2013

Fatty acid and polyketide synthesis enzymes: sequences, structures, and mechanisms

David Cantu
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons, Chemical Engineering Commons, and the Chemistry Commons

Recommended Citation
Cantu, David, "Fatty acid and polyketide synthesis enzymes: sequences, structures, and mechanisms" (2013). Graduate Theses and Dissertations. 13020.
http://lib.dr.iastate.edu/etd/13020

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Fatty acid and polyketide synthesis enzymes: sequences, structures, and mechanisms

by

David C. Cantu

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements of the degree of DOCTOR OF PHILOSOPHY

Major: Chemical Engineering

Program of Study Committee
Peter J. Reilly, Major Professor
Laura R. Jarboe
Robert L. Jernigan
Monica H. Lamm
Basil J. Nikolau

Iowa State University
Ames, Iowa
2013

Copyright © David C. Cantu, 2013. All rights reserved.
# TABLE OF CONTENTS

Acknowledgments......................................................................................................................v

Chapter 1: Introduction..........................................................................................................1

Chapter 2: Overview of methods.............................................................................................3

Chapter 3: ThYme: a database for thioester-active enzymes
   Abstract..........................................................................................................................9
   Introduction....................................................................................................................9
   Identifying and populating families..............................................................................12
   Content..........................................................................................................................12
   Database organization and features..............................................................................13
   Updates..........................................................................................................................14
   Experimentally characterized sequences....................................................................14
   Sequences with multiple domains................................................................................14
   Similarity to other enzyme databases..........................................................................15
   Conclusion.....................................................................................................................16
   March 2013 update.......................................................................................................16
   References.....................................................................................................................16
   Figures.............................................................................................................................18

Chapter 4: Thioesterases: A new perspective based on their primary and tertiary structures
   Abstract..........................................................................................................................19
   Introduction....................................................................................................................19
   Identification of thioesterase families..........................................................................20
   Identification of thioesterase clans..............................................................................24
   ThYme database...........................................................................................................25
   Results and discussion..................................................................................................26
Chapter 7: Structural classification and properties of ketoacyl reductases, hydroxyacyl dehydratases, and enoyl reductases

Abstract........................................................................................................................88
Introduction..................................................................................................................88
Computational methods...............................................................................................90
Results and discussion...............................................................................................90
Conclusions..................................................................................................................99
References....................................................................................................................99
Tables........................................................................................................................102
Figures........................................................................................................................107

Chapter 8: Acyl carrier protein structural classification and normal mode analysis

Abstract......................................................................................................................109
Introduction..............................................................................................................109
Results and discussion.............................................................................................110
Computational methods...........................................................................................118
Concluding remarks.................................................................................................120
Acknowledgments....................................................................................................121
References................................................................................................................121
Tables.........................................................................................................................125
Figures........................................................................................................................129
Supplementary Information.......................................................................................132

Chapter 9: Conclusion...............................................................................................144

Appendix: Theory and computation show that Asp463 is the catalytic proton donor in a mannosidase..............................................................................................................145
ACKNOWLEDGMENTS

National Science Foundation via the Engineering Research Center for Biorenewable
Chemicals, CBiRC

Major Professor: Peter J. Reilly

POS Committee: Laura R. Jarboe, Robert L. Jernigan, Monica H. Lamm, Basil J. Nikolau

Reilly Group:
Past and current graduate students: Blake Mertz, Luis Petersen, Yingfei Chen
Undergraduate students: Jay P. Chipman, Michael J. Forrester, Katherine Charov, Tingsong
Dai, Zachary Beversdorf, Julia Blum

Parc Cientific de Barcelona:
Albert Ardevol and Carme Rovira

Nikolau Group:
Fuyuan Jing, Marna Yandeau-Nelson, Shivani Garg, Bryon Upton

Information Technology Services:
Matthew L. Lemons, Daniel S. Black, Jeffrey Sorensen
CHAPTER 1
INTRODUCTION

The overall aim of this work is to further the understanding of the enzymes involved with fatty acid and polyketide synthesis. Although scientific questions remain to be resolved about them, they are significant as well due to the human use of their products. Fatty acids are major components of detergents, cosmetics, foods, and fuels. Polyketides are complex chemicals with medicinal applications, such as antibiotics and cancer-combating drugs. Additionally, the fatty acid and polyketide synthetic pathways have been proposed as alternative sources for the chemical industry’s platform compounds. It is of great interest to be able to design these enzymes to produce desired fatty acids or polyketides.

The initial goal of this work was to classify and organize fatty acid and polyketide synthesis enzymes based on sequence and structure, providing a framework to access and understand the existing information and knowledge about them. Many sequences with extensive experimental information exist, as well as some with resolved tertiary structures. However, a significantly greater number of similar sequences present in nature have been identified but have no confirmed activity or biochemical information. This framework assists and promotes ways to rationally discover the diversity of the enzymes’ natural specificity and activity. Most of the work included in this report is directed to such an aim.

Being able to engineer enzymes to produce compounds not found in nature or existing ones in greater quantities is an ultimate goal of studying enzymes. Additional efforts of this work are directed towards understanding the mechanisms of enzyme catalysis in atomic-level resolution. This provides fundamental knowledge that may contribute towards enzyme and product design.

Fatty acid and polyketide synthesis enzymes include acyl-coenzyme A (CoA) synthases, acyl-CoA carboxylases, acyl transferases, ketoacyl synthases, ketoacyl reductases, hydroxyacyl dehydratases, enoyl reductases, and thioesterases. The substrates, most commonly acyl chains, are covalently bound to either CoA or acyl carrier proteins (ACPs) during the course of reactions. This work reports the efforts and findings resulting from studying these enzymes, focusing mainly on thioesterases.
The next chapter presents an overview of the computational methods used throughout. Bioinformatics techniques were used to study sequences and structures, and molecular modeling and simulation methods were employed to study structures and mechanisms.

Chapter 3 describes the ThYme database, where almost all sequences and structures of the enzyme groups in fatty acid and polyketide synthesis are gathered and classified into families based on amino acid sequence similarity. This chapter has a detailed description of the enzyme groups present, as well as an overview of the database’s organization and use.

Chapter 4 describes work with thioesterases. A computational protocol to identify and populate enzyme families was developed with thioesterases and is described in this chapter. An overall and detailed review of thioesterases appears, including classification into families and clans, structural analysis, known catalytic residues and mechanisms, and the pathways and products involved.

Chapter 5 focuses on a single acyl-ACP thioesterase family, presenting my contribution to a collaborative effort to characterize and explore the substrate specificities of this family’s acyl-ACP thioesterases that determine free fatty acid chain length. Statistical and phylogenetic studies identified representative sequences for biochemical characterization, so that the entire substrate specificity and enzyme activity space was explored.

In Chapter 6, mixed quantum mechanics/molecular mechanics simulations, done to find the mechanism of a thioesterase in atomic resolution, along with its reaction energy barrier and transition-state geometry, are presented.

Chapter 7 is a review of ketoacyl reductases, hydroxyacyl dehydratases, and enoyl reductases, including their classification into families, structural analysis, known catalytic residues and mechanisms, and pathways and products.

ACP findings are reported in Chapter 8. Besides presenting ACP families and reviewing the pathways and products involved, normal vibrational mode analysis on ACP structures was conducted, and unknown ACP structures were predicted.

The Appendix presents work started during a summer research program at ISU and completed before graduate school. The proton donor residue in a glycoside hydrolase was identified with theory and automated docking calculations.
CHAPTER 2
OVERVIEW OF METHODS

This chapter presents an overview of the computational methods used, falling into three categories: sequences, structures, and molecular simulation. An overview of the fatty acid and polyketide synthesis enzymes is included in Chapter 3. More detailed reviews of thioesterases appear in Chapters 4 and 5, thioesterase mechanisms in Chapter 6, ketoacyl reductases, hydroxyacyl dehydratases, and enoyl reductases in Chapter 7, and acyl carrier proteins in Chapter 8.

Sequences

Alignments and phylogeny

Proteins and enzymes are composed of amino acid residues arranged in a sequence. When two sequences are highly similar, their functions may be same and/or come from a common ancestor. Before comparing two or more sequences, they need to be aligned. When two sequences are aligned, each paired site may have a match with same residues at the same site, a mismatch when different residues occupy the same site, or a gap when one sequence does not have a residue at such position. A scoring matrix rewards points for matches and penalties for mismatches and gaps. There are different scoring matrices; the most reliable ones for proteins are the PAM and BLOSUM matrices that are based on observed mutations in nature. Alignment algorithms try to find the optimal pairing between sequences that maximizes the alignment score.

A pairwise alignment is when two sequences are aligned, while aligning more than two sequences gives a multiple sequence alignment (MSA). MSAs are more challenging, as the best alignment between a single pair may not be the ideal for the whole set. Several strategies have been developed to solve MSAs. Most have either a progressive approach where additional sequences are added to an initial pair, or iterative approaches where MSAs are made multiple times by changing the order of sequences, or clustering approaches. Sequence alignments can be done with nucleotide or amino acid sequences. In this work, all alignments and sequence comparisons are of amino acid sequences, and they appear in Chapters 3, 4, 5,
A phylogenetic tree is a visual representation of an MSA. They can be used to make evolutionary inferences. However, in this work phylogenetic trees are made to visually identify groups of similar sequences within a larger set, and to confirm that groups of sequences are more similar to each other than to those outside the group. Phylogenetic work appears mainly in Chapter 5, but also in Chapter 8.

**Searching**

There are many identified sequences in nature, most if not all deposited at the National Center for Biotechnology Information. Most sequences have not been expressed and characterized experimentally, and their putative functions are inferred from sequence similarities. Given the large number of existing sequences, making pairwise alignments between a query sequence and all existing ones would be very time-consuming. BLAST (Basic Local Alignment Search Tool) is a heuristic algorithm that allows finding all sequences in a library or database similar to a query sequence. It does not produce optimal alignments, but it is computationally very efficient and is widely used. BLAST in this work was used to find and populate enzyme families, appearing in Chapters 3, 4, 7, and 8.

**Structures**

**Superimposition and comparison**

Proteins have three-dimensional structures that include α-helices, β-strands, and loops, known as secondary structure elements. Primary sequence determines secondary structure. The way that secondary structure elements are arranged is known as a fold. Protein folds are very conserved, and not much primary sequence similarity is required for different proteins to maintain the same structural fold.

To compare and quantify the structural difference between two proteins with the same fold, they are superimposed, and the distance between corresponding atoms is calculated, typically reported as root mean square deviations (RMSDs). A correct superimposition, or structural alignment, is required to properly identify corresponding residues and atoms between two structures. Existing methods to superimpose protein structures seek to minimize
the RMSD between two structures. Most are based solely on geometric criteria, others seek to align secondary structure elements, or require identification of key residues. In this work, protein structure superimpositions, comparisons, or RMSD calculations appear in Chapters 3–8.

Normal mode analysis

The dynamic behavior of equilibrium protein structures can be described by their vibrational normal modes. The normal modes of a structure include three translational, three rotational, and various vibrational ones. The slowest vibrational modes of a protein represent their dynamic equilibrium behavior. Although molecular dynamics or first-principles calculations (see the molecular simulation section) methods can be used to find the normal modes of any molecule, more efficient methods exist for proteins such as elastic network models. In such models, nodes connected by springs with force constants represent a protein. A node is placed for every α-carbon atom, and the system is allowed to oscillate. Parameters such as force constant and spring equilibrium distances can be changed so that average residue fluctuations agree with residue fluctuations from experimentally resolved three-dimensional structures. Normal mode analyses done with elastic network models appear in Chapter 8.

Predictions

There are two main approaches for protein three-dimensional structure prediction: comparative modeling and physics-based modeling. Predicting the structure of a protein solely from its sequence with a physics-based model, or a de novo approach, is challenging and time-consuming. Comparative modeling methods were used in this work, in Chapter 8, to predict unknown protein structures, followed by molecular dynamics (a physics-based approach, discussed in the molecular simulation section) to refine predicted structures. Comparative modeling uses experimentally resolved three-dimensional structures and sequence similarity to predict the three-dimensional structure of a protein. Comparative modeling methods assume that structure is more conserved than is its sequence. Two types were used in this work, homology modeling and threading.
In homology modeling, when the sequence of a protein with known structure (template structure) is similar to the sequence of a protein with unknown structure (target structure), the template structure is used to predict the target structure. Sequence similarity needs to be above ~25% to have a reliable prediction using homology modeling. Loops and sections with no defined secondary structure, as well as the orientation of amino acid side chains, are difficult to predict. Energy minimizations with physics-based methods can be used to refine the predicted structures. Threading, or fold recognition, like homology modeling uses a template based on known structures to predict the structure of a target protein. Threading differs from homology modeling in that while homology modeling strictly uses a single structure as a template, threading uses many structures within a conserved fold to build the template. Threading is used over homology modeling when no template structures have sequence similarity above ~25% to the target structure.

**Molecular simulation**

*Classical molecular dynamics*

Molecular dynamics (MD) is a simulation method used to study the time-resolved behavior of a system. A model, or force field, describes the energy of a system separated into units. From the energy, the forces acting on each unit can be obtained, and using Newton’s equations of motion, the velocities of each unit can be calculated. For each time step, the simulation allows each unit of the system to move according to its mass and velocity, the energy is recalculated in the new conformation, and the new velocities are recalculated. The behavior of the system will tend toward equilibrium and energy minima.

In the case of proteins, usually the units of the system are atoms, and the system is a protein solvated in water at physiological conditions. However, the system can vary according to what is desired to be studied. Several force fields have been developed specially for proteins, and they are known to accurately model their behavior. In this work, the Amber force field was used to account for covalent bond stretching, bending, and turning, as well as for van der Waals and electrostatic non-bonded interactions. If the system is solvated in water, their atoms also require a force field. MD allows studying the behavior of proteins and enzymes, solvated in water, time-resolved, and in atomic resolution. Atom-resolved MD
methods with elaborate force fields such as Amber are limited to systems in the order to $10^6$ atoms and in nanosecond time scales to conform with available computational resources. MD is used in Chapters 6 and 8.

*First-principles molecular dynamics*

In *ab initio*, or first-principles, MD the electronic configuration of the system is taken into account to calculate the energy of the system, while the nuclei are treated with classical MD. Traditional *ab initio* MD is known as Born-Oppenheimer molecular dynamics (BOMD). In BOMD, the electron wave functions are optimized in each time step to calculate the energy of the system at each nuclear configuration. Car-Parrinello molecular dynamics (CPMD) differs from BOMD in that the electron wave functions are optimized once at the start of the simulation, and fictitious electron variables allow the wave functions to evolve with the nuclei along the simulation. However, shorter time steps are required due to electronic degrees of freedom: time steps in BOMD are typically around ~1 fs, while those of CPMD are ~0.10 fs. However, CPMD is computationally cheaper. CPMD uses density functional theory (DFT) for electronic calculations. CPMD simulations are limited to systems with less than 500 atoms and picosecond time scales with available computational resources. All first-principles MD calculations in this work appear in Chapter 6 and used CPMD.

Mixed quantum mechanics/molecular mechanics (QM/MM) methods can be used to study enzyme reactions. Classical MD methods model the behavior of solvated proteins well; however, since classical force fields do not account for electrons explicitly, *ab initio* methods are needed to simulate chemical reactions. In QM/MM methods, atoms in the QM region are treated with *ab initio* MD, and atoms in the MM region with classical MD. In this study, QM/MM simulations were done where the enzyme active site was the QM region and the rest of the enzyme and water were in the MM region. The developed method, where CPMD determines the behavior of the QM region was used in Chapter 6.

*Metadynamics*

While the QM/MM approach allows first-principles calculations on a relatively large system, such as a solvated enzyme, to simulate enzymatic reaction a time-scale barrier needs
to be overcome. A QM/MM simulation would need to run for at least $1/k_{\text{cat}}$ to spontaneously observe a reaction. This is in the millisecond time scale and is very much outside the range of *ab initio* MD methods. Therefore, rare-event, or non-equilibrium, simulation methods are needed.

Metadynamics\(^8\) is used to simulate rare events by reducing the dimensions of a system, and has been used to solve problems in chemistry, materials science, and biophysics.\(^9\)

Collective variables (CVs) are chosen, and repulsive energy terms (Gaussian-like potentials) are added in CV space of the system explored by MD. This allows the system to escape free energy minima and to explore non-equilibrium conformations. CVs need to clearly describe the event desired to be simulated, and to differentiate clearly between initial and final states. Further, the free energy surface (FES) of the simulated event as a function of CV space can be reconstructed by summing the Gaussian-like repulsive terms added during the simulation. The metadynamics technique was used in Chapter 6 to induce the enzymatic reaction and to reconstruct the FES of the process.

References

CHAPTER 3

THYME: A DATABASE FOR THIOESTER-ACTIVE ENZYMES

David C. Cantu¹, Yingfei Chen¹, Matthew L. Lemons² and Peter J. Reilly¹

Department of Chemical and Biological Engineering¹, and Information Technology Services², Iowa State University

Modified from a paper published in Nucleic Acids Research, 39, D342-D346, with the same title and authors.

