University of Nevada, Reno

Comparison of protein structures to infer enzyme function and identify structurally significant amino acids

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Chemical Engineering

by
Benjamin T. Caswell

Dr. David C. Cantu
Thesis Advisor

May, 2022

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

Benjamin Caswell

entitled

Comparison of protein structures to infer enzyme function and identify structurally significant amino acids

be accepted in partial fulfillment of the
requirements for the degree of

Master of Science

David C. Cantu
Advisor

Maryam Raeeszadeh-Sarmazdeh
Committee Member

Tin Nguyen
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

May, 2022

ABSTRACT

Biomolecules, particularly proteins, are increasingly taking center stage as an area of research in myriad industries and disciplines. Deepening our understanding of how protein structure affects function will impact our approaches to modern medicine, environmental treatment, and chemical production. In the first part of this work, primary sequence similarity was used to classify all available thioesterases into families whose similarities were confirmed by sequence alignment and structure superimposition. Thirty-five thioesterase families were identified, analyzed, and made available in the updated thioester-active enzyme (ThYme) database. In the second part of this work, a method was developed to determine spatially correlated residues between two protein structures without formal structural superimposition. This method was used to determine the mutation space of four sets of protein structures by correlating residues that occupy the same spatial position within their structures. A target structure was selected from each set to serve as a comparison point, and a scoring system was developed to quantify, for each residue in the target protein, the mutations in that spatial position. Results demonstrate that residues located on the exterior of a protein tend to have a greater diversity of composition than those located on interior surfaces, and that this contrast seems uniform across different secondary structures.

ACKNOWLEDGEMENTS

Dr. David C. Cantu, I could not have been luckier than for you, of all professors, to take me under your wing. Your kindness and approachability are revered among students and your eminent intelligence and knowledge were incredibly present as you taught me and guided me through this program. You are responsible for so much of the growth that I have experienced throughout the last two years; for that and all the rest, I am truly grateful.

Mike Kivistik, acting as a teaching assistant in your Unit Operations lab was a pleasure, and a great learning experience to boot! Thank you for all the assistance and guidance that you have provided and for the wonderful conversation along the way.

Dr. Tin Nguyen, I am glad to have had the chance to collaborate with you on the ThYme database and thank you for sitting on my advisory committee.

Dr. Maryam Raeeszadeh-Sarmazdeh, your instruction in my favorite course, advanced reactor design, was very impactful. Also, thank you for sitting on my advisory committee.

Lisa Caswell, you have been there through thick and thin. You have helped me through the rough times and celebrated the good; without you, I do not know where I would be. Thank you for being a loving wife and amazing friend, I love you and cannot wait to see where we go from here!

Mom, Mother Dearest, thank you for molding me into the capable person that I am. You taught me the skills that have been critical to my success, skills that some are not lucky

enough to have been taught. Your success is an inspiration that helps drive me forward. Thank you and I love you.

Ian, Devin, Josh, Marissa, Patrick, Rebecca, Harlen, Vanessa, Nathan, Emily, Brennan, Andy, Caitlin, Derek, and any others that I hopefully did not forget, you are all responsible for the person that I am today. Without a family like you around me I doubt I would be as happy as I am with who and where I am. Thank you for your love, support, and friendship.

Ethan, Sarah, Ryan, Dylan, Paul, Huckabee, Sam, Liana, Shane, Erika, Colton, Sierra, Connor, Jayem, B, Matt, Erik, and Morgan. You welcomed me into Reno with open arms and made me feel at home. Thank you for all your support and for all the amazing times.

TABLE OF CONTENTS

CHAPTER 1

Introduction

Proteins were identified to be polypeptides by Emil Fischer and Franz Hofmeister in 1902, and the first crystal structures were determined by Max Perutz and John Cowdery in 1958. Since then, there has been a nearly exponential growth in the number of known protein sequences and resolved protein structures, allowing us to glean some understanding of their inner workings. Knowledge of protein structure has allowed researchers to improve the stability and effectiveness of insulin,[1] design custom hyper-stable central immune cytokine mimics that support development of therapeutic candidates, and improve production of various industrial products.[2] Amino acid sequences have a large three-dimensional conformational space, yet each protein converges on a specific structure that defines its function. Further, protein structures are dynamic and can undergo conformational changes as reactions occur or solvent conditions shift.

Thioesterases catalyze the hydrolysis of thioester bonds and are present in many biochemical pathways including polyketide synthesis, lipid metabolism, and non-ribosomal peptide synthesis. Thioesterases commonly act on acyl-acyl carrier protein (acyl-ACP) or acyl-coenzyme A (acyl-CoA) substrates and their substrate specificity plays an essential role in determining the composition of fatty acids produced by fatty acid synthases. These roles can be medically important, affecting obesity, diabetes, and nonalcoholic liver disease.[3] Fatty acid metabolism is also relevant to the production of fatty acids from renewable feedstocks, where modification of thioesterases or their

expression can generate higher fatty acid yields.[4] Enzymes with many different structural folds and structures can catalyze thioesterase activity.[5]

Catalytic residues and mechanisms of thioester hydrolysis are different according to the structural fold of the thioesterase enzyme, and can even be vary within some structural folds. Thioesterases with the α/β-hydrolase fold have conserved catalytic residues resulting in a consistent mechanism where the nucleophile forms a substrate intermediate after donating a proton to the histidine which is stabilized by a nucleophile.[6] However, HotDog fold enzymes lacks conserved catalytic residues and a defined non-solvated binding pocket,[7] resulting in a variety of potential mechanisms.

Given that the number of known thioesterase primary sequences is rising to $>10^5$, an exhaustive study of enzymes is impractical, so it is important to develop methods to predict their structure, function, and mechanisms. The relationship between primary and tertiary structure is known to produce accurate predictions of tertiary structure from protein sequence, and prediction of catalytic residues through direct structural comparison is reliable. Application of these principles is essential to understand and predict enzymatic function when literature is unavailable, therefore the thioesterase enzymes were predicted into families according to the primary structure. Chapter 3 presents the updated thioesterase families in the thioester active enzyme (ThYme) database. The ThYme database contains all available sequences and structures of enzymes that hydrolyze thioester-containing substrates and groups them into 35 families based on the similarity of their primary structure, as a high degree of primary structure similarity also gives a high degree of tertiary structure similarity. The structural similarity of these families is confirmed, and catalytic residue predictions are made. The families

are available in the ThYme database, which allows users to make detailed searches and download data for local use.

Effective protein engineering does not necessarily require knowledge of protein structure; directed evolution does not require knowledge of protein sequence or structure and is an option when high-throughput screening is available, though this can be cost and time intensive. Rational and semi-rational design, where mutations are selected through knowledge of protein structure and sequence, can be more cost and time efficient.[2] The efficiency of these approaches is directly affected by the quantity and quality of information available; increasing our understanding of protein structure directly benefits effective protein design. Despite an abundance of tools available to align protein sequences and methods for protein structural superimposition, an approach that performs a spatial alignment to identify residues in different protein structures that spatially correspond is lacking. Sequence alignments do not take into account three dimensional structures, and protein superimposition approaches focus on superimposing the structures as nearly as possible, generally prioritizing the alignment of secondary structures. Here, we compare proteins by examining the relative positions of amino acids in three-dimensional space within a protein structure to quantify the mutation space of an amino acid residue. Therefore, the Mutation Space Structural Comparison method, in which the relative positions of amino acids in three-dimensional space within a protein structure are analyzed to quantify the mutation space of an amino acid residue, was developed and is presented in Chapter 4. This tool takes a set of protein structures with a selected target structure as an input. Each residue in the target structure is scored with respect to the position and composition of residues from the other proteins in the set. This results in a

heat map of the target structure that differentiates residues by their degree of spatial and compositional conservation with respect to the set.

Chapter 1 References

1. Rege NK, Wickramasinghe NP, Tustan AN, Phillips NFB, Yee VC, Ismail-Beigi F, Weiss MA (2018) Structure-based stabilization of insulin as a therapeutic protein assembly via enhanced aromatic–aromatic interactions. J. Biol. Chem. 293:10895–10910.

2. Kapoor S, Rafiq A, Sharma S (2017) Protein engineering and its applications in food industry. Crit. Rev. Food Sci. Nutr. 57:2321–2329.

3. Tillander V, Alexson SEH, Cohen DE (2017) Deactivating Fatty Acids: Acyl-CoA Thioesterase-Mediated Control of Lipid Metabolism. Trends Endocrinol. Metab. 28:473–484.

4. Lennen RM, Pfleger BF (2012) Engineering Escherichia coli to synthesize free fatty acids. Trends Biotechnol. 30:659–667.

5. Caswell BT, de Carvalho CC, Nguyen H, Roy M, Nguyen T, Cantu DC (2022) Thioesterase enzyme families: Functions, structures, and mechanisms. Protein Sci. 31:652–676.

6. Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, et al. (1992) The α/β hydrolase fold. Protein Eng. Des. Sel. 5:197–211.

7. Zhuang Z, Song F, Zhao H, Li L, Cao J, Eisenstein E, Herzberg O, Dunaway-Mariano D (2008) Divergence of Function in the Hot Dog Fold Enzyme Superfamily: The Bacterial Thioesterase YciA. Biochemistry 47:2789–2796.

CHAPTER 2

Methods Overview

This work relies on access to reliable protein sequences and structures. To update the thioesterase (TE) families in the ThYme database, it was necessary to identify all known thioesterases that have been experimentally confirmed as having thioesterase function; these were obtained from Uniprot,[1] where reviewed entries (SWISS-PROT) of protein sequences are denoted to have "Evidence at Protein Level". The SWISS-PROT database is well curated and annotated, is designed to have a minimum level of redundancy, and was built to integrate thoroughly with other databases, providing users with a dependable source of information. All known thioesterase sequences that have been experimentally verified were obtained from SWISS-PROT and were grouped into possible families using the basic local alignment search tool (BLAST).[2]

BLAST is a highly efficient sequence comparison tool that finds regions of local similarity between a target sequence and all sequences present in the database that the user searches against. Highly similar sequences to the target are assigned high bit-scores and low expectation values. The expectation value, or E-value, given to a sequence represents the likelihood that a user will find a hit with that score by chance in searching a database of a similar size. Each TE family in the ThYme database is based on a representative protein sequence that has been experimentally confirmed to have thioesterase function, and the members of each family are identified based on the results of a BLAST search, with the representative sequence as the query searching in the NCBI GenBank nr peptide sequence database[3] using the protein-protein algorithm. BLAST

searches were performed with a downloaded version of BLAST-2.9.0-2 against the GenBank nr database. The Max Target Sequences parameter was maximized, the E-value cutoff was set to $1x10^{-7}$, and other parameters were left on default settings. An E-value of $1x10^{-7}$ was used due to the significant increase in the number of sequences in the nr database, which includes all known protein sequences.

The BLAST results of each of the experimentally characterized thioesterases from SWISS-PROT were compared to each other to identify an initial set of possible families and their representative sequences. A manual review was done to exclude sequences that lacked experimentally confirmed thioesterase activity or included improper structural models. From ~200 experimentally characterized TE sequences a minimal set of representative sequences was selected that resulted in possible TE families, which should have minimal overlap. An optimal set of representative sequences is the smallest subset of sequences that contain all other experimentally characterized sequences in their collective BLAST results. A random sampling method was applied, due to the high number of permutations in a set this size, to identify the optimal representative sequences; ~50 sequences were selected to represent possible families. A representative sequence was chosen at random and that sequence and any sequences remaining in the set that were present in its BLAST result were removed from the set. Families were selected in this manner until no sequences remained in the set, and the representative sequences were recorded. This process was repeated approximately 10,000 times, and the most optimal results were found to contain ~50 families. Thirty-five TE families were identified and assessed for redundancy, completeness, and structural similarity, and are described in Chapter 3.

The Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB)[4] has been instrumental in the completion of this work thanks to their expansive and well annotated library of resolved protein structures, mostly from crystallography. PDB structures were used to confirm the TE families identified in the ThYme database (Chapter 3), as well as for developing the quantification of mutation space (Chapter 4).

To confirm structural similarity in TE families, all available PDB structures within each family were downloaded and superimposed using MultiProt, a highly efficient multiple structural alignment tool that can reliably superimpose a large numbers of protein structures.[5] Multiprot was used with OnlyRefMol set to 1, scoring set to 2, and all other parameters left on default settings. The output of this program provided modified PDB files for each structure, with their cartesian coordinates appropriately translated and rotated to be superimposed with the reference molecule.

The structural similarity within each family was quantified with the root mean square distance (RMSD) between each known PDB structure in a family and the reference structure of that family. For these calculations, the distances between corresponding α-carbon atoms for two superimposed structures were calculated. The average distance between adjacent α-carbon atoms in the reference structure was used as a cutoff distance, the maximum distance that could separate corresponding α-carbon atoms between the two structures while still considering them to be correlated. If the distance between these was greater than the cutoff, then they were not spatially correlated. The ratio of successfully correlated residues to the total number of reference structure residues was calculated as *P* for that comparison. This comparison was

completed for each family, each time comparing the representative structure to all available PDB structures found in that family. For each family, the RMSD and *P* for each of these comparisons was calculated as $RMSD_{ave}$ and P_{ave} and serve as quantification and confirmation of the structural similarity within these protein families.

A method of structural comparison was developed to study the diversity of amino acid residues that reside in the same relative spatial position in a set of structures. Four protein families were studied in this way, two thioesterase families, a glycoside hydrolase family, and a ketoacyl synthase family. For each family, a target three-dimensional structure was selected, and all other known tertiary structures from that family were the subject structures. The positions of the α-carbon atoms were extracted from the available PDB structures and used to represent the location of each residue. For each family, the sequences of the proteins that have resolved three-dimensional structures underwent a multiple sequence alignment using the MUSCLE.[6] The sequence-conserved residues indicated by MUSCLE were identified for each structure. For each structure, the cartesian coordinates of these sequence-conserved residues were averaged and used as a center of mass reference point for that structure. The two sequence-conserved residues that were separated by the greatest spatial distance were selected as two other reference points, with care taken to select the same two residues for all structures in the set. Each molecule in the set was then rotated about its central reference point in a consistent manner with respect to the two other selected reference points such that all molecules in the set were oriented in nearly the same direction, regardless of their original location and orientation in cartesian space. Following this, the position of each residue in each protein structure was redefined from cartesian coordinates to a set of three vectors: one from the sequence-

conserved center of mass to that residue, and one from each of the two selected reference points to that residue. This conversion is crucial for our intended method of study: by redirecting the structures in a consistent manner and defining each point as vectors in relation to common reference points, we can compare the spatial positions of residues by similarity of their vector fields, without requiring structural superimposition. This allows us to study the space that amino acids in proteins inhabit and the shape of the structure in a manner that sequence alignment does not directly offer.

A measure of how many different amino acid residues in a target structure can be found in a spatial site is quantified by the mutation space of spatially conserved residues (MSSC) score, defined by a formula of our design. The MSSC score utilizes Grantham's distance to compare the physicochemical properties (composition, polarity, and molecular volume) of corresponding amino acid residues.[7] A low Grantham's distance indicates a greater degree of physicochemical similarity, and a high Grantham's distance indicates a lower degree of similarity between those two amino acids. The MSSC formula also considers the number of residues that occupy the same spatial location to each target residue, and the number of unique, non-identical amino acids that were correlated to each target residue. The MSSC score quantifies the mutation space of each residue in the target structure based on the overall similarity of residues from structures in the comparison set that were located in the same spatial position as that target residue. These scores were used to create a normalized heat-map for each target structure, visualizing which residue positions had the most and least consistent composition with respect to the set of comparison structures.

Data analysis, when not specifically done using external software, was completed using custom scripts that were written primarily in python 3.8. Few external libraries were necessary for our analyses; standard libraries like *os* and *regex* were used for their expected functions, and *numpy* was used for non-trivial mathematical functions.

Chapter 2 References

1. Bateman A (2019) UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 47:D506–D515.

2. Altschup SF, Gish W, Miller W, Myers EW, Lipman DJ Basic Local Alignment Search Tool. 1990.

3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic Acids Res. 37:D26--D31.

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res. 28:235–242.

5. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins Struct. Funct. Genet. 56:143–156.

6. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

7. Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 185:862–864.

CHAPTER 3

Thioesterase enzyme families: functions, structures, and mechanisms

Benjamin T. Caswell, Caio C. de Carvalho, Hung Nguyen, Monikrishna Roy, Tin Nguyen, David C. Cantu

Modified from a manuscript that was published in *Protein Science* with the same title and authors

3.1. Introduction

Thioesterases (TEs) hydrolyze thioester bonds and catalyze reactions in many different pathways such as fatty acid synthesis, polyketide synthesis, and non-ribosomal peptide synthesis. TEs are enzymes used in the biological production of tailored fatty acids and other medically relevant compounds such as macrolide antibiotics.[1–4] TEs catalyze the hydrolysis of a wide variety of thioesters; for example acyl-coenzyme A (CoA) hydrolysis occurs in the biological production of 3-hydroxybutyrate,[5] in fatty acid β-oxidation,[6,7] in vitamin K biosynthesis,[8] and in 4-chlorobenzoate dehalogenation,[9] among multiple pathways. TEs are also medically important, for example protein palmitoylation plays a role in malaria pathogenesis,[10] and acyl-CoA thioesterases are involved with fatty acid metabolism that affects obesity, diabetes, and nonalcoholic fatty liver disease in humans.[11]

Classifying enzymes by primary structure (amino acid sequence) into families allows to predict the tertiary structure of all enzymes in a family as well as to identify catalytic

residues and mechanisms. In 2010 the TE enzymes were classified into twenty-three families,[12] and placed in the publicly available ThYme (Thioester-active enzYmes) database.[13] This is particularly useful since known protein sequences vastly outnumber enzymes whose function has been experimentally characterized or whose structure has been experimentally determined.

Enzyme family classification allows to infer the structure and function of an uncharacterized sequence in an organism of interest, based on a single enzyme with a known function and structure in a family. For example, structural knowledge of bacterial enzymes in thioesterase family 14 (TE14) led to understanding substrate-protein interactions in algal thioesterases,[14] as well as to structure prediction and analysis of plant sequences in the same family.[15] Further, structural predictions and analysis of plant sequences in TE14, combined with site-directed mutagenesis, resulted in identifying the catalytic residues of the *C. viscosissima* acyl-ACP TE, relevant for the biological production of tailored fatty acids.[16] More recently, knowledge of enzyme sequences and their substrate specificity was used to predict function from structure, as recently done with acyl-ACP TEs.[17]

Since we first classified the TEs into families, the number of known protein sequences has increased by about three orders of magnitude, and more thioesterases have been experimentally characterized. New thioesterase substrate specificities have been determined: as examples, i) in TE4, a preference toward short chain fatty acids was observed in acyl-CoA thioesterases;[18] ii) RpaL, a TesB-like TE4 enzyme from *R. palustris*, was found to be active on aromatic and long and short aliphatic molecules bound to CoA;[19] iii) in TE6, YciA enzymes from *M. extorquens* were shown to be

hydrolyze ethylmalonyl–CoA for dicarboxylic acid production;[20] and , iv) aryl-CoA substrate specificity was observed for enzymes in TE13.[21]

More thioesterases have been identified since we first classified TEs into families, some which form part of existing families. As examples, i) guanosine diphosphate regulation thioesterases from *N. meningitidis* appear in TE6;[22] ii) acyl-lipid thioesterase (ALT) from *A. thaliana* in TE9;[23] iii) methylketone synthases,[24] which were originally characterized from tomato prior to the ThYme database, have also been found in *S. melongena* and *G. max* and form part of TE9;[25,26] iv) *S. oneidensis* YbgC, which was found to primarily hydrolyze short chain acyl-CoA thioesters, also forms part of TE9;[27] v) BorB, required for borrelidin biosynthesis, is a member of TE18;[28] and vi) the *I. galbana* thioesterase/carboxylesterase (IgTeCe) is in TE21.[29]

Structural knowledge about how enzymes perform thioester hydrolysis has increased; an insightful, recent review describes thioesterase structures, with a particularly useful and clear connection of catalytic residues with enzyme topology.[30] Since we first classified TEs into families, new thioesterase structures have been resolved, as examples: i) in TE4 the TesB enzyme in *Y. pestis* was crystallized;[31] as were ii) the TesB enzymes in mycobacteria;[32] iii) in TE6 the human ACOT12 enzyme structure was obtained;[33] iv) in TE11 the tertiary structure of the thioesterase involved with azinomycin biosynthesis was determined;[34] and v) in TE12, the *Synechocystis* 1,4-dihydroxy-2-naphthoyl-CoA thioesterase was crystallized.[35]

Given the increase in known sequences, structures, and experimental characterization, TE families were updated. In this work, we report thirty-five TE families: their functions and mechanisms described, their structures analyzed, catalytic residues predicted, as well

as showing the phylogenetic analysis of TE enzymes with the main structural folds. The updated TE families are available in the new, updated ThYme database (http://thyme.engr.unr.edu).

3.2. Results and discussion

Based on sequence similarity, following the approach described in the Methods section, we identified thirty-five thioesterase families almost completely unrelated by primary structure. In the following sections we discuss their functions (3.2.1) and tertiary structures and catalytic residues (3.2.2). All the TE families are based on experimentally characterized enzymes, and most include tertiary structures from crystallization.

