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A.R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B.R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G.J., Ideker, T., and Bader, G.D. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols 2,* 2366-2382.

83. Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciufo, S., Fedorov, B., Kiryutin, B., O'Neill, K., Resch, W., Resenchuk, S., Schafer, S., Tolstoy, I., and Tatusova, T. (2009) The national center for biotechnology information's protein cluster database. *Nuc. Acids Res. 37,* D216-233.

84. Hall, R.S., Agarwal, R., Hitchcock, D., Sauder, J.M., Burley, S.K., Swaminathan, S., and Raushel, F.M. (2010) Discovery and structure

determination of the orphan enzyme isoxanthopterin deaminase. *Biochemistry 49,* 4374-4382.

85. Porter, T.N., Li, Y., and Raushel, F.M. (2004) Mechanism of the dihydroorotase reaction. *Biochemistry 43, 16285-16292.*

86. Wang, C.-C., Tsau, H.-W., Chen, W.-T., Huang, C.-Y. (2010) Identification and characterization of a putative dihydroorotase, KPN01074, from *Klebsiella pneumoniae. Protein J. 29,* 445-452.

87. Cummings, J.A., Fedorov, A.A., Xu, C., Brown, S., Fedorov, E., Babbitt, P.C., Almo, S.C., and Raushel, F.M. (2009) Annotating enzymes of uncertain function: The deacylation of D-amino acids by members of the amidohydrolase superfamily. *Biochemistry 48,* 6469-6481.

88. Hermann, J.C., Ghanem, E., Li, Y., Raushel, F.M., Irwin, J.J., and Shoichet, B.K. (2006) Predicting substrates by docking high energy intermediates to enzyme structures. *J. Am. Chem. Soc. 128,* 15882-15891.

89. Hermann, J.C., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K., and Raushel, F.M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature 448,* 775-780.

90. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999) The use of gene clusters to functional coupling. *Proc. Natl. Acad. Sci. 96,* 2896-2901.

91. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. 96,* 4285-4288.

92. Galperin, M.Y., and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nature Biotech. 18,* 609-613.

93. Huynen, M., Snel, B., Lathe, W. III, and Bork, P. (2000) Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome 10,* 1204-1210.

94. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998) Predicting function: From genes to genomes and back. *J. Mol. Biol. 283,* 707-725.

95. Su, D., Hohn, M.J., Palioura, S., Sherrer, R.L., Yuan, J., Söll, D., and O'Donoghue, P. (2009) How an obscure archaeal gene inspired the discovery of selenocysteine biosynthesis in humans. *IUBMB Life 61,* 35-39.

96. Tormay, P., Wilting, R., Lottspreich, F., Mehta, P.K., Christen, P., and Böck, A. (1998) Bacterial selenocysteine synthase: Structural and functional properties. *Eur. J. Biochem. 254,* 655-661.

97. Kaiser, J.T., Gromadski, K., Rother, M., Engelhardt, H., Rodnina, M.V., and Wahl, M.C. (2005) Structural and functional investigation of a putative archeal selenocysteine synthase. *Biochemistry 44,* 13315- 13327.

98. Marti-Arbona, R., Xu, C., Steele, S., Weeks, A., Kuty, G.F., Seibert, C.M., and Raushel, F.M. (2006) Annotating enzymes of unknown function: *N*-formimino-L-glutamate deiminase is a member of the amidohydrolase superfamily. *Biochemistry 45,* 1997-2005.

99. Fewson, C.A. (1988) Microbial metabolism of mandelate: A microcosm of diversity. *FEMS Microbiol. Rev. 54,* 85-110.

100. Fewson, C.A., Allison, N., Hamilton, I.D., Jardine, J., and Scott, A.J. (1988) Comparison of mandelate dehydrogenases for various strains of *Actinobacter calcoaceticus:* Similarity of natural and 'evolved' forms. *J. Gen. Microbiol. 134,* 967-991.

101. Baker, D.P., Kleanthous, C., Neen, J.N., Weinhold, E., and Fewson, C.A. (1992) Mechanistic and active-site studies on D-(-)-mandelate dehydrogenase from *Rhodotorula graminis. Biochem. J. 281,* 211-218.