Abstract

The ThYme (Thioester-active enzYme; http://www.enzyme.cbirc.iastate.edu) database has been constructed to bring together amino acid sequences and three-dimensional (tertiary) structures of all the enzymes constituting the fatty acid synthesis and polyketide synthesis cycles. These enzymes are active on thioester-containing substrates, specifically those that are parts of the acyl-CoA synthase, acyl-CoA carboxylase, acyl transferase, ketoacyl synthase, ketoacyl reductase, hydroxyacyl dehydratase, enoyl reductase, and thioesterase enzyme groups. These groups have been classified into families, members of which are similar in sequences, tertiary structures, and catalytic mechanisms, implying common protein ancestry. ThYme is updated as sequences and tertiary structures become available.

Introduction

The ThYme (Thioester-active enzYme, http://www.enzyme.cbirc.iastate.edu) database presents enzymes acting on thioester-containing substrates, especially those involved in fatty acid and polyketide synthesis.

There are different ways to classify enzymes and proteins. The Enzyme Commission (EC) scheme classifies enzymes by the reactants or substrates that they primarily attack and by the reactions that they catalyze.¹ Another way is by three-dimensional (tertiary) structure, as found in the SCOP database.² A third method is to classify enzymes by primary (amino acid sequence) structure similarity. We have done so for thioesterases³ (TEs) and now for the other enzyme groups in the fatty acid synthesis cycle. Previously, this has been done with
glycoside hydrolases and other carbohydrate enzymes\(^4\) and with peptidases.\(^5\) Also, Pfam\(^6\) has done the same in a more universal way.

The fatty acid synthesis cycle (Figure 3.1) is the main pathway used by organisms to form lipids. The constituent members of this cycle are activated by the presence of thioester groups binding either coenzyme A (CoA) or acyl carrier protein (ACP). First, catalyzed by acyl-CoA synthases (ACSs), an acyl group is joined with CoA to make acyl-CoA, also called the priming substrate. Second, the priming substrate is carboxylated by acyl-CoA carboxylases (ACCs) to make the elongating substrate. The elongating substrate’s carrier molecule may be changed from CoA to ACP by acyl transferases (ATs). Then ketoacyl synthases (KSs) join the priming and elongating substrates, releasing a carbon dioxide and making ketoacyl-ACPs. The ketoacyl-ACP molecule then passes through a series of reduction, dehydration, and reduction steps catalyzed by ketoacyl reductases (KRs), hydroxyacyl dehydratases (HDs), and enoyl reductases (ERs), respectively, to create an acyl-ACP molecule two carbon atoms longer than the priming substrate. This new longer acyl-ACP molecule is then joined by a KS to another elongating substrate. This cycle elongates the acyl chain by two carbon atoms each turn until TEs hydrolyzes the CoA or ACP from the acyl group, effectively terminating fatty acid biosynthesis. Also, methylketone synthases (MKSs) can release molecules from the cycle before the reduction-dehydration-reduction steps. These enzymes first hydrolyze the thioester bond and then decarboxylate the carboxyl group of a 3-oxoacyl-ACP molecule, leaving a terminal methyloxo group.\(^7\) They have a TE domain, which appears in ThYme with other TEs; they do not form a large enzyme group.

More specifically, the enzyme groups involved in the fatty acid synthesis cycle and that appear in ThYme are the following:

(1) ACSs (part of EC 6.2.1, acid-thiol ligases). These enzymes add coenzyme A (CoA) or occasionally an acyl-carrier protein (ACP) to acetate or longer acceptors, powered by ATP or occasionally by GTP. This yields the activated compound and usually AMP, but in some cases ADP or GDP. ACSs are described by EC 6.2.1.1 to EC 6.2.1.36, with two entries having been deleted.

(2) ACCs (part of EC 6.4.1, ligases that form carbon-carbon bonds). In this step, the activated acceptor is elongated by the addition of a keto group derived from CO\(_2\),
yielding malonyl-CoA, malonyl-ACP, or a longer activated molecule. Seven ACCs with EC designations from 6.4.1 to 6.4.7 are listed.

(3) ATs (part of EC 2.3.1, acyl transferases transferring groups other than amino-acyl groups). These enzymes catalyze the transfer of an acyl chain from a CoA to an ACP or vice versa.

(4) KSs (part of EC 2.3.1, acyl transferases transferring groups other than amino-acyl groups). Here the activated malonyl or longer moiety is joined to an activated cycle constituent, releasing CO$_2$ and HSX, where SX is CoA or ACP. The growing chain is elongated by generally two, but occasionally more, carbon atoms. This EC category contains 190 entries, of which three has been deleted. Twenty EC entries out of 187 are KSs.

(5) KRs (part of EC 1.1.1, oxidoreductases acting on the CH–OH group of donors with NAD$^+$ or NADP$^+$ as acceptor, describing the reverse reaction). In those fatty acid synthesis cycle reactions, 3-oxo groups are reduced to 3-hydroxy groups by NADH or NADPH. EC 1.1.1.– contains at present 300 entries, 15 having been deleted.

(6) HDs (part of EC 4.2.1, carbon-oxygen hydro-lyases). Here the 3-hydroxy group is removed as water, yielding a double bond linking the 2- and 3-carbon atoms. There are 120 listings in this EC group, 16 having been deleted.

(7) ERs (part of EC 1.3.1, oxidoreductases acting on the CH–CH group of donors with NAD$^+$ or NADP$^+$ as acceptor). The 2,3-ene bond is reduced to a single bond. This EC group has 84 listings, of which four have been deleted.

(8) TEs (part of EC 3.1.2, thioester hydrolases). The thioester group is cleaved with water, leaving a fatty acid and HSX. The 27 EC entries have lost three members by deletion.

Polyketide biosynthesis is similar to fatty acid biosynthesis, yet it is more flexible and complex. Here the condensation-reduction-dehydration-reduction cycle is not completed at every turn; the KS-catalyzed reaction can occur between an intermediate in the cycle and an elongating substrate. This allows carbonyl, hydroxyl, and/or ethylene groups into the acyl chain. The TE will either hydrolyze acyl-CoA or acyl-ACP with a water molecule, or cyclize the chain using an alcohol on the chain itself for hydrolysis. Also, different compounds can be used for priming and elongating substrates.
These processes can be carried out by individual independent enzymes, or by large multi-modular fatty acid synthases (FASs) or polyketide synthases (PKSs) that contain the number of domains necessary, and in a specific order, to produce the desired molecule.

Among other uses, fatty acids have been recently proposed as biofuel feedstocks,⁸ while short-chain fatty acids could become feedstocks for biorenewable platform chemicals.⁹ Polyketides are a diverse family of chemicals, with some having medicinal applications such as erythromycin and tetracycline as antibiotics and doxorubicin and mithramycin in chemotherapy. Tailoring these molecules is of great interest; for that effort ThYme can be a useful tool in finding naturally occurring enzymes and in facilitating enzyme design.

Identifying and populating families

Family members must have strong sequence similarity and near-identical tertiary structures, and they must share general mechanisms as well as catalytic residues located in the same position. Methods for identifying and populating families were developed with TEs and later applied to other sequence groups. They were detailed in our previous work.³ (1) Experimentally confirmed enzyme sequences were used as queries. They were gathered from UniProt¹⁰, using only reviewed entries noted as having “Evidence at protein level”. (2) A series of successive Basic Local Alignment Search Tool (BLAST)¹¹ searches and comparison among results reduced query sequences to a few representative ones. (3) The catalytic domains of representative query sequences were subjected to BLAST to populate the families. These domains were selected by referring to Pfam-A,⁶ or by constructing a hidden Markov model profile¹² from a multiple sequence alignment (MSA) based on the initial BLAST result. (4) Experimentally confirmed enzymes were surveyed to search for missing potential enzyme families. (5) The uniqueness of the families was confirmed by MSAs, by tertiary structure superposition and comparison, and by catalytic residue positions.

Content

At present, ACSs are divided into five families, ATs into one, KSs into five, KRs into
four, HDs into six, ERs into six, and TEs into 23. ACCs are multidomain proteins first shown organized into domains followed by each domain divided into families: one family of the biotin carboxylase (BC) domain, one family of the biotin carboxyl carrier protein (BCCP), and two families of the carboxyl transferase (CT) domain appear. These enzyme groups’ annotation and sequences in each family appear in ThYme organized in the way mentioned below.

Database organization and features

The home page gives links to every enzyme group, as well as general information for viewers and citing and contact information. In each enzyme group’s main page, all families are listed in a table with “Names of enzymes and genes present”, which presents a non-exhaustive overview of the sequences found. This is meant to guide new users to the family that contains their enzymes of interest.

At the top of each enzyme family’s page (Figure 3.2), a table gives general information about the family, describing protein folds (if known from crystal structures), the names of enzymes and genes present (the list is not exhaustive), EC numbers (the most common ones), the catalytic residues (if they are known from the literature), and other notes. Also shown is the total number of Protein Data Bank \(^\text{13}\) (PDB) structures, and enzymes with “Evidence at protein level” and “Evidence at transcript level” (See “Experimentally characterized proteins” section below). This annotation might not be complete for all families.

Within an enzyme family’s page, all sequences appear by rows ordered into archaea, bacteria, and eukaryota, and alphabetically by producing species. All sequences in a row are identical and come from only one species. Identical sequences from different species are separated into different rows; however, identical sequences from different strains of the same species are not separated. If more than 500 rows exist, they are shown in multiple pages for a single family. The information is organized into the following columns: a) names or designations given to the proteins; b) EC numbers assigned to them, with a link to the ExPASy \(^\text{14}\) proteomics server; c) genus and species names along with strain designations of the organisms that produced them, with a link to the National Center for Biotechnology Information (NCBI) taxonomy browser; \(^\text{15}\) d) their GenBank identification, with a link to the
NCBI’s protein database; their RefSeq identification, with a link also to the NCBI’s protein database; their UniProt identification, with a link to the UniProt database; and their PDB identification, with a link to the PDB, if their known tertiary structure is available. All sequence names and EC numbers are taken from either UniProt or NCBI’s protein database; we do not assign sequence names or EC numbers.

Three features make navigating and retrieving information in ThYme easier. A search tool allows keywords, EC numbers, and GenBank, RefSeq, UniProt, or PDB accession codes to be searched. Furthermore, each family can be downloaded into a comma-separated value (csv) file, which can be viewed in a spreadsheet. Also, on each family’s page, only rows that include a PDB link or a UniProt link marked with “Evidence at transcript level” or “Evidence at protein level” can be viewed.

Updates

The content of existing families is updated continuously as NCBI’s protein database, UniProt, and PDB databases are updated; if a new sequence belongs in an existing family, it will appear there. To delete or merge existing families, as well as to define new families, the authors’ inspection and judgment is necessary; this cannot be automated.

Experimentally characterized sequences

Most sequences have no underlying specific experimental work, as they come from large genomic sequencing projects. The UniProt database, under the field “Protein existence” marks their entries with either “Evidence at protein level” or “Evidence at transcript level” if some experimental work has been done on the sequence. In ThYMe, we mark UniProt accessions with “Evidence at Protein Level” with a [P], and those with “Evidence at Transcript Level” with a [T]. The UniProt link or its equivalent in GenBank shows the experimental work’s literature. This should help users identify previous work on enzymes of interest.

Sequences with multiple domains

Some enzymes that appear in ThYme are multidomain FASs, PKSs, or non-ribosomal
peptide synthases. Each domain in these enzymes has its specific function, but all appear in a single sequence under the same GenBank, RefSeq, UniProt, or PDB accession. When the accession code of a multidomain enzyme appears in a family, only the domain of the enzyme group in which the family appears belongs in the family. (Example: UniProt P12785 is a rat fatty acid synthase. Its AT domain appears in AT2, its KS domain appears in KS3, its HD domain appears in HD4, and its TE domain appears in TE16.) A single multidomain sequence can have different PDB structures for each domain. Only the structure related to each family’s domain is shown. (Example: UniProt P49327 has several PDB structures. Among them, TE domain 1XKT appears in a TE family, AT domain 2JFD appears in an AT family, and so forth.)

**Similarity to other enzyme databases**

ThYme is most similar to CAZy in appearance and structure, in that both are interactive lists of enzyme primary and tertiary structures. However they are different in content, as ThYme shows enzymes active on substrates with thioester groups and CAZy shows enzymes active on carbohydrates. ThYme encompasses eight enzyme groups; CAZy on the other hand brings together four enzyme groups as well as different families of carbohydrate-binding modules.

ThYme is somewhat similar to MEROPS, which classifies peptidases and therefore has many more different enzyme groups and total number of listings. MEROPS and ThYme are also different in appearance and in the method by which listings are accessed.

The ESTHER database and the Lipase Engineering Database report sequences of the a/b hydrolase superfamily and lipases, respectively. In both databases, some of their families correspond with some TE families in ThYme, although the exact content and format differ.

Finally, Pfam has identified many protein families. Most ThYme families have an equivalent in Pfam. Our differences in methodology lead to different family content: Pfam families are more inclusive, covering a wide range of sequences, while ThYme families are smaller, with all sequences within a family having strong sequence similarity. Also, the purpose and format of the two databases are different; we focus on thioester-active enzymes and provide sequences and structures in families, while Pfam covers all proteins and, given a
query, it identifies the family or domain.

Conclusion

The ThYme database should provide a useful source of information on these enzymes that can help predict active sites, catalytic residues, and mechanisms of individual sequences, as well as providing a standardized nomenclature.

March 2013 update

Since the publication of this manuscript, the following changes have been done to the database: ACP families were added, two new TE families were found and added, one ER family was merged with a KR family, and two HD families were added.

References


Figures

Figure 3.1: The fatty acid synthesis cycle and the enzyme groups that comprise it

Figure 3.2: A family homepage in ThYme
CHAPTER 4
THIOESTERASES: A NEW PERSPECTIVE BASED ON THEIR PRIMARY AND TERTIARY STRUCTURES
David C. Cantu, Yingfei Chen, and Peter J. Reilly

Department of Chemical and Biological Engineering, Iowa State University

Modified from a paper published in Protein Science, 19, 1281-1295, with the same title and authors.

Abstract
Thioesterases (TEs) are classified into EC 3.1.2.1 through EC 3.1.2.27 based on their activities on different substrates, with many remaining unclassified (EC 3.1.2.–). Analysis of primary and tertiary structures of known TEs casts a new light on this enzyme group. We used strong primary sequence conservation based on experimentally proved proteins as the main criterion, followed by verification with tertiary structure superpositions, mechanisms, and catalytic residue positions, to accurately define TE families. At present TEs fall into 23 families almost completely unrelated to each other by primary structure. It is assumed that all members of the same family have essentially the same tertiary structure; however, TEs in different families can have markedly different folds and mechanisms. Conversely, the latter sometimes have very similar tertiary structures and catalytic mechanisms despite being only slightly or not at all related by primary structure, indicating that they have common distant ancestors and can be grouped into clans. At present four clans encompass 12 TE families. The new constantly updated ThYme (Thioester-active enzYmes) database contains TE primary and tertiary structures, classified into families and clans that are different from those currently found in other databases.

Introduction
The thioesterases (TEs), or thioester hydrolases, comprise a large enzyme group whose members hydrolyze the thioester bond between a carbonyl group and a sulfur atom. They are classified by the Nomenclature Committee of the International Union of Biochemistry and
Molecular Biology (NC-IUBMB) into EC (Enzyme Commission) 3.1.2.1 to EC 3.1.2.27, as well as EC 3.1.2.-- for unclassified TEs.\(^1\) Substrates of 15 of these 27 groupings contain coenzyme A (CoA), two contain acyl carrier proteins (ACPs), four have glutathione or its derivatives, one has ubiquitin, and two contain other moieties. In addition, three groupings have been deleted.

The EC classification system is based on enzyme function and substrate identity, and it was first formulated when very few amino acid sequences (primary structures) and three-dimensional (tertiary) structures of enzymes were available. Another way to classify enzymes is by primary structure into families and by tertiary structure into clans or superfamilies. Some databases are built this way: Pfam\(^2\) has a collection of protein families and domains, and SCOP\(^3\) classifies protein structures into classes, folds, families, and superfamilies. Other databases treat certain enzyme groups more specifically. For instance, MEROPS\(^4\) is a major database for peptidases, and CAZy\(^5\) covers carbohydrate-active enzymes.

It is common to observe that members of more than one EC grouping are found in one enzyme family based on similar amino acid sequences, implying that they have a common ancestor, mechanism, and tertiary structure. Conversely, members of a single EC grouping may be located in more than one enzyme family, being totally or almost totally unrelated in primary structure and potentially in mechanism and tertiary structure.

A further observation is that members of two different enzyme families may have very similar tertiary structures and mechanisms even though their primary structures are very different. This may imply that they are members of the same clan or superfamily, descended from a more distant common ancestor.

In this work, TE primary and tertiary structures will be analyzed to conclude how TEs are divided (and united) into families and clans. Structures, mechanisms, and catalytic residues are compared between families and clans. We compare our findings with existing databases such as Pfam and SCOP. Results also appear in a new continuously-updated database, ThYme (Thioester-active enzYmes, http://www.enzyme.cbirc.iastate.edu) that includes families and clans of enzyme groups that are part of the fatty acid synthesis cycle, TEs among them.
Identification of thioesterase families

Family members must have strong sequence similarity and near-identical tertiary structures, and they must share general mechanisms as well as catalytic residues located in the same position.