3.2.1. Thioesterase families and their functions

Enzymes in families TE1 to TE13, TE24 to TE26, TE28, TE31 to TE35 hydrolyze substrates with various functionalities bound by a thioester to CoA. Those in TE14 to TE19, and TE30 add a water to break the thioester bonds between acyl groups and an acyl carrier protein (ACP). The enzymes in TE20, TE21, TE27, and TE29 cleave the bonds between acyl groups and other proteins. Members of TE22 and TE23 break bonds between acyl groups and glutathione and its derivatives. The thioester-carrying moiety in CoA and ACP is a pantetheine residue, while glutathione itself carries the sulfur moiety, and in non-ACP proteins the sulfur-carrying moiety is built up mainly from a cysteine residue.

For most TE families, the main function of their enzymes is thioester hydrolysis; however, TE is not the main activity for TE33 – TE35. All the reported TE families have

at least one member that has been experimentally confirmed to have thioesterase function, however some families have members that catalyze other reactions besides TE.

Some TE families include enzymes that are the TE domains of larger, multimodular proteins such as fatty acid synthases (FASs), polyketide synthases (PKSs), or non-ribosomal peptide synthases (NRPs). FASs, PKSs, and NRPs are large enzymes with multiple domains each having different functions. Only the TE domains were used to identify TE family members.

The functions of enzymes in families TE1 – TE23 are described in detail in our previous work[12], and those of families TE24 – TE35 are described here. Table 3.1 includes common names and genes, their overall function, known substrate specificities, and references for all TE families.

Enzymes in family TE24, assigned to EC 3.1.2.2, are able to hydrolyze fatty acyl-CoA molecules with varying chain lengths (C_4-C_{18}), but they usually show a preference for long chain fatty acyl groups.[36] TE24 members from *M. tuberculosis* are involved in the synthesis of mycolic acids, which are used by the organism to form a protective layer around pathogens.[37]

Members of TE25, which include EC 3.1.2.29 among others, are able to breakdown fluoroacetyl-CoA, suggesting a key metabolic step in the resistance mechanism of *S. cattley* to fluoroacetate, a well know toxic substance produced by plants as a biodefense.[38,39]

Family TE26 includes structures ybfF enzymes that hydrolyze palmitoyl-CoA and malonyl-CoA.[40] TE26 also includes alcohol acetyl transferases which could produce industrially relevant esters. The yeast *W. anomalus* showed alcohol acetyltransferase

activity with ethanol and acetyl-CoA, releasing free CoA under high acetyl-CoA concentration. Although thioester hydrolysis is not the main function of the alcohol acetyltransferases in TE26, free CoA in absence of ethanol was also reported, confirming thioesterase activity by acetyl-CoA hydrolysis.[41]

Enzymes in TE27 (EC 3.1.2.22), described as mitochondrial palmitoyl-protein thioesterases, present in mammals, include the α/β hydrolases 10 (ABHD10) enzymes. ABHD10 enzymes are related with S-palmytoilation, a reversible lipid post-translational modification.[42]

Enzymes in TE28 include mpaH, responsible for making mycophenolic acid from mycophenolyl-CoA, a natural antibiotic produced in the *P. brevicompactum* peroxisome. These enzymes have a C-terminal cyclase/thioesterase domain that catalyzes the cyclization and release of the polyketide.[43,44]

Family TE29 (EC 3.1.2.22) includes acyl-protein thioesterases (APTs). APT enzymes are known to remove palmitate from cytosolic cysteine residues, such as S-hexadecanoyl-L-cysteinyl, in the Golgi complex of *H. sapiens.*[45]

Enzymes in TE30 (EC 3.1.2.-) are known to be involved in the biosynthesis of citrinin, a mycotoxin, in *Penicillium* and *Monascus* species. Multi-domain polyketide synthases (PKSs) are associated in citrinin biosynthesis. Type I and type VII PKS enzymes have a TE domain (CitA) involved in hydrolysis of thioester bond tethered with an acyl carrier protein (ACP), releasing a free ACP and an aldehyde.[46]

Family TE31 (EC 3.1.2.2) has thioesterases that break down long-chain acyl-CoA molecules, releasing acyl chains used for reacylation of precursors of cardiolipin, a mitochondrial phospholipid found in *H. sapiens* and other mammalians.[47]

Among enzymes from TE32 (EC 3.1.2.32), those from *P. aeruginosa* can hydrolyze 2-aminobenzoylacetyl-CoA to form 2-aminobenzoylacetate and CoA, a reaction in the signaling system for the expression of virulence genes that affect the cell density.[48,49]

TE33 (EC 2.3.1.84 and EC 3.1.2.20) includes alcohol acetyltransferase (AATase) enzymes, also known as alcohol-O-acyltransferase, that in *S. cerevisiae* hydrolyze thioesters, but whose main function is not thioesterase activity. These enzymes promote the esterification of isoamyl alcohol by acetyl-CoA. TE33 members, which prefer long- and straight-chain alcohol substrates over those with short and branched-chains, transfer the acyl group from an acyl-CoA donor to an acceptor alcohol, releasing acyl esters that can be applied in the food and beverage industry as flavoring agents. Some acetate ester products are: ethyl acetate, isoamyl acetate, isobutyl acetate, butyl acetate, hexyl acetate, heptyl acetate and octyl acetate.[50,51]

Family TE34 includes citramalyl-CoA lyase (EC 2.3.3.9 or EC 3.1.2.30), a human mitochondrial enzyme involved in vitamin B_{12} metabolism that is expressed from polymorphic human genes known as CLYBL, which turns malyl-CoA into malate and free coenzyme A.[52] Also present in TE34 are malyl-CoA lyase enzymes, which are structurally similar to CitE enzymes,[53] were described as a multifunctional enzyme that plays a role in autotrophic CO_2 fixation by *C. aurantiacus*. These enzymes catalyze steps to generate (S)-malyl-CoA and β-methylmalyl-CoA in the 3-hydroxipropionate pathway.

Family TE35 (EC 3.1.1.4 and EC 3.1.2.2) includes enzymes encoded by the PLA2G6 human gene. Also known as VIA calcium-independent phospholipase A2 (iPLA$_2$β), they perform SN-2 acyl chain hydrolysis, producing free fatty acids and lysophospholipids. Also, although not their main function, these enzymes can hydrolyze the thioester bonds

from saturated long-chain fatty acyl-CoAs.[54,55]

Other enzymes that have thioesterase function, but were not classified into a family, include human mitochondrial 3-ketoacyl-CoA thiolases that are active on short, medium, or long-chain substrates to release free CoA, with the fastest rate being attributed to butyryl-CoA.[56] The main function of thiolases is a condensation of acyl groups, and not thioesterase. Ubiquitin carboxyl-terminal hydrolases[57] were not classified into TE families because peptidase activity is their main function, and they can be found in the MEROPS database.[58]

3.2.2. Thioesterase families and their structures, catalytic residues, and mechanisms

The tertiary structures in each TE family were superimposed to confirm structural similarity. Each family that underwent this analysis exhibits members very highly similar in tertiary structure; their cores are nearly identical and their overall resemblance is high. This structural similarity is shown by $RMSD_{ave}$ values of <1.4 Å and P_{ave} values of >77% in all families (see Methods section for definitions). Table 3.2 reports the structural fold of the enzymes in each family, as well as the $RMSD_{ave}$ and P_{ave} values for families with more than two known tertiary structures. Table 3.3 describes the catalytic residues, and their corresponding literature, of the structures in each TE family. We predicted catalytic residues from tertiary structure superimposition as those which spatially correspond with known catalytic residues in superimposed structures, also reported in Table 3.3. Figures 3.1 and 3.2 show how catalytic residues were predicted, based on structure superimposition and spatial correspondence, for TEs with HotDog fold (TE25) and an α/β-Hydrolase fold (TE20), respectively. Enzymes in TE23 and TE32 have available

tertiary structures, however their catalytic residues have not been proposed, and therefore predictions based on structural superimpositions were not done. Other families do not have any known tertiary structures: TE7, TE28, TE29, TE30, and TE33. Predicting catalytic residues was not necessary for TE13, TE14, TE17, TE18, TE19, TE24, TE26, and TE31 as every structure in these families has published literature indicating the catalytic residues, see Table 3.3. Within each of these families the catalytic residues are suitably conserved between structures, with the exception of TE19 and TE26, which each only have single known structures.

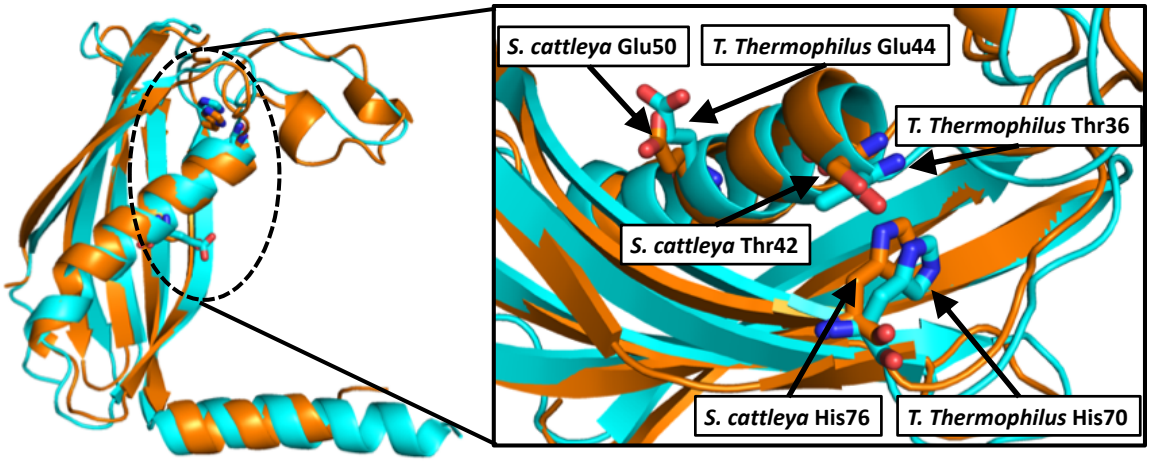

Figure 3.1: The catalytic residues of a HotDog fold enzyme in TE25 from *T. thermophilus* (cyan) were predicted based on known residues from another TE25 enzyme from *S. cattleya* (orange).

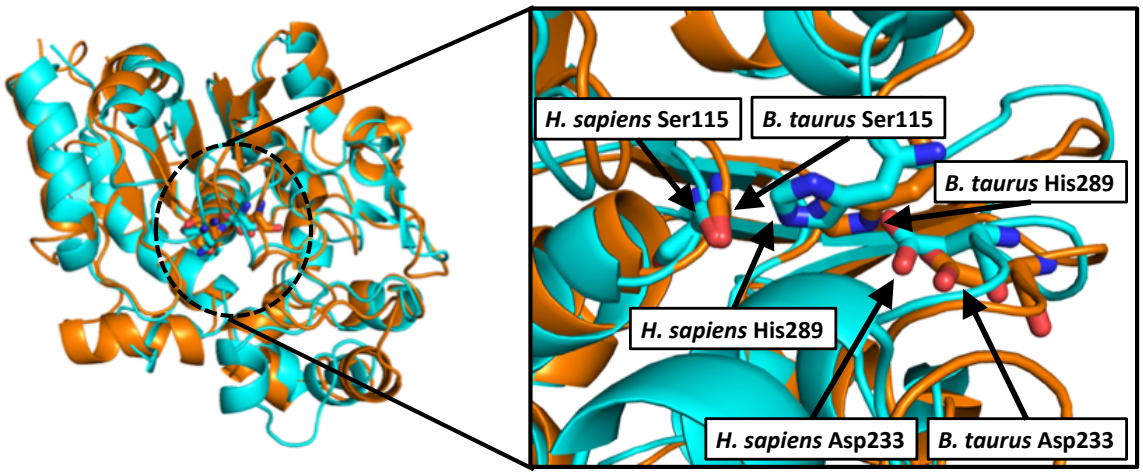

Figure 3.2: The catalytic residues of an α/β-Hydrolase fold enzyme in TE20 from *H. sapiens* (cyan) were predicted based on known residues from another TE20 enzyme from *B. taurus* (orange).

3.2.2.1. HotDog catalytic residues and mechanisms

Families with HotDog[59,60] fold structures (TE4 - TE15, TE24, TE25, TE31) have highly similar tertiary structures, indicated by the consistently low $RMSD_{ave}$ and high P_{ave} values.

HotDog-fold enzymes lack defined non-solvated binding pockets and conserved catalytic residues,[61] thus a variety of catalytic residues and mechanisms exist.

In TE4, *M. marinum* TesB2 (3U0A) catalytic residues were predicted to be Asp194–Ser216–Gln266, based on comparison to an *E. coli* thioesterase II enzyme (1C8U) in which Asp204–Gln278–Thr228 orient a water molecule for nucleophilic attack on the substrate.[62]. This is consistent with the catalytic residues found in *Y. pestis* TesB (4QFW, 4R4U); a structure that presents an octameric quaternary structure, unique among HotDog families.[31] A *S. cerevisiae* thioesterase I structure (1TBU) contains only residues

from the N-terminal domain that does not include the residues that could be compared to the catalytic triad. Catalytic residues for the remaining family members, were predicted, see Table 3.3. Of note in these predictions are *M. avium* MAV2540 (3RD7) and MAP1729c (4R9Z); these inactive TesB enzymes contain a mutation in which the highly conserved Asp residue is substituted for an Ala residue. Within TesB thioesterases, this mutation appears to be unique to *Mycobacterium* species.[32]

In TE6, *M. musculus* Acot7 N-terminal domain (2V1O) and C-terminal domain (2Q2B) catalytic residues are reported as Asn24 and Asp213 respectively.[63] The structures for human Acot12 (3B7K, 4MOB, 4MOC) and *M. musculus* Acot7 (4ZV3, 6VFY) contain both N and C-terminal domains. Our alignment placed both 2V1O and 2Q2B over the C-terminal of these structures confirm catalytic residues in the C-terminal domain. Using this molecular symmetry, the N-terminal catalytic residues were predicted as well. This follows with literature which indicates that these structures form a functioning active site when joined as a dimer.[33] A study on *N. meningitidis* thioesterase 12 (5SZU) supported these findings, pointing to a covalent disulfide bond dimer linkage that is requisite for enzymatic activity.[22] The Asn-Asp catalytic motif is highly consistent in this family, recently supported by findings on a *B. cereus* thioesterase (7CZ3).[64] Unique among the family is a *S. aureus* thioesterase (4NCP) that also relies on a Thr residue for catalysis.[65] Also in TE6, YciA structures have and aspartic acid catalytic residues in the same structural position as those in *C. jejuni* Cj0915 (3D6L) and *H. influenzae* Rd KW20 HI0827 (1YLI, 3BJK).[66,67]

Although TE7 has no known crystal structures, sequence analysis with other acyl-CoA thioesterase (ACOT) enzyme suggests that Asp120 and Asn305 are catalytic

residues in the mouse ACOT9 enzyme.[68]

It was proposed for TE8 enzymes, based on the crystal structure of a human Them2 enzyme, that Gly57 and Asn50 bind and polarize the thioester carbonyl group while Asp65 and Ser85 orient and activate the water nucleophile.[69,70] It was later proposed, based on mixed quantum mechanics / molecular mechanics simulations of the same human enzyme, that a His-Ser pair acts as the acid proton donor in a concerted mechanism where the Asp residue activates the water molecule.[71] Based on superimposition with the crystal structure of the human Them2, the structures for *M. musculus* Acot13 (2CY9) and *D. renio* Acot13 (4ORD) are predicted to have the same Asn50, His56, Gly57, Asp65 catalytic structure.[69] The position of these catalytic residues seem to be extremely highly conserved in this family; the position of the catalytic residues in 2CY9 and 2F0X are exactly the same and are only shifted by one position in 4ORD (e.g. Asp65 to Asp66).

In TE9, an *E. coli* enzyme (1S5U) is predicted to have catalytic residues Tyr14–Asp18–His25, based on a strong spatial correlation with the catalytic structure (Tyr7–Asp11–His18) of an *H. pylori* enzyme (2PZH) in the superimposed structures.[72]

It was proposed for TE10 4-hydroxybenzoyl-CoA TEs (1LO7, 1LO8, 1LO9) that a helix dipole moment make the thioester carbonyl group more susceptible to a nucleophilic attack by Asp17.[73] We predict that Asp16 in an *A. halodurans* enzyme (5WH9) is catalytic, based on the Asp17 residue of a *Pseudomonas* thioesterase (1BVQ).[74]

TE11 thioesterases in *Arthrobacter* (1Q4S, 1Q4T, 1Q4U), *E. coli* K-12 (4K49), and *A. thaliana* At2g48320 (4K02) all have nearly identically positioned glycine and glutamic

acid catalytic residues.[75,35] The crystal structures of other members of this family spatially align well, and are predicted to have the same Gly–Glu catalytic residues (Table 3.3). Members of TE11 may also act as chain elongation and cyclization domains in certain synthetic pathways.[34]

TE12 enzymes from *Synechocystis* (4K00) and *Prochlorococcus* (2HX5) bacteria have been crystallized. In 4K00, Asp16 was proposed to act as a nucleophile, while it is also possible that it acts as a base to attack the thioester through activation of a water molecule. The thioester oxygen atom could be stabilized by the amide hydrogen on Phe23. Also, Pro57, which positioned above the substrate moiety, may contribute to substrate specificity.[35]

From the structures 1WLU, 1J1Y, 1WM6, 1WLV, and 1WN3, a study proposed that TE13 *T. thermophilus* PaaI thioesterase hydrolyze substrates with an Asp48-activated water nucleophile.[76] By comparison an *E. coli* PaaI structure (2FS2) with the *Arthrobacter* TE11 structures and site-directed mutagenesis, a mechanism similar to that in TE11 was proposed: Gly53 prepares the thioester for a nucleophilic attack from Asp61.[77]

TE14, which has many bacterial sequences that have been less characterized than their plant counterparts, has a surprising breath of substrate specificity.[78] In TE14, a site-directed mutagenesis study on a FatB enzyme from *A. thaliana* pointed to a Cys264, His229, and Asn227 papain-like catalytic triad.[79] Another site-directed mutagenesis study on a FatB enzyme, from *U. californica,* proposed a catalytic network of Asp281, Asn283, His285, and Glu319.[80] More recently, structural predictions and site-directed mutagenesis resulted in identifying the catalytic residues of the *C. viscosissima* acyl-ACP TE.[16]

In TE15, a mechanism based on CalE7 enzyme (2W3X), which has no acidic residues in the catalytic region, was proposed: Asn19 and Arg37 hold the substrate while a water molecule or hydroxide anion acts as a nucleophile, and Tyr29 assists in decarboxylation.[81] Asn, Arg, and Tyr residues in a *M. chersina* tebC (2XEM, 2XFL), as well as *S.s globisporus* (4I4J) and *A. verrucosospora* (5VPJ) thioesterases are predicted to be catalytic based on spatial correspondence with the superimposed *M. echinospora* structure (2W3X).

The crystal structure for TE24 is represented by PDB 2PFC and 3B18. The quaternary structure is formed by three dimers and has a long and narrow substrate-binding site. The catalytic site is formed by Asn83, Tyr87, Tyr33 and Met118 for subunit A and Tyr66, Thr70, His72 and Asn74 for subunit B.[36] Notably, the active site lacks acidic residues common to HotDog thioesterases, which is also observed in a TE24 *Streptomyces* enzyme.[82]

In TE25, a *T. thermophilus* thioesterase (2CWZ) is predicted to have Thr36, Glu44, and His70 as catalytic residues, see Figure 3.1, based on the spatial superimposition with the catalytic residues in *S. cattleya* fIK (3KUV).[83] The specificity for fluorine-containing compounds could arise from substrate binding through a hydrophobic pocket formed by a helical lid structure (side chains of Val46 and Val54), as well as by Val23, Leu26, Phe33, and Phe36 in *S. cattleya* fIK.[39]

Family TE31 has Them4 and Them5 isoforms, which have been crystalized and are reported by the 4AE8 and 4AE7 structures respectively, forming a homodimer unity. Their structures consist of a long central alpha helix surrounded by a six stranded curved antiparallel beta-sheets.[69,84] Both isoforms are formed by two active sites per homodimer

at the end of each HotDog helix: His152, Gly153, Gly154 / His158, Gly159, Gly160 (active site one), and Asp161, Thr177 / Asp167, Thr183 (active site two).[47]

3.2.2.2. α/β hydrolase catalytic residues and mechanisms

The α/β-hydrolase fold,[85] found in TE2, TE16 to TE22, and TE26 to TE28, shows higher variation in $RMSD_{ave}$ and P_{ave} values than the HotDog fold. Most α/β-hydrolase fold proteins, not only TEs, are present in the ESTHER database.[86] Two families, TE29 and TE30, based on sequence similarity, are likely to have α/β-hydrolase-like folds, however, there are no available structures to confirm. α/β hydrolases have conserved catalytic residues: a nucleophile–histidine–acid triad.[85] Serine, cysteine, or aspartate can act as the nucleophile. There is a large variation of fold architecture and binding sites in α/β hydrolases.[87] In their catalytic mechanism, the acid stabilizes the histidine, which acts as a base by accepting a proton from the nucleophile, which forms a substrate intermediate that attacked by water. In PKSs or NRPs that make cyclic products, for example in erythromycin biosynthesis,[88] a hydroxyl group from the substrate chain is used instead of a water molecule. Different cyclization mechanisms lead to a wide variety of PKS or NRP products.[89]

The structure of TE2 is represented by 3HLK, which comes from human ACOT2, and 3K2I, which comes from human ACOT4. These structures are somewhat unique for this fold: in the primary structure for these enzymes the Asp residue precedes the His residue, where in all other α/β hydrolase thioesterases the His residue precedes the Asp residue.[90] The catalytic residues of 3K2I (Table 3.3) are predicted based on alignment

with 3HLK.