102. Chen, Y.P., Dilworth, M.J., and Glenn, A.R. (1989) Degradation of mandelate and 4-hydroxymandelate by *Rhizobium leguminosarum* biovar *trifolii* TAI. *Arch. Microbiol. 151,* 520-525.

103. Fewson, C.A. (1992) Function, properties, and evolution of mandelate dehydrogenases and other enzymes of the mandelate pathway. In *The evolution of metabolic function.* Mortlock, R.P. CRC press, Boca Raton FL, 115-141.

104. Tsou, A.Y., Ransom, S.C., and Gerlt, J.A. (1990) Mandelate pathway of *Pseudomonase putida:* Sequence relationships involving mandelate

racemase, (*S*)-mandelate dehydrogenase, and benzoyl formate decarboxylase and expression of benzoylformate decarboxylase in *E. coli. Biochemistry 29,* 9856-9862.

105. Hegeman, G.D. (1966) Synthesis of the enzymes of the mandelate pathway by *Pseudomonas putida*. *J. Bacteriol. 91,* 1140-1154.

106. Ju, X., Yu, H.-L., Pan, J., Wei, D.-Z., Xu, J.-H. (2010) Bioproduction of chiral mandelate by enantioselective deacylation of α-acetoxyphenylacetic acid using whole cells for newly isolated *Pseudomonas* sp. ECU1011. *Appl. Microbiol. Biotechnol. 86,* 83-91.

107. Kaiser, E.T., and Carson, F.W. (1964) Studies on the esterase action of carboxypeptidase A. Kinetics of the hydrolysis of acetyl-L-mandelate. *J. Am. Chem. Soc. 86,* 2922-2926.

108. Carson, E.W., and Kaiser, E.T. (1966) pH dependence of the hydrolysis of O-acetyl-L-mandelate catalyzed by carboxypeptidase A. A critical examination. *J. Am. Chem. Soc. 88,* 1212-1223.

109. Roberts, R.J. (2004) Identifying protein function – A call for community action. *PLoS Biol. 2,* E42.

110. Rost, B., Liu, J., Nair, R., Wrezeszczynski, K.O., and Ofran, Y. (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci. 60,* 2637-2650.

111. Freidberg, I. (2006) Automated protein function prediction – The genomic challenge. *Brief. Bioinform. 7,* 225-242.

112. Glasner, M., Gerlt, J.A., and Babbitt, P.C. (2006) Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol. 10,* 492-497.

113. Newton, R.J., Griffin, L.E., Bowles, K.M., Meile, C., Gifford, S., Givens, C.E., Howard, E.C., King, E., Oakley, C.A., Risch, C.R., Rinta-Kanto, J., Sharma, S., Sun, S., Varaljay, V., Vila-Costa, M., Westrich, J.R., and Moran, M.A. (2010) Genome characteristics of a generalist marine bacterial lineage. *ISME Journal: Multidisciplinary journal of microbial ecology 4,* 784-798.

114. Sweetlove, L.J., and Fernie, A.R. (2005) Regulation of metabolic networks: Understanding metabolic complexity in the systems biology era. *New Phytol. 168.* 9-24.

115. Kim, S.-H. (2000) Structural genomics of microbes: An objective. *Curr. Opin. Struct. Biol. 10,* 380-383.

116. Kolb, P., Ferreira, R.S., Irwin, J.J., and Shoichet, B.K. (2009) Docking and chemoinformatic screens for new ligand and targets. *Curr. Opin. Biotechnol. 20,* 429-436.

117. Senisterra, G.A., Markin, E., Yamazaki, K., Hui, R., Vedadi, M., and Awrey, D.E. (2006) Screening for ligands using a generic and high-throughput light-scattering based assay. *J. Biomol. Screen. 11,* 940-948.

118. Galperin, M.Y., and Koonin, E.V. (2010) From complete genome sequence to 'complete' understanding? *Trends in biotech. 28,* 398-406.

119. Pace, C.N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) How to measure and predict the molar absorption coefficient for a protein. *Protein Sci. 4,* 2411-2423.

120. Doi, E., Shibata, D., and Matoba, T. (1981) Modified colorimetric ninhydrin methods for peptidase assay. *Anal. Biochem. 118,* 173-184.