In general, TE families were identified in the following way: 1) experimentally confirmed TE sequences were used as queries; 2) a series of successive Basic Local Alignment Search Tool (BLAST) searches and comparison among results reduced query sequences to a few representative ones; 3) the catalytic domains of representative query sequences were subjected to BLAST to populate the families, 4) experimentally confirmed TEs were surveyed to search for missing potential TE families; 5) the uniqueness of the families was confirmed by multiple sequence alignments (MSAs), by tertiary structure superposition and comparison, and by catalytic residue positions. Methods are detailed below.

BLAST searches

TE families were identified by using BLAST on the catalytic domains of query TE sequences. Query sequences were taken from EC 3.1.2.1 to EC 3.1.2.27 and EC 3.1.2.– entries in the UniProt database. Only reviewed Swiss-Prot entries, which have a higher level of annotation, and those noted by “Evidence at Protein Level”, were used. Those described by “Inferred from Homology” or “Evidence at Transcript Level”, as well as fragments and putative or probable enzymes, were excluded from the query sequences. The sequences that met these criteria at that time made up the query sequence list. This list contained only 11 of the 27 TE EC numbers, as well as sequences having EC 3.1.2.– numbers, whereas 16 EC numbers did not meet these criteria, had no sequences, or had been deleted by the NC–IUBMB. Literature searches were also done to search for confirmed TEs whose information was not yet in UniProt or that were labeled under another EC number.

The first entry from the query sequence list was compared to the National Center for Biotechnology Information’s (NCBI) GenBanknr peptide sequence database using BLAST. The protein–protein BLAST algorithm was used, the cutoff E-value was set to 0.001, and Max Target Sequences was set to 10,000 to ensure that all results with \( E = 0.001 \) or less were
reported. All other parameters were default values. Later query sequences found in the BLAST results were deleted from the list. The remaining query sequences were subjected to BLAST in turn until all the listed sequences were either deleted or used. To automate this step, blast-2.2.19 was downloaded from NCBI’s webpage (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/unix_setup.html) and installed on a Unix platform. The nr peptide sequence database was also downloaded from this URL and updated before using BLAST. A script was written to successively use BLAST, compare results, and delete retrieved query sequence list members automatically. This identified representative query sequences.

After the first BLAST run, catalytic domains of representative query sequences were identified from Pfam-A (http://pfam.sanger.ac.uk/). Only these domains were subjected to BLAST to populate the families. If Pfam-A did not identify a domain with TE activity, MSAs using ClustalX 2.0\textsuperscript{10} or MUSCLE 3.6\textsuperscript{11} were performed on ~50 random sequences from the BLAST output file of the corresponding query sequence. Then profile hidden Markov models\textsuperscript{12} using hmmer-2.3.3 (http://hmmer.janelia.org) were used to identify a conserved domain in those sequences that was likely to be the catalytic domain.

The completeness of the representative query sequences was ascertained by searching all original query sequences in the BLAST output files, and checking for catalytic domains not included in the representative query sequences.

**Multiple sequence alignments**

The above procedures yielded BLAST output files making up each family. MSAs were then constructed to confirm or disallow each BLAST output file as a single family, or to check whether two or more output files should be merged into one family.

MSAs were constructed by randomly selecting ~50 sequences from one or more BLAST output files. If a sequence entry appeared in more than one BLAST output file, suggesting that these output files comprise one family, equal numbers of sequences were taken from each output file to obtain ~50 sequences in total. The sequences in FASTA format were obtained from Batch Entrez (http://www.ncbi.nlm.nih.gov/sites/batchentrez?db=Protein) and all sequences of protein fragments were deleted. ClustalX and MUSCLE default parameters
were used. Sufficient conservation of residues and residue chemical character confirms the presence of a single family. Lack of conservation suggests that more than one family exists, and if so, MSAs were performed on smaller sets of sequences.

Comparison of tertiary structures

All tertiary structures from the BLAST output files were obtained from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB) webpage (http://www.rcsb.org) and compared within one family and among families. PyMOL version 1.1 (DeLano Scientific, Palo Alto, CA; http://www.pymol.org/) was used to select the monomer of each PDB file containing the TE catalytic domain. Those structures were then superimposed with MultiProt\textsuperscript{13} version 1.0. MultiProt scoring and biocore parameters were both set to 2, with all other parameters remaining at their default values. All the monomer structures within one family were superimposed in MultiProt, and α-carbon atoms of all superimposed monomer structures were extracted to calculate the root mean square deviations (RMSDs) using the methods described below. Also, the positions of the catalytic residues in all tertiary structures within a family were determined to verify that they coincided.

RMSD calculations

An RMSD between as many α-carbon atoms as possible in two structures was calculated using. When two structures are superimposed, the RMSD is between the distances of corresponding α-carbon atoms. The shortest molecule was chosen as a pivot. The RMSD was found by calculating the distances between each α-carbon atoms in the compared molecule to those of the pivot molecule. An \( n \times m \) distance matrix was built, where \( n \) is the number of residues of the pivot molecule and \( m \) is the number of residues of the structure being compared. The lowest value in each row was considered for the RMSD calculation if it was lower than a cutoff value: the mean distance between two sequential α-carbon atoms within the pivot molecule. The number of residues in the molecule compared to the pivot that are used to calculate the RMSD was compared to the total number of residues in that molecule. Its ratio gives a percentage value (\( P \)) of residues used to calculate the RMSD. This \( P \) value is
a measure of the significance of the RMSD value. However, when more than two superimposed structures were being compared, more than one molecule was compared to the pivot and each RMSD was calculated. The RMSD calculation between a molecule and the pivot was done \( s - 1 \) times, where \( s \) is the number of superimposed structures. When \( s > 2 \), average RMSD (RMSD\(_{\text{ave}}\)) and \( P (P_{\text{ave}}) \) values were calculated. In addition, minimal (RMSD\(_{\text{min}}\) and \( P_{\text{min}} \)) and maximal (RMSD\(_{\text{max}}\) and \( P_{\text{max}} \)) values, for single comparisons between the pivot molecule and any other molecule, can be obtained.

**Identification of thioesterase clans**

Two or more TE families are grouped into a clan if all the sequences within them show some (not strong) sequence similarity, if their structures are strongly similar (narrowing the search to families with the same fold), and if they share similar active sites and general mechanisms. To consider all aspects of clan classification criteria, several methods are used to combine sequence and structural analysis. In addition, catalytic mechanisms of members of each family were gathered from the literature, and positions of catalytic residues were determined to verify that they coincided.

**Sequence analysis**

PSI–BLAST\(^6\) is an iterative method that builds an amino acid sequence profile and compares that profile to a primary structure database. It was conducted on each TE family; the sequence profile was built out of sequences that have known tertiary structures, and only sequences within a family were used for the profile. Two iterations were done with an \( E \)-value of 0.001 for the first BLAST search and 0.005 for the iterations. PSI–BLAST results were searched for sequences of structures that are part of other TE families. If sequences of one family appeared in the PSI–BLAST results of another family, then those two families have some sequence similarity.

**Secondary structure element analysis**

The arrangement of secondary structure elements (SSEs) (\( \alpha \)-helices and \( \beta \)-sheets only) in the core of TE tertiary structures was considered. This was done for only one structure per
family, since all structures within a family share nearly identical protein cores. Structures were viewed in PyMOL, and the SSE arrangement from N- to C-terminus was recorded for every TE family. Two families are considered to have similar structures if they have the same SSE order and arrangement.

**Visual inspection**

Adding to SSE analysis, one can view tertiary structures and visually decide which ones are similar. They can be viewed in PyMOL using the ‘align’ command to quickly see two or more superimposed structures.

**Superpositions and RMSDs**

All the previous methods can point to suspected clans, but a more quantitative comparison of tertiary structures is needed. Monomers of all structures in a suspected clan were superimposed with MultiProt. The superposition was verified visually. MultiProt reports an RMSD between matched residues in all the superimposed structures. Because few residues are matched among all superimposed structures within a clan, the reported RMSDs thus represent how those residues are superimposed, but not the whole similarity between the structures. For that, an RMSD more representative of structural similarity needs to be calculated: α-carbon atoms of all superimposed monomer structures were extracted to calculate RMSDs between them using the method detailed previously.

**ThYme database**

All the sequences in each family are displayed on the ThYme database website (http://www.enzyme.cbirc.iastate.edu). These sequences are taken, using a series of scripts, from the BLAST results of the catalytic domains of the representative query sequences. Matching accessions, taxonomical data, protein names, and EC numbers are taken from UniProt and GenBank databases, and in some cases, EFetch requests. Each TE family is shown on a page where sequences are arranged into archaea, bacteria, and eukaryota, then alphabetically by species. In each row, a single sequence or group of sequences with 100% identical catalytic domains are shown with their protein name and UniProt and/or GenBank
accession codes. EC numbers are shown only when they appear in a sequence’s UniProt or GenBank annotation. If a crystal structure is known, the PDB accession code also appears. ThYme will be continuously updated: the content of each family will grow as GenBank, UniProt, and PDB do; however, to create a new family, or to merge or delete existing ones, human judgment and manual changes will be necessary.

Results and discussion

Thioesterase family classification

Use of BLAST with TE query sequences followed by construction of MSAs and superposition of tertiary structures yielded 23 families almost completely unrelated by primary structure (Table 4.1).

Enzymes in families TE1 to TE13 hydrolyze substrates with various acyl moieties and CoA, those in TE14 to TE19 attack bonds between acyl groups and ACP, and those in TE20 and TE21 cleave the bonds between acyl groups and proteins. Members of TE22 and TE23 break bonds between acyl groups and glutathione and its derivatives (Table 4.2). The sulfur-carrying moiety in CoA and ACP is a pantethiene 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.

All tertiary structures within each family have almost identical cores and very strong overall resemblance (Table 4.3) shown by RMSD$_{ave}$ values of <1.8 Å and $P_{ave}$ values of >75%, with two exceptions. TE4 has a $P_{ave}$ value of 33.3% because it has only two crystal structures, of which one monomer (1C8U) is a double HotDog, while another monomer (1TBU) is incomplete with only a single HotDog. Similarly, in TE16 the $P_{ave}$ value is 65.8% because the TE domain of one structure (2VSQ) is smaller than the rest.

Of the families whose members hydrolyze acyl-CoAs, all have HotDog$^{14,15}$ folds (Table 4.3, Figure 4.1) except for TE1, TE2, and TE3. TE1 enzymes have NagB folds, and they have acetyl-CoA hydrolase (EC 3.1.2.1) activity as well as acetate or succinate-CoA transferase (EC 2.8.3.–) activity. They are found mainly in bacteria and fungi, although they are also present in archaea. Enzymes coded by the acetyl-CoA hydrolase $ACHI$ gene from Saccharomyces cerevisiae are present in TE1.$^{16}$ Fungal enzymes in this family are involved
with acetate levels and CoA transfer in mitochondria.\textsuperscript{17}

TE2 enzymes have $\alpha/\beta$-hydrolase\textsuperscript{18} folds (Figure 1). They are mainly found in eukaryotes (animals), but they are also present in bacteria. They have mostly palmitoyl (EC 3.1.2.2) and bile acid-CoA:amino acid N-acyl transferase (BAT) (EC 2.3.1.65) activities. The acyl-CoA TE (Acot) genes ACOT1, ACOT2, ACOT4, and ACOT6 from \textit{Homo sapiens} are present in this family, as well as the Acot1 through Acot6 genes from \textit{Mus musculus}, \textit{Rattus norvegicus}, and similar species.\textsuperscript{19} Also in TE2 are the BAAT TEs that transfer bile acid from bile acid-CoA to amino acids in the liver; these conjugates later solvate fatty acids in the gastrointestinal tract.\textsuperscript{20}

Enzymes in TE3 are part of the SGNH hydrolase superfamily with a flavodoxin-like fold. They are mainly found in bacteria and have acyl-CoA hydrolase (EC 3.1.2.20), arylesterase (EC 3.1.1.2) and lysophospholipase (EC 3.1.1.5) activities. Some TE3 enzymes come from the \textit{tesA} gene, and they are located in the periplasm and are involved in fatty acid synthesis.\textsuperscript{21} TE3 enzymes are also called acyl-CoA thioesterase I, protease I, and lysophospholipase L\textsubscript{1}, and the genes that code for them, \textit{tesA}, \textit{apeA}, and \textit{pldC}, respectively, are nearly identical.\textsuperscript{22}

The rest of the acyl-CoA hydrolase families have HotDog folds. TE4 enzymes, present in bacteria and eukaryotes, are acyl-CoA hydrolases as well as palmitoyl-CoA (EC 3.1.2.2) and choloyl-CoA (EC 3.1.2.27) hydrolases. The Acot8 gene encodes for peroxisomal TEs,\textsuperscript{23} which are found in TE4. Also in this family are acyl-CoA thioesterase II enzymes, encoded by the \textit{tesB} gene, that can hydrolyze a broad range of short- to long-chain acyl-CoA thioesters, but whose physiological function is not known.\textsuperscript{24}

TE5 acyl-CoA enzymes, also known as thioesterase III\textsubscript{s}, are present in bacteria. They are encoded by the \textit{tesC} (or \textit{ybaW}) gene, and are long-chain acyl-CoA TEs preferring 3,5-tetradecadienoyl-CoA as a substrate.\textsuperscript{25}

TE6 members, present in eukaryotes, bacteria, and archaea, have acyl-CoA hydrolase activities with various specificities. Acot enzymes 7, 11, and 12, present in eukaryotes, are found in TE6. Acot7 enzymes (also known as BACH: brain acyl-CoA hydrolases) are expressed mainly in brain tissue and preferentially attack C\textsubscript{8} to C\textsubscript{18} acyl-CoA chains.\textsuperscript{26} Acot11 (also known as BFIT: brown fat inducible thioesterase, or Them1: thioesterase superfamily member 1) enzymes are specific toward medium- and long-chain acyl-CoA
molecules, and they may be involved with obesity in humans. Acot12 (also known as CACH: cytoplasmic acyl-CoA hydrolase) enzymes in humans hydrolyze acetyl-CoA. Many bacterial TE6 sequences are YciA TE5s that hydrolyze a wide range of acyl-CoA thioesters and may help to form membranes. They preferentially attack butyryl, hexanoyl, lauroyl, and palmitoyl-CoA substrates.

TE7 enzymes are acyl-CoA TE7s found in eukaryota and bacteria. In this family are the Acot9 and Acot10 enzymes (previously known as MT-ACT48), which are expressed in the mitochondria and have short- to long-chain acyl-CoA TE activity, showing preference for C14 chains.

Most TE8 members, mainly present in eukaryota but also in bacteria, are acyl-CoA thioesterase 13 (Acot13) enzymes, also known as thioesterase superfamily member 2 (Them2). Enzymes in this family hydrolyze short to long acyl-CoA (C6 to C18) chains, preferring the latter.

TE9 members are found only in bacteria, and they have acyl-CoA hydrolase activity, mostly unclassified (3.1.2.--), but ADP-dependent short-chain acyl-CoA hydrolases (EC 3.1.2.18), and 4-hydroxybenzoyl-CoA hydrolases (EC 3.1.2.23) are also found. The YbgC TEs are found in this family; some hydrolyze primarily short-chain acyl-CoA thioesters, while others prefer long-chain acyl-CoA thioesters. Also, the thioesterase domain of methylketone synthase, MKS2, recently discovered in tomato, is found in TE9.

The enzymes in TE10 and TE11 are found only in bacteria, and most have 4-hydroxybenzoyl-CoA hydrolase (EC 3.1.2.23) activity. They, along with other enzymes, convert 4-chlorobenzoate to 4-hydroxybenzoate in soil-dwelling bacteria. Also in TE11 are the EntH (YbdB) TEs, involved with enterobactin (an iron chelator) biosynthesis in Escherichia coli. This is a unique example of a HotDog-fold enzyme involved in non-ribosomal peptide biosynthesis.

Most TE12 enzymes are 1,4-dihydroxy-2-napthoyl (DNHA)-CoA hydrolases, involved in vitamin K1 biosynthesis, and they are found mostly in bacteria. TE13 enzymes occur in archaea and bacteria. Most are either PaaI or PaaD enzymes in the phenylacetic acid degradation pathway, and they are part of the paa gene cluster.

TE14 to TE19 enzymes hydrolyze acyl-ACP thioesters, with those in TE14 and TE15
having HotDog folds while the rest have $\alpha/\beta$-hydrolase folds. TE14 enzymes are found in bacteria and plants; they have acyl-ACP hydrolase (EC 3.1.2.14) activity. Many plant enzymes in this family have been experimentally characterized: they contain FatA and FatB genes and can hydrolyze C$_8$ to C$_{18}$ acyl-ACP thioesters.\textsuperscript{40} All TE14 bacterial sequences come from genomic or structural genomic studies.

TE15 is a small family whose enzymes are present mainly in bacteria. Among them is the thioesterase CalE7 involved with enediyne biosynthesis. After substrate-ACP hydrolysis, these enzymes decarboxylate the product before release.\textsuperscript{41} Enzymes in this family are the few TEs with HotDog domains involved with polyketide biosynthesis.