In TE16 most structures show a consistent Ser–Asp–His catalytic triad: seen in the human fatty acid synthase TE domain,[91–94] the TE domain in *Bacillus* NRPSs surfactin and fengycin synthetases,[95,96] the TE domain of the *Aspergillus* aflatoxin polyketide synthase,[97] the TE domain of *Mycobacterium* polyketide synthases involved in making mycolic acids,[98] and in the TE domain of NocB enzyme in *Nocardia*.[99] However, based on structural superimposition with TE16 structures with identified catalytic residues, we predict that the thioesterase domain of an *A. baumannii* NRPS enzyme (4ZXH, 4ZXI)[100] has a Cys–Asp–His catalytic triad (Table 3.3).

TE17 has enzymes, which are the TE domain of macrocycle-forming polyketide synthases, such as of 6-deoxyerythronolide B synthase from *S. erythraea*,[101–104] picromycin synthase from *S. venezuelae*,[101,105,106] and tautomycin synthase.[107] They all show a consistent Ser–Asp–His catalytic triad.

Member of TE18 with crystal structures are type II thioesterases, a class of enzyme responsible for a variety of functions, primarily maintenance of biosynthetic pathways through release of undesired intermediates from carrier protein domains.[28,108–114] A lid-flip conformational change is present in these enzymes and the Ser-Asp-His catalytic triad is conserved. This can be seen in the surfactin synthase from *B. subtilis*,[115] from the rifamycin biosynthetic cluster from *A. mediterranei*,[114] the borrelidin biosynthetic cluster from *Streptomyces*,[28] in the prodiginine biosynthetic pathway in *S. coelicolor*,[112] and in ClbQ and YbtT enzymes in *E. coli*.[109,110] This also holds true in a human thioesterase II and in a TesA from *M. tuberculosis.*[108,111]

In family TE19 a single structure is known, that of a *V. harveyi* thioesterase, which

also has the Ser–Asp–His catalytic triad.[116]

Families TE20, TE21, and TE22 all share the characteristic Ser–Asp–His catalytic triad. Comparison of tertiary structures within each family leads us to predict that this Ser–Asp–His catalytic triad is consistent for all structures, see Table 3.3 and Figure 3.2.

TE21 includes mainly eukaryotic acyl-protein hydrolases, as well as enzymes with different functions. The carboxylesterase from *P. fluorescens* has very little activity on triacylglycerides with fatty acids longer than four carbons, likely due to the loops constraining the active-site cleft.[117] A closely related human enzyme, hAPT1, originally thought to be a lysophospholipase, has been shown to have stronger thioesterase activity.[118] Another acyl protein thioesterase (APT), from *F. tularensis,* has a similar substrate specificity profile to both of the aforementioned enzymes, though unlike *P. fluorescens* it lacks a lid domain.[119] This was confirmed by another study that examined the mechanism of isoform-selective inhibitors on human APT1.[120] The carboxylesterase from *P. aeruginosa* was shown to have no activity on triacylglycerols, and a preference for eight-carbon acyl substrates. The human lypophospohlipase A2 is a cystolic serine hydrolase partially responsible for lysophospholipid metabolism.[121] All of these structures follow the Ser-Asp-His catalytic motif.

Members of TE22 are involved in glutathione-dependent formaldehyde detoxification, and many of the crystal structures in this family are of S-formylglutathione hydrolase (SFGH) enzymes. These have been studied in a variety of species: *S. cerevisiae*,[122] *A. fabrum str. C58*,[123] *P. translucida TAC125*,[124] S. frigidimarina,[125] and *N. meningitidis MC58*.[126] Other functions are present in this family as well: i) a human esterase has been studied because it is relevant to retinoblastoma,[127]

and ii) an oil-degrading bacterium, *O. antarctica*, expresses an enzyme with carboxylesterase and thioesterase activity.[128] TE22 enzymes have the characteristic Ser-Asp-His catalytic triad. Based on this, the catalytic structure of a *S. cerevisiae* SFGH (4FLM) is predicted as Ser161–Asp241–His276 (Table 3.3).

A study on the only crystal structures found for this family, ybfF from *E. coli* (3BF7, 3BF8), suggests that this family is unique within the α/β hydrolase thioesterases: rather than the typical Ser–Asp–His catalytic triad, this family seems to have a Ser89–Asp113–Ser206–His234 catalytic tetrad. The α/β hydrolase domain of these structures gives good alignment with other canonical α/β hydrolases. However, the Asp113 residue, which normally lies above or parallel to the His234 imidizole rings, is located in the lower section of the His imidazole ring. The expected position for the Asp113 residue is instead occupied by Ser206, which is well conserved in the ybfF enzymes.[40]

The structure of TE27 enzymes is described by a *M. musculus* ABHD10, which shows a Ser–His–Asp catalytic triad. The location of the catalytic serine residue suggests a hydrophobic interaction between the lipid substrate and the interior surface of the protein. A "cap domain" above the catalytic triad forms a binding pocket and affects substrate accessibility.[42] We predict that Ser113–Asp216–His246 is the catalytic triad in an *A. vitis* enzyme based on comparison to the *M. musculus* ABHD10 enzyme.[42]

Families TE28 and TE29 have no crystal structures. TE28 shows sequence similarity with a putative α/β hydrolase fold enzyme, and their structure and mechanisms still unknow despite a close relationship with fatty acid synthases.[43] TE29 may also have an α/β hydrolase fold, as was predicted from gene ABHD17C.[44]

The structure of an CitA enzyme in TE30, predicted by homology from a co-

expression of the PKS gene, suggests a Ser122-His235-Asp207 as catalytic triad.[46]

3.2.2.3. Catalytic residues and mechanisms in other folds

TEs are found in the NagB (TE1) and SGNH (TE3) folds.[129–133] In TE1, which also includes acyl-CoA transferases, we predict that the catalytic residues of a putative acetyl-CoA hydrolase from *P. givgivalis* (2NVV) and a CoA transferase from *P. aeruginosa* (2G39) are Val259–Glu284–Asn337–Gly378 and Ile264–Glu288–Asn341–Gly382, respectively, based on those known from *A. aceti* AarCH6 structures (4EU3, 5DDK).[134,135]

In TE3, comparison to available structures – *E. coli* tesA (e.g., 1IVN, 1JRL)[136] and *Pseudoalteromonas* estA (3HP4)[130] – reveals the likely catalytic residues for an *E. coil* thioesterase (6LFB, 6LFC) and *A. indicum* AlinE4 esterase (6IQ9, 6IQA, 6IQB) are Ser10–Asp154–His157 and Ser13–Asp162–His165 respectively. TesA enzymes were found to have a Ser–His–Asp catalytic triad similar to those in α/β hydrolases,[136] and a switch loop movement that occurs during catalysis.[137] The crystal structure of TesA from *E. coli* was found to be particularly compact and rigid, which likely pushes the substrate specificity toward smaller chain lengths.[131] It has also proved to be a useful candidate for attempts at engineering thioesterases to produce specific lengths of free fatty acids.[132] Other SGNH fold thioesterases, CrmE10 and AlinE4 were similarly susceptible to engineering for increased enzymatic activity.[133]

Two families have the β-lactamase fold: TE23 and TE32. The structures in TE23 are significantly less well conserved than those in TE32. TE23 hydroxyglutathione hydrolases, which include glyoxalase II enzymes, have a metallo-β-lactamase fold, and

their mechanisms are very different from the rest of TEs that do not have catalytic metal ions. Crystal structures of human glyoxalase II (1QH3, 1QH5) reveal two zinc ions with octahedral coordination, interacting with His and Asp residues. Based on this, a study proposed that a hydroxide ion bonded with both ions attacks the carbonyl carbon atom of the glutathione thioester substrate, forming a tetrahedral intermediate, followed by breakage of the C–S bond.[138] In mitochondrial glyoxalase II from *A. thaliana* (1XM8, 2Q42) the zinc ions were also coordinated by His and Asp residues, but were in trigonal bipyramidal and tetrahedral geometries.[139] Another glyoxylase II enzyme, from *S. typhimurium* (2QED), was proposed to have an uncommon metal affinity: a diiron, dimanganese, or hybrid Fe/Mn.[140] A unique member of the family, a persulfide dioxygenase from *M. xanthus* (4YSB), has a single ion in the active site with a two-His and one-carboxylate triad coordination pattern.[141]

Enzymes in TE32 have monomeric metallo-β-lactamase fold structures, with an Fe(II)Fe(III) center in the active site and an αβ/αβ sandwich core. All the resolved structures in this family are PqsE enzymes from *P. aeruginosa*, a human pathogen of particular interest due to its tendency for antibiotic resistance.[142] The active center of the enzyme is covered by a lid formed by two α-helices in the C-terminal region, affecting substrate access.[49] It has also been demonstrated that PqsE has a role in alkylquinolone biosynthesis.[143]

Although TE33 includes no crystal structures, a mechanism has been proposed, which shows an active site His acting as a base, with the substrate hydroxyl forming a hydrogen bond with a histidinde residue.[144–146] A nucleophilic attack from a deprotonated hydroxyl at the carbonyl of an acyl-CoA thioester was described, as was the involvement of an Asp

residue in the stabilization of the structure within the active site.[51,144,145,147]

Crystal structure 5VXS represents a member from TE34 and reveals a homotrimer with a substrate-bound cavity located between the N-terminal from one subunit and the C-terminal from the subsequent subunit. The N-terminal forms a $\beta_8\alpha_8$ -TIM barrel fold and the C-terminal is characterized by a lid-domain consisting of two helices connect by a β-hairpin loop. The β-hairpin loop presents a highly conserved Asp320 that removes a proton from the substrate during the catalytic activity.[52,53,148,149] In TE34, the catalytic residues for a *M. tuberculosis* (6AQ4), *C. sphaeroides* (4L9Y, 4L9Z), and *M. extorquens* (5UGR) enzymes are predicted to be Asp261, Asp299, and Asp304, respectively, based on comparison to human CLYBL structure (5VXS).[52] The catalytic residues for the remaining family members could not be confidently predicted by structural comparison. Two of these are CitE proteins from *M. tuberculosis*: one study (1U5H) predicts that the catalytic site is in a hydrophobic cavity formed by the C-terminal tips of the TIM β-barrel,[150] while another study (6AQ4) shows that the active site contains an Mg^{2+} ion coordinated by the ligand, Glu112, Asp138, and two water molecules.[151] Closely related to 1U5H is *Y. pestis* RipC (3QLL), for which the active site is similarly predicted. However, it is also suggested that the active site for 3QLL may be formed through an intermonomer interaction.[152]

The structure 6AUN in TE35 is characterized by the presence of an Ankyrin domain, a 33-residue helix-turn-helix structure followed by a hairpin-like loop, and a catalytic domain. Regarding the catalytic mechanisms, a dyad formed by Ser-Asp is responsible for the lipid hydrolysis.[153,154]

3.2.4. Updated ThYme database

All the sequences and structures in the TE families described here appear in the ThYme database,[13] which is in the process of being completely updated and has a new home at the University of Nevada, Reno (http://thyme.engr.unr.edu). Families, their member sequences, taxonomical data, accession codes, and protein names can be viewed using the ThYme database online interface. The database has links to UniProt[155], GenBank[156], and Protein Data Bank[157] databases. Although the content of families will be updated automatically, human judgement will still be necessary for adding, merging, or deleting families.

In the new ThYme website, each enzyme class (e.g., TEs) will have an interactive interface where users can narrow to viewing content of a single family or multiple families. Each unique sequence is displayed as a row containing: the family, the organism, protein names, protein identifiers, protein evidence information, crystal structures, gene names, and finally gene and pathway identifiers. Each entry will display, at the minimum, the family and a protein identifier; all other fields will be populated if suitable data is available. The content has multiple search fields such as name or identifier, and results can be narrowed to show only entries with evidence at protein level or known crystal structures.

3.3. Conclusions

Thioesterase families have been updated through analysis of the primary structures of all known thioesterase sequences. New families have been proposed, and all sequences and structures are classified into new, or previously identified, families. This system of

classification provides a standardized nomenclature and a means to predict the tertiary structure, function, and mechanism of a thioesterase sequence that has not been experimentally characterized. These assertions are supported by family members displaying a high degree of primary and tertiary structural similarity, highly conserved active sites and catalytic residues, and consistent mechanisms. Examination of families that share a fold reveals some similarity in primary and tertiary structures, catalytic residues and active sites, and mechanisms. Convergent and divergent evolution is suggested from phylogenetic analyses of thioesterases whose structures have the two main structural folds.

3.4. Methods

For a sequence to be considered a member of a family it must have a strong sequence similarity (~30%), a nearly identical tertiary structure to other structures in the family, and catalytic residues in the same locations as the other members of that family.

The protocol by which the new thioesterase families were identified is described: i) enzyme sequences experimentally confirmed to have thioesterase activity are gathered and those present in a previously existing family (TE1 – TE23) were discarded; ii) each of the remaining thioesterase sequences are independently processed by the Basic Local Alignment Search Tool (BLAST)[158] and results were compared with the other sequences' results to identify the representative sequences that will originate new families; iii) the catalytic domains of the representative sequences were processed by BLAST to populate potential new families; iv) the number of shared sequences were counted for all permutations of pairs of potential new families, highly similar families (>15% sequences

in common) were merged; v) intra-family congruity and inter-family uniqueness were confirmed by tertiary structure superimposition, comparison of catalytic residue position and identity, multiple sequence alignments (MSAs), and final examination of shared sequences between all possible pairs of families; vi) sequences common to multiple families are assigned to the family with the highest sequence similarity.

3.4.1. Sequence selection and BLAST searches

Enzyme sequences experimentally confirmed to have thioesterase activity were extracted from the Swiss-Prot database in Uniprot[159] which contains only reviewed sequences and has a higher level of annotation. Possible thioesterases were identified by a label of EC 3.1.2.1 to EC 3.1.2.32, EC 3.1.2.–, or having "thioesterase" in the description, as well as having "Evidence at Protein Level". Less stringently verified sequences, like those with "Evidence at Transcript level" or "Inferred from Homology", as well as fragments or theoretical proteins, were disregarded. The primary sequences meeting the criteria, and not in TE1 – TE23, were collected, resulting in ~200 new query sequence candidates.

Each of these sequences was subjected to a BLAST search against the National Center for Bio-technology Information's (NCBI) GenBank nr peptide sequence database using the protein-protein algorithm.[160] These BLAST searches were completed using a local instance of blast-2.9.0-2 and the nr database, both downloaded from NCBI on a Unix system. Previously, an E-value cutoff of $1x10^{-3}$ was used;[12] however, due to the growth of the nr database by ~3 orders of magnitude, an E-value of $1x10^{-7}$ was used to capture as many sequences with the required similarity as possible while minimizing the number of redundant sequences. The highest Max Target Sequences was used to capture

all sequences within an E-value of $1x10^{-7}$. Other parameters were left at default settings.

BLAST results were compared against each other to check for common sequences and identify the representative sequences that results in the lowest number of BLAST results with no overlapping, common sequences. The query sequences of unique, non-redundant BLAST results become the representative sequences that will originate new families from all confirmed thioesterase sequences. The referenced literature in Uniport is checked to confirm experimental thioesterase activity. The catalytic domain of each of the new representative sequences, identified in Pfam-A,[161] were used to populate the prospective new families with BLAST as described above.

3.4.2. Comparison of tertiary structures

All known tertiary structures in each family was obtained from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB).[157] Enzyme tertiary structures were reviewed to exclude fragments, putative proteins, and non-TE domains from multidomain proteins from any structural comparisons.

All monomer structures were extracted, and for each family a reference structure was selected, which served as the pivot around which other monomers were superimposed. The shortest monomer in each family was selected as the pivot to ensure consistent alignment of the core structure and allow for uniform structural similarity calculations. All monomers within each family were superimposed using MultiProt[162] with OnlyRefMol set to 1, Scoring set to 2, and all other parameters left at default.

A root mean square distance (RMSD) of the superimposed tertiary structures in each family with more than one structure was done to quantify structural similarity. For

RMSD calculations, the distances between corresponding alpha carbon atoms (C_α) from two superimposed structures (pivot and subject) were calculated. A cutoff distance, calculated as the average distance between sequential C_αs in the pivot structure, was used to determine corresponding C_αs between the pivot and subject structures. Any pairs more distant than the cutoff were not considered to be corresponding and were not used in the RMSD calculation. The percentage value (P) of C_αs used to calculate the RMSD implies the significance of the RMSD calculation. For a given family, the pivot structure was superimposed to all other structures, resulting in $n - 1$ calculations, where n is the number of monomers being compared within that family. For families where $n > 2$, the average RMSD and P values ($RMSD_{ave}$ and P_{ave} respectively) were calculated.

Chapter 3 Tables

Table 3.1. *Thioesterase Families, Gene and Enzyme Names, Functions, and Substrate Specificities*

Family	Genes and/or Enzyme Names	General Function	Known Substrate Specificities	References
TE1	Ach1	Acyl-CoA hydrolase	Acetyl-CoA	163,164
TE2	Acot1–Acot6 BAAT thioesterase	Acyl-CoA hydrolase	Palmitoyl-CoA Bile-acid-CoA	165,166
TE3	*tesA* *estA* acyl-CoA thioesterase I protease I lysophospholipase L1	Acyl-CoA hydrolase	Medium- to long-chain acyl-CoA	167,168
TE4	*tesB* acyl-CoA thioesterase II Acot8	Acyl-CoA hydrolase	Short-chain acyl-CoA Short- to long-chain acyl-CoA Palmitoyl-CoA Choloyl-CoA	18,169,170
TE5	*tesC* (ybaW) acyl-CoA thioesterase III	Acyl-CoA hydrolase	Long-chain acyl-CoA 3,5-tetradecadienoyl-CoA	171
TE6	Acot7 (BACH) Acot11 (BFIT, Them1) Acot12 (CACH) YciA	Acyl-CoA hydrolase	Short- to long-chain acyl-CoA, C_4 to C_{18} Ethylmalonyl-CoA	20,33,61,67,172–174
TE7	Acot9, Acot10	Acyl-CoA hydrolase	Short- to long-chain acyl-CoA	68,175
TE8	Acot13 (Them2)	Acyl-CoA hydrolase	Short- to long-chain acyl-CoA, C_6 to C_{18}	176
TE9	YbgC acyl-lipid thioesterase (ALT) metylketone synthases (MKS)	Acyl-CoA hydrolase	Short-chain acyl-CoA Short- to long-chain acyl-CoA 4-hydroxybenzoyl-CoA	23,27,72,177,178
TE10	4HBT-I	Acyl-CoA hydrolase	4-Hydroxybenzoyl-CoA	74
TE11	4HBT-II EntH (YbdB) menI DHNAT1 1,4-Dihydroxy-2-naphthoyl-CoA hydrolase AziG	Acyl-CoA hydrolase	4-Hydroxybenzoyl-CoA	34,179
TE12	1,4-Dihydroxy-2-naphthoyl-CoA hydrolase	Acyl-CoA hydrolase	1,4-Dihydroxy-2-napthoyl-CoA	180
TE13	paaI paaD	Acyl-CoA hydrolase	Short and medium-chain acyl-CoA, Hydroxyphenylacetyl-CoA Aryl-CoA	21,181
TE14	FatA FatB	Acyl-ACP hydrolase	Short- to long-chain acyl-ACP	78,182