121. Chapman, E., and Wong, C.H. (2002) A pH sensitive colorimetric assay for the high-throughput screening of enzyme inhibitors and substrates: A case study using kinases. *Bioorg. Med. Chem. 10,* 551-555.

122. Otwinowski, Z., and Minor, W. (1997) Processing of x-ray diffraction data collected in oscillation mode. *Methods Enzymol. 276,* 307-326.

123. Schneider, T.R., and Sheldrick, G.M. (2002) Substructure solution with SHELXD. *Acta Cryst. D58,* 1772-1779.

124. CCP4 (1994) CCP4 Suite: Programs for protein crystallography. *Acta Cryst. D50,* 760-763.

125. De-La-Fortelle, E., and Bricogne, G. (1997) Maximum-likelihood heavy atom parameter refinement in the MIR and MAD methods. *Methods Enzymol. 276,* 472-493.

126. Perrakis, A., Morris, R., and Lamzin, V.S. (1999) Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol. 6,* 458-463.

127. Brunger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, F.S., Kuszwewski, J., Nilges, M., Pannu, N.S., Read,

R.J., Rice, L.M., Somonsom, T., and Warren, G.L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Cryst. D54,* 905-921.

128. Jones, T.A., Zou, J.-Y., Cowan, S.W., and Kjeldgaard, M. (1991) Improved methods in building protein models in electron density map and the location of errors in these models. *Acta Cryst. A47,* 110-119.

129. Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nuc. Acids Res. 28,* 27-30.

130. Irwin, J.J., Raushel, F.M., and Shoichet, B.K. (2005) Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry 44,* 12316-12328.

131. Wlodek, S., Skillman, A.G., and Nicholls, A. (2010) Ligand entropy in gas-phase, upon salvation and protein complexation. Fast estimation with quasi-newton hessian. *J. Chem. Theory Comput. 6,* 2140-2152.

132. Waterhouse, A.M., Procter, J.B., Martin, D.MA. Clamp, M., and Barton, G.J. (2009) Jalview version 2 – A multiple sequence alignment editor and analysis workbench. *Bioinformatics 25,* 1189-1191.

133. Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics 5,* 113-132.

134. Northrop, D.B. (1998) On the meaning of K_m and V/K in enzyme kinetics. *J. Chem. Ed. 75,* 1153-1157.

135. Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D.,, Huang, K.H., Keller, K., Novichkov, P.S., Dubchak, I.L., Alm, E.J., and Arkin, A.P. (2009) MicrobesOnline: An integrated portal for comparative and functional genomics. *Nuc. Acids Res. 38*, D396-400.

136. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nuc. Acids Res. 22*, 4673-4680.

137. Abendroth, J., Niefind, K., May, O., Siemann, M., Syldatk, C., and Schomburg, D. (2002) The structure of L-hydantoinase from *Arthrobacter aurescens* leads to an understanding of dihydropyrimidinase substrate and enantio specificity. *Biochemistry 41,* 8589-8597.

138. Lohkamp, B., Andersen, B., Piškur, J., and Dobritzsch, D. (2006) The crystal structures of dihydropyrimidinases reaffirm the close relationship between cyclic amidohydrolases and explain their substrate specificity. *J. Biol. Chem. 281,* 13762-13776.

139. Schübel, U., Kraut, M., Morsdorf, G., and Meyer, O. (1995) Molecular characterization of the gene cluster *coxMSL* encoding the molybdenum-containing carbon monoxide dehydrogenase of *Oligotropha carboxidovorans. J. Bacteriol. 177,* 2197-2203.

140. Meyer, O., Frunzke, K., and Mörsdorf, G. (1993) Biochemistry of the aerobic utilization of carbon monoxide. In Murrell, J.C., and Kelly, D.P. (ed) *Microbial growth on C1 compounds. Intercept, Ltd.* San Diego, CA, USA. 433-459.

141. Eitinger, T., Rodionov, D.A., Grote, M., and Schneider, E. (2011) Canonical and ECF-type ATP-binding cassette importers in prokaryotes: Diversity in modular organization and cellular functions. *FEMS Microbiol. Rev. 35,* 3-67.