TE16 enzymes occur in both eukaryotes and bacteria, and they have oleoyl-ACP hydrolase (EC 3.1.2.14) activity. They include the TE domains of fatty acid synthases (FASs), also known as Thioesterase I, that terminate fatty acid synthesis,\textsuperscript{42} and the TE domain of polyketide synthases (PKSs) and non-ribosomal peptide synthases (NRPs), also known as Type I thioesterases (TE I), that terminate polyketide biosynthesis,\textsuperscript{43} or non-ribosomal peptide biosynthesis.\textsuperscript{44} In the case of NRPs, instead of an ACP as the carrier molecule, a polypeptide carrier protein (PCP) is used. TE17 enzymes are only found in bacteria, mainly in \textit{Streptomyces}. They are the TE domains of various PKSs. FASs, PKSs, and NRPs are large multimodular enzymes with many domains having different functions. Only the TE domains were used to identify these family members.

Enzymes in TE18 are present in eukaryotes and bacteria and mainly have oleoyl-ACP hydrolase (EC 3.1.2.14) activity. Some enzymes in this family are S-acyl fatty acid synthetases/thioester hydrolases (Thioesterase II).\textsuperscript{45} They work with FASs to produce medium-chain (C$_8$–C$_{12}$) fatty acids in milk.\textsuperscript{46} The Type II thioesterases (TE IIs) are found in TE18; these enzymes play an important role in polyketide and non-ribosomal peptide biosynthesis by removing aberrant acyl chains from multimodular polyketide synthases and non-ribosomal peptide synthases.\textsuperscript{47,48} TE18 enzymes are independent TEs, not integrated to the multimodular FASs, PKSs or NRPs.

TE19 enzymes are classified as acyltransferases (EC 2.3.1.–), but they hydrolyze acyl-ACP molecules, mainly myristoyl-ACP.\textsuperscript{49} These enzymes divert fatty acids to the luminescent system in certain bacteria.
TE20 members, found only in eukaryotes, are palmitoyl-protein TEs (EC 3.1.2.22) encoded by PPT genes. They hydrolyze the thioester bond between a palmitoyl group and a cysteine residue in proteins.\textsuperscript{50} Mutations in PPT enzymes have been linked to neuronal ceroid lipofuscinosis, a genetic neurodegenerative disorder.\textsuperscript{51}

TE21 enzymes were originally identified as lysophospholipases,\textsuperscript{52} but they are also acyl-protein APT1 TEs.\textsuperscript{53} They hydrolyze thioester bonds between acyl chains and cysteine residues on proteins. Many proteins in this family also have carboxyesterase (EC 3.1.1.1) activity.

Among TE22 enzymes are S-formylglutathione hydrolases (EC 3.1.2.12) catalyzing formaldehyde detoxification; they hydrolyze S-formylglutathione into formate and glutathione.\textsuperscript{54} Also in TE22 are enzymes with acetyl esterase (EC 3.1.1.6) and carboxyesterase (EC 3.1.1.1) activity.

TE23 members are hydroxyglutathione hydrolases (EC 3.1.2.6), also known as glyoxalase II enzymes, that hydrolyze S-D-lactoyl-glutathione to glutathione and lactic acid in methylglyoxal detoxification.\textsuperscript{55} TE23 enzymes occur in archaea, bacteria, and eukaryotes and have a metallo-b-lactamase fold.\textsuperscript{56}

\textit{Correspondence to EC groupings}

These TE families bear rather limited resemblance to EC numbers representing TEs. For instance, acetyl-CoA hydrolases (3.1.2.1) occur in TE1, TE6, and TE7; palmitoyl-CoA hydrolases (EC 3.1.2.2) are found in TE2, TE4, TE6, and TE7; oleoyl-ACP hydrolases (EC 3.1.2.14) occur in TE14 and TE16 to TE18, and acyl-CoA hydrolases (EC 3.1.2.20) are found in TE3, TE6, and TE7. Conversely, of the 24 EC numbers remaining after three deletions, only 11 of them (EC 3.1.2.1, 3.1.2.2, 3.1.2.6, 3.1.2.12, 3.1.2.14, 3.2.1.18, 3.1.2.19, 3.1.2.20, 3.1.2.22, 3.1.2.23, and 3.1.2.27, along with unclassified thioesterases (EC 3.1.2.--)) occur in significant numbers among the 23 TE families. Of course, further EC numbers characteristic of TEs will likely appear as more TEs are sequenced and characterized.

\textit{Other thioesterases}

Ubiquitin carboxyl-terminal hydrolases (EC 3.1.2.15) cleave a wide variety of products
from the C-terminal glycine residue of ubiquitin. They were first identified as thiolesterases because they cleave dithiothreitol from ubiquitin, and they were thought to also hydrolyze ubiquitin-glutathione and other ubiquitin thioesters. \(^{57}\) It was later shown that they hydrolyze amides and other groups from ubiquitin. \(^{58}\) These enzymes belong to a larger class of peptidases called deubiquitinating enzymes that hydrolyze lysine-glycine amide bonds in ubiquitinated proteins. \(^{59}\) Several families of these enzymes can be found in MEROPS, the peptidase database. We identified eleven ubiquitin thiolesterase families by the methods described above, but we have not included them here or in the ThYme database, as peptidase activity is their main function, and they can be found in MEROPS.

Certain acyl transferases (EC 2.3.1.--), for example 2.3.1.9, 2.3.1.16, 2.3.1.38, and 2.3.1.39 among others, can hydrolyze acyl-CoA or acyl-ACP substrates and later join the liberated acyl group to another acyl-CoA or acyl-ACP molecule. Although they hydrolyze thioesters, this is not their main function, and therefore we also decided not to include these enzymes here.

**Thioesterase clan classification**

TE families 4 to 6 and 8 to 15, all with members having HotDog crystal structures, were subjected to the methods described above and two clans were found: TE-A comprising families TE5, TE9, TE10, and TE12; and TE-B with TE8, TE11, and TE13.

PSI–BLAST analysis suggested that TE5, TE9, TE10, and TE12 should be grouped into one clan and TE8, TE11, and TE13 into another, since slight sequence similarities among these families were found. SSE analysis of the structures pointed to TE5, TE6, TE10, TE12, and TE15 (having five \(\beta\)-strands) being placed in one clan and TE8, TE11, and TE13 (having six \(\beta\)-strands) being placed in another (Table 4.4); visual inspection suggested the same two groupings, with the first also including TE9. All crystal structures in candidate families of both possible clans were tested with superpositions and RMSD analysis (Figure 4.1, Table 4.5). These different tests led to the two clans being defined. Members of TE-A are all acyl-CoA hydrolases active on many substrates including short, long, branched, and aromatic acyl chains. Catalytic residues (see below) in TE6 are placed differently than those of other TE-A families, and TE6 was therefore not included in this clan. The different substrate specificities,
catalytic residues, and mechanism (see below) of TE15 members suggested that it also be excluded from TE-A. TE-B enzymes are also acyl-CoA hydrolases, except for the YbdB TEs in TE11 involved with enterobactin biosynthesis. TE4, TE7 (which has no known tertiary structure), and TE14 enzymes are sufficiently different from members of TE-A and TE-B that they were not considered for placement in either clan; the first two are acyl-CoA hydrolases, while the third is an acyl-ACP hydrolase.

TE families 2 and 16 through 22, whose members all have α/β-hydrolase crystal structures, belong to two clans: TE-C comprising TE16, TE17, and TE18, and TE-D with TE20 and TE21.

Both sequence analysis and SSE arrangement suggested only one clan of TE16, TE17, and TE18 (Table 4.4). Visual inspection suggested the two clans described above, and they were confirmed by superpositions, RMSD analysis, and the position of catalytic residues (Figure 4.1, Table 4.4). Families in TE-C contain acyl-ACP hydrolases present in multidomain FASs, PKSs, and NRPs, as well as independent acyl-ACP TEs involved in those pathways. TE-D enzymes hydrolyze palmitoyl and other acyl groups from protein surfaces. TE2, an acyl-CoA hydrolase, TE19, a myristoyl-ACP hydrolase, and TE22, active on glutathione-activated molecules, are not part of either clan.

**TE tertiary structures, catalytic residues, and mechanisms**

Catalytic mechanisms and residues of each TE family were gathered from crystal structure articles. The PDB files, proposed catalytic residues, and producing organisms of the relevant TEs are listed in Table 4.6.

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

In TE-A, only TE9 and TE10 can be further analyzed, as TE5 and TE12 at present have only one crystal structure each with no corresponding refereed article. In TE9 the YbgC structure 2PZH is a tetramer of two dimers. After comparing this structure to 1LO9 in TE10 and other YbgC crystals, the authors proposed that His18, Tyr7, and Asp11 play important roles in catalysis.

TE10 4-hydroxybenzoyl-CoA TEs have homotetrameric quaternary structures. It was
suggested from structures 1LO7, 1LO8, and 1LO9 that hydrogen bonds and the positive end of a helix dipole moment make the thioester carbonyl group more susceptible to a nucleophilic attack by Asp17 through an acyl-enzyme intermediate.60

TE-B families include TE8, TE11, and TE13. Members of TE8 are tetramers composed of two HotDog dimers. Based on a crystal structure of a human Them2 enzyme (3F5O), it was proposed that Gly57 and Asn50 bind and polarize the thioester carbonyl group while Asp65 and Ser85 orient and activate the water nucleophile.61

In TE11, Arthrobacter sp. strain SU 4-hydroxybenzoyl-CoA TE crystal structures reveal a tetrameric enzyme with a dimer of dimers. Structures 1Q4S, 1Q4T, and 1Q4U led to the proposal that Gly65 polarizes the carbonyl group for a nucleophilic attack carried out by Glu73.62

Both TE10 and TE11 are 4-hydroxybenzoyl-CoA TEs of similar substrate specificities and metabolic functions; however, their tertiary and quaternary structures are different and they use different active-site regions and residues for catalysis. This supports placing these two families in two different clans.

TE13 PaaI thioesterase from Thermus thermophilus HB8 yielded homotetrameric quaternary structures 1WLU, 1J1Y, 1WM6, 1WLV, and 1WN3. From those structures, a study proposed that these enzymes use an induced-fit mechanism to hydrolyze the substrate via an Asp48-activated water nucleophile.63 Comparison of the structure of another PaaI, from E. coli (2FS2) with the Arthrobacter TE11 structures, as well as site-directed mutagenesis, pointed to a mechanism similar to that in TE11: Gly53 prepares the thioester for a nucleophilic attack from Asp61.64 4-Hydroxybenzoyl-CoA enzymes from TE11 and the PaaI enzymes from TE13 catalyze two different reactions in different organisms, and their primary sequences are not related; yet their tertiary structures, catalytic residues, and mechanisms are similar, supporting the conclusion that both TE11 and TE13 are part of TE-B.

Of TE families with HotDog structures not placed in clans, TE4 catalytic residues and mechanisms have been identified based on structure 1C8U and site-directed mutagenesis: an Asp204–Gln278–Thr228 triad orients a water molecule for nucleophilic attack on the substrate.65 The double HotDog structure of 1C8U and its catalytic residues and mechanism
differ from other HotDog enzymes, supporting the exclusion of TE4 from TE-A and TE-B.

In TE6, Acot7 structure 2Q2B is a trimer of HotDog dimers; both domains are required for activity and Asn24 from the N-domain and Asp213 from the C-domain have been identified as catalytic residues through site-directed mutagenesis. Also in TE6, YciA structures 1YLI and 3D6L have a trimer of dimers with two binding sites across the dimer interface. An D44A mutation (3BJK) of 1YLI eliminated activity.

The two known tertiary structures in TE14 come from structural genomic studies and do not have supporting literature. However, based on a bioinformatics-guided site-directed mutagenesis study on a FatB enzyme from Arabidopsis, a study proposed that Cys264, His229, and Asn227 make up a papain-like catalytic triad.

Structure 2W3X of TE15 CalE7 has no acidic residues in the catalytic region. Based on site-directed mutagenesis, a mechanism different from other HotDogs was proposed: Asn19 and Arg37 anchor the substrate with hydrogen bonds while a water molecule or hydroxide anion acts as a nucleophile, attacking the substrate carbonyl group. Arg37 also acts as an oxyanion hole to stabilize the intermediate, and Tyr29 facilitates decarboxylation following hydrolysis.

Unlike HotDogs, α/β hydrolases have very conserved catalytic residues: a nucleophile–histidine–acid triad. The nucleophile can be serine, cysteine, or aspartate; the histidine is always conserved, and the third residue is always acidic. Fold architecture, substrate specificities, and binding sites vary within this fold. Generally, the acid stabilizes a histidine residue that acts as a base, accepting a proton from the nucleophile that forms an intermediate with the substrate, which is then attacked by water. In some PKSs and NRPs that make cyclic products, a hydroxyl group from the substrate chain is used for lactonization instead of a water molecule.

TE-C encompasses TE16, TE17, and TE18. Members of all three families have the same Ser–His–Asp triad. The position of the acid residue differs from what is usually found in α/β hydrolases, as was seen in human FAS TE domain 1XKT. In TE17, a substrate channel was found in the TE domain of 6-deoxyerythronolide B synthase (DEBS-TE) from Saccharopolyspora erythraea (1KEZ), unique among TE α/β hydrolases. The substrate channel depends on pH and supports the prediction that all macrocycle-forming TEs from PKSs have a similar
catalytic mechanism. Many macrocycle-forming TE domains from PKSs were found in TE17, again supporting this prediction.

Members of the two families, TE20 and TE21, in TE-D also have Ser–His–Asp catalytic triads. TE21 structure 1AUO confirmed that some bacterial carboxylesterases with broad substrate specificity have an α/β hydrolase fold. These enzyme structures are very similar to the mainly eukaryotic acyl-protein hydrolases, even though they have very different specificities.

TE2, TE19, and TE22 are other α/β hydrolases that are not part of any clan, but they still maintain the characteristic catalytic triad.

TEs are found in NagB (TE1), SGNH (TE3) and lactamase (TE23) folds. TE1 has two structures with no refereed literature, so their mechanism and catalytic residues remain unknown.

From TE3 structures 1IVN, 1JRL and 1J00, the existence of a Ser–His–Asp catalytic triad similar to those in the TE α/β hydrolases was proposed. The later crystal structures 1U8U and 1V2G suggested that a conformational change, described as a switch loop movement, occurs during catalysis.

Since TE23 hydroxyglutathione hydrolases (glyoxalase IIs) have metallo-β-lactamase folds, their mechanisms are very different from the rest of TEs that do not have catalytic metal ions. Two crystal structures, IQH3 and IQH5, have seven His and Asp residues and a water molecule interacting with two zinc ions. 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.

Predicted catalytic residues

Some TE tertiary structures, including all those in TE1, TE5, TE12, and TE14, are not supplemented by refereed literature. Furthermore, TE7 does not have a known tertiary structure. Therefore the catalytic residues and mechanisms for TEs in these families are not securely known.

Since the positions of catalytic residues must be conserved within families and clans,
unknown identities of catalytic residues in TEs can be predicted by viewing the positions of the catalytic residues in all superimposed structures in a family or clan.

In TE2, we predict that in the human ACOT4 structure 3K2I the catalytic triad is Ser232–His360–Asp326, based on the human ACOT2 structure 3HLK. In TE6, based on structures from mouse Acot7 (2Q2B), *Haemophilus influenzae* YciA (3BJK), and *Campylobacter jejuni* (3D6L), Asp245 in human ACOT7 and Asp36 in human ACOT12 appear to be catalytic residues. In TE8, study of the structures of human Them2 (3F5O and 2F0X) leads to the prediction that human TE structure 2H4U and mouse TE structure 2CY9 have the same catalytic residues: Asn50, Asp65, Ser83, and Gly57. In TE9, based on *Helicobacter pylori* YbgC enzyme 2PZH, we predict that in *E. coli* 1S5U the catalytic residues are Tyr14, Asp18, and His25, in *Bartonella henselae* 3HM0 they are Tyr23, Asp28, and His35, and in *Thermus thermophilus* they are 1Z54 Tyr11, Asp15, and His22. All three residues in the three structures superimposed very closely and in the same position with the exception of Tyr35 in 3HM0, where the ring is facing in another direction and the residue is one amino acid position displaced. This might lead to a difference in enzyme substrate specificity. In TE11, based on TE structures 1Q4S, 1Q4T, and 1Q4U from *Arthrobacter* sp., we predict that the catalytic residues are Gly55 and Glu63 in both structure 1VH9 of the putative TE from *E. coli* and in structures 2B6E, 1SC0, and 3LZ7 of a hypothetical *H. influenzae* TE. In TE13, *E. coli* structure 1PSU shares the same Gly53 and Asp61 catalytic residues with *E. coli* PaaI 2FS2, and *T. thermophilus* 2DSL shares the same Gly40 and Asp48 catalytic residues with the other *T. thermophilus* PaaI structures 1J1Y, 1WLU, 1WLV, 1WM6, and 1WN3.