TE15	Thioesterase CalE7	Acyl-ACP hydrolase	—	[81]
TE16	Thioesterase I type I thioesterase TE domain of FAS TE domain of PKS or NRP	Acyl-ACP hydrolase	Long-chain acyl-ACP Polyketides Non-ribosomal peptides	[183–185]
TE17	TE domain of PKS	Acyl-ACP hydrolase	Polyketides	[184]
TE18	Thioesterase II type II thioesterase (TE II) tesA rifR OLAH SAST	Acyl-ACP hydrolase	Medium-chain acyl-ACP Polyketides Non-ribosomal peptides	[108,114,186–188]
TE19	luxD	Acyl-ACP hydrolase	Myristoyl-ACP	[189]
TE20	ppt1 ppt2 palmitoyl-protein thioesterase	Acyl-protein hydrolase	Palmitoyl-protein	[190–192]
TE21	apt1 apt2 acyl-protein thioesterase phospholipase carboxylesterase	Acyl-protein hydrolase	Thioacylate proteins Palmitoyl-protein	[193,194]
TE22	S-formylglutathione hydrolase esterase A esterase D	Glutathione hydrolase	S-formylglutathione	[195]
TE23	Hydroxyglutathione hydrolase glyoxalase II	Glutathione hydrolase	D-lactoylglutathione	[196,197]
TE24	Fcot-like thioesterase Type III thioesterase CmiS1	Acyl-CoA hydrolase	Palmitoyl-CoA Stearoyl-CoA Lauroyl-CoA Hexanoyl-CoA	[36,37,82]
TE25	Fluoroacetyl-CoA thioesterase	Acyl-CoA hydrolase	Fluoroacetyl-CoA	[38,39]
TE26	EAT1 ybfF	Acyl-CoA hydrolase	Acetyl-CoA Palmitoyl-CoA Malonyl-CoA	[40,41]
TE27	ABHD10 Palmitoyl-protein thioesterase	Acyl-protein hydrolase	S-palmitoyl-protein	[42]
TE28	mpaH Type I acyl-CoA thioesterase	Acyl-CoA hydrolase	Malonyl-CoA	[43,44]
TE29	ABHD17A, ABHD17B ABHD17C	Acyl-protein hydrolase	S-hexadecanoyl-L-cysteinyl	[45]
TE30	citA lovG mlcF mpL1	Acyl-ACP hydrolase	Malonyl-ACP Acetoacetyl-ACP	[46]

	afoC mokD			
TE31	Them4 Them5	Acyl-CoA hydrolase	Long-chain acyl-CoA	[47]
TE32	ACAA2 3-ketoacyl-CoA thiolase	Acyl-CoA hydrolase	2-aminobenzoylacetyl-CoA	[48,49]
TE33	ATF1 Alcohol O-acetyltransferase	Alcohol acetyl transferases	Acyl-CoA	[50,51]
TE34	CLYBL Citramalyl-CoA lyase citE RipC	Citramalyl-CoA lyase	Malyl-CoA	[52,84,151,152]
TE35	PLA2G6 Calcium-independent phospholipase A2	Calcium-independent phospholipase	Long-chain fatty acyl-CoAs	[54,55]

Table 3.2. *Thioesterase Folds and Structure Superimposition*

Family	Fold	$RMSD_{ave}$, Å	P_{ave}, %	Structures in the Protein Data Bank
TE1	NagB	0.92	95.7	2G39, 2NVV, 4EU3, 4EU4, 4EU5, 4EU6, 4EU7, 4EU8, 4EU9, 4EUA, 4EUB, 4EUC, 4EUD, 5DDK, 5DW4, 5DW5, 5DW6, 5E5H
TE2	α/β-Hydrolase	0.86	94.6	3HLK, 3K2I
TE3	SGNH	0.92	87.4	1IVN, 1J00, 1JRL, 1U8U, 1V2G, 3HP4, 4JGG, 5TIC, 5TID, 5TIE, 5TIF, 6IQ9, 6IQA, 6IQB, 6LFB, 6LFC, 7C23, 7C29, 7C2A, 7C82, 7C84
TE4	HotDog	1.09	81.6	1C8U, 1TBU, 3RD7, 3U0A, 4QFW, 4R4U, 4R9Z
TE5	HotDog	—	—	1NJK
TE6	HotDog	1.09	86.9	1YLI, 2EIS, 2G6S, 2Q2B, 2QQ2, 2V1O, 3B7K, 3BJK, 3D6L, 4IEN, 4MOB, 4MOC, 4ZV3, 5DM5, 5SZU, 5SZV, 5SZY, 5SZZ, 5T02, 5V3A, 4NCP, 5EGJ, 5EGK, 5EGL, 5HWF, 5HZ4, 6VFY, 7CZ3
TE7	Putative HotDog	—	—	—
TE8	HotDog	0.56	97.7	2CY9, 2F0X, 2H4U, 3F5O, 4ORD
TE9	HotDog	0.48	96.7	1S5U, 2PZH, 5KL9, 5T06, 5T07
TE10	HotDog	1.01	94.2	1BVQ, 1LO7, 1LO8, 1LO9, 1Z54, 5WH9
TE11	HotDog	0.90	98.4	1Q4S, 1Q4T, 1Q4U, 1SBK, 1SC0, 1VH5, 1VH9, 1VI8, 2B6E, 3LZ7, 3R32, 3R34, 3R35, 3R36, 3R37, 3R3A, 3R3B, 3R3C, 3R3D, 3R3F, 3S4K, 3TEA, 4K02, 4K49, 4K4A, 4K4B, 4K4C, 4K4D, 4M20, 4QD7, 4QD8, 4QD9, 4QDA, 4QDB, 4YBV, 5EP5, 5HMB, 5HMC
TE12	HotDog	0.92	88.3	2HX5, 4K00
TE13	HotDog	0.49	98.8	1J1Y, 1PSU, 1WLU, 1WLV, 1WM6, 1WN3, 2DSL, 2FS2
TE14	HotDog	1.36	81.3	2ESS, 2OWN, 4GAK, 5X04
TE15	HotDog	0.85	96.2	2W3X, 2XEM, 2XFL, 4I4J, 5VPJ
TE16	α/β-Hydrolase	1.40	64.5	1JMK, 1XKT, 2CB9, 2CBG, 2K2Q, 2PX6, 3ILS, 3TJM, 4Z49, 4ZXH, 4ZXI, 5V3W, 5V3X, 5V3Y, 5V3Z, 5V40, 5V41, 5V42, 6OJC, 6OJD
TE17	α/β-Hydrolase	1.23	79.2	1KEZ, 1MN6, 1MNA, 1MNQ, 1MO2, 2H7X, 2H7Y, 2HFJ, 2HFK, 3LCR, 5D3K, 5D3Z, 6MLK
TE18	α/β-Hydrolase	1.16	77.0	3FLA, 3FLB, 3QMV, 3QMW, 4XJV, 5UGZ, 6BA8, 6BA9, 6FVJ, 6FW5, 6VAP
TE19	α/β-Hydrolase	—	—	1THT
TE20	α/β-Hydrolase	0.69	90.6	1EH5, 1EI9, 1EXW, 1PJA, 3GRO
TE21	α/β-Hydrolase	1.03	85.6	1AUO, 1AUR, 1FJ2, 3CN7, 3CN9, 3U0V, 4F21, 4FHZ, 4FTW, 5KRE, 5SYM, 5SYN, 6AVV, 6AVW, 6AVX, 6AVY, 6BJE, 6QGN, 6QGO, 6QGQ, 6QGS
TE22	α/β-Hydrolase	0.90	95.6	1PV1, 3C6B, 3E4D, 3FCX, 3I6Y, 3LS2, 3S8Y, 4B6G, 4FLM, 4FOL, 6JZL
TE23	Lactamase	1.24	82.6	1QH3, 1QH5, 1XM8, 2Q42, 2QED, 3TP9, 4YSB,

				6RZ0, 6S0I
TE24	HotDog	0.85	93.1	2PFC, 3B18, 5WSX, 5WSY
TE25	HotDog	0.71	99.5	2CWZ, 3KUV, 3KUW, 3KV7, 3KV8, 3KVI, 3KVU, 3KVZ, 3KW1, 3KX7, 3KX8, 3P2Q, 3P2R, 3P2S, 3P3F, 3P3I
TE26	α/β-Hydrolase	—[a]	—	3BF7, 3BF8
TE27	α/β-Hydrolase	1.06	85.2	3LLC, 6NY9
TE28	Putative α/β-Hydrolase	—	—	—
TE29	Putative α/β-Hydrolase	—	—	—
TE30	Putative α/β-Hydrolase	—	—	—
TE31	HotDog	0.54	98.5	4AE7, 4AE8, 4GAH
TE32	Lactamase	0.31	1.00	2Q0I, 2Q0J, 2VW8, 3DH8, 5HIO, 5HIP, 5HIQ, 5HIS
TE33	—	—	—	—
TE34	Beta-hairpin (C-terminal) TIM barrel (N-termina)	1.15	87.4	1SGJ, 1U5H, 1U5V, 1Z6K, 3QLL, 4L9Y, 4L9Z, 5UGR, 5VXC, 5VXO, 5VXS, 6AQ4
TE35	—	—	—	6AUN

[a] RMSD and P_{ave} for TE26 were not calculated because the two PDB entries are of the same protein structure.

Table 3.3. *Thioesterase Families and Catalytic Residues*

Family	Catalytic Residues	Corresponding Structure	Producing Organism	Reference
TE1	Val270, Glu294, Asn347, Gly388	4EU3, 4EU4, 4EU5, 4EU6, 4EU7, 4EU8, 4EU9, 4EUA, 4EUB, 4EUC, 4EUD	*Acetobacter aceti*	[134]
	Val270, Glu294, Asn347, Gly388	5DDK, 5DW4, 5DW5, 5DW6, 5E5H	*Acetobacter aceti*	[135]
	Val259, Glu284, Asn337, Gly378	2NVV	*Porphyromonas givgivalis*	Predicted in this work
	Ile264, Glu288, Asn341, Gly382	2G39	*Pseudomonas aeruginosa*	Predicted in this work
TE2	Ser294, His422, Asp388	3HLK	*Homo sapiens*	[90]
	Ser232, His360, Asp326	3K2I	*Homo sapiens*	Predicted in this work
TE3	Ser10, Asp154, His157	1IVN, 1JRL, 1J00, 1U8U, 1V2G	*Escherichia coli*	[136]
	Ser11, Asp158, His161	3HP4	*Pseudoalteromonas sp.*	[130]
	Ser9, Asp156, His159	4JGG	*Pseudomonas aeruginosa*	[131]
	Ser10, Asp154, His157	5TIC, 5TID, 5TIE, 5TIF	*Escherichia coli*	[132]
	Ser10, Asp154, His157	6LFB, 6LFC	*Escherichia coli*	Predicted in this work
	Ser29, Asp178, His181	7C23, 7C29, 7C2A, 7C82, 7C84	*Croceicoccus marinus*	[133]
	Ser13, Asp162, His165	6IQ9, 6IQA, 6IQB	*Altericroceibacterium indicum*	Predicted in this work
TE4	Asp204, Thr228, Gln278	1C8U	*Escherichia coli*	[62]
	Asp194, Ser216, Gln266	3U0A	*Mycobacterium marinum M*	Predicted in this work
	Asp204, Thr228, Gln278	4QFW, 4R4U	*Yersinia pestis*	[31]
	—	1TBU	*Saccharomyces cerevisiae*	—
	Ala202, Leu225, Gln275[a]	3RD7	*Mycobacterium avium 104*	Predicted in this work
	Ala197, Gln 216, Gln266[a]	4R9Z	*Mycobacterium avium subsp. paratuberculosis K-10*	Predicted in this work
TE5	—	1NJK	*Escherichia coli*	—
TE6	Asp213	2Q2B	*Mus musculus*	[63]
	Asn24	2V1O		
	Asp44	1YLI, 3BJK	*Haemophilus influenzae Rd KW20*	[66]
	Asp34	3D6L	*Campylobacter jejuni*	[67]
	Asp36, Asn195	3B7K, 4MOB, 4MOC	*Homo sapiens*	Predicted in this work
	Asp245	2QQ2	*Homo sapiens*	Predicted in this work

	Asp46	5DM5	*Yersinia pestis*	Predicted in this work
	Asp31	2EIS	*Thermus thermophilus*	Predicted in this work
	Asn70, Asp259	4ZV3, 6VFY	*Mus musculus*	Predicted in this work
	Asn24, Asp39	4IEN, 5SZU, 5SZV, 5SZY, 5SZZ, 5T02, 5V3A	*Neisseria meningitidis*	[22]
	Asn28, Asp43, Thr60	4NCP, 5EGJ, 5EGK, 5EGL, 5HWF, 5HZ4	*Staphylococcus aureus, subsp. Aureus Mu50*	[65]
	Asn23, Asp38	7CZ3	*Bacillus cereus ATCC 14579*	[64]
TE7	—	—	—	—
TE8	Asn50, His56, Gly57, Asp65	2F0X, 3F5O, 2H4U	*Homo sapiens*	[69,70]
	Asn50, His56, Gly57, Asp65	2CY9	*Mus musculus*	Predicted in this work
	Asn51, His57, Gly58, Asp66	4ORD	*Danio rerio*	Predicted in this work
	Asp65, Ser83, His134	Simulation[%]	*Homo sapiens*	[71]
TE9	Tyr7, Asp11, His18	2PZH	*Helicobacter pylori*	[72]
	Tyr14, Asp18, His25	1S5U, 5KL9, 5T06, 5T07	*Escherichia coli*	Predicted in this work
TE10	Asp17	1BVQ, 1LO7, 1LO8, 1LO9	*Pseudomonas sp.*	[73]
	Asp16	5WH9	*Alkalihalobacillus halodurans C-125*	Predicted in this work
TE11	Gly65, Glu73	1Q4S, 1Q4T, 1Q4U	*Arthrobacter sp.*	[75]
	Gly55, Glu63	1VH9, 1VH5, 1VI8, 1SBK	*Escherichia coli*	Predicted in this work
	Gly55, Glu63	2B6E, 1SC0, 3LZ7	*Haemophilus influenzae*	Predicted in this work
	Gly39, Glu47	4M20, 4YBV, 5EP5	*Staphylococcus aureus, subsp. Aureus Mu50*	Predicted in this work
	Gly65, Ala73	3R32, 3R34, 3R35, 3R36, 3R37, 3R3A, 3R3B, 3R3C, 3R3D, 3R3F, 3TEA	*Arthrobacter sp.*	Predicted in this work
	Gly52, Glu60	3S4K	*Mycobacterium tuberculosis*	Predicted in this work
	Gly55, Glu63	4K49, 4K4A, 4K4B, 4K4C, 4K4D	*Escherichia coli K-12*	[198]
	Gly56, Glu64	4QD7, 4QD8, 4QD9, 4QDA, 4QDB	*Pseudomonas aeruginosa*	Predicted in this work
	Gly49, Glu57	5HMB, 5HMC	*Streptomyces sahachiroi*	Predicted in this work
	Gly49, Glu57	4K02	*Arabidopsis thaliana*	[35]
TE12	Asp16	2HX5	*Prochlorococcus marinus*	Predicted in this work
	Asp16	4K00	*Synechocystis sp. PCC*	[35]

			6803 substr. Kazusa	
TE13	Gly40, Asp48	1WLU, 1J1Y, 1WM6, 1WLV, 1WN3, 2DSL[b]	*Thermus thermophilius*	[76]
	Gly53, Asp61	2FS2, 1PSU	*Escherichia coli*	[77]
TE14	Asp281, Asn283, His285, Glu319	2ESS	*Bacteroides thetaiotaomicron VPI-5482*	[80]
	Asp281, Asn283, His285, Glu319	2OWN	*Lactiplantibacillus plantarum*	[80]
	Asp281, Asn283, His285, Glu319	4GAK	*Spirosoma linguale DSM 74*	[80]
	Asp281, Asn283, His285, Glu319	5X04	*Umbellulaia californica*	[80]
TE15	Asn19, Tyr29, Arg37	2W3X	*Micromonospora echinospora*	[81]
	Asn23, Tyr33, Arg41	2XEM, 2XFL	*Micromonospora chersina*	Predicted in this work
	Asn21, Tyr31, Arg39	4I4J	*Streptomyces globisporus*	Predicted in this work
	Asn17, Tyr27, Arg35	5VPJ	*Actinomadura verrucosospora*	Predicted in this work
TE16	Ser2308, Asp2338, His2481	1XKT, 2PX6, 3TJM, 4Z49	*Homo sapiens*	[91]
	Ser80, Asp107, His207	1JMK	*Bacillus subtilis*	[95]
	Ser84, Asp 111, His201	2CB9, 2CBG	*Bacillus subtilis*	[96]
	Ser1937, Asp1964, His2088	3ILS	*Aspergillus parasiticus*	[97]
	Cys1135, Asp1162, His1295	4ZXH, 4ZXI	*Acinetobacter baumannii AB307-0294*	Predicted in this work
	Ser1533, Asp1560, His1699	5V3W, 5V3X, 5V3Y, 5V3Z, 5V40, 5V41, 5V42	*Mycobacterium tuberculosis*	[98]
	Ser1790, Asp1806, His1901	6OJC, 6OJD	*Nocardia uniformis subsp. tsuyamanensis*	[99]
TE17	Ser142, Asp169, His259	1KEZ, 1MO2, 5D3K, 5D3Z, 6MLK	*Saccaropolyspora erythaea*	[102]
	Ser148, Asp176, His268	1MN6, 1MNA, 1MNQ, 2H7X, 2H7Y, 2HFJ, 2HFK	*Streptomyces venezuelae*	[101]
	Ser132, Asp159, His255	3LCR	*Streptomyces sp. CK4412*	[107]
TE18	Ser86, Asp189, His216	2K2Q, 2RON	*Brevibacillus parabrevis, Bacillus subtilis*	[199]
	Ser94, Asp200, His228	3FLA, 3FLB	*Amycolatopsis mediterranei*	[114]
	Ser107, Asp213, His241	3QMV, 3QMW	*Streptomyces coelicolor*	[112]
	Ser101, Asp212, His237	4XJV	*Homo sapiens*	[111]
	Ser78, Asp186, His215	5UGZ	*Escherichia coli*	[110]

	Ser89, Asp197, His225	6BA8, 6BA9	*Escherichia coli*	[109]
	Ser104, Asp208, His236	6FVJ, 6FW5	*Mycobacterium tuberculosis*	[108]
	Ser98, Asp204, His232	6VAP	*Streptomyces sp. WAC02707*	[28]
TE19	Ser114, Asp211, His241	1THT	*Vibrio harveyi*	[116]
TE20	Ser115, Asp233, His289	1EH5, 1EI9, 1EXW	*Bos taurus*	[200]
	Ser111, Asp228, His283	1PJA, 3GRO	*Homo sapiens*	Predicted in this work
TE21	Ser114, Asp168, His199	1AUO, 1AUR	*Pseudomonas fluorescens*	[117]
	Ser114, Asp169, His203	1FJ2	*Homo sapiens*	[118]
	Ser113, Asp166, His197	3CN7, 3CN9	*Pseudomonas aeruginosa*	[201]
	Ser124, Asp179, Glu212	3U0V	*Homo sapiens*	Predicted in this work
	Ser116, Asp170, His 202	4F21	*Francisella tularensis subsp. tularensis SCHU S4*	[119]
	Ser165, Asp216, His248	4FHZ, 4FTW	*Cereibacter sphaeroides*	Predicted in this work
	Ser119, Asp174, His209	5SYM	*Homo sapiens*	[120]
	Ser122, Asp176, His210	5SYN	*Homo sapiens*	
	Ser106, Asp160, His192	6AVV, 6AVW, 6AVX	*Arabidopsis thaliana*	Predicted in this work
	Ser126, Asp197, His230	6AVY	*Zea mays*	Predicted in this work
	Ser122, Asp176, His210	6BJE	*Homo sapiens*	[121]
	Ser119, Asp174, His208	6QGN, 6QGO, 6QGQ, 6QGS	*Homo sapiens*	Predicted in this work
TE22	Ser161, Asp241, His276	1PV1, 3C6B	*Saccharomyces cerevisia*	[122]
	Ser147, Asp223, His256	3E4D	*Agrobacterium fabrum str. C58*	[123]
	Ser153, Asp230, His264	3FCX	*Homo sapiens*	[127]
	Ser148, Asp224, His257	3I6Y, 3S8Y	*Oleispira antarctica*	[128]
	Ser147, Asp225, His258	3LS2	*Pseudoalteromonas translucida TAC125*	[124]
	Ser145, Asp221, His254	4B6G	*Neisseria meningitidis MC58*	[126]
	Ser161, Asp241, His276	4FLM, 4FOL	*Saccaromyces cerevisia*	Predicted in this work
	Ser148, Asp224, His257	6JZL	*Shewanella frigidimarina*	[125]

TE23	—	—[b]	—	—
TE24	Asn83, Tyr87, Tyr33, and Met118 (subunit A) and Tyr66, Thr70, His72, and Asn74 (subunit B)	2PFC, 3B18	*Mycobacterium tuberculosis*	[36]
	Tyr53, Ile54, His59, Asn61, and Ser62 (subunit A) and Tyr20, Asn70, Met73, Tyr74, and Ile107 (subunit B)	5WSX, 5WSY	*Streptomyces avermitilis MA-4680 = NBRC 14893*	[82]
TE25	Thr42, Glu50, His76, and a water molecule	3KUV, 3KUW, 3KV7, 3KV8, 3KVI, 3KVU, 3KVZ, 3KW1, 3KX7, 3KX8	*Streptomyces cattleya*	[83]
	Thr36, Glu44, His70	2CWZ	*Thermus thermophilus HB8*	Predicted in this work
	Thr42, Glu50, His76	3P2Q, 3P2R, 3P2S, 3P3F, 3P3I	*Streptomyces cattleya*	[39]
TE26	Ser89, Asp113, Ser206, His234	3BF7, 3BF8	*Escherichia coli*	[40]
TE27	Ser100, Asp197, His227	6NY9	*Mus musculus*	[42]
	Ser113, Asp216, His246	3LLC	*Agrobacterium vitis S4*	Predicted in this work
TE28	—	—	—	—
TE29	—	—	—	—
TE30	—	—	—	—
TE31	Thr308, Ser473	4AE7, 4AE8, 4GAH	*Homo sapiens*	[47]
TE32	—	—[b]	—	—
TE33	—	—	—	—
TE34	Asp320	5VXS, 5VXC, 5VXO	*Homo sapiens*	[52]
	—	1SGJ	*Deinococcus radiodurans*	—
	—	1U5H, 1U5V, 1Z6K	*Mycobacterium tuberculosis*	—
	Glu49	6AQ4	*Mycobacterium tuberculosis*	[151]
	—	3QLL	*Yersinia pestis*	—
	Asp299	4L9Y, 4L9Z	*Cereibacter sphaeroides 2.4.1*	Predicted in this work
	Asp304	5UGR	*Methylorubrum extorquens AM1*	Predicted in this work
TE35	Ser465, Asp598	6AUN	*Cricetulus griseus*	[153]

[%]Predicted from mixed quantum mechanics / molecular mechanics simulations based on the 3F5O crystal structure

[a] Catalytic residue prediction for 3RD7 was based purely on their high degree of spatial correlation with the catalytic residues of 1C8U and 4QFW. It is noted that these residues do not have a high degree of chemical similarity.