142. Koyanagi, T., Katayama, T., Suzuki, H., and Kumagai, H. (2004) Identification of the LIV-I/LS system as the third phenylalanine transporter in *Escherichia coli* K12. *J. Bacteriol. 186,* 343-350.

143. Boyington, J.C., Gladyshev, V.N., Khangulov, S.V., Stadtman, T.C., and Sun P.D. (1997) Crystal structure of formate dehydrogenase H: Catalysis involving Mo, molybdopterin, selenocysteine and Fe4S4 cluster. *Science 275,* 1305-1308.

144. Gladyshev, V.N., Khangulov, S.V., Axley, M.J., and Stadtman, T.C. (1994) Coordination of selenium to molybdenum in formate dehydrogenase H from *Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 91,* 7708-7711.

145. Ehrenreich, A., Forchhammer, K., Tormay, P., Veprek, B., and Böck, A. (1992) Selenoprotein synthesis in *E. coli*: Purification and characterization of the enzyme catalyzing selenium activation. *Eur. J. Biochem. 206,* 767-773.

146. Leinfelder, W., Forchhammer, K., Veprek, B., Zehelein, E., and Böck, A. (1990) *In vitro* synthesis of selenocysteinyl-$tRNA_{UCA}$ from seryl-$tRNA_{UCA}$:

Involvement and characterization of the *selD* gene product. *Proc. Natl. Acad. Sci. U.S.A. 87,* 543-547.

147. Itoh, Y., Sekine, S.-I., Matasumoto, E., Akasaka, R., Takemoto, C., Shirouzu, M., and Yokoyama, S. (2009) Structure of selenophosphate synthetase essential for selenium incorporation into proteins and RNAs. *J. Mol. Biol. 385,* 1456-1469.

148. Hulsen, T., de Vlieg, J., and Groenen, P.M.A. (2006) PhyloPat: Phylogenetic pattern analysis of eukaryotic genes. *BMC Bioinformatics 7,* 398-405/

149. Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., and Kasif, S. (2002) Computational identification of operons in microbial genomes. *Genome Res. 12,* 1221-1230.

150. Hsiao, T.-L., Revelles, O., Chen, L., Sauer, U., and Vitkup, D. (2010) Automatic policing of biochemical annotations using genomic correlations. *Nature Chem. Biol. 6,* 34-40.

151. Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000) Protein function in the post-genomic era. *Nature 405,* 823-826.

152. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature 403,* 623-627.

153. Brown, P.O., and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature 21,* 33-37.

154. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., and Staudt, L.M. (2002) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature 403,* 503-511.

155. Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science 285,* 751-753.

156. Dandekar, T., Snel, B., Huynen, M., and Pork, P. (1998) Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci. 23*, 324-328.

157. Mehta, P.K., and Christen P. (1998) The molecular evolution of pyridoxal-5'-phosphate dependent enzymes. *Adv. Enzy. and Related Areas of Mol. Biol. 74*, 129-184.

158. John, R.A. (1995) Pyridoxal phosphate enzymes. *Biochimica et Biophysica acta. 1248,* 81-96.

159. Giepmans, B.N.G., Adams, S.R., Ellisman, M.H., and Tsien, R.Y. (2006) The fluorescent toolbox for assessing protein location and function. *Science 312,* 217-224.

160. Niesen, F.H., Berglund, H., and Vedadi, M. (2007) The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols 2,* 2212-2221.

161. Engelhardt, H., Forchhammer, K., Müller, S., Goldie, K.N., and Böck, A. (1992) Structure of selenocysteine synthase from *Escherichia coli* and location of tRNA in the seryl-tRNA^{Sec}-enzyme complex. *Molec. Microbiol. 6,* 3461-3467.

162. Ganichkin, O.M., Xu, X.-M., Carlson, B.A., Mix, H., Hatfield, D.L., Gladyshev, V.N., and Wahl, M.C. (2008) Structure and catalytic mechanism of eukaryotic selenocysteine synthase. *J. Biol. Chem. 283,* 5849-5865.

163. Paliora, S., Sherrer, R.L., Steitz, T.A., Soll, D., and Simonovic, M. (2009) The human SepSecS-tRNA^{Sec} complex reveals the mechanism of selenocysteine formation. *Science 325,* 321-325