Comparing TE20 bovine PPT1 (1EI9) and human PPT2 (1PJ A) suggests that human PPT1 structure 3GRO has the catalytic triad Ser115–His289–Asp233, the same residues as the bovine PPT1. In TE21, *Pseudomonas fluorescens* carboxylesterases 1AUO and 1AUR, and human acyl-protein TE 1FJ2 lead to predicting that the catalytic residues in *Pseudomonas aeruginosa* structures 3CN7 and 3CN9 should be Ser113, His197, and Asp166. In TE22, esterase A 2UZ0, S-formylglutathione hydrolase 1PV1, human esterase D 3FCX, and a putative enzyme (3E4D), suggest that in *Oleispira antartica* esterase 3I6Y the catalytic triad is Ser148–His257–Asp224.
Within TE-A, two families, TE5 and TE12, only have one crystal structure each (1NJK and 2HX5, respectively) with no supporting refereed article. Superimposition of structures within TE-A lead us to predict that the catalytic residue in TE5 YbaW probable enzyme is Asp13, while in TE12 putative thioesterase (2HX5) the catalytic residue is Asp19.

**Convergent and divergent evolution**

These results show the effects of both convergent and divergent evolution. The former is exemplified by the fact that members in TE families from different clans and/or folds, descended from different ancestors, are active on substrates, many of them the same and all containing the thioester group, attacking the bond between its carbonyl carbon atom and its adjacent sulfur atom. Divergent evolution is shown by the presence in many TE families of enzymes with separate names and EC numbers (beyond the undifferentiated number EC 3.1.2.--), signifying activities on different substrates, even though they have similar primary and tertiary structures and mechanisms. A more profound indication of divergent evolution is the presence of clans containing families with very different primary structures but common tertiary structures and mechanisms.

**Thioesterases in existing databases and previous classifications**

Some TEs have been previously classified by sequence or structure in different databases. The subjective boundaries between subfamilies, families, and clans or superfamilies, as well as the use of different methodologies, give rise to differences among databases. A summary of these differences can be found on Table 4.7.

The SCOP classification system\(^3\) is based on protein structure, and the SCOP database was assembled mainly by visual inspection and comparison, with some automation. We used it to search for the structures of TE family members. Our classes and folds correspond with SCOP; however, some differences arise at the superfamily and family levels. For example, structures in TE2, TE15, and TE18 are not classified, nor are those in TE7, which does not have a known structure. Some of our clans correspond to SCOP families: TE-B corresponds to the Paal/YdiI-like SCOP family, and TE-A is somewhat similar to the 4HBT-like SCOP family.
Pfam has identified and classified many protein families. Pfam-A is manually curated and its families are identified with hidden Markov model profiles built from carefully chosen seed multiple alignments. The main difference in methodology is that in Pfam, sequences that fit a profile from a seed alignment are part of a family, while in ThYme sequences strongly similar to a query sequence that has TE function and evidence at protein level are part of a TE family. This leads to differences in the families: Pfam families are more inclusive and cover a wider range of sequences, while ThYme families are smaller, with all sequences having a very strong sequence similarity. For example, Pfam-A family 4HBT includes our families TE5 to TE13 and TE15. We have shown that these families have varied structures and overall functions, as well as different mechanisms and catalytic residues. Furthermore, some of these ThYme families appear in two clans, TE-A and TE-B, and some are not part of any clan.

Dillon and Bateman in 2004 defined subfamilies for the HotDog fold. Among HotDog fold enzymes are TEs, and some of the subfamilies presented correspond with our families. For instance, TE4 corresponds to the TesB-like subfamily, TE5 to the YbaW subfamily, TE6 to the Acyl-CoA thioesterase subfamily, TE9 to the YbgC-like subfamily, TE10 and TE11 to the 4HBT-I and 4HBT-II subfamilies, TE13 to the PaaI subfamily, and TE14 to the Fat subfamily. However, the exact content, sequences, and structures of these subfamilies might not correspond to the corresponding families in ThYme.

Two other more specific databases, ESTHER, a database for the a>b hydrolase fold, and the Lipase Engineering Database, contain some TEs. Both have families that resemble TE19, TE20, and TE21, and ESTHER also has one family similar to TE16. Again, the exact content, sequences, and structures in these families might not correspond to those in ThYme.

ThYme has detailed structural and biological annotation for TEs and other thioester-active enzymes. It does not compete with but rather supplements existing databases such as Pfam, as it is focused on thioester-active enzymes in the same way that CAZy is focused on carbohydrate-active enzymes and MEROPS is focused on peptidases. It can accurately point out differences and similarities in structure, mechanisms, and catalytic residues. The differences in methodology and purpose lead to differences with other databases. We follow the SCOP classification system, but differences in the databases arise because SCOP’s
purpose is to present a tertiary structure classification, while ThYme classifies TEs and other thioester-active enzymes mainly by primary structure. Pfam has a wide scope, and it is an extremely useful tool to identify domains within sequences, but it is not specific to a particular enzyme group. Dillon and Bateman’s article,\textsuperscript{15} as well as ESTHER and the Lipase Engineering Database, have different focuses, and their overall contents do not correspond to that in ThYme.

**Concluding remarks**

Analysis of primary and tertiary structures has led to classifying all TEs into families and clans. Within families all enzymes share very similar primary and tertiary structures, active sites, catalytic residues, and mechanisms. Within clans, enzymes share some sequence similarity and similar structures, active sites, catalytic residues, and mechanisms. This classification system can help to predict an individual enzyme’s structure, function, or mechanism. It also provides a standardized nomenclature.

**Acknowledgment**

The authors thank Luis Petersen for his help in automating BLAST.

**References**


## Tables

Table 4.1: Thioesterase Families and Common Names of their Members

<table>
<thead>
<tr>
<th>Family</th>
<th>Producing organisms</th>
<th>Genes and/or other names of family members</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE1</td>
<td>A, B, E&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Ach1</td>
</tr>
<tr>
<td>TE2</td>
<td>A, B, E</td>
<td>Acot1–Acot6, BAAT thioesterase</td>
</tr>
<tr>
<td>TE3</td>
<td>A, B</td>
<td><em>tesA</em>, acyl-CoA thioesterase I, protease I, lysophospholipase L1</td>
</tr>
<tr>
<td>TE4</td>
<td>B, E</td>
<td><em>tesB</em>, acyl-CoA thioesterase II, Acot8</td>
</tr>
<tr>
<td>TE5</td>
<td>B</td>
<td><em>tesC</em> (<em>ybaW</em>), acyl-CoA thioesterase III</td>
</tr>
<tr>
<td>TE6</td>
<td>A, B, E</td>
<td>Acot7 (BACH), Acot11 (BFIT, Them1), Acot12 (CACH), YciA</td>
</tr>
<tr>
<td>TE7</td>
<td>B, E</td>
<td>Acot9, Acot10</td>
</tr>
<tr>
<td>TE8</td>
<td>A, B, E</td>
<td>Acot13 (Them2)</td>
</tr>
<tr>
<td>TE9</td>
<td>B</td>
<td>YbgC</td>
</tr>
<tr>
<td>TE10</td>
<td>B</td>
<td>4HBT-I</td>
</tr>
<tr>
<td>TE11</td>
<td>B</td>
<td>4HBT-II, EntH (YbdB)</td>
</tr>
<tr>
<td>TE12</td>
<td>B, E</td>
<td>DNHA-CoA hydrolase</td>
</tr>
<tr>
<td>TE13</td>
<td>A, B</td>
<td>pal1, palD</td>
</tr>
<tr>
<td>TE14</td>
<td>B, E</td>
<td>FatA, FatB</td>
</tr>
<tr>
<td>TE15</td>
<td>B</td>
<td>Thioesterase CalE7</td>
</tr>
<tr>
<td>TE16</td>
<td>A, B, E</td>
<td>TE domain of FAS (Thioesterase I), TE domain of PKS or NRP (type I thioesterase (TE I))</td>
</tr>
<tr>
<td>TE17</td>
<td>B</td>
<td>TE domain of PKS</td>
</tr>
<tr>
<td>TE18</td>
<td>B, E</td>
<td>Thioesterase II, type II thioesterase (TE II)</td>
</tr>
<tr>
<td>TE19</td>
<td>B</td>
<td>luxD</td>
</tr>
<tr>
<td>TE20</td>
<td>E</td>
<td>ppt1, ppt2, palmitoyl-protein thioesterase</td>
</tr>
<tr>
<td>TE21</td>
<td>A, B, E</td>
<td>apt1, apt2, thioesterase, phospholipase, carboxylesterase</td>
</tr>
<tr>
<td>TE22</td>
<td>A, B, E</td>
<td>S-formylglutathione hydrolase, esterase A, esterase D</td>
</tr>
<tr>
<td>TE23</td>
<td>A, B, E</td>
<td>Hydroxyglutathione hydrolase, glyoxalase II</td>
</tr>
</tbody>
</table>

<sup>a</sup> A: archaea, B: bacteria, E: eukaryota. Most prevalent producers bolded.
Table 4.2: Thioesterase Functions and Substrate Specificities

<table>
<thead>
<tr>
<th>Family</th>
<th>General function</th>
<th>EC number</th>
<th>Preferred substrate specificity (if known)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE1</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.1, 2.8.3.–</td>
<td>Acetyl-CoA</td>
</tr>
<tr>
<td>TE2</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–, 3.1.2.2, 2.3.1.65</td>
<td>Palmitoyl-CoA, bile-acid-CoA</td>
</tr>
<tr>
<td>TE3</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–, 3.1.2.20, 3.1.1.2, 3.1.1.5</td>
<td>Medium- to long-chain acyl-CoA</td>
</tr>
<tr>
<td>TE4</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–, 3.1.2.2, 3.1.2.27</td>
<td>Short- to long-chain acyl-CoA, palmitoyl-CoA, choloyl-CoA</td>
</tr>
<tr>
<td>TE5</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–</td>
<td>Long-chain acyl-CoA, 3,5-tetradecadienoyl-CoA</td>
</tr>
<tr>
<td>TE6</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–, 3.1.2.1, 3.1.2.2, 3.1.2.18, 3.1.2.19, 3.1.2.20</td>
<td>Short- to long-chain acyl-CoA, C₄ to C₁₈</td>
</tr>
<tr>
<td>TE7</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–, 3.1.2.1, 3.1.2.2, 3.1.2.20</td>
<td>Short- to long-chain acyl-CoA</td>
</tr>
<tr>
<td>TE8</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–</td>
<td>Short- to long-chain acyl-CoA, C₆ to C₁₈</td>
</tr>
<tr>
<td>TE9</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–, 3.1.2.18</td>
<td>Short- to long-chain acyl-CoA, 4-hydroxybenzoyl-CoA</td>
</tr>
<tr>
<td>TE10</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.23</td>
<td>4-Hydroxybenzoyl-CoA</td>
</tr>
<tr>
<td>TE11</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–</td>
<td>4-Hydroxybenzoyl-CoA</td>
</tr>
<tr>
<td>TE12</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–</td>
<td>1,4-Dihydroxy-2-napthoyl-CoA</td>
</tr>
<tr>
<td>TE13</td>
<td>Acyl-CoA hydrolase</td>
<td>3.1.2.–</td>
<td>Short and medium-chain acyl-CoA, several hydroxyphenylacyl-CoA substrates</td>
</tr>
<tr>
<td>TE14</td>
<td>Acyl-ACP hydrolase</td>
<td>3.1.2.–, 3.1.2.14</td>
<td>Short- to long-chain acyl-ACP, C₈ to C₁₈</td>
</tr>
<tr>
<td>TE15</td>
<td>Acyl-ACP hydrolase</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>TE16</td>
<td>Acyl-ACP hydrolase</td>
<td>3.1.2.14a</td>
<td>Long-chain acyl-ACP, various polyketides and non-ribosomal peptides</td>
</tr>
<tr>
<td>TE17</td>
<td>Acyl-ACP hydrolase</td>
<td>3.1.2.14b</td>
<td>Several polyketides</td>
</tr>
<tr>
<td>TE18</td>
<td>Acyl-ACP hydrolase</td>
<td>3.1.2.–, 3.1.2.14</td>
<td>Medium-chain acyl-ACP, various polyketides and non-ribosomal peptides</td>
</tr>
<tr>
<td>TE19</td>
<td>Acyl-ACP hydrolase</td>
<td>2.3.1.–</td>
<td>Myristoyl-ACP</td>
</tr>
<tr>
<td>TE20</td>
<td>Protein-palmitoyl hydrolase</td>
<td>3.1.2.–, 3.1.2.22</td>
<td>Palmitoyl-protein</td>
</tr>
<tr>
<td>TE21</td>
<td>Protein-acyl hydrolase</td>
<td>3.1.2.–, 3.1.1.1</td>
<td>—</td>
</tr>
<tr>
<td>TE22</td>
<td>Glutathione hydrolase</td>
<td>3.1.2.12, 3.1.1.1, 3.1.1.6</td>
<td>S-formylglutathione</td>
</tr>
<tr>
<td>TE23</td>
<td>Glutathione hydrolase</td>
<td>3.1.2.6</td>
<td>D-lactoylglutathione</td>
</tr>
</tbody>
</table>

a TE domain. FASs, PKSs and NRPs can have several EC numbers such as 2.3.1.85, 2.3.1.94, 2.3.1.–, 2.7.7.–, and 5.1.1.–.
b TE domain of PKSs.
Table 4.3: Thioesterase Folds

<table>
<thead>
<tr>
<th>Family</th>
<th>Fold</th>
<th>RMSD&lt;sub&gt;ave&lt;/sub&gt;, Å</th>
<th>P&lt;sub&gt;ave&lt;/sub&gt;, %</th>
<th>PDB files</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE1</td>
<td>NagB</td>
<td>1.25</td>
<td>96.4</td>
<td>2G39, 2NVV</td>
</tr>
<tr>
<td>TE2</td>
<td>α/β-Hydrolase</td>
<td>1.00</td>
<td>96.6</td>
<td>3HLK, 3K2I</td>
</tr>
<tr>
<td>TE3</td>
<td>Flavodoxin-like</td>
<td>0.58</td>
<td>96.6</td>
<td>11V, 1J00, 1JRL, 1U8U, 1V2G, 3HP4</td>
</tr>
<tr>
<td>TE4</td>
<td>HotDog</td>
<td>0.90</td>
<td>33.3</td>
<td>1C8U, 1TBU</td>
</tr>
<tr>
<td>TE5</td>
<td>HotDog</td>
<td>—</td>
<td>—</td>
<td>1NJ3</td>
</tr>
<tr>
<td>TE6</td>
<td>HotDog</td>
<td>1.39</td>
<td>75.9</td>
<td>3B7K, 2Q2B, 2V10, 2QQ2, 1YLI, 3BJK, 3D6L</td>
</tr>
<tr>
<td>TE7</td>
<td>HotDog</td>
<td>—</td>
<td>—</td>
<td>2H4U, 3F50, 2F0X, 2CY9</td>
</tr>
<tr>
<td>TE8</td>
<td>HotDog</td>
<td>0.58</td>
<td>88.3</td>
<td>2PZH, 1SSU, 3HM0, 1Z54</td>
</tr>
<tr>
<td>TE9</td>
<td>HotDog</td>
<td>1.19</td>
<td>88.8</td>
<td>1BVQ, 1LO7, 1LO8, 1LO9</td>
</tr>
<tr>
<td>TE10</td>
<td>HotDog</td>
<td>0.67</td>
<td>97.1</td>
<td>1Q4S, 1Q4T, 1Q4U, 1VH9, 2B6E, 1SC0, 3LZ7</td>
</tr>
<tr>
<td>TE11</td>
<td>HotDog</td>
<td>0.87</td>
<td>93.9</td>
<td>2H5X</td>
</tr>
<tr>
<td>TE12</td>
<td>HotDog</td>
<td>—</td>
<td>—</td>
<td>2FS2, 1PSU, 2DSL, 1J1Y, 1WL, 1WLV, 1WM6, 1WN3</td>
</tr>
<tr>
<td>TE13</td>
<td>HotDog</td>
<td>1.65</td>
<td>87.7</td>
<td>2OWN, 2ESS</td>
</tr>
<tr>
<td>TE14</td>
<td>HotDog</td>
<td>—</td>
<td>—</td>
<td>2W3X</td>
</tr>
<tr>
<td>TE15</td>
<td>HotDog</td>
<td>—</td>
<td>—</td>
<td>2VZ8&lt;sup&gt;a&lt;/sup&gt;, 2VZ9&lt;sup&gt;a&lt;/sup&gt;, 2PX6, 1XKT, 2ROQ&lt;sup&gt;b&lt;/sup&gt;, 2CB9, 2CBG, 2VSQ, 1JMK</td>
</tr>
<tr>
<td>TE16</td>
<td>α/β-Hydrolase</td>
<td>1.51</td>
<td>66.9</td>
<td>1MO2, 1KEZ, 1MN6, 2H7X, 2H7Y, 2HK, 2HFJ, 1MNA, 1MNQ</td>
</tr>
<tr>
<td>TE17</td>
<td>α/β-Hydrolase</td>
<td>1.67</td>
<td>82.4</td>
<td>3FLA, 3FLB, 2RON&lt;sup&gt;b&lt;/sup&gt;, 2K2Q&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>TE18</td>
<td>α/β-Hydrolase</td>
<td>0.83</td>
<td>97.2</td>
<td>1THT</td>
</tr>
<tr>
<td>TE19</td>
<td>α/β-Hydrolase</td>
<td>—</td>
<td>—</td>
<td>1EH5, 3GRO, 1EI9, 1Ex, 1PIA</td>
</tr>
<tr>
<td>TE20</td>
<td>α/β-Hydrolase</td>
<td>1.41</td>
<td>91.2</td>
<td>1FJ2, 1AUO, 1AUR, 3CN7, 3CN9</td>
</tr>
<tr>
<td>TE21</td>
<td>α/β-Hydrolase</td>
<td>0.82</td>
<td>96.7</td>
<td>3FCX, 3C6B, 2UZ0, 1VP1, 1I6Y, 3E4D, 3LS2</td>
</tr>
<tr>
<td>TE22</td>
<td>α/β-Hydrolase</td>
<td>1.69</td>
<td>78.9</td>
<td>2QED, 1XM8, 2P18, 2GCU, 2Q42, 1QH3, 1QH5, 2P1E</td>
</tr>
</tbody>
</table>

<sup>a</sup> 2VZ8 and 2VZ9 have TE domains in their FASTA format. Therefore these were picked up by BLAST, but their PDB files do not include the TE domain, and they were not included in the RMSD calculation.