[b] Even though structures are known, catalytic residues have not been determined, so none are predicted

Chapter 3 References

1. Lennen RM, Pfleger BF (2012) Engineering Escherichia coli to synthesize free fatty acids. Trends Biotechnol. 30:659–667.

2. Zhang X, Li M, Agrawal A, San K-Y (2011) Efficient free fatty acid production in Escherichia coli using plant acyl-ACP thioesterases. Metab. Eng. 13:713–722.

3. Tang M-C, Fischer CR, Chari J V, Tan D, Suresh S, Chu A, Miranda M, Smith J, Zhang Z, Garg NK, et al. (2019) Thioesterase-Catalyzed Aminoacylation and Thiolation of Polyketides in Fungi. J. Am. Chem. Soc. 141:8198–8206.

4. Paiva P, Medina FE, Viegas M, Ferreira P, Neves RPP, Sousa JPM, Ramos MJ, Fernandes PA (2021) Animal Fatty Acid Synthase: A Chemical Nanofactory. Chem. Rev. 121:9502–9553.

5. Guevara-Martínez M, Perez-Zabaleta M, Gustavsson M, Quillaguamán J, Larsson G, van Maris AJA (2019) The role of the acyl-CoA thioesterase "YciA" in the production of (R)-3-hydroxybutyrate by recombinant Escherichia coli. Appl. Microbiol. Biotechnol. 103:3693–3704.

6. De Marcos Lousa C, van Roermund CWT, Postis VLG, Dietrich D, Kerr ID, Wanders RJA, Baldwin SA, Baker A, Theodoulou FL (2013) Intrinsic acyl-CoA thioesterase activity of a peroxisomal ATP binding cassette transporter is required for transport and metabolism of fatty acids. Proc. Natl. Acad. Sci. 110:1279 LP – 1284.

7. Hunt MC, Siponen MI, Alexson SEH (2012) The emerging role of acyl-CoA thioesterases and acyltransferases in regulating peroxisomal lipid metabolism. Biochim. Biophys. Acta - Mol. Basis Dis. 1822:1397–1410.

8. Widhalm JR, Ducluzeau A-L, Buller NE, Elowsky CG, Olsen LJ, Basset GJC (2012) Phylloquinone (vitamin K1) biosynthesis in plants: two peroxisomal thioesterases of lactobacillales origin hydrolyze 1,4-dihydroxy-2-naphthoyl-coa. Plant J. 71:205–215.

9. Song F, Thoden JB, Zhuang Z, Latham J, Trujillo M, Holden HM, Dunaway-Mariano D (2012) The Catalytic Mechanism of the Hotdog-fold Enzyme Superfamily 4-Hydroxybenzoyl-CoA Thioesterase from Arthrobacter sp. Strain SU. Biochemistry 51:7000–7016.

10. Jones ML, Collins MO, Goulding D, Choudhary JS, Rayner JC (2012) Analysis of Protein Palmitoylation Reveals a Pervasive Role in Plasmodium Development and Pathogenesis. Cell Host Microbe 12:246–258.

11. Tillander V, Alexson SEH, Cohen DE (2017) Deactivating Fatty Acids: Acyl-CoA Thioesterase-Mediated Control of Lipid Metabolism. Trends Endocrinol. Metab. 28:473–484.

12. Cantu DC, Chen Y, Reilly PJ (2010) Thioesterases: A new perspective based on their primary and tertiary structures. Protein Sci. 19:1281–1295.

13. Cantu DC, Chen Y, Lemons ML, Reilly PJ (2011) ThYme: a database for thioester-active enzymes. Nucleic Acids Res. 39:D342–D346.

14. Blatti JL, Beld J, Behnke CA, Mendez M, Mayfield SP, Burkart MD (2012) Manipulating Fatty Acid Biosynthesis in Microalgae for Biofuel through Protein-Protein Interactions. PLoS One 7:1–12.

15. Marika Z, Nathan R, Aashna S, Brendon D, Gordon W, Silver PA, Way JC (2018) Chimeric Fatty

Acyl-Acyl Carrier Protein Thioesterases Provide Mechanistic Insight into Enzyme Specificity and Expression. Appl. Environ. Microbiol. 84:e02868-17.

16. Jing F, Yandeau-Nelson MD, Nikolau BJ (2018) Identification of active site residues implies a two-step catalytic mechanism for acyl-ACP thioesterase. Biochem. J. 475:3861–3873.

17. Banerjee D, Jindra MA, Linot AJ, Pfleger BF, Maranas CD (2021) EnZymClass: Substrate specificity prediction tool of plant acyl-ACP thioesterases based on Ensemble Learning. bioRxiv [Internet]:2021.07.06.451235. Available from: http://biorxiv.org/content/early/2021/07/06/2021.07.06.451235.abstract

18. McMahon MD, Prather KL (2014) Functional Screening and In Vitro Analysis Reveal Thioesterases with Enhanced Substrate Specificity Profiles That Improve Short-Chain Fatty Acid Production in Escherichia coli. Appl. Environ. Microbiol. 80:1042–1050.

19. Hickman TWP, Baud D, Benhamou L, Hailes HC, Ward JM (2020) Characterisation of four hotdog-fold thioesterases for their implementation in a novel organic acid production system. Appl. Microbiol. Biotechnol. 104:4397–4406.

20. Sonntag F, Buchhaupt M, Schrader J (2014) Thioesterases for ethylmalonyl–CoA pathway derived dicarboxylic acid production in Methylobacterium extorquens AM1. Appl. Microbiol. Biotechnol. 98:4533–4544.

21. Sánchez-Reyez A, Batista-García RA, Valdés-García G, Ortiz E, Perezgasga L, Zárate-Romero A, Pastor N, Folch-Mallol JL (2017) A family 13 thioesterase isolated from an activated sludge metagenome: Insights into aromatic compounds metabolism. Proteins Struct. Funct. Bioinforma. 85:1222–1237.

22. Khandokar YB, Srivastava P, Cowieson N, Sarker S, Aragao D, Das S, Smith KM, Raidal SR, Forwood JK (2017) Structural insights into GDP-mediated regulation of a bacterial acyl-CoA thioesterase. J. Biol. Chem. 292:20461–20471.

23. Pulsifer IP, Lowe C, Narayaran SA, Busuttil AS, Vishwanath SJ, Domergue F, Rowland O (2014) Acyl-lipid thioesterase1-4 from Arabidopsis thaliana form a novel family of fatty acyl-acyl carrier protein thioesterases with divergent expression patterns and substrate specificities. Plant Mol. Biol. 84:549–563.

24. Yu G, Nguyen TTH, Guo Y, Schauvinhold I, Auldridge ME, Bhuiyan N, Ben-Israel I, Iijima Y, Fridman E, Noel JP, et al. (2010) Enzymatic Functions of Wild Tomato Methylketone Synthases 1 and 2. Plant Physiol. 154:67–77.

25. Khuat VLU, Bui VTT, Tran HTD, Truong NX, Nguyen TC, Mai PHH, Dang TLA, Dinh HM, Pham HTA, Nguyen TTH (2019) Characterization of Solanum melongena Thioesterases Related to Tomato Methylketone Synthase 2. Genes 10.

26. Tran HT, Le NT, Khuat VL, Nguyen TT (2019) Identification and Functional Characterization of a Soybean (Glycine max) Thioesterase that Acts on Intermediates of Fatty Acid Biosynthesis. Plants 8.

27. Gao T, Meng Q, Gao H (2017) Thioesterase YbgC affects motility by modulating c-di-GMP levels in Shewanella oneidensis. Sci. Rep. 7:3932.

28. Curran SC, Pereira JH, Baluyot MJ, Lake J, Puetz H, Rosenburg DJ, Adams P, Keasling JD (2020) Structure and Function of BorB, the Type II Thioesterase from the Borrelidin Biosynthetic Gene Cluster. Biochemistry 59:1630–1639.

29. Kerviel V, Hérault J, Dumur J, Ergan F, Poisson L, Loiseau C (2014) Cloning and expression of a gene

from Isochrysis galbana modifying fatty acid profiles in Escherichia coli. J. Appl. Phycol. 26:2109–2115.

30. Swarbrick CMD, Nanson JD, Patterson EI, Forwood JK (2020) Structure, function, and regulation of thioesterases. Prog. Lipid Res. 79:101036.

31. Swarbrick CMD, Perugini MA, Cowieson N, Forwood JK (2015) Structural and functional characterization of TesB from Yersinia pestis reveals a unique octameric arrangement of hotdog domains. Acta Crystallogr. Sect. D Biol. Crystallogr. 71:986–995.

32. Swarbrick CMD, Bythrow G V., Aragao D, Germain GA, Quadri LEN, Forwood JK (2017) Mycobacteria Encode Active and Inactive Classes of TesB Fatty-Acyl CoA Thioesterases Revealed through Structural and Functional Analysis. Biochemistry 56:1460–1472.

33. Swarbrick CMD, Roman N, Cowieson N, Patterson EI, Nanson J, Siponen MI, Berglund H, Lehtiö L, Forwood JK (2014) Structural basis for regulation of the human acetyl-CoA thioesterase 12 and interactions with the steroidogenic acute regulatory protein-related lipid transfer (START) domain. J. Biol. Chem. 289:24263–24274.

34. Mori S, Simkhada D, Zhang H, Erb MS, Zhang Y, Williams H, Fedoseyenko D, Russell WK, Kim D, Fleer N, et al. (2016) Polyketide Ring Expansion Mediated by a Thioesterase, Chain Elongation and Cyclization Domain, in Azinomycin Biosynthesis: Characterization of AziB and AziG. Biochemistry 55:704–714.

35. Furt F, Allen WJ, Widhalm JR, Madzelan P, Rizzo RC, Basset G, Wilson MA (2013) Functional convergence of structurally distinct thioesterases from cyanobacteria and plants involved in phylloquinone biosynthesis. Acta Crystallogr. Sect. D Biol. Crystallogr. 69:1876–1888.

36. Wang F, Langley R, Gulten G, Wang L, Sacchettini JC (2007) Identification of a Type III Thioesterase Reveals the Function of an Operon Crucial for Mtb Virulence. Chem. Biol. 14:543–551.

37. Gurvitz A, Hiltunen JK, Kastaniotis AJ (2009) Heterologous expression of mycobacterial proteins in Saccharomyces cerevisiae reveals two physiologically functional 3-hydroxyacyl-thioester dehydratases, HtdX and HtdY, in addition to HadABC and HtdZ. J. Bacteriol. 191:2683–2690.

38. Huang F, Haydock SF, Spiteller D, Mironenko T, Li TL, O'Hagan D, Leadlay PF, Spencer JB (2006) The Gene Cluster for Fluorometabolite Biosynthesis in Streptomyces cattleya: A Thioesterase Confers Resistance to Fluoroacetyl-Coenzyme A. Chem. Biol. 13:475–484.

39. Weeks AM, Coyle SM, Jinek M, Doudna JA, Chang MCY (2010) Structural and biochemical studies of a fluoroacetyl-CoA-specific thioesterase reveal a molecular basis for fluorine selectivity. Biochemistry 49:9269–9279.

40. Park SY, Lee SH, Lee J, Nishi K, Kim YS, Jung CH, Kim JS (2008) High-resolution Structure of ybfF from Escherichia coli K12: A Unique Substrate-binding Crevice Generated by Domain Arrangement. J. Mol. Biol. 376:1426–1437.

41. Kruis AJ, Levisson M, Mars AE, van der Ploeg M, Garcés Daza F, Ellena V, Kengen SWM, van der Oost J, Weusthuis RA (2017) Ethyl acetate production by the elusive alcohol acetyltransferase from yeast. Metab. Eng. 41:92–101.

42. Cao Y, Qiu T, Kathayat RS, Azizi SA, Thorne AK, Ahn D, Fukata Y, Fukata M, Rice PA, Dickinson BC (2019) ABHD10 is an S-depalmitoylase affecting redox homeostasis through peroxiredoxin-5. Nat. Chem. Biol. 15:1232–1240.

43. Regueira TB, Kildegaard KR, Hansen BG, Mortensen UH, Hertweck C, Nielsen J (2011) Molecular basis for mycophenolic acid biosynthesis in Penicillium brevicompactum. Appl. Environ. Microbiol. 77:3035–3043.

44. Zhang W, Cao S, Qiu L, Qi F, Li Z, Yang Y, Huang S, Bai F, Liu C, Wan X, et al. (2015) Functional characterization of MpaG′, the O-methyltransferase involved in the biosynthesis of mycophenolic acid. ChemBioChem 16:565–569.

45. Lin DTS, Conibear E (2015) ABHD17 proteins are novel protein depalmitoylases that regulate N-Ras palmitate turnover and subcellular localization. Elife 4:1–14.

46. Storm PA, Townsend CA (2017) In trans hydrolysis of carrier protein-bound acyl intermediates by CitA during citrinin biosynthesis. Chem. Commun. 54:50–53.

47. Zhuravleva E, Gut H, Hynx D, Marcellin D, Bleck CKE, Genoud C, Cron P, Keusch JJ, Dummler B, Esposti MD, et al. (2012) Acyl Coenzyme A Thioesterase Them5/Acot15 Is Involved in Cardiolipin Remodeling and Fatty Liver Development. Mol. Cell. Biol. 32:2685–2697.

48. Drees SL, Fetzner S (2015) PqsE of Pseudomonas aeruginosa acts as pathway-specific thioesterase in the biosynthesis of alkylquinolone signaling molecules. Chem. Biol. 22:611–618.

49. Yu S, Jensen V, Seeliger J, Feldmann I, Weber S, Schleicher E, Häussler S, Blankenfeldt W (2009) Structure elucidation and preliminary assessment of hydrolase activity of PqsE, the Pseudomonas quinolone signal (PQS) response protein. Biochemistry 48:10298–10307.

50. Minetoki T, Bogaki T, Iwamatsu A, Fujii T, Hamachi M (1993) The Purification, Properties and Internal Peptide Sequences of Alcohol Acetyltransferase Isolated from Saccharomyces cerevisiae Kyokai No. 7. Biosci. Biotechnol. Biochem. 57:2094–2098.

51. Nancolas B, Bull ID, Stenner R, Dufour V, Curnow P (2017) Saccharomyces cerevisiae Atf1p is an alcohol acetyltransferase and a thioesterase in vitro. Yeast 34:239–251.

52. Shen H, Campanello GC, Flicker D, Grabarek Z, Hu J, Luo C, Banerjee R, Mootha VK (2017) The Human Knockout Gene CLYBL Connects Itaconate to Vitamin B12. Cell 171:771--782.e11.

53. Zarzycki J, Kerfeld CA (2013) The crystal structures of the tri-functional Chloroflexus aurantiacus and bi-functional Rhodobacter sphaeroides malyl-CoA lyases and comparison with CitE-like superfamily enzymes and malate synthases. BMC Struct. Biol. 13.

54. Engel LA, Jing Z, O'Brien DE, Sun M, Kotzbauer PT (2010) Catalytic function of PLA2G6 is impaired by mutations associated with infantile neuroaxonal dystrophy but not dystonia-parkinsonism. PLoS One 5.

55. Jenkins CM, Yan W, Mancuso DJ, Gross RW (2006) Highly selective hydrolysis of fatty acyl-CoAs by calcium-independent phospholipase A2$β$: Enzyme autoacylation and acyl-CoA-mediated reversal of calmodulin inhibition of phospholipase A2 activity. J. Biol. Chem. 281:15615–15624.

56. Kiema TR, Harijan RK, Strozyk M, Fukao T, Alexson SEH, Wierenga RK (2014) The crystal structure of human mitochondrial 3-ketoacyl-CoA thiolase (T1): Insight into the reaction mechanism of its thiolase and thioesterase activities. Acta Crystallogr. Sect. D Biol. Crystallogr. 70:3212–3225.

57. Rose IA, Warms JVB (1983) An enzyme with ubiquitin carboxy-terminal esterase activity from reticulocytes. Biochemistry 22:4234–4237.

58. Rawlings ND, Barrett AJ, Bateman A (2010) MEROPS: the peptidase database. Nucleic Acids Res.

38:D227–D233.

59. Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL (1996) Structure of a dehydratase–isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. Structure 4:253–264.

60. Dillon SC, Bateman A (2004) The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. BMC Bioinformatics 5:109.

61. Zhuang Z, Song F, Zhao H, Li L, Cao J, Eisenstein E, Herzberg O, Dunaway-Mariano D (2008) Divergence of Function in the Hot Dog Fold Enzyme Superfamily: The Bacterial Thioesterase YciA. Biochemistry 47:2789–2796.

62. Li J, Derewenda U, Dauter Z, Smith S, Derewenda ZS (2000) Crystal structure of the Escherichia coli thioesterase II, a homolog of the human Nef binding enzyme. Nat. Struct. Biol. 7:555–559.

63. Forwood JK, Thakur AS, Guncar G, Marfori M, Mouradov D, Meng W, Robinson J, Huber T, Kellie S, Martin JL, et al. Structural basis for recruitment of tandem hotdog domains in acyl-CoA thioesterase 7 and its role in inflammation. 2007.

64. Park J, Kim YJ, Lee D, Kim KJ (2021) Structural basis for nucleotide-independent regulation of acyl-CoA thioesterase from Bacillus cereus ATCC 14579. Int. J. Biol. Macromol. 170:390–396.

65. Khandokar Y, Srivastava P, Raidal S, Sarker S, Forwood JK (2020) Structural basis for disulphide-CoA inhibition of a butyryl-CoA hexameric thioesterase. J. Struct. Biol. 210.

66. Willis MA, Zhuang Z, Song F, Howard A, Dunaway-Mariano D, Herzberg O (2008) Structure of YciA from Haemophilus influenzae (HI0827), a Hexameric Broad Specificity Acyl-Coenzyme A Thioesterase,. Biochemistry 47:2797–2805.

67. Yokoyama T, Choi K-J, Bosch AM, Yeo H-J (2009) Structure and function of a Campylobacter jejuni thioesterase Cj0915, a hexameric hot dog fold enzyme. Biochim. Biophys. Acta - Proteins Proteomics 1794:1073–1081.

68. Tillander V, Arvidsson Nordström E, Reilly J, Strozyk M, Van Veldhoven PP, Hunt MC, Alexson SEH (2014) Acyl-CoA thioesterase 9 (ACOT9) in mouse may provide a novel link between fatty acid and amino acid metabolism in mitochondria. Cell. Mol. Life Sci. 71:933–948.

69. Cao J, Xu H, Zhao H, Gong W, Dunaway-Mariano D (2009) The Mechanisms of Human Hotdog-fold Thioesterase 2 (hTHEM2) Substrate Recognition and Catalysis Illuminated by a Structure and Function Based Analysis,. Biochemistry 48:1293–1304.

70. Cheng Z, Song F, Shan X, Wei Z, Wang Y, Dunaway-Mariano D, Gong W (2006) Crystal structure of human thioesterase superfamily member 2. Biochem. Biophys. Res. Commun. 349:172–177.

71. Cantu DC, Ardèvol A, Rovira C, Reilly PJ (2014) Molecular mechanism of a hotdog-fold Acyl-CoA thioesterase. Chem. - A Eur. J. 20:9045–9051.

72. Angelini A, Cendron L, Goncalves S, Zanotti G, Terradot L (2008) Structural and enzymatic characterization of HP0496, a YbgC thioesterase from Helicobacter pylori. Proteins Struct. Funct. Bioinforma. 72:1212–1221.

73. Thoden JB, Holden HM, Zhuang Z, Dunaway-Mariano D (2002) X-ray Crystallographic Analyses of Inhibitor and Substrate Complexes of Wild-type and Mutant 4-Hydroxybenzoyl-CoA Thioesterase. J. Biol.

Chem. 277:27468–27476.

74. Benning MM, Wesenberg G, Liu R, Taylor KL, Dunaway-Mariano D, Holden HM (1998) The Three-dimensional Structure of 4-Hydroxybenzoyl-CoA Thioesterase from Pseudomonas sp. Strain CBS-3. J. Biol. Chem. 273:33572–33579.

75. Thoden JB, Zhuang Z, Dunaway-Mariano D, Holden HM (2003) The Structure of 4-Hydroxybenzoyl-CoA Thioesterase from Arthrobacter sp. strain SU. J. Biol. Chem. 278:43709–43716.

76. Kunishima N, Asada Y, Sugahara M, Ishijima J, Nodake Y, Sugahara M, Miyano M, Kuramitsu S, Yokoyama S, Sugahara M (2005) A novel induced-fit reaction mechanism of asymmetric hot dog thioesterase PaaI. J. Mol. Biol. 352:212–228.

77. Song F, Zhuang Z, Finci L, Dunaway-Mariano D, Kniewel R, Buglino JA, Solorzano V, Wu J, Lima CD (2006) Structure, Function, and Mechanism of the Phenylacetate Pathway Hot Dog-fold Thioesterase PaaI. J. Biol. Chem. 281:11028–11038.

78. Jing F, Cantu DC, Tvaruzkova J, Chipman JP, Nikolau BJ, Yandeau-Nelson MD, Reilly PJ (2011) Phylogenetic and experimental characterization of an acyl-ACP thioesterase family reveals significant diversity in enzymatic specificity and activity. BMC Biochem. 12:44.

79. Mayer KM, Shanklin J (2007) Identification of amino acid residues involved in substrate specificity of plant acyl-ACP thioesterases using a bioinformatics-guided approach. BMC Plant Biol. 7:1.

80. Feng Y, Wang Y, Liu J, Liu Y, Cao X, Xue S (2017) Structural Insight into Acyl-ACP Thioesterase toward Substrate Specificity Design. ACS Chem. Biol. 12:2830–2836.

81. Kotaka M, Kong R, Qureshi I, Ho QS, Sun H, Liew CW, Goh LP, Cheung P, Mu Y, Lescar J, et al. (2009) Structure and Catalytic Mechanism of the Thioesterase CalE7 in Enediyne Biosynthesis. J. Biol. Chem. 284:15739–15749.

82. Chisuga T, Miyanaga A, Kudo F, Eguchi T (2017) Structural analysis of the dual-function thioesterase SAV606 unravels the mechanism of Michael addition of glycine to an α,β-unsaturated thioester. J. Biol. Chem. 292:10926–10937.

83. Dias MVB, Huang F, Chirgadze DY, Tosin M, Spiteller D, Dry EFV, Leadlay PF, Spencer JB, Blundell TL (2010) Structural basis for the activity and substrate specificity of fluoroacetyl-CoA thioesterase FIK. J. Biol. Chem. 285:22495–22504.

84. Pidugu LS, Maity K, Ramaswamy K, Surolia N, Suguna K (2009) Analysis of proteins with the “hot dog” fold: prediction of function and identification of catalytic residues of hypothetical proteins. BMC Struct. Biol. 9:37.

85. Ollis, David L, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, et al. (1992) The α/β hydrolase fold. Protein Eng. Des. Sel. 5:197–211.

86. Lenfant N, Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A (2013) ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to explore diversity of functions. Nucleic Acids Res. 41:D423–D429.

87. Nardini M, Dijkstra BW (1999) α/β Hydrolase fold enzymes: the family keeps growing. Curr. Opin. Struct. Biol. 9:732–737.

88. Chen X-P, Shi T, Wang X-L, Wang J, Chen Q, Bai L, Zhao Y-L (2016) Theoretical Studies on the

Mechanism of Thioesterase-Catalyzed Macrocyclization in Erythromycin Biosynthesis. ACS Catal. 6:4369–4378.

89. Adrover-Castellano ML, Schmidt JJ, Sherman DH (2021) Biosynthetic Cyclization Catalysts for the Assembly of Peptide and Polyketide Natural Products. ChemCatChem 13:2095–2116.

90. Mandel CR, Tweel B, Tong L (2009) Crystal structure of human mitochondrial acyl-CoA thioesterase (ACOT2). Biochem. Biophys. Res. Commun. 385:630–633.

91. Chakravarty B, Gu Z, Chirala SS, Wakil SJ, Quiocho FA (2004) Human fatty acid synthase: Structure and substrate selectivity of the thioesterase domain. Proc. Natl. Acad. Sci. U. S. A. 101:15567 LP – 15572.

92. Pemble CW, Johnson LC, Kridel SJ, Lowther WT (2007) Crystal structure of the thioesterase domain of human fatty acid synthase inhibited by Orlistat. Nat. Struct. Mol. Biol. 14:704–709.

93. Zhang W, Chakravarty B, Zheng F, Gu Z, Wu H, Mao J, Wakil SJ, Quiocho FA (2011) Crystal structure of FAS thioesterase domain with polyunsaturated fatty acyl adduct and inhibition by dihomo-γ-linolenic acid. Proc. Natl. Acad. Sci. 108:15757–15762.

94. Park I-H, Venable JD, Steckler C, Cellitti SE, Lesley SA, Spraggon G, Brock A (2015) Estimation of Hydrogen-Exchange Protection Factors from MD Simulation Based on Amide Hydrogen Bonding Analysis. J. Chem. Inf. Model. 55:1914–1925.

95. Bruner SD, Weber T, Kohli RM, Schwarzer D, Marahiel MA, Walsh CT, Stubbs MT (2002) Structural basis for the cyclization of the lipopeptide antibiotic surfactin by the thioesterase domain SrfTE. Structure 10:301–310.

96. Samel SA, Wagner B, Marahiel MA, Essen LO (2006) The Thioesterase Domain of the Fengycin Biosynthesis Cluster: A Structural Base for the Macrocyclization of a Non-ribosomal Lipopeptide. J. Mol. Biol. 359:876–889.

97. Korman TP, Crawford JM, Labonte JW, Newman AG, Wong J, Townsend CA, Tsai SC (2010) Structure and function of an iterative polyketide synthase thioesterase domain catalyzing Claisen cyclization in aflatoxin biosynthesis. Proc. Natl. Acad. Sci. U. S. A. 107:6246–6251.

98. Aggarwal A, Parai MK, Shetty N, Wallis D, Woolhiser L, Hastings C, Dutta NK, Galaviz S, Dhakal RC, Shrestha R, et al. (2017) Development of a Novel Lead that Targets M. tuberculosis Polyketide Synthase 13. Cell 170:249-259.e25.

99. Patel KD, d'Andrea FB, Gaudelli NM, Buller AR, Townsend CA, Gulick AM (2019) Structure of a bound peptide phosphonate reveals the mechanism of nocardicin bifunctional thioesterase epimerase-hydrolase half-reactions. Nat. Commun. 10.

100. Drake EJ, Miller BR, Shi C, Tarrasch JT, Sundlov JA, Leigh Allen C, Skiniotis G, Aldrich CC, Gulick AM (2016) Structures of two distinct conformations of holo-non-ribosomal peptide synthetases. Nature [Internet] 529:235–238. Available from: https://doi.org/10.1038/nature16163

101. Tsai S-C, Lu H, Cane DE, Khosla C, Stroud RM (2002) Insights into Channel Architecture and Substrate Specificity from Crystal Structures of Two Macrocycle-Forming Thioesterases of Modular Polyketide Synthases,. Biochemistry 41:12598–12606.

102. Tsai SC, Miercke LJW, Krucinski J, Gokhale R, Chen JCH, Foster PG, Cane DE, Khosla C, Stroud RM (2001) Crystal structure of the macrocycle-forming thioesterase domain of the erythromycin polyketide synthase: Versatility from a unique substrate channel. Proc. Natl. Acad. Sci. U. S. A. 98:14808–

14813.

103. Li X, Sevillano N, La Greca F, Hsu J, Mathews II, Matsui T, Craik CS, Khosla C (2018) Discovery and Characterization of a Thioesterase-Specific Monoclonal Antibody That Recognizes the 6-Deoxyerythronolide B Synthase. Biochemistry 57:6201–6208.

104. Argyropoulos P, Bergeret F, Pardin C, Reimer JM, Pinto A, Boddy CN, Schmeing TM (2016) Towards a characterization of the structural determinants of specificity in the macrocyclizing thioesterase for deoxyerythronolide B biosynthesis. Biochim. Biophys. Acta - Gen. Subj. 1860:486–497.

105. Giraldes JW, Akey DL, Kittendorf JD, Sherman DH, Smith JL, Fecik RA (2006) Structural and mechanistic insights into polyketide macrolactonization from polyketide-based affinity labels. Nat. Chem. Biol. 2:531–536.

106. Akey DL, Kittendorf JD, Giraldes JW, Fecik RA, Sherman DH, Smith JL (2006) Structural basis for macrolactonization by the pikromycin thioesterase. Nat. Chem. Biol. 2:537–542.

107. Scaglione JB, Akey DL, Sullivan R, Kittendorf JD, Rath CM, Kim ES, Smith JL, Sherman DH (2010) Biochemical and structural characterization of the tautomycetin thioesterase: Analysis of a stereoselective polyketide hydrolase. Angew. Chemie - Int. Ed. 49:5726–5730.

108. Nguyen PC, Nguyen VS, Martin BP, Fourquet P, Camoin L, Spilling CD, Cavalier JF, Cambillau C, Canaan S (2018) Biochemical and Structural Characterization of TesA, a Major Thioesterase Required for Outer-Envelope Lipid Biosynthesis in Mycobacterium tuberculosis. J. Mol. Biol. 430:5120–5136.

109. Ohlemacher SI, Xu Y, Kober DL, Malik M, Nix JC, Brett TJ, Henderson JP (2018) YbtT is a low-specificity type II thioesterase that maintains production of the metallophore yersiniabactin in pathogenic enterobacteria. J. Biol. Chem. 293:19572–19585.

110. Sandhya Guntaka N, Healy AR, Crawford JM, Herzon SB, Bruner SD (2017) Structure and Functional Analysis of ClbQ, an Unusual Intermediate-Releasing Thioesterase from the Colibactin Biosynthetic Pathway HHS Public Access. ACS Chem Biol 12:2598–2608.

111. Ritchie MK, Johnson LC, Clodfelter JE, Pemble CW, Fulp BE, Furdui CM, Kridel SJ, Lowther WT (2016) Crystal structure and substrate specificity of human thioesterase 2: Insights into the molecular basis for the modulation of fatty acid synthase. J. Biol. Chem. 291:3520–3530.

112. Whicher JR, Florova G, Sydor PK, Singh R, Alhamadsheh M, Challis GL, Reynolds KA, Smith JL (2011) Structure and function of the RedJ protein, a thioesterase from the prodiginine biosynthetic pathway in Streptomyces coelicolor. J. Biol. Chem. 286:22558–22569.

113. Linne U, Schwarzer D, Schroeder GN, Marahiel MA (2004) Mutational analysis of a type II thioesterase associated with nonribosomal peptide synthesis. Eur. J. Biochem. 271:1536–1545.

114. Claxton HB, Akey DL, Silver MK, Admiraal SJ, Smith JL (2009) Structure and Functional Analysis of RifR, the Type II Thioesterase from the Rifamycin Biosynthetic Pathway. J. Biol. Chem. 284:5021–5029.

115. Koglin A, Löhr F, Bernhard F, Rogov V V, Frueh DP, Strieter ER, Mofid MR, Güntert P, Wagner G, Walsh CT, et al. (2008) Structural basis for the selectivity of the external thioesterase of the surfactin synthetase. Nature [Internet] 454:907–911. Available from: https://doi.org/10.1038/nature07161

116. Lawson DM, Derewenda U, Serre L, Ferri S, SzittnerJI Y Wei IR, Meighen EA, Derewenda ZS Structure of a Myristoyl-ACP-Specific Thioesterase from Vibrio harveyi. 1994.

117. Kim KK, Song HK, Shin DH, Hwang KY, Choe S, Yoo OJ, Suh SW (1997) Crystal structure of carboxylesterase from Pseudomonas fluorescens, an α/β hydrolase with broad substrate specificity. Structure 5:1571–1584.

118. Devedjiev Y, Dauter Z, Kuznetsov SR, Jones TLZ, Derewenda ZS (2000) Crystal structure of the human acyl protein thioesterase I from a single X-ray data set to 1.5 Å. Structure 8:1137–1146.

119. Filippova E V., Weston LA, Kuhn ML, Geissler B, Gehring AM, Armoush N, Adkins CT, Minasov G, Dubrovska I, Shuvalova L, et al. (2013) Large scale structural rearrangement of a serine hydrolase from francisella tularensis facilitates catalysis. J. Biol. Chem. 288:10522–10535.

120. Won SJ, Davda D, Labby KJ, Hwang SY, Pricer R, Majmudar JD, Armacost KA, Rodriguez LA, Rodriguez CL, Chong FS, et al. (2016) Molecular Mechanism for Isoform-Selective Inhibition of Acyl Protein Thioesterases 1 and 2 (APT1 and APT2). ACS Chem. Biol. 11:3374–3382.

121. Wepy JA, Galligan JJ, Kingsley PJ, Xu S, Goodman MC, Tallman KA, Rouzer CA, Marnett LJ (2019) Lysophospholipases cooperate to mediate lipid homeostasis and lysophospholipid signaling. J. Lipid Res. 60:360–374.

122. Legler PM, Kumaran D, Swaminathan S, Studier FW, Millard CB (2008) Structural characterization and reversal of the natural organophosphate resistance of a D-type esterase, Saccharomyces cerevisiae S-formylglutathione hydrolase. Biochemistry 47:9592–9601.

123. Van Straaten KE, Gonzalez CF, Valladares RB, Xu X, Savchenko A V., Sanders DAR (2009) The structure of a putative S-formylglutathione hydrolase from Agrobacterium tumefaciens. Protein Sci. 18:2196–2202.

124. Alterio V, Aurilia V, Romanelli A, Parracino A, Saviano M, D'Auria S, de Simone G (2010) Crystal structure of an S-formylglutathione hydrolase from pseudoalteromonas haloplanktis TAC125. Biopolymers 93:669–677.

125. Lee CW, Yoo W, Park SH, Le LTHL, Jeong CS, Ryu BH, Shin SC, Kim HW, Park H, Kim KK, et al. (2019) Structural and functional characterization of a novel cold-active S-formylglutathione hydrolase (SfSFGH) homolog from Shewanella frigidimarina, a psychrophilic bacterium. Microb. Cell Fact. 18.

126. Chen NH, Couñago RM, Djoko KY, Jennings MP, Apicella MA, Kobe B, McEwan AG (2013) A glutathione-dependent detoxification system is required for formaldehyde resistance and optimal survival of neisseria meningitidis in biofilms. Antioxidants Redox Signal. 18:743–755.

127. Wu D, Li Y, Song G, Zhang D, Shaw N, Liu Z (2009) Crystal structure of human esterase D: a potential genetic marker of retinoblastoma. FASEB J. 23:1441–1446.

128. Lemak S, Tchigvintsev A, Petit P, Flick R, Singer AU, Brown G, Evdokimova E, Egorova O, Gonzalez CF, Chernikova TN, et al. (2012) Structure and activity of the cold-active and anion-activated carboxyl esterase OLEI01171 from the oil-degrading marine bacterium Oleispira antarctica. Biochem. J. 445:193–203.

129. Lo Y-C, Lin S-C, Shaw J-F, Liaw Y-C (2003) Crystal Structure of Escherichia coli Thioesterase I/Protease I/Lysophospholipase L1: Consensus Sequence Blocks Constitute the Catalytic Center of SGNH-hydrolases through a Conserved Hydrogen Bond Network. J. Mol. Biol. 330:539–551.

130. Brzuszkiewicz A, Nowak E, Dauter Z, Dauter M, Cieśliski H, Dlługolłęcka A, Kur J (2009) Structure of esta esterase from psychrotrophic pseudoalteromonas sp. 643A covalently inhibited by monoethylphosphonate. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 65:862–865.

131. Kovačić F, Granzin J, Wilhelm S, Kojić-Prodić B, Batra-Safferling R, Jaeger KE (2013) Structural and Functional Characterisation of TesA - A Novel Lysophospholipase A from Pseudomonas aeruginosa. PLoS One 8.

132. Grisewood MJ, Hernández-Lozada NJ, Thoden JB, Gifford NP, Mendez-Perez D, Schoenberger HA, Allan MF, Floy ME, Lai RY, Holden HM, et al. (2017) Computational Redesign of Acyl-ACP Thioesterase with Improved Selectivity toward Medium-Chain-Length Fatty Acids. ACS Catal. 7:3837–3849.

133. Li Z, Li L, Huo Y, Chen Z, Zhao Y, Huang J, Jian S, Rong Z, Wu D, Gan J, et al. (2020) Erratum: Structure-guided protein engineering increases enzymatic activities of the SGNH family esterases. Biotechnol. Biofuels 13.

134. Mullins EA, Kappock TJ (2012) Crystal structures of Acetobacter aceti succinyl-coenzyme A (CoA):Acetate CoA-transferase reveal specificity determinants and illustrate the mechanism used by class i CoA-transferases. Biochemistry 51:8422–8434.

135. Murphy JR, Mullins EA, Kappock TJ (2016) Functional dissection of the bipartite active site of the class I Coenzyme A (CoA)-transferase succinyl-CoA: Acetate CoA-transferase. Front. Chem. 4.

136. Lo YC, Lin SC, Shaw JF, Liaw YC (2003) Crystal structure of Escherichia coli thioesterase I/protease I/lysophospholipase L1: Consensus sequence blocks constitute the catalytic center of SGNH-hydrolases through a conserved hydrogen bond network. J. Mol. Biol. 330:539–551.

137. Lo Y-C, Lin S-C, Shaw J-F, Liaw Y-C (2005) Substrate Specificities of Escherichia coli Thioesterase I/Protease I/Lysophospholipase L1 Are Governed by Its Switch Loop Movement. Biochemistry 44:1971–1979.

138. Cameron AD, Ridderström M, Olin B, Mannervik B (1999) Crystal structure of human glyoxalase II and its complex with a glutathione thiolester substrate analogue. Structure 7:1067–1078.

139. Marasinghe GPK, Sander IM, Bennett B, Periyannan G, Yang KW, Makaroff CA, Crowder MW (2005) Structural studies on a mitochondrial glyoxalase II. J. Biol. Chem. 280:40668–40675.

140. Campos-Bermudez VA, Leite NR, Krog R, Costa-Filho AJ, Soncini FC, Oliva G, Vila AJ (2007) Biochemical and structural characterization of Salmonella typhimurium glyoxalase II: New insights into metal ion selectivity. Biochemistry 46:11069–11079.

141. Sattler SA, Wang X, Lewis KM, DeHan PJ, Park CM, Xin Y, Liu H, Xian M, Xun L, Kang C (2015) Characterizations of two bacterial persulfide dioxygenases of the metallo-β-lactamase superfamily. J. Biol. Chem. 290:18914–18923.

142. Oke M, Carter LG, Johnson KA, Liu H, McMahon SA, Yan X, Kerou M, Weikart ND, Kadi N, Sheikh MA, et al. (2010) The scottish structural proteomics facility: Targets, methods and outputs. J. Struct. Funct. Genomics 11:167–180.

143. Zender M, Witzgall F, Drees SL, Weidel E, Maurer CK, Fetzner S, Blankenfeldt W, Empting M, Hartmann RW (2016) Dissecting the Multiple Roles of PqsE in Pseudomonas aeruginosa Virulence by Discovery of Small Tool Compounds. ACS Chem. Biol. 11:1755–1763.

144. Morales-Quintana L, Nuñez-Tobar MX, Moya-León MA, Herrera R (2013) Molecular dynamics simulation and site-directed mutagenesis of alcohol acyltransferase: A proposed mechanism of catalysis. J. Chem. Inf. Model. 53:2689–2700.

145. Navarro-Retamal C, Gaete-Eastman C, Herrera R, Caballero J, Alzate-Morales JH (2016) Structural

and affinity determinants in the interaction between alcohol acyltransferase from F. x ananassa and several alcohol substrates: A computational study. PLoS One 11:1–14.

146. Galaz S, Morales-Quintana L, Moya-León MA, Herrera R (2013) Structural analysis of the alcohol acyltransferase protein family from Cucumis melo shows that enzyme activity depends on an essential solvent channel. FEBS J. 280:1344–1357.

147. Kleanthous C, Shaw W V. (1984) Analysis of the mechanism of chloramphenicol acetyltransferase by steady-state kinetics. Evidence for a ternary-complex mechanism. Biochem. J. 223:211–220.

148. Bracken CD, Neighbor AM, Lamlenn KK, Thomas GC, Schubert HL, Whitby FG, Howard BR (2011) Crystal structures of a halophilic archaeal malate synthase from Haloferax volcanii and comparisons with isoforms A and G. BMC Struct. Biol. 11.

149. Strittmatter L, Li Y, Nakatsuka NJ, Calvo SE, Grabarek Z, Mootha VK (2014) CLYBL is a polymorphic human enzyme with malate synthase and β-methylmalate synthase activity. Hum. Mol. Genet. 23:2313–2323.

150. Goulding CW, Bowers PM, Segelke B, Lekin T, Kim C-Y, Terwilliger TC, Eisenberg D (2007) The Structure and Computational Analysis of Mycobacterium tuberculosis Protein CitE Suggest a Novel Enzymatic Function. J. Mol. Biol. [Internet] 365:275–283. Available from: https://www.sciencedirect.com/science/article/pii/S0022283606013210

151. Wang H, Fedorov AA, Fedorov E V., Hunt DM, Rodgers A, Douglas HL, Garza-Garcia A, Bonanno JB, Almo SC, de Carvalho LPS (2019) An essential bifunctional enzyme in Mycobacterium tuberculosis for itaconate dissimilation and leucine catabolism. Proc. Natl. Acad. Sci. U. S. A. 116:15907–15913.

152. Torres R, Chim N, Sankaran B, Pujol C, Bliska JB, Goulding CW (2012) Structural insights into RipC, a putative citrate lyase beta subunit from a Yersinia pestis virulence operon. Acta Crystallogr. Sect. F 68:2–7.