<sup>b</sup> NMR-resolved structures not included in RMSD calculation.
Table 4.4: Thioesterase Core Secondary Structure Elements

<table>
<thead>
<tr>
<th>Clan</th>
<th>Family</th>
<th>Secondary structural element</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HotDog</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TE-A</td>
<td>TE5</td>
<td>β-α-β-β-β-β</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE9</td>
<td>α-β-β-β-β</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE10</td>
<td>β-α-β-β-β-β</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE12</td>
<td>β-α-β-β-β-β</td>
</tr>
<tr>
<td>TE-B</td>
<td>TE8</td>
<td>β-β-α-β-β-β-β</td>
</tr>
<tr>
<td>TE-B</td>
<td>TE11</td>
<td>β-β-α-β-β-β-β</td>
</tr>
<tr>
<td>—</td>
<td>TE4</td>
<td>α-β-β-β-β-β-β-α-β-β-β</td>
</tr>
<tr>
<td>—</td>
<td>TE6</td>
<td>β-α-β-β-β</td>
</tr>
<tr>
<td>—</td>
<td>TE14</td>
<td>β-α-β-β-β-β-β-α-β-β-β</td>
</tr>
<tr>
<td>—</td>
<td>TE15</td>
<td>β-α-β-β-β-β</td>
</tr>
<tr>
<td><strong>α/β-Hydrolase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TE-C</td>
<td>TE16</td>
<td>β-α-β-α-β-α-β-β-α-β-α</td>
</tr>
<tr>
<td>TE-C</td>
<td>TE17</td>
<td>β-α-β-α-β-α-β-β-α-β-α</td>
</tr>
<tr>
<td>TE-C</td>
<td>TE18</td>
<td>β-β-α-β-α-β-β-α-β-α-β-α</td>
</tr>
<tr>
<td>TE-D</td>
<td>TE20</td>
<td>β-α-β-α-β-α-β-α-β-α-β-α-β-α</td>
</tr>
<tr>
<td>TE-D</td>
<td>TE21</td>
<td>β-α-β-β-α-β-α-β-α-β-α-β-α</td>
</tr>
<tr>
<td>—</td>
<td>TE2</td>
<td>β-β-α-β-α-β-α-β-α-β-α-β-α</td>
</tr>
<tr>
<td>—</td>
<td>TE19</td>
<td>β-β-β-α-β-α-β-α-β-α-β-α-β-α-β-α</td>
</tr>
<tr>
<td>—</td>
<td>TE22</td>
<td>β-β-β-α-β-α-β-α-β-α-β-α-β-α-β-α</td>
</tr>
</tbody>
</table>

Table 4.5: RMSD Analysis of TE Clan Members

<table>
<thead>
<tr>
<th>Clan</th>
<th>RMSD_{min}, Å</th>
<th>RMSD_{ave}, Å</th>
<th>RMSD_{max}, Å</th>
<th>P_{min}, %</th>
<th>P_{ave}, %</th>
<th>P_{max}, %</th>
<th>Cutoff, Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE-A</td>
<td>1.14</td>
<td>1.33</td>
<td>1.53</td>
<td>77.5</td>
<td>87.1</td>
<td>90.9</td>
<td>3.81</td>
</tr>
<tr>
<td>TE-B</td>
<td>0.11</td>
<td>0.97</td>
<td>2.02</td>
<td>72.3</td>
<td>86.8</td>
<td>100.0</td>
<td>3.80</td>
</tr>
<tr>
<td>TE-C</td>
<td>1.81</td>
<td>1.94</td>
<td>2.13</td>
<td>52.6</td>
<td>58.3</td>
<td>75.2</td>
<td>3.82</td>
</tr>
<tr>
<td>TE-D</td>
<td>0.44</td>
<td>1.45</td>
<td>2.00</td>
<td>67.0</td>
<td>80.9</td>
<td>100.0</td>
<td>3.79</td>
</tr>
</tbody>
</table>
Table 4.6: TE Family and Clan Mechanisms

<table>
<thead>
<tr>
<th>Clan</th>
<th>Family</th>
<th>Catalytic residues*</th>
<th>PDB file</th>
<th>Producing organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>TE1</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>—</td>
<td>TE2</td>
<td>Ser294, His422, Asp388&lt;br&gt;Ser232&lt;sup&gt;b&lt;/sup&gt;, His360&lt;sup&gt;b&lt;/sup&gt;, Asp326&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3HLK&lt;br&gt;3K2I</td>
<td>Homo sapiens&lt;br&gt;H. sapiens</td>
</tr>
<tr>
<td>—</td>
<td>TE3</td>
<td>Ser10, His157, Asp154</td>
<td>IIVN, 1JRL, 1J00, 1U8U, 1V2G</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>—</td>
<td>TE4</td>
<td>Asp204, Thr228, Gln278</td>
<td>1C8U</td>
<td>E. coli</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE5</td>
<td>Asp13&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1NJK</td>
<td>E. coli</td>
</tr>
<tr>
<td>—</td>
<td>TE6</td>
<td>Asn24, Asp213&lt;br&gt;Asp44&lt;br&gt;Asp34&lt;br&gt;Asp36&lt;sup&gt;b&lt;/sup&gt;&lt;br&gt;Asp245&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2Q2B, 2V1O&lt;br&gt;1YLI, 3BJK&lt;br&gt;3D6L&lt;br&gt;3B7K&lt;br&gt;2QQ2</td>
<td>Mus musculus&lt;br&gt;Haemophilus influenzae&lt;br&gt;Campylobacter jejuni&lt;br&gt;H. sapiens&lt;br&gt;H. sapiens</td>
</tr>
<tr>
<td>—</td>
<td>TE7</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>TE-B</td>
<td>TE8</td>
<td>Asn50, Asp65, Ser83, Gly57&lt;br&gt;Asn50&lt;sup&gt;b&lt;/sup&gt;, Asp65&lt;sup&gt;b&lt;/sup&gt;, Ser83&lt;sup&gt;b&lt;/sup&gt;, Gly57&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2F0X, 3F5O, 2H4U&lt;sup&gt;b&lt;/sup&gt;&lt;br&gt;2CY9</td>
<td>H. sapiens&lt;br&gt;M. musculus&lt;br&gt;Helicobacter pylori&lt;br&gt;E. coli&lt;br&gt;Bartonella henselae&lt;br&gt;Thermus thermophilus&lt;br&gt;Pseudomonas sp.&lt;br&gt;Arthrobacter sp.&lt;br&gt;E. coli&lt;br&gt;H. influenzae&lt;br&gt;Prochlorococcus marinus&lt;br&gt;Thermus thermophilus</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE9</td>
<td>Tyr7, Asp11, His18&lt;br&gt;Tyr14&lt;sup&gt;b&lt;/sup&gt;, Asp18&lt;sup&gt;b&lt;/sup&gt;, His25&lt;sup&gt;b&lt;/sup&gt;&lt;br&gt;Tyr23&lt;sup&gt;b&lt;/sup&gt;, Asp28&lt;sup&gt;b&lt;/sup&gt;, His35&lt;sup&gt;b&lt;/sup&gt;&lt;br&gt;Tyr11&lt;sup&gt;b&lt;/sup&gt;, Asp15&lt;sup&gt;b&lt;/sup&gt;, His22&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2PZH&lt;br&gt;1S5U&lt;br&gt;3HM0&lt;br&gt;1Z54</td>
<td>E. coli&lt;br&gt;H. influenzae&lt;br&gt;THERMUS THERMOPHILUS&lt;br&gt;E. coli&lt;br&gt;H. influenzae&lt;br&gt;Prochlorococcus marinus&lt;br&gt;Thermus thermophilus</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE10</td>
<td>Asp17</td>
<td>1BVQ, 1LO7, 1LO8, 1LO9</td>
<td>H. sapiens&lt;br&gt;M. musculus&lt;br&gt;Helicobacter pylori&lt;br&gt;E. coli&lt;br&gt;Bartonella henselae&lt;br&gt;Thermus thermophilus&lt;br&gt;Pseudomonas sp.&lt;br&gt;Arthrobacter sp.&lt;br&gt;E. coli&lt;br&gt;H. influenzae&lt;br&gt;Prochlorococcus marinus&lt;br&gt;Thermus thermophilus</td>
</tr>
<tr>
<td>TE-B</td>
<td>TE11</td>
<td>Gly65, Glu73&lt;br&gt;Gly55&lt;sup&gt;b&lt;/sup&gt;, Glu63&lt;sup&gt;b&lt;/sup&gt;&lt;br&gt;Gly55&lt;sup&gt;b&lt;/sup&gt;, Glu63&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1Q4S, 1Q4T, 1Q4U&lt;br&gt;1VH9&lt;br&gt;2B6E, 1SC0, 3LZ7</td>
<td>H. sapiens&lt;br&gt;M. musculus&lt;br&gt;Helicobacter pylori&lt;br&gt;E. coli&lt;br&gt;Bartonella henselae&lt;br&gt;Thermus thermophilus&lt;br&gt;Pseudomonas sp.&lt;br&gt;Arthrobacter sp.&lt;br&gt;E. coli&lt;br&gt;H. influenzae&lt;br&gt;Prochlorococcus marinus&lt;br&gt;Thermus thermophilus</td>
</tr>
<tr>
<td>TE-A</td>
<td>TE12</td>
<td>Asp19&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2HX5</td>
<td>E. coli&lt;br&gt;Arabidopsis thaliana&lt;br&gt;Micromonospora echinospora&lt;br&gt;H. sapiens&lt;br&gt;Bacillus subtilis&lt;br&gt;B. subtilis</td>
</tr>
<tr>
<td>—</td>
<td>TE13</td>
<td>Gly40&lt;sup&gt;b&lt;/sup&gt;, Asp48&lt;br&gt;Gly53, Asp61</td>
<td>1WLU, 1J1Y, 1WM6, 1WLV, 1WN3, 2DSL&lt;sup&gt;b&lt;/sup&gt;&lt;br&gt;2FS2, 1PSU&lt;sup&gt;b&lt;/sup&gt;</td>
<td>E. coli&lt;br&gt;Arabidopsis thaliana&lt;br&gt;Micromonospora echinospora&lt;br&gt;H. sapiens&lt;br&gt;Bacillus subtilis&lt;br&gt;B. subtilis</td>
</tr>
<tr>
<td>—</td>
<td>TE14</td>
<td>Cys264, His229, Asn227</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>—</td>
<td>TE15</td>
<td>Asn19, Tyr29, Arg37</td>
<td>2W3X</td>
<td>—</td>
</tr>
<tr>
<td>TE-C</td>
<td>TE16</td>
<td>Ser2308, His2481, Asp2338&lt;br&gt;Ser80, His207, Asp107&lt;br&gt;Ser84, His201, Asp111</td>
<td>1XKT, 2PX6&lt;br&gt;1JMK&lt;br&gt;2CB9, 2CBG</td>
<td>E. coli&lt;br&gt;Arabidopsis thaliana&lt;br&gt;Micromonospora echinospora&lt;br&gt;H. sapiens&lt;br&gt;Bacillus subtilis&lt;br&gt;B. subtilis</td>
</tr>
<tr>
<td>TE-C</td>
<td>TE17</td>
<td>Ser142, His259, Asp169</td>
<td>1KEZ, 1MO2</td>
<td>Saccharopolyspora erythraea</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>------------------------</td>
<td>--------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td></td>
<td>TE18</td>
<td>Ser148, His268, Asp176</td>
<td>1MN6, 1MNA, 1MNQ</td>
<td>S. erythraea</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>Ser148, His268, Asp176</td>
<td>2HFK, 2HFJ, 2H7X, 2H7Y</td>
<td>Streptomyces venezuelae</td>
</tr>
<tr>
<td>TE-D</td>
<td>TE19</td>
<td>Ser94, His228, Asp200</td>
<td>3FLA, 3FLB</td>
<td>Amycolatopsis mediterranei</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser114, His241, Asp211</td>
<td>1THT</td>
<td>Vibrio harveyi</td>
</tr>
<tr>
<td></td>
<td>TE20</td>
<td>Ser115, His289, Asp233</td>
<td>1EI9, 1EH5, 1EXW</td>
<td>S. erythraea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser115, His289, Asp233</td>
<td>3GRO</td>
<td>Streptomyces venezuelae</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser111, His283, Asp228</td>
<td>1PJA</td>
<td>Vibrio harveyi</td>
</tr>
<tr>
<td></td>
<td>TE21</td>
<td>Ser114, His199, Asp168</td>
<td>1AUO, 1AUR</td>
<td>H. sapiens</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser114, His203, Asp169</td>
<td>1FJ2</td>
<td>Pseudomonas fluorescens</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser113&lt;sup&gt;b&lt;/sup&gt;, His197&lt;sup&gt;b&lt;/sup&gt;, Asp166&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3CN7, 3CN9</td>
<td>H. sapiens</td>
</tr>
<tr>
<td></td>
<td>TE22</td>
<td>Ser120, His231, Asp202</td>
<td>2UZ0</td>
<td>Pseudomonas aeruginosa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser161, His276, Asp241</td>
<td>1PV1, 2C6B</td>
<td>Streptococcus pneumoniae</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser149, His260, Asp226</td>
<td>3FCX</td>
<td>S. cerevisiae</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser147, His256, Asp223</td>
<td>3E4D</td>
<td>H. sapiens</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser147, His258, Asp255</td>
<td>3LS2</td>
<td>Agrobacterium tumefaciens</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser148&lt;sup&gt;b&lt;/sup&gt;, His257&lt;sup&gt;b&lt;/sup&gt;, Asp224&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3I6Y</td>
<td>Pseudoalteromonas h</td>
</tr>
<tr>
<td></td>
<td>TE23</td>
<td>Various His and Asp residues, along with Zn, Fe, or Mn ions</td>
<td>1QH3</td>
<td>Oleispira antarctica</td>
</tr>
</tbody>
</table>

<sup>a</sup> As proposed by the authors in the literature cited, except as noted.

<sup>b</sup> Predicted by the authors of this paper based upon the position of catalytic residues in superimposed structures within families.

<sup>c</sup> Predicted by the authors of this paper based upon the position of catalytic residues in superimposed structures within clan TE-A.
<table>
<thead>
<tr>
<th>Table 4.7: Comparison of Family and Clan Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThYme clan</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td><strong>TE1</strong></td>
</tr>
<tr>
<td><strong>TE2</strong></td>
</tr>
<tr>
<td><strong>TE3</strong></td>
</tr>
<tr>
<td><strong>TE4</strong></td>
</tr>
<tr>
<td><strong>TE-A</strong></td>
</tr>
<tr>
<td><strong>TE6</strong></td>
</tr>
<tr>
<td><strong>TE7</strong></td>
</tr>
<tr>
<td><strong>TE-B</strong></td>
</tr>
<tr>
<td><strong>TE-A</strong></td>
</tr>
<tr>
<td><strong>TE-A</strong></td>
</tr>
<tr>
<td><strong>TE-B</strong></td>
</tr>
<tr>
<td><strong>TE-A</strong></td>
</tr>
<tr>
<td><strong>TE-B</strong></td>
</tr>
<tr>
<td><strong>TE14</strong></td>
</tr>
<tr>
<td><strong>TE15</strong></td>
</tr>
<tr>
<td><strong>TE-C</strong></td>
</tr>
<tr>
<td><strong>TE-C</strong></td>
</tr>
<tr>
<td><strong>TE-C</strong></td>
</tr>
<tr>
<td><strong>TE-C</strong></td>
</tr>
<tr>
<td><strong>TE-C</strong></td>
</tr>
<tr>
<td><strong>TE-D</strong></td>
</tr>
<tr>
<td><strong>TE-D</strong></td>
</tr>
<tr>
<td><strong>TE22</strong></td>
</tr>
<tr>
<td><strong>TE23</strong></td>
</tr>
</tbody>
</table>
**Figures**

Figure 4.1: Superimposed tertiary structures of single representatives of each TE family in a clan. A: TE-A acyl-CoA hydrolases from *Escherichia coli* (TE5) (green), *Helicobacter pylori* (TE9) (red), *Pseudomonas sp.* (TE10) (yellow), and *Prochlorococcus marinus* (TE12) (blue). B: TE-B acyl-CoA hydrolases from *Homo sapiens* (TE8) (blue), *Arthrobacter sp.* (TE11) (red), and *E. coli* (TE13) (yellow). C: TE-C acyl-ACP hydrolases from *H. sapiens* (TE16) (blue), *Saccharopolyspora erythraea* (TE17) (red), and *Amycolatopsis mediterranei* (TE18) (yellow). D: TE-D protein-acyl hydrolases from *Bos taurus* (TE20) (blue) and *H. sapiens* (TE21) (yellow).
CHAPTER 5
PHYLOGENETIC CHARACTERIZATION OF AN ACYL-ACP THIOESTERASE FAMILY
David C. Cantu, Jay P. Chipman, and Peter J. Reilly

An excerpt of a paper published in BM C Biochemistry, 12, 44.