153. Malley KR, Koroleva O, Miller I, Sanishvili R, Jenkins CM, Gross RW, Korolev S (2018) The structure of iPLA2$β$ reveals dimeric active sites and suggests mechanisms of regulation and localization. Nat. Commun. 9:1–11.

154. Tang J, Kriz RW, Wolfman N, Shaffer M, Seehra J, Jones SS (1997) A novel cytosolic calcium-independent phospholipase A2 contains eight ankyrin motifs. J. Biol. Chem. 272:8567–8575.

155. Consortium TU (2008) The Universal Protein Resource (UniProt). Nucleic Acids Res. 36:D190–D195.

156. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic Acids Res. 37:D26--D31.

157. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res. 28:235–242.

158. Altschup SF, Gish W, Miller W, Myers EW, Lipman DJ Basic Local Alignment Search Tool. 1990.

159. Bateman A (2019) UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 47:D506–D515.

160. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res.

25:3389–3402.

161. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. (2019) The Pfam protein families database in 2019. Nucleic Acids Res. 47:D427–D432.

162. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins Struct. Funct. Genet. 56:143–156.

163. Lee FJ, Lin LW, Smith JA (1990) A glucose-repressible gene encodes acetyl-CoA hydrolase from Saccharomyces cerevisiae. J. Biol. Chem. 265:7413–7418.

164. Fleck CB, Brock M (2009) Re-characterisation of Saccharomyces cerevisiae Ach1p: Fungal CoA-transferases are involved in acetic acid detoxification. Fungal Genet. Biol. 46:473–485.

165. Hunt MC, Rautanen A, Westin MAK, Svensson LT, Alexson SEH (2006) Analysis of the mouse and human acyl-CoA thioesterase (ACOT) gene clusters shows that convergent, functional evolution results in a reduced number of human peroxisomal ACOTs1. FASEB J. 20:1855–1864.

166. Johnson MR, Barnes S, Kwakye JB, Diasio RB (1991) Purification and characterization of bile acid-CoA:amino acid N-acyltransferase from human liver. J. Biol. Chem. 266:10227–10233.

167. Cho H, Cronan JE (1995) Defective Export of a Periplasmic Enzyme Disrupts Regulation of Fatty Acid Synthesis. J. Biol. Chem. 270:4216–4219.

168. Karasawa K, Yokoyama K, Setaka M, Nojima S (1999) The Escherichia coli pldC Gene Encoding Lysophospholipase L1 Is Identical to the apeA and tesA Genes Encoding Protease I and Thioesterase I, Respectively1. J. Biochem. 126:445–448.

169. Jones JM, Nau K, Geraghty MT, Erdmann R, Gould SJ (1999) Identification of Peroxisomal Acyl-CoA Thioesterases in Yeast and Humans. J. Biol. Chem. 274:9216–9223.

170. Naggert J, Narasimhan ML, DeVeaux L, Cho H, Randhawa ZI, Cronan JE, Green BN, Smith S (1991) Cloning, sequencing, and characterization of Escherichia coli thioesterase II. J. Biol. Chem. 266:11044–11050.

171. Nie L, Ren Y, Schulz H (2008) Identification and Characterization of Escherichia coli Thioesterase III That Functions in Fatty Acid β-Oxidation. Biochemistry 47:7744–7751.

172. Yamada J, Furihata T, Tamura H, Watanabe T, Suga T (1996) Long-Chain Acyl-CoA Hydrolase from Rat Brain Cytosol: Purification, Characterization, and Immunohistochemical Localization. Arch. Biochem. Biophys. 326:106–114.

173. Adams SH, Chui C, Schilbach SL, Yu XX, Godard AD, Grimaldi JC, Lee J, Dowd P, Colman S, Lewin DA (2001) BFIT, a unique acyl-CoA thioesterase induced in thermogenic brown adipose tissue: cloning, organization of the human gene and assessment of a potential link to obesity. Biochem. J. 360:135–142.

174. Suematsu N, Isohashi F (2006) Molecular cloning and functional expression of human cytosolic acetyl-CoA hydrolase. Acta Biochim. Pol. 53:553–561.

175. Poupon V, Bègue B, Gagnon J, Dautry-Varsat A, Cerf-Bensussan N, Benmerah A (1999) Molecular Cloning and Characterization of MT-ACT48, a Novel Mitochondrial Acyl-CoA Thioesterase. J. Biol. Chem. 274:19188–19194.

176. Wei J, Kang HW, Cohen DE (2009) Thioesterase superfamily member 2 (Them2)/acyl-CoA thioesterase 13 (Acot13): a homotetrameric hotdog fold thioesterase with selectivity for long-chain fatty acyl-CoAs. Biochem. J. 421:311–322.

177. Zhuang Z, Song F, Martin BM, Dunaway-Mariano D (2002) The YbgC protein encoded by the ybgC gene of the tol-pal gene cluster of Haemophilus influenzae catalyzes acyl-coenzyme A thioester hydrolysis. FEBS Lett. 516:161–163.

178. Ben-Israel I, Yu G, Austin MB, Bhuiyan N, Auldridge M, Nguyen T, Schauvinhold I, Noel JP, Pichersky E, Fridman E (2009) Multiple Biochemical and Morphological Factors Underlie the Production of Methylketones in Tomato Trichomes. Plant Physiol. 151:1952–1964.

179. Damien L, Aurélia B, Emmanuelle B (2007) The Hotdog Thioesterase EntH (YbdB) Plays a Role In Vivo in Optimal Enterobactin Biosynthesis by Interacting with the ArCP Domain of EntB. J. Bacteriol. 189:7112–7126.

180. Widhalm JR, van Oostende C, Furt F, Basset GJC (2009) A dedicated thioesterase of the Hotdog-fold family is required for the biosynthesis of the naphthoquinone ring of vitamin K_1. Proc. Natl. Acad. Sci. 106:5599 LP-- 5603.

181. Ferrández A, Miñambres B, García B, Olivera ER, Luengo JM, García JL, Díaz E (1998) Catabolism of Phenylacetic Acid in Escherichia coli: Characterization of a New Aerobic Hybrid Pathway. J. Biol. Chem. 273:25974–25986.

182. Jones A, Davies HM, Voelker TA (1995) Palmitoyl-acyl carrier protein (ACP) thioesterase and the evolutionary origin of plant acyl-ACP thioesterases. Plant Cell 7:359–371.

183. Wakil SJ (1989) Fatty acid synthase, a proficient multifunctional enzyme. Biochemistry 28:4523–4530.

184. Gokhale RS, Hunziker D, Cane DE, Khosla C (1999) Mechanism and specificity of the terminal thioesterase domain from the erythromycin polyketide synthase. Chem. Biol. 6:117–125.

185. Kohli RM, Takagi J, Walsh CT (2002) The thioesterase domain from a nonribosomal peptide synthetase as a cyclization catalyst for integrin binding peptides. Proc. Natl. Acad. Sci. 99:1247 LP – 1252.

186. Libertini LJ, Smith S (1978) Purification and properties of a thioesterase from lactating rat mammary gland which modifies the product specificity of fatty acid synthetase. J. Biol. Chem. 253:1393–1401.

187. Mikkelsen J, Witkowski A, Smith S (1987) Interaction of rat mammary gland thioesterase II with fatty acid synthetase is dependent on the presence of acyl chains on the synthetase. J. Biol. Chem. 262:1570–1574.

188. Heathcote ML, Staunton J, Leadlay PF (2001) Role of type II thioesterases: evidence for removal of short acyl chains produced by aberrant decarboxylation of chain extender units. Chem. Biol. 8:207–220.

189. Ferri SR, Meighen EA (1991) A lux-specific myristoyl transferase in luminescent bacteria related to eukaryotic serine esterases. J. Biol. Chem. 266:12852–12857.

190. Camp LA, Hofmann SL (1993) Purification and properties of a palmitoyl-protein thioesterase that cleaves palmitate from H-Ras. J. Biol. Chem. 268:22566–22574.

191. Vesa J, Hellsten E, Verkruyse LA, Camp LA, Rapola J, Santavuori P, Hofmann SL, Peltonen L (1995) Mutations in the palmitoyl protein thioesterase gene causing infantile neuronal ceroid lipofuscinosis.

Nature 376:584–587.

192. Das AK, Bellizzi JJ, Tandel S, Biehl E, Clardy J, Hofmann SL (2000) Structural basis for the insensitivity of a serine enzyme (Palmitoyl-protein thioesterase) to phenylmethylsulfonyl fluoride. J. Biol. Chem. 275:23847–23851.

193. Sugimoto H, Hayashi H, Yamashita S (1996) Purification, cDNA Cloning, and Regulation of Lysophospholipase from Rat Liver. J. Biol. Chem. 271:7705–7711.

194. Duncan JA, Gilman AG (1998) A Cytoplasmic Acyl-Protein Thioesterase That Removes Palmitate from G Protein α Subunits and p21RAS. J. Biol. Chem. 273:15830–15837.

195. Gonzalez CF, Proudfoot M, Brown G, Korniyenko Y, Mori H, Savchenko A V, Yakunin AF (2006) Molecular Basis of Formaldehyde Detoxification: Characterization of Two S-Formylglutathione Hydrolases From Escherichia coli, FrmB and YeiG. J. Biol. Chem. 281:14514–14522.

196. Vander Jagt DL (1993) Glyoxalase II: Molecular characteristics, kinetics and mechanism. Biochem. Soc. Trans. 21:522–527.

197. Carfi A, Pares S, Duée E, Galleni M, Duez C, Frère JM, Dideberg O (1995) The 3-D structure of a zinc metallo-beta-lactamase from Bacillus cereus reveals a new type of protein fold. EMBO J. 14:4914–4921.

198. Wu R, Latham JA, Chen D, Farelli J, Zhao H, Matthews K, Allen KN, Dunaway-Mariano D (2014) Structure and catalysis in the Escherichia coli hotdog-fold thioesterase paralogs YdiI and YbdB. Biochemistry 53:4788–4805.

199. Koglin A, Löhr F, Bernhard F, Rogov V V, Frueh DP, Strieter ER, Mofid MR, Güntert P, Wagner G, Walsh CT, et al. (2008) Structural basis for the selectivity of the external thioesterase of the surfactin synthetase. Nature 454:907–911.

200. Bellizzi JJ, Widom J, Kemp C, Lu J-Y, Das AK, Hofmann SL, Clardy J (2000) The crystal structure of palmitoyl protein thioesterase 1 and the molecular basis of infantile neuronal ceroid lipofuscinosis. Proc. Natl. Acad. Sci. 97:4573–4578.

201. Pesaresi A, Lamba D (2010) Insights into the fatty acid chain length specificity of the carboxylesterase PA3859 from Pseudomonas aeruginosa: A combined structural, biochemical and computational study. Enferm. Infecc. Microbiol. Clin. 28:1787–1792.

CHAPTER 4

Mutation space of spatially conserved amino acid sites in proteins

Benjamin T. Caswell and David C. Cantu

Modified from a manuscript in preparation that will be submitted to *Protein Science* with the same title and authors

4.1. Introduction

An important question about protein structure, that affects protein engineering, is how the three-dimensional structure of a protein will be affected by specific mutations. In this work, we present a method to quantify the likelihood that a specific point mutation will affect the tertiary structure of a protein.

Proteins can be compared by different approaches. A well-established approach is to predict mutation effects on protein function through examination of genetic and protein sequences and comparison to function, largely focusing on single nucleotide polymorphism. Amino acid sequence-based approaches tend to not rely on direct protein tertiary structure comparisons, instead utilizing the comparatively massive amount of available sequencing data to attain higher precision and resolution in their results.[1–6] These are certainly not the only means through which sequences are used to compare proteins,[7,8] but generally sequence data alone is not sufficient to make specific structural predictions.

Proteins can also be compared by superimposing their three-dimensional structures that focus on attaining the best global fit by minimizing the distance between residues or substructures in different proteins. Tertiary structure superimposition methods provide consistent and useful information for comparing protein structure through direct tertiary structural comparison,[9–11] and allow to make inferences on the how specific amino acids affect function. To compare highly divergent proteins that may share only a small conserved core or region, structural comparisons can be performed through superimposing only highly similar structural fragments.[12] Both three-dimensional comparison approaches directly compares structures through minimizing the distance between structural elements to obtain the best superimposition. Tertiary structure comparison is desirable as it provides data that sequence driven methods struggle to attain, but, like sequence-based comparison methods that seek to optimize the local or global alignment, tertiary structure superimposition methods minimize distances between structures to optimally superimpose protein structures.

In this work, we present the mutation space of spatially conserved (MSSC) amino acid sites in proteins. The MSSC compares protein tertiary structures without any formal superimposition. This method quantifies how many mutations occur in each spatially conserved amino site in a target protein in a group of multiple protein structures. The MSSC examines each residue in a target protein, compares it to the residues present in the same relative position in other protein structures from that group, and uses the physicochemical criteria of mutations found in each conserved spatial site to quantify the mutation space of that residue. The MSSC provides a unique perspective because it does

not seek to identify the best sequence fit or the best structural superimposition, but rather only inform how many amino acid mutations occur in a spatial site in a protein structure.

4.2. Results

The mutation space of a spatially conserved amino acid site is a quantification of the overall conservation of a specific spatial location for a group of similar proteins, accounting for the diversity of amino acids found in a particular site and the degree of spatial and physicochemical conservation of that site. This is analogous to scoring conserved sites in a multiple sequence alignment, but in three-dimensional space considering the spatial location of residues, instead of solely the order in which they appear in a protein sequence. To be able to quantify the mutation space for a site in three-dimensional space for several proteins, corresponding residues for a spatial site must be identified. At most one residue from each protein in a group of protein structures can occupy a site in three-dimensional space: these are the spatially corresponding residues, and how they are identified is described in Section 4.2.1. Once spatially corresponding residues are identified the mutation space is calculated based on the residues that appear in each spatial site in three-dimensional space; this is described in Section 4.2.2.

4.2.1. Identifying spatially corresponding residues in the tertiary structure of proteins

A method to identify spatially corresponding residues was developed to avoid relying on protein structure superimposition approaches and external software, since the goal is not to superimpose structures, but rather to identify spatially corresponding residues between

two protein structures based on the relative position of each residue in its protein structure. Within a set of structures (*i.e.*, group of proteins), a target protein structure is selected, while the remaining ones are the subject protein structures. All the amino acid sequences in the group of proteins are used to obtain a multiple sequence alignment, and sequence-conserved residues in the whole group are identified. For each three-dimensional structure in the group of proteins, the average position of sequence-conserved residues is labeled as the center of mass of conserved residues for that protein structure (***COM***), which is then defined as its origin in cartesian space. For each three-dimensional structure in the group of proteins, the two sequence-conserved residues in that set that are separated by the greatest spatial distance are then selected as reference points, $\boldsymbol{CR^1}$ and $\boldsymbol{CR^2}$, keeping assignments consistent between each protein structure in the group. Each protein structure is then rotated about its ***COM*** origin such that $\boldsymbol{CR^1}$ is aligned with the z-axis and $\boldsymbol{CR^2}$ is on the $\{(x, 0, z)|x \geq 0, z \in \mathbb{R}\}$ plane.

Following this spatial realignment, the position of each residue ***j*** within a protein structure ***s***, $\boldsymbol{r_{j,s}}$, is redefined by the vectors:

$$\overrightarrow{v_{j,s}^{1}} = \overrightarrow{CR_s^1 r_{j,s}} \quad \text{Eq. 1}$$

$$\overrightarrow{v_{j,s}^{2}} = \overrightarrow{CR_s^2 r_{j,s}} \quad \text{Eq. 2}$$

$$\overrightarrow{v_{j,s}^{com}} = \overrightarrow{COM_s r_{j,s}} \quad \text{Eq. 3}$$

The initial spatial realignment and vector conversion results in that protein structures that are highly similar will have highly similar vector fields, regardless of the absolute position and orientation of each protein structure in cartesian space. Each residue from a subject protein structure ***s***, $\boldsymbol{r_{j,s}}$, is compared to each residue from the target protein

structure t, $r_{i,t}$.

The similarity of the spatial positions of two residues is determined through examination of related defining vectors. The position of each residue is defined by three vectors which originate at the selected reference points, CR^1, CR^2, and COM. These reference points, for the groups of similar proteins studied, have highly conserved spatial positions in all structures. Therefore, if the vectors defining a residue in a subject structure, $\overrightarrow{v^1_{J,s}}$, $\overrightarrow{v^2_{J,s}}$, and $\overrightarrow{v^{com}_{J,s}}$, are all oriented in the same direction and have the same magnitudes as the vectors defining a residue in the target structure, $\overrightarrow{v^1_{\iota,t}}$, $\overrightarrow{v^2_{\iota,t}}$, and $\overrightarrow{v^{com}_{\iota,t}}$ respectively, then the two residues must occupy the same spatial position relative to their reference points. To determine the similarity of orientation of related defining vectors, their cosine similarity is determined:

$$csim^1_{i,j} = \frac{\overrightarrow{v^1_{J,s}} \cdot \overrightarrow{v^1_{\iota,t}}}{\left\|\overrightarrow{v^1_{J,s}}\right\| \left\|\overrightarrow{v^1_{\iota,t}}\right\|} \quad \text{Eq. 4}$$

$$csim^2_{i,j} = \frac{\overrightarrow{v^2_{J,s}} \cdot \overrightarrow{v^2_{\iota,t}}}{\left\|\overrightarrow{v^2_{J,s}}\right\| \left\|\overrightarrow{v^2_{\iota,t}}\right\|} \quad \text{Eq. 5}$$

$$csim^{com}_{i,j} = \frac{\overrightarrow{v^{com}_{J,s}} \cdot \overrightarrow{v^{com}_{\iota,t}}}{\left\|\overrightarrow{v^{com}_{J,s}}\right\| \left\|\overrightarrow{v^{com}_{\iota,t}}\right\|} \quad \text{Eq. 6}$$

which are subsequently averaged as:

$$csim^{avg}_{i,j} = \frac{csim^1_{i,j} + csim^2_{i,j} + csim^{com}_{i,j}}{3} \quad \text{Eq. 7}$$

Then, the similarity of magnitudes of defining vectors is determined by comparing the root mean squared difference (RMSD) of the vector magnitudes from the subject residue:

$$RMSD_{j,s} = \sqrt{\frac{\left\|\overrightarrow{v_{J,s}^{1}}\right\|^2 + \left\|\overrightarrow{v_{j,s}^{2}}\right\|^2 + \left\|\overrightarrow{v_{J,s}^{com}}\right\|^2}{3}} \quad \text{Eq. 8}$$

with the RMSD from the target residue. The difference in their magnitudes is determined as:

$$mdif_{i,j} = \left|RMSD_{j,s} - RMSD_{i,t}\right| \quad \text{Eq. 9}$$

For a residue in a subject protein structure $\boldsymbol{r_{j,s}}$ to be spatially correlated with a residue in the target protein structure $\boldsymbol{r_{i,t}}$, the following condition must be met:

$$mdif_{i,j} \leq 2.6 * csim_{i,j}^{avg} - 1.35 \quad \text{Eq. 10}$$

This criterion is selected based on two points,

$$\left(csim_{i,j}^{avg}, mdif_{i,j}\right) = (0.90, 1.00) \quad \text{Exp. 1}$$

$$\left(csim_{i,j}^{avg}, mdif_{i,j}\right) = (0.99, 1.25) \quad \text{Exp. 2}$$

The point shown in Exp. 1, when used as the sole condition for correlation:

$$mdif_{i,j} \leq 1.00 \text{ and } csim_{i,j}^{avg} < 0.90 \quad \text{Exp. 3}$$

was sufficient to identify most possible spatially correlated residues. Examination of the oriented structures confirmed this, but also revealed that some correlations were not identified properly using only this condition. Incorporation of the latter point provides the model with flexibility to trade a degree of vector magnitude similarity for orientation similarity, generating more complete spatial correlation sets.

This analysis is run for every target-subject protein structure pair, where the subject proteins are proteins in the group that are not the target. For each residue in the target protein, a list of possible subject residue matches is generated for each subject protein. For a given subject-target protein pair, each target residue may be possibly

correlated to zero or more subject residues.

For a given subject-target pair, only one subject residue can be considered to occupy the same spatial position as a given target residue. The following algorithm was developed to, for all residues in a target structure, make a final assignment of at most one subject residue from each subject structure to each target residue. First, for a given subject-target pair, target residues that have only one possible spatially correlated subject residue, according to the criteria described, are identified, and those subject residues are assigned as being spatially correlated to their respective target residues. Then, the subject residues in these 1:1 correlations are removed as possible matches from the possible-correlations set. This process is repeated until no more 1:1 (subject residue: target residue) matches are found. The second step identifies 2:2 correlations, meaning when two adjacent target residues are possibly spatially correlated to two adjacent subject residues, *e.g.* $\boldsymbol{r_{15,t}}$ and $\boldsymbol{r_{16,t}}$ are both possibly correlated to $\boldsymbol{r_{23,s}}$ and $\boldsymbol{r_{24,s}}$. These are assigned to minimize the total $\boldsymbol{mdif_{i,j}}$, *e.g.*, $\boldsymbol{r_{23,s}}$ is assigned to $\boldsymbol{r_{15,t}}$ and $\boldsymbol{r_{24,s}}$ is assigned to $\boldsymbol{r_{16,t}}$ if:

$$mdif_{15,23} + mdif_{16,24} < mdif_{16,23} + mdif_{15,24} \quad \text{Exp. 4}$$

These subject residues are assigned as spatially correlated to target residues and removed from the possible correlation set as in the 1:1 cycle. This cycle is repeated until no more 2:2 matches are found. The 1:1 and 2:2 cycles are alternately repeated until no more changes are made to the assignment of spatially correlated residues. This macrocycle assigns most of the final target-subject residue pairs; the remaining residues are assigned in a manner which retains the sequential alignment between the target and subject protein structures.