Title of the full paper: Phylogenetic and experimental characterization of an acyl-ACP thioesterase family reveals significant diversity in enzymatic specificity and activity.

Authors of the full paper: Fuyuan Jing+2, David C. Cantu+1, Jarmila Tvaruzkova2, Jay P. Chipman1, Basil J. Nikolau2, Marna D. Yandeau-Nelson2, and Peter J. Reilly1

Departments of Chemical and Biological Engineering1, and Biochemistry, Biophysics & Molecular Biology2, Iowa State University; *Equal contributions

Abstract
Thioesterases (TEs) have been recently classified into families based on sequence and three-dimensional structure similarity. TE14 is a family that includes the plant enzymes FatA and FatB as well as many uncharacterized bacterial sequences. The FatA and FatB enzymes terminate fatty acid biosynthesis in plants and are key in determining fatty acid length and overall composition in seed oil. We have separated TE14 into subfamilies based on phylogenetic and statistical analyses, giving a more detailed classification that allows the better inference of biochemical properties, and an experimental biochemical characterization of individual enzymes in each subfamily. TE14 is made up of ten subfamilies, with four composed of plant sequences and six of bacterial sequences.

Introduction
The thioesterases (TEs) are a large enzyme group whose members cleave the bond linking the sulfur and carbonyl carbon atoms of a thioester group. They have been classified into families,1 where all sequences in a family have strongly related amino acid sequences
and very similar three-dimensional (tertiary) structures and positions of catalytic residues. This implies that they are descended from a common protein ancestor and that they use similar catalytic mechanisms. Virtually all TE amino acid sequences are found, classified into families, in the constantly-updated ThYme database (www.enzyme.cbirc.iastate.edu), separated into those of archaeal, bacterial, and eukaryotic origin, with information on their EC numbers (if available), GenBank, RefSeq, and UniProt identifiers, and Protein Data Bank (PDB) locators of any with known tertiary structures.

At present 23 TE families exist of which 13 families (TE1–TE13) are composed of acyl-CoA hydrolases, six (TE14–TE19) of acyl-acyl carrier protein (ACP) hydrolases, two (TE20 and TE21) of protein-acyl hydrolases, and two (TE22 and TE23) of glutathione hydrolases.1

This work focuses on one acyl-ACP TE family, TE14, which is made up of sequences produced by bacteria and plants. Among the plant-derived members in TE14 are the FatA and FatB enzymes.2 All bacterial sequences come from genomic projects, including PDB structures 2ESS and 2OWN. No existing bacterial sequence in this family has been experimentally characterized. TE14 members have HotDog folds and their catalytic residues have been proposed to be cysteine, histidine, and asparagine, arranged in a papain-like catalytic triad.3

Several FatA and FatB enzymes have been characterized biochemically, with some examples below. FatA enzymes act on long-chain acyl-ACPs with a preference for 18:1 fatty acids.2,4-6 FatB enzymes preferably hydrolyze acyl-ACPs with saturated fatty acid chains.2 Many Cuphea sequences have been characterized: Cuphea hookeriana FatB enzymes are specific for acyl-ACPs with 8- and 10-carbon-long chains,2 Cuphea palustris enzymes for 8-, 10-, 14-, and 16-carbon-long chains,7 and Cuphea wrightii enzymes for 12-, 14-, and 16-carbon-long chains.8 Sequences from Arabidopsis thaliana are specific for chains of 14 to 18 carbon atoms.9 Myristica fragans FatB enzymes hydrolyze acyl-ACPs with chains of 14 to 18 carbon atoms, and Ulmus americana enzymes hydrolyze acyl-ACP with chains of mainly 8, 10, 14, and 16 carbon atoms.10

TE14 plant acyl-ACP TEs are plastid-targeted enzymes encoded in the nucleus11 that terminate fatty acid synthesis in plants.12 Produced fatty acids later become part of membranes or are stored as triacylglycerols.13 Since they are key in determining fatty acid
length and composition in seed oils, TE14 enzymes become a target for producing oils and other carbon-chain-based molecules. Plant triacylglycerols have been proposed as biofuel feedstocks, while short-chain fatty acids could become feedstocks for biorenewable platform chemicals.

This study has two main objectives: 1) to identify TE14 subfamilies and to experimentally characterize their members, and 2) to search for naturally occurring TEs that hydrolyze short-chain acyl-ACP molecules. To do that, we have conducted a phylogenetic and statistical study of TE14, dividing it into subfamilies, allowing researchers to better infer biochemical properties of closely related sequences.

**Phylogenetic and statistical analysis methods**

TE sequences that make part of TE14 in the ThYme database to date were taken from the GenBank and UniProt databases. Fragments and incomplete sequences were removed, yielding 360 amino acid sequences. The catalytic domains of these sequences were aligned using MUSCLE 3.6 with default parameters to create a multiple sequence alignment (MSA).

An unrooted phylogenetic tree based on the MSA was built using Molecular Evolutionary Genetics Analysis 4 (MEGA4). The minimum evolution algorithm was used due to its high effectiveness with large data sets, gaps were treated with pairwise deletion, and an amino acid Jones-Taylor-Thornton (JTT) distance model was chosen. The phylogenetic tree was further verified by a bootstrap test with 1000 replicates.

The bootstrapped consensus tree was qualitatively analyzed and broken down into apparent subfamilies. A statistical analysis was done to show that all sequences within a subfamily are more closely related to each other than to those in other subfamilies. Based on the MSA, JTT distances between all sequences were calculated and arranged into a \( j \times j \) matrix, where \( j \) is the total number of sequences. Inter-subfamily distances and variances were determined using this matrix. For each apparent subfamily a smaller \( k \times k \) matrix, where \( k \) is the number of sequences in a given subfamily, was calculated. From this, intra-subfamily mean distances and variances can be determined. These values were applied to the equation below to determine \( z \):
where \(x_{avg,ij} \), \(x_{avg,ii} \), and \(x_{avg,jj} \) are the inter- and intra-subfamily mean JTT distances, \(n_{ij} \), \(n_{ii} \), and \(n_{jj} \) are the total number of taxa used for each \(x_{avg} \) value, and \(\sigma^2_i \), \(\sigma^2_{ii} \), and \(\sigma^2_{jj} \) are the pooled inter- and intra-subfamily variances.\(^{22}\)

A \(z\)-value > 3.3 between two subfamilies shows that the difference between the subfamilies is statistically significant to \(p < 0.001\). If a \(z\)-value between two apparent subfamilies were < 3.3, different apparent subfamilies were chosen and/or individual sequences were removed, and the statistical calculations were repeated. Subfamilies were finally defined with a phylogenetic tree in which all \(z\)-values exceeded 3.3, sometime leaving some sequences outside any subfamily (i.e. non-grouped sequences).

All sequences within individual subfamilies were aligned using MUSCLE 3.6 and rooted phylogenetic trees were built in MEGA4 with the same tree and bootstrap parameters as described above. A few sequences from another subfamily (that with the highest \(z\)-value) were chosen to root individual subfamily’s trees.

**Results and discussion**

A subfamily must have at least five sequences from different species, and it must pass the statistical tests. Ten subfamilies were found (Figure 5.1) in which 326 sequences appear; in addition 34 sequences could not be grouped into a subfamily. All \(z\)-values are > 3.4, ranging from 3.41 to 29.7 (Table 5.1), and mean distances between different subfamilies are larger than those within subfamilies (Table 5.1). Individual trees of each subfamily appear in the Supplementary Information at the end of the chapter.

All 81 Subfamily A (SubA) members are produced by angiosperms, many encoded by various FatB genes, of which some have been previously experimentally characterized (Supplementary Information, Figure 5S-A). Several genes were chosen for experimental characterization: two *Iris* and two *Sorghum* genes, also one from *Ulmus* and one from *Cuphea* that have been previously characterized as controls. Three coconut genes previously
identified appear in this subfamily.

Of the 21 SubB members, most come from angiosperms with some from the moss species *Physcomitrella patens* (Figure 5S-B). No previous experimental work has been done on any sequence appearing in SubB. Three genes were experimentally characterized, two from *Sorghum* and one from *Physcomitrella*.

SubC has 32 members, all produced by angiosperms, and a significant number of them are encoded by FatA genes act on long-chain acyl-ACP molecules (Figure 5S-C). No genes were chosen for experimental characterization in this family, as many have already been characterized.

SubD has six algal sequences, with none having any previous experimental work, found in *Chlamydomonas, Ostreococcus*, and *Micromonas* (Figure 5S-D). One *Micromonas* sequence was selected for further experimental characterization.

The 17 sequences in SubE all come from Gram-negative bacteria: halophilic (*Salinibacter* and *Rhodothermus*), sulfate-reducing (*Desulfovibrio, Desulfohalobium*, and *Desulfonatronospira*), chemoorganotrophic (*Spirosoma*), metal-reducing (*Geobacter, Pelobacter*, and *Anaeromyxobacter*), and marine (*Microscilla*). (Figure 5S-E) No subfamily member has been experimentally characterized. One *Desulfovobrio* sequence was experimentally studied.

SubF consists of 24 sequences, mainly from *Bacteroides* but also from other related bacteria (Figure 5S-F). No member of this subfamily has been experimentally characterized, but PDB structure 2ESS (Figure 5.2), obtained from a structural genomic effort, is part of this subfamily. One *Bacteroides* and one *Parabacteroides* genes were experimentally characterized.

SubG has 31 sequences, primarily from *Clostridium* (Figure 5S-G). No subfamily G member had been subjected to experimental work. One *Clostridium* gene was experimentally characterized.

SubH has 27 bacterial sequences with no previous experimental information (Figure 5S-H). All members are from bacterial genera, mainly from *Clostridium*. One *Clostridium* gene and one *Bryantella* gene were chosen for experimental characterization.

SubI has eight bacterial sequences with no previous experimental information (Figure 5S-
Members come from various bacterial genera, mainly from *Clostridium*. One *Geobacillus* gene was experimentally studied.

SubJ has 79 members, all Gram-positive lactic acid bacteria, almost completely from the genera *Lactobacillus*, *Enterococcus*, and *Streptococcus* (Figure 5S-J). No previous experimental characterization has been done on any sequence in this family, but PDB structure 2OWN (Figure 5.2), obtained from a structural genomic effort, appears in this family. One *Streptococcus* gene and two *Lactobacillus* genes were experimentally characterized.

Some sequences were not grouped into any subfamily. These include two plant, four moss, and 28 bacterial sequences. They appear throughout the overall phylogenetic tree, with most more closely related to subfamilies SubF, SubG, SubH, SubI, and a few closely related to SubA, SubC, and SubE. No experimental work has been done in any of these sequences.

**Comments on published paper and conclusion**

Thirty-one genes were chosen based on this phylogenetic work, cloned in *E. coli*, and the *in vivo* substrate specificities found. Based on their specificities, these enzymes were clustered into three classes: 1) Class I acyl-ACP TEs act primarily on 14- and 16-carbon acyl-ACP substrates; 2) Class II acyl-ACP TEs have broad substrate specificities, with major activities toward 8- and 14-carbon acyl-ACP substrates; and 3) Class III acyl-ACP TEs act predominantly on 8-carbon acyl-ACPs. Several novel acyl-ACP TEs act on short-chain and unsaturated acyl-ACP or 3-ketoacyl-ACP substrates, indicating the diversity of enzymatic specificity in this enzyme family.

This study revealed that bacterial sequences provide access to additional functional diversity, both relative to acyl chain length specificity (shorter acyl chains, as short as four carbon atoms), as well as acyl chains that contain additional chemical functionalities (unsaturated acyl chains and acyl chains containing carbonyl groups). This additional functional diversity in acyl-ACP TEs can potentially be used to diversify the fatty acid biosynthesis pathway to produce biorenewable chemicals. Also, it is important to highlight that subfamily and substrate specificity did not correlate. Two very similar sequences in the same subfamily can show different substrate specificities, and vice versa.
Acknowledgments
The authors thank Professor Derrick Rollins for the equation to establish the statistical justification for separating the subfamilies.

References

### Table 5.1: Mean JTT distances and z values (bolded) within and between different subfamilies

<table>
<thead>
<tr>
<th></th>
<th>SubA</th>
<th>SubB</th>
<th>SubC</th>
<th>SubD</th>
<th>SubE</th>
<th>SubF</th>
<th>SubG</th>
<th>SubH</th>
<th>SubI</th>
<th>SubJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubA</td>
<td></td>
<td>0.36 ± 0.13&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubB</td>
<td>0.68 ± 0.13</td>
<td>0.58 ± 0.20</td>
<td>4.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubC</td>
<td>0.94 ± 0.10</td>
<td>0.97 ± 0.13</td>
<td>0.25 ± 0.10</td>
<td>25.37</td>
<td>10.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubD</td>
<td>1.40 ± 0.08</td>
<td>1.42 ± 0.11</td>
<td>1.29 ± 0.14</td>
<td>0.95 ± 0.30</td>
<td>6.14</td>
<td>5.03</td>
<td>5.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubE</td>
<td>1.80 ± 0.17</td>
<td>1.85 ± 0.19</td>
<td>1.76 ± 0.20</td>
<td>1.77 ± 0.18</td>
<td>9.28</td>
<td>7.67</td>
<td>9.10</td>
<td>3.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubF</td>
<td>2.09 ± 0.12</td>
<td>2.19 ± 0.20</td>
<td>2.00 ± 0.17</td>
<td>2.04 ± 0.16</td>
<td>1.98 ± 0.20</td>
<td>0.46 ± 0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubG</td>
<td>2.04 ± 0.19</td>
<td>2.08 ± 0.19</td>
<td>2.01 ± 0.17</td>
<td>2.13 ± 0.18</td>
<td>2.02 ± 0.25</td>
<td>2.05 ± 0.23</td>
<td>1.13 ± 0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubH</td>
<td>2.12 ± 0.13</td>
<td>2.23 ± 0.18</td>
<td>2.08 ± 0.14</td>
<td>2.12 ± 0.19</td>
<td>2.05 ± 0.24</td>
<td>2.03 ± 0.22</td>
<td>1.81 ± 0.26</td>
<td>0.96 ± 0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubI</td>
<td>2.24 ± 0.16</td>
<td>2.32 ± 0.17</td>
<td>2.28 ± 0.19</td>
<td>2.37 ± 0.22</td>
<td>2.28 ± 0.26</td>
<td>2.37 ± 0.23</td>
<td>2.22 ± 0.26</td>
<td>2.38 ± 0.26</td>
<td>1.95 ± 0.34</td>
<td></td>
</tr>
<tr>
<td>SubJ</td>
<td>2.21 ± 0.13</td>
<td>2.24 ± 0.15</td>
<td>2.23 ± 0.15</td>
<td>2.36 ± 0.20</td>
<td>2.20 ± 0.22</td>
<td>2.16 ± 0.17</td>
<td>1.91 ± 0.17</td>
<td>2.19 ± 0.25</td>
<td>2.26 ± 0.21</td>
<td>1.23 ± 0.40</td>
</tr>
</tbody>
</table>

<sup>a</sup>Standard deviation
Figures

Figure 5.1: Unrooted phylogenetic tree of TE14. Stars denote sequences to be experimentally characterized.

Figure 5.2: Superimposed PDB structures 2OWN (red) from Lactobacillus plantarum (SubJ) and 2ESS (blue) from Bacteroides thetaiotaomicron (SubF).
Supplementary Information for Chapter 5
Phylogenetic characterization of an acyl-ACP thioesterase family

Figure 5S-A: Rooted phylogenetic tree of SubA. Black diamonds mark genes that were synthesized for experimental characterization, and black circles mark sequences isolated.

Figure 5S-B: Rooted phylogenetic tree of SubB. Black diamonds mark genes that were synthesized for experimental characterization.

Figure 5S-C: Rooted phylogenetic tree of SubC.

Figure 5S-D: Rooted phylogenetic tree of SubD. Black diamonds mark genes that were synthesized for experimental characterization.

Figure 5S-E: Rooted phylogenetic tree of SubE. Black diamonds mark genes that were synthesized for experimental characterization.

Figure 5S-F: Rooted phylogenetic tree of SubF. Black diamonds mark genes that were synthesized for experimental characterization, and the black square marks a sequence with a known PDB structure.