4.2.2. Quantifying the mutation space in spatially conserved sites

For each residue in the target protein structure, the spatially correlated residues of each subject protein structure are identified (Section 4.2.1); now the mutation space of each residue in the target protein structure needs to be quantified. In a set of structures (*i.e.*, group of proteins), there are $\boldsymbol{x}$+1 structures in a set: one target structure against which $\boldsymbol{x}$ subject structures are compared. A residue in any subject protein structure is a spatially conserved residue if it occupies the same location in three-dimensional space as a corresponding target protein residue, for both relative to their respective structures since structures are not superimposed. Each subject protein residue can be spatially correlated, or conserved, to at most one residue from the target protein structure. Also, it is possible that a target protein residue does not spatially correlate with any residues from a subject structure: if the criteria described in Section 4.2.1 is not met, then a target protein residue does not have a spatially corresponding subject protein residue.

To quantify the mutation space of each residue $\boldsymbol{i}$ in the target protein structure we developed the mutation space of spatially correlated residues ($\boldsymbol{MSSC_i}$):

$$MSSC_i = \frac{n_{\text{unique}}}{c} \sum_{r=1}^{19} n_{i,r} D_{i,r} \quad \text{Eq. 12}$$

which is used to score each residue in the target structure. Each target protein residue can have at most $\boldsymbol{x}$ correlated subject protein residues, where $\boldsymbol{x}$ is the number of subject protein structures. When a target residue is found to be spatially correlated to $\boldsymbol{c}$ subject residues, at most one per subject protein structure, then there are $\boldsymbol{x} - \boldsymbol{c}$ instances of subject structures having no spatially correlated residue for that specific residue of the target protein. At the core of this equation is the Grantham's distance[13], $\boldsymbol{D_{i,r}}$, indicating

the distance between target residue $\boldsymbol{i}$ and one of the $\boldsymbol{r} = 19$ other standard amino acids. Grantham's distance is based purely on physicochemical criteria. For scoring purposes, any non-standard amino acids are treated as their nearest relative, *e.g*., selenomethionine is scored as if it were methionine. In the Grantham's distance matrix, there are $\boldsymbol{n_{i,r}}$ instances of a given mutation, such as Cys-Trp or Cys-Glu, on each target residue such that:

$$c = \sum_{r=1}^{19} n_{i,r} \quad \text{Eq. 13}$$

the number of unique mutations, $\boldsymbol{n_{\text{unique}}}$, gives the count of unique, non-equivalent amino acid correlations. That is, given a set of 11 proteins with one target structure and $\boldsymbol{x} = 10$ subject structures, for a Cys residue in the target protein structure that is spatially correlated to 3 Cys in three different subject structures, 1 Ser in another subject structure, and 6 Gly in six different subject structures, it has $\boldsymbol{c} = 10$ spatially correlated residues and $\boldsymbol{n_{\text{unique}}} = 2$ unique mutations (Cys – Gly, Cys – Ser). The mutational space score of residue $\boldsymbol{i}$ in the target protein structure, $\boldsymbol{MSSC_i}$, is low for target residue that has fewer and more similar mutations among its spatially correlated subject residues and high for a target residue that has many different mutations and more dissimilar mutations. The $\boldsymbol{MSSC_i}$, can also be low if $\boldsymbol{c}$ is small (approximately $\frac{\boldsymbol{c}}{\boldsymbol{x}} = 0.5$) since the scores for the $\boldsymbol{x}$ - $\boldsymbol{c}$ non-correlated subject residues are counted as zero. A convenient value to observe is the sample occurrence ratio:

$$SOR = \frac{c}{x} \quad \text{Eq. 14}$$

which indicates the proportion of the subject protein structures that were found to have a

correlated residue for a given target residue.

4.3. Discussion

To illustrate how MSSC analysis can lead to valuable structural insights in proteins, we present four example cases from different protein folds and enzyme functions: two thioesterase enzymes, one ketoacyl synthase enzyme, and one glycoside hydrolase enzyme. Examination of the mutations space of the four test cases reveals a common trend: mutations are more common and pronounced on the protein surface than the interior. This is observed for all cases with different structural folds, and enzymatic functions, suggesting that interior residues (*e.g.*, not in contact with solvent) play a role in maintaining the three-dimensional structure of proteins. Exterior α helices display this phenomenon clearly; positions on an α helix that lie closer to the main bulk of the protein are consistently more conserved than positions that are more exposed to the solvent. This pattern is less pronounced in α helices that are 'buried' in the structure, suggesting that the likelihood of solvent interaction may play a role in selecting/promoting mutations. In β sheets amino acid that are exposed to solvent have a higher mutation space than those exposed to protein structure core. Even in loops, their inherent disorder results in this pattern being less clear, but it is still present. Therefore, regardless of secondary structure, solvent exposed residues have a higher mutation space than internal ones.

Enzyme family TE11[9] protein is a HotDog fold thioesterase enzyme that hydrolyses acyl-CoA thioester bonds in many pathways, for example enterobactin biosynthesis. PDB 1SC0 is a TE11 *Haemophilus influenzae* enzyme and is the target structure for which $\boldsymbol{MSSC_i}$ is calculated with respect to the other thirty-seven TE11

structures. (Figure 4.1). The MSSC of each residue in the target structure was calculated. The scores, both absolute and relative, can vary depending on the size and composition of the set of proteins studied, so it is convenient to visualize the results in relative terms. Therefore, in Figure 4.1 the mutation space scores of residues in 1SC0 are expressed as a heatmap, where the darkest red represents the highest MSSC for that target structure. Inspection of the mutation space of 1SC0 shows a notable asymmetry: there is an orientation to the MSSC scores in secondary structures. Residues that lie on the exterior of protein, or those that are more solvent exposed, tend to have higher MSSC scores than those facing the interior. This phenomenon is clear in the beta sheet present in 1SC0: the 26 residues residing on the exterior surface have an average MSSC of 210 and SOR of 0.978, while the 23 residues interior residues had an average MSSC of 110 and SOR of 0.998.

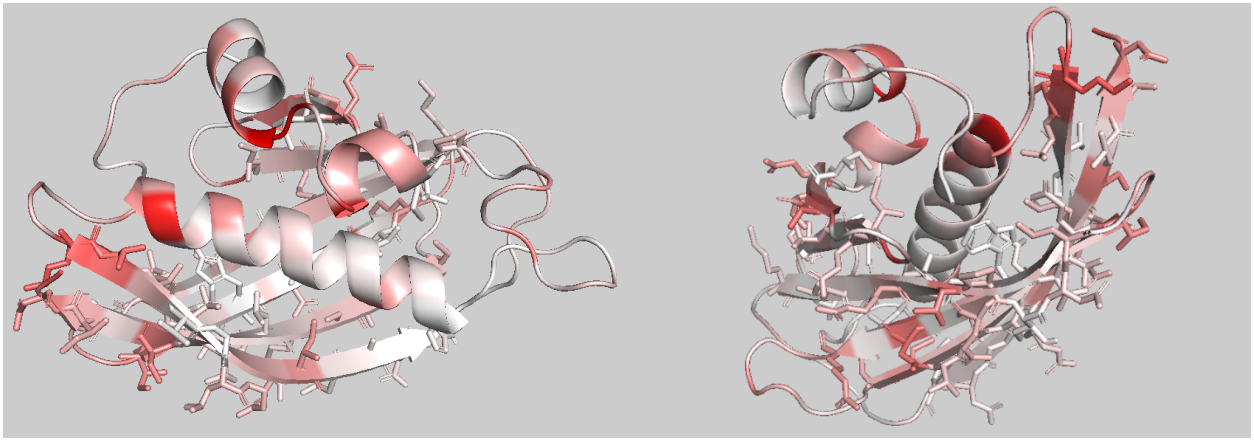

Figure 4.1. TE11 structure 1SC0 with a HotDog fold. Darker red indicates higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein. Residue side chains as sticks are shown to highlight the difference between interior and exterior residues.

TE21[9] is also a thioesterase enzyme family, however their proteins have an α/β

hydrolase fold. The second example is the twenty-one structures in TE21 with the *Homo sapiens* enzyme (PDB 1FJ2) as the target structure (Figure 4.2). As in the TE11 example, secondary structures have higher MSSC scores (*i.e.*, greater mutation space) in positions that are more solvent exposed, as shown in the α helices present in 1FJ2. Residues side chains on an α helix have been represented as sticks to demonstrate the consistency of this phenomenon across secondary structures.

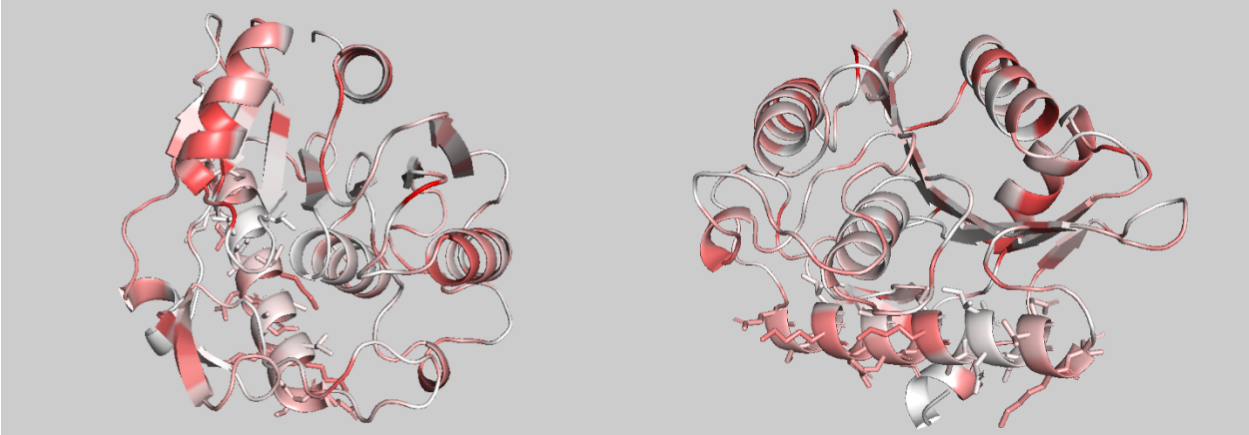

Figure 4.2. TE21 structure 1FJ2 with an α/β hydrolase fold. Darker red indicates higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein. Residue side chains as sticks are shown to highlight the difference between interior and exterior residues.

Ketoacyl synthase enzyme family KS1[14] has a Thiolase-like fold, and currently has 46 resolved protein structures. The third example is the 46 KS1 structures with the *Paraburkholderia xenovorans* structure (PDB 4EFI) as the target protein (Figure 4.3). Of the four examples studied, KS1 enzymes show the most homogeneous mutation space with less variation between solvent-exposed and protein-exposed residues. Protein conformational changes and flexibility may play a role in the consistency with which some residues interact with one another, homogenizing the mutation space distribution.

Further study may reveal that homogeneity of mutation space distribution is fold dependent.

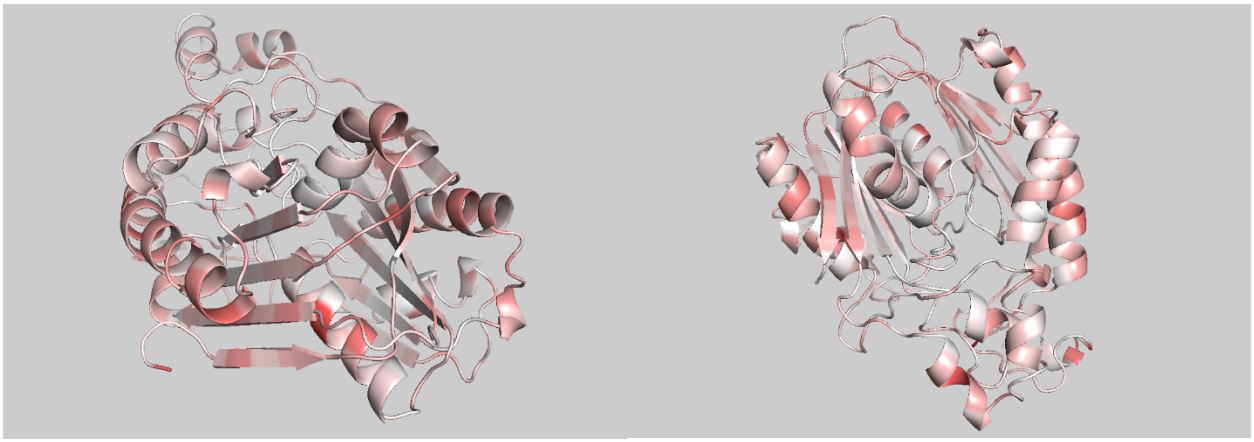

Figure 4.3. KS1 structure 4EFI. Darker red indicates higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein.

GH8, a glycoside hydrolase enzyme family[15] and has 44 known structures with an $(\alpha/\alpha)_6$ barrel fold. The fourth example is the 44 GH8 structures with PDB 1H12 from *Pseudoalteromonas haloplanktis* as the target protein (Figure 4.4). The mutation space of residues lying in closer proximity to other residues is generally lower than those that are more solvent exposed. There is not a significant difference in mutation space between secondary structures. It is worth noting that the 'core' of this structure, likely the substrate binding site, is a region of very low mutation space (very low MSSC scores). This suggests that residues involved with substrate specificity may also have low MSSC scores even if they are solvent exposed.

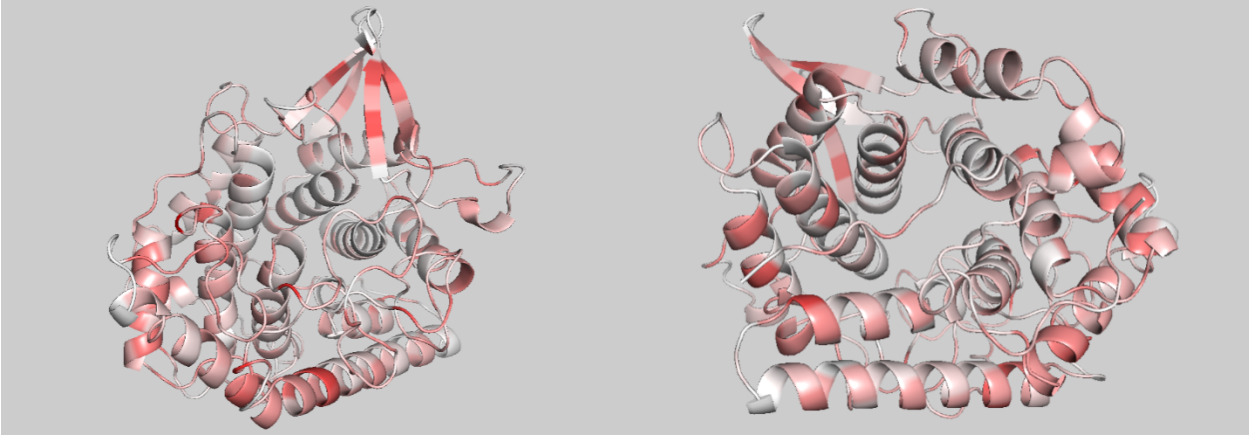

Figure 4.4. GH8 structure 1H12. Darker red indicates higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein.

MSSC has a limitation since correlated residues are determined based on the similarity of their locations relative to common reference points, the position of these reference points relative to each other and relative to their respective protein structures must be highly conserved. In this work we studied protein structures within protein families, *i.e.*, structures that have the same fold and a high degree of structural similarity; therefore sequence-conserved residues are very highly conserved in their spatial positions, which gives consistent structural realignments and comparisons. If the set of proteins studied were more varied, the sequence-conserved residues would likely not be spatially conserved and the approach of identifying spatially conserved residues (Section 4.2.1) may not accurately determine spatially correlated residues.

It is worth noting that some residues that are fully conserved in a multiple sequence alignment did not receive an MSSC of zero (i.e., fully conserved spatially as well), as one might expect. As an example, KS1 Gln274 is fully conserved in the multiple sequence alignment, however, its MSSC score in the *Paraburkholderia xenovorans* target protein is 1.69, which means that, in at least one other structure, a different residue

occupies the spatial position that the target Gln274 residue does in its structure. MSSC focuses on the positions of residues relative to other residues and reference points within the structure. Therefore, minor variations in bond angles and residue size can shift residues from their expected spatial position, resulting in an MSSC that points to a different mutation space than expected. However, this is a natural consequence of our intent; we aim to examine and compare proteins through the space that their amino acid residues occupy, offering a different perspective than current methods. The function and stability of a protein are both intimately tied to its structure, and MSSC provides a novel lens through which to compare and examine protein structures.

4.4. Conclusions

The mutation space of spatially conserved amino acid sites (MSSC) is a quantity for protein structural analysis that was developed in this work. The mutation space of four protein structures in their respective families have been analyzed and presented. This analysis has revealed that mutations are not uniform throughout the protein structure; rather, the composition of amino acid positions in a target structure vary in ordered ways. Regardless of secondary structure, residue positions in closer proximity to other residues are more highly conserved, and those that lie further or are more solvent exposed are more commonly mutated. These results demonstrate how quantifying the MSSC of residues in a protein can be used for protein structural comparison to make informed decisions when selecting point mutations for their proteins, or to understand the structural similarity of a protein more thoroughly within a set of related proteins.

4.5. Methods

Multiple sequence alignments were performed using MUSCLE[7] using the default settings and output for processing in the CLUSTALW format. Other than when specifically mentioned, data processing was performed with Python 3.8. Custom scripts were written for all data processing and the NumPy math package was utilized for non-trivial math functions. Throughout this work, the spatial position of residues within a protein structure are defined by the cartesian coordinates of or reference vectors ending at the α-carbon atom of that residue.

Chapter 4 References

1. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: A web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21:3176–3178.

2. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum. Mutat. 17:263–270.

3. Lee W, Yue P, Zhang Z (2009) Analytical methods for inferring functional effects of single base pair substitutions in human cancers. Hum. Genet. 126:481–498.

4. Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J. Mol. Biol. 353:459–473.

5. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. 7:61–80.

6. Krebs FS, Zoete V, Trottet M, Pouchon T, Bovigny C, Michielin O (2021) Swiss-PO: a new tool to analyze the impact of mutations on protein three-dimensional structures for precision oncology. npj Precis. Oncol. 5.

7. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

9. Caswell BT, de Carvalho CC, Nguyen H, Roy M, Nguyen T, Cantu DC (2022) Thioesterase enzyme families: Functions, structures, and mechanisms. Protein Sci. 31:652–676.

10. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins Struct. Funct. Genet. 56:143–156.

11. Holm L (2020) DALI and the persistence of protein shape. Protein Sci. 29:128–140.

12. Ortiz AR, Strauss CEM, Olmea O (2009) MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Sci. 11:2606–2621.

13. Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 185:862–864.

14. Cantu DC, Chen Y, Lemons ML, Reilly PJ (2011) ThYme: a database for thioester-active enzymes. Nucleic Acids Res. 39:D342–D346.

15. Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. 42:490–495.

CHAPTER 5

Conclusions

This work presented how protein structure analysis plays a role in classifying proteins by defining the thioesterase families, and presented a new quantity, the mutation space of spatially conserved sites in a protein structure, to analyze which amino acids are more structurally significant.

All known thioesterase sequences were classified into thirty-five distinct, non-redundant families based on the similarity of their primary structures. This method is shown to result in families whose members share a high degree of structural and enzymatic similarity, from which structures, catalytic residues, and enzymatic mechanisms can be predicted. This similarity was confirmed through multiple sequence alignments and superimposition of three-dimensional structures. The thioesterase families are available in the ThYme database, along with other enzyme groups (e.g. ketoacyl synthases) that act on thioester-containing substrates. The approach and methods developed with the thioesterases will be instrumental to update all the enzyme groups in the ThYme database.

The development of a structural comparison method that forgoes superimposition to identify spatially corresponding residues is reported in this work. The positions of amino acids are described in terms of their relative positions within a protein structure, and are used to quantify the similarity of the relative positions of amino acid residues in a different protein structure, resulting in an examination of the mutation space that each amino acid in a structure. This method provides meaningful information about a target

protein structure within a set of similar protein structures. Four example target structures from different protein folds and with different enzymatic functions were scored by the similarity of residues found in the same relative spatial position in each respective set of comparison structures. A heat map for each target structure was generated, showing the relative compositional conservation of residue positions with respect to the set. In the examples that the MSSC were applied show that residue positions in closer proximity to other residues are more highly conserved, and those that lie further or are more solvent exposed are more commonly mutated.

Mutation space analysis could benefit researchers in related fields by providing information that could help them make more informed decisions when selecting point mutations. Work is ongoing to improve the ability to meaningfully compare sets of more dissimilar structures, as well as to improve consistency of structural realignment and vectorization.