Figure 5S-G: Rooted phylogenetic tree of SubG. Black diamonds mark genes that were synthesized for experimental characterization.

Figure 5S-H: Rooted phylogenetic tree of SubH. Black diamonds mark genes that were synthesized for experimental characterization.

Figure 5S-I: Rooted phylogenetic tree of SubI. Black diamonds mark genes that were synthesized for experimental characterization.

Figure 5S-J: Rooted phylogenetic tree of SubJ. Black diamonds mark genes that were synthesized for experimental characterization, and the black square marks a sequence with a known PDB structure.
Figure 5S-B

Figure 5S-C
Figure 5S-D

Figure 5S-E

Figure 5S-F
CHAPTER 6
INVESTIGATION OF A THIOESTERASE MECHANISM BY QM/MM METADYNAMICS SIMULATION

David C. Cantu¹, Albert Ardevol², Carme Rovira², and Peter J. Reilly¹

Department of Chemical and Biological Engineering¹, Iowa State University, and Computer Simulation and Modeling Laboratory², Parc Científic de Barcelona

Modified from a manuscript submitted to the Journal of Physical Chemistry B.

Abstract

Many thioesterases have a structural HotDog fold. Based on crystal structures, many putative mechanisms have been proposed for HotDog fold thioesterases. The reaction of the human thioesterase enzyme (hTHEM2) is explored by first-principles methods to elucidate atomic and electronic details of the mechanism, transition-state conformation, and energy of the process. Mixed quantum mechanics/molecular mechanics simulations using the metadynamics technique are used. An acid-base-like mechanism is found. The activation of nucleophilic water by an aspartate residue acting as a base, as well as the tetrahedral-like intermediate, are confirmed as previously proposed. Additionally, new evidence pointing toward thioester substrate protonation from a histidine residue via a serine residue is presented. To our knowledge, this is the first time a thioesterase has been studied by first-principles methods.

Introduction

Thioesterases (TEs) (EC 3.1.2.–) hydrolyze the thioester bond between a sulfur atom and a carbonyl group. At present TEs are classified into 25 families¹ by statistically significant differences in amino acid sequence, and are present in the ThYme database.² Families with strong structural similarities may be grouped into clans. Most TEs hydrolyze acyl-acyl carrier protein (ACP) or acyl-coenzyme A (CoA) molecules. Acyl-ACP TEs mainly occur in fatty acid synthesis, while acyl-CoA TEs are found in numerous biochemical pathways. Most acyl-ACP and acyl-CoA TEs have one of the two major structural folds: the α/β hydrolase
fold, and the HotDog fold. The mechanism of α/β hydrolase-fold TEs has been resolved; it is very similar to that of serine proteases that employ a serine/histidine/aspartate catalytic triad that forms a covalent intermediate with the substrate that is hydrolyzed by water.

HotDog-fold enzymes lack defined binding pockets and conserved catalytic residues. A variety of catalytic residues and mechanisms have been proposed for TEs with this structural fold. Given the sequence and structural variety displayed and the wide array of substrates that HotDog-fold TEs act upon, several mechanisms exist in different TEs of this type. They can be roughly divided into two main groups, Clan TE-A and Clan TE-B by structure and active-site geometry Arrangement Ψ and Arrangement R. In most cases, Clan TE-A TEs display an active-site Arrangement Ψ, while Clan TE-B TEs have an active-site Arrangement R. There are other HotDog TEs that do not fall into these two clans, such as acyl-ACP TEs of type II fatty acid synthesis, as well as CalE7 TEs, and fluoroacyl-CoA TEs.

The 4-hydroxybenzoyl-CoA thioesterase 1 (4HBT1) from Family TE10 shows the mechanism present in Clan TE-A enzymes. In 4HBT1, an aspartate residue attacks the thioester carbon atom, forming an anhydride acyl-enzyme intermediate. Then the cleaved CoASH group activates a water nucleophile that hydrolyzes the intermediate, the second step being rate-limiting.

Several mechanisms for Clan TE-B enzymes have been proposed. Two mechanisms have been found for the 4-hydroxybenzoyl-CoA thioesterase 2 (4HBT2) enzyme, present in TE11. In one mechanism, a glutamate residue acts as a nucleophile that attacks the thioester carbon atom in the substrate, forming a substrate-enzyme intermediate. This is followed by hydrolysis by a water molecule activated by a threonine residue, while glutamine and glycine residues stabilize the thioester. Interestingly, in the second proposed mechanism, when the glutamate residue is substituted by an aspartate, the aspartate acts as a base that activates a water nucleophile that attacks and hydrolyzes the thioester carbon atom. The proposed mechanisms of TE8 and TE13 are similar to the second proposed mechanism of TE11. In thioesterase PaaI, present in TE13, it has been proposed that an aspartate residue acts as a base, activating a water molecule that acts as a nucleophile attacking the thioester carbon atom.

In the human TE superfamily member 2 (hTHEM2) enzyme, from TE8, the proposed
mechanism is a nucleophilic attack by a water molecule activated by Asp65, and oriented by Asp65 and Ser83, on the thioester carbon atom.\textsuperscript{14} A tetrahedral-like intermediate is formed where Asn50 and Gly57 polarize the thioester, and Asn50 assists the leaving thiolate group. No residue was proposed to protonate the thioester substrate’s sulfur. The physiological function of hTHEM2 is unknown; however, it acts on a range of acyl-CoA substrates \textit{in vitro}.\textsuperscript{14} Mutagenesis studies reveal that only mutating Asp65 stops enzymatic activity completely. Other active-site mutations reduce activity significantly.\textsuperscript{14} The enzyme and active site are shown in Figure 6.1.

This work is an effort to further study HotDog-fold TE mechanisms. The hTHEM2 enzyme in TE8 was chosen because its Protein Data Bank (PDB) structure 3F5O\textsuperscript{14} was crystallized with an inhibitor (undecan-2-one-CoA, in which O=C-CH\textsubscript{2}-S substitutes for thioester O=C-S) in the active site, which allows for first-principles molecular simulation. The reaction of the hTHEM2 enzyme is explored by first-principles methods to elucidate atomic and electronic details of the mechanism, transition-state conformation, and energy of the process. The activation of a nucleophilic water by Asp65, acting as a base, as well as the tetrahedral intermediate, are confirmed. Additionally, new evidence pointing toward thioester substrate protonation from a histidine residue via a serine residue is presented. To our knowledge, no TE reaction has been studied by first-principles methods. Procedures closely follow previous work done with hydrolytic enzymes.\textsuperscript{15-18}

\textbf{Computational methods}

\textit{Classical Molecular Dynamics}

Molecular dynamics (MD) simulations were done on crystal structure 3F5O\textsuperscript{14} with hexanoyl-CoA as the substrate. Chains A and B of 3F5O were taken from the PDB with the undecan-2-one-CoA inhibitor removed. Protonation states of histidine residues (HIP, double-protonated; HID, N\textsubscript{D} protonated; or HIE, N\textsubscript{E} protonated) were chosen based on the surrounding environment and the potential hydrogen bonds with neighboring residues. Amino acid residues were modeled with the FF99SB Amber\textsuperscript{19} force field. Active-site crystallographic water molecules were kept, and the enzyme was solvated by adding a water box with a 12-Å cushion around the protein surface. Water molecules were treated using the
TIP3P force field. Chlorine ions were added to neutralize the charge. The thioester substrate, hexanoyl-CoA, was prepared by editing the inhibitor undecan-2-one-CoA molecules using MacMolPlt software. It was placed in the enzyme active site keeping crystallographic CoA coordinates. The antechamber module in Amber and the Generalized Amber Force Field with RESP charges were used to obtain substrate parameters. Topology and coordinate files of the neutralized and solvated substrate-enzyme system were created for MD in Amber11.

The system was minimized in steps. First the enzyme and substrates were restrained and water and ions were allowed to move freely, followed by restraining only the substrate and α-carbon atoms of enzyme residues, followed by minimization with no restrictions. The system was then taken slowly over 100 ps to 300K with substrate and enzyme restrained. Then the system was simulated for 1 ns with active-site atom restraints. The structure and simulation were analyzed with the Visual Molecular Dynamics (VMD) package. The final frame was chosen as the starting geometry for QM/MM simulations.

**QM/MM Molecular Dynamics**

The method developed by Laio and coworkers that combines first-principles Car-Parrinello MD (CPMD) method with classical MD was used to perform quantum mechanics/molecular mechanics (QM/MM) simulations. It couples a QM region of the system, whose atomic dynamics are ruled by electron density computed with density functional theory, with an MM region whose atomic dynamics are described by an empirical force field. The QM/MM interface is modeled by link-atom pseudopotentials that saturate the QM region. The electrostatic interactions between QM and MM regions were treated with a Hamiltonian coupling scheme where short-range electrostatic interactions between QM and MM regions were explicitly taken into account. A modified Coulomb potential ensures no unphysical escape of the electronic density from the QM to the MM region. Distant electrostatic interactions are treated via a multipole expansion.

The QM region (Figure 6.2) includes the side chains of active-site residues Asp65, Ser83, and His134 from chain A, Asn50 from chain B, the nucleophilic water molecule, and a section of the hexanoyl-CoA thioester substrate (COH). Amino acid side chains were capped
at carbon atoms with a link atom: Asn50, Asp65, and Ser83 at their α-carbon atoms and His134 at its β-carbon atom. The substrate was capped at the C4 atom of the hexanoyl-CoA molecule and at the second carbon atom from the thioester sulfur of the CoA group. The QM region was enclosed in a 27-Å x 40-Å x 27-Å cell. The enzyme was studied in the three possible protonation states of His134. The HIP134 system included a total of 49 atoms in the QM region with a net zero charge, and for the HID134 system 48 atoms and a –1 charge. The HIE134 case was not considered after classical MD (Results section). Kohn-Sham orbitals were expanded in a plane-wave basis set with a kinetic energy cutoff of 70 Ry. Ab initio pseudopotentials generated by the Troullier-Martins scheme were used, as was the Perdew-Burke-Ernzerhoff functional in the generalized gradient-correction of density functional theory.

CPMD simulations used a time step of 0.145 fs (6 atomic units of time) and a fictitious electron mass of 700 atomic units. The system reached a constant temperature of 300K by coupling it to a Nosé-Hoover thermostat of 3500 cm⁻¹ frequency. Running QM/MM MD while annealing the ionic velocities until the maximal component of the nuclear gradient was <10⁻² optimized the geometry. Active-site distance constraints were placed, and the system was simulated for 0.87 ps, or 6,000 time steps, of QM/MM MD. The final QM/MM MD frame is the starting geometry for metadynamics QM/MM simulations. The CPMD program (CPMD–Program) was used for all QM/MM simulations.

Metadynamics

The metadynamics technique was used to induce the enzymatic reaction and to reconstruct the free energy surface (FES) of the process. Metadynamics is used to simulate rare events by reducing the dimensions of a system, and it has been used to solve problems in chemistry, materials science, and biophysics. Collective variables (CVs) are chosen, and repulsive energy terms (Gaussian-like potentials) are added in CV space of the system explored by MD. This allows the system to escape free energy minima and to explore non-equilibrium conformations. CVs need to clearly describe the event desired to be explored and to differentiate between initial and final states. Further, the FES of the simulated event as a function of CV space can be reconstructed by summing the Gaussian-like repulsive terms.
added during the simulation. Metadynamics was first used to simulate an enzymatic reaction in 2006.33

Coordination numbers (CNs) indicate the presence of a covalent bond. The CN between atoms a and b is given by:

\[
CN_{ab} = \frac{1 - \left(\frac{d_{ab}}{d^*}\right)^p}{1 - \left(\frac{d_{ab}}{d^*}\right)^{p+q}}
\]  

(1)

where \(d_{ab}\) is the distance between atoms a and b, \(d^*\) is the threshold distance of bonding, and \(p\) and \(q\) are exponents that determine the curve of CN as a function of \(d_{ab}\). \(CN_{ab}\) values range from 0 (no bond) to 1 (bond present).

The CVs chosen in this work are differences of CNs between two bonds (Figure 6.3).34 CVnucl describes the nucleophilic attack of the water molecules, being the difference between the CN of the thioester carbon atom and nucleophilic water molecule and the CN between the thioester carbon and sulfur atoms (Equation 2). CVprot-ser describes the protonation of the thioester from active-site serine residue, being the difference between the CN of the thioester sulfur and the Ser83 hydrogen atom and the CN between the Ser83 hydrogen and oxygen atoms (Equation 3). Similarly, CVprot-asn describes the protonation of the thioester sulfur atom by the active-site Asn50 residue, being the difference between the CN of the thioester sulfur and the Asn50 hydrogen atoms and the CN between the Asn50 hydrogen and nitrogen atoms (Equation 4). The values of all CVs range from –1 in the reactant side of the reaction to +1 in the product side of the reaction.

\[
CV_{nucl} = CN(Owat-CCoh) - CN(Ccoh-Scoh)
\]  

(2)

\[
CV_{protser} = CN(Scoh-Hser) - CN(Hser-Oser)
\]  

(3)

\[
CV_{protasn} = CN(Scoh-Hasn) - CN(Hasn-Nasn)
\]  

(4)

Metadynamics simulations were done with the Car-Parrinello approach where the Lagrangian includes extra terms that describe the fictitious dynamics of the CVs.34 These are coupled by a harmonic potential to the real CV values. Mass and force constants of the harmonic potential were chosen to ensure the best agreement between real and fictitious particles. A mass constant of 20 amu and force constant of 7 au were chosen. The height of
the repulsive Gaussian-like terms used was either 0.005 or 0.002 Hartree atomic units of energy (3 or 1.25 kcal/mol), and the width 0.08 or 0.05 Å. Gaussian-like potentials were added every 200 MD steps.

Results and discussion

Classical and QM/MM Molecular Dynamics

Due to its active-site location, all possible protonation states of His134 were explored. The classical MD protocol was performed independently, without restraints in the equilibration nanosecond, on the system with HID134, HIE134, and HIP134 protonation states. Crystallographic His134 coordinates were compared to the His134 coordinates after classical MD (Supplementary Information, Figure 6S-A), and the RMSD between corresponding non-hydrogen atoms was calculated, yielding 0.78 Å for the HID134 system, 1.45 Å for the HIE134 system, and 0.85 Å for the HIP134 system. The histidine residue side chain changed conformation significantly in the HIE134 case, but not so for HID134 and HIP134. The HIE case therefore was not explored further.

The system and active-site coordinates were analyzed after performing the classical MD and QM/MM MD protocols on the HID134 and HIP134 systems. The final frames after classical MD and after QM/MM MD were compared to crystallographic coordinates. The RMSD values of active-site residues and the whole enzyme between crystallographic coordinates corresponding with final classical MD and QM/MM MD coordinates are shown in Table 6.1 for both the HID134 and HIP134 systems. Also, active-site residues of the final frames from classical MD and QM/MM MD are shown with their corresponding crystal-structure residues superimposed (Figure 6S-B). Low RMSD values and similar structures reveal the final classical MD frame as a good starting frame for QM/MM MD, and the final QM/MM MD frame as a good starting geometry for QM/MM metadynamics simulations.

QM/MM Metadynamics: Reaction Pathway and Free Energy Surface

QM/MM metadynamics simulations on the HIP134 system used $CV_{nucl}$ and $CV_{prot-ser}$ as collective variables, while the HID134 system used $CV_{nucl}$ and $CV_{prot-asn}$ to explore the reaction. The HIP134 system simulation achieved enzymatic reaction (discussed in this
section), while the HID134 system simulation did not (next section).

The QM/MM metadynamics simulation of the HIP134 system ran for 8.7 ps (60,000 time steps). The values of $CV_{nucl}$ and $CV_{prot-ser}$ over the simulation time can be seen in Figure 6.4A. Relevant bond lengths are also shown as a function of simulation time. Figure 6.4B shows two distances, $O_{asp-H_{wat}}$ and $H_{wat-O_{wat}}$, that represent the activation of the water nucleophile by Asp65 acting as a base. Figure 6.4C shows the distance between the nucleophilic water oxygen atom to the thioester carbon atom, $O_{wat-C_{coh}}$, and the thioester bond length, $C_{coh-S_{coh}}$, representing thioester hydrolysis. In Figure 6.4D the distance between the thioester sulfur atom and the Ser83 hydrogen atom, $S_{coh-H_{ser}}$, and the length between the Ser83 hydrogen and oxygen atoms, $H_{ser-O_{ser}}$, appear, showing substrate protonation. Figure 6.4E shows the distances between the Ser83 oxygen atom and the HIP134 hydrogen atom, $O_{ser-H_{hip}}$, and between the HIP134 hydrogen and nitrogen atoms, $H_{hip-N_{hip}}$, which represent the protonation of Ser83 by HIP134.

The reactant state was explored during the first 1 ps of simulation time. The values of the CVs remain between $-1$ and $-0.05$, and the water molecule O–H bond, the Ser83 O–H bond, and the HIP134 N–H bond oscillate around their average bond lengths. Between 1 ps and 3 ps of simulation time, $CV_{nucl}$ approaches zero, while $CV_{prot-ser}$ remains close to $-1$. This represents the nucleophilic water molecule approaching the thioester carbon atom, and the thioester bond length increasing (Figure 6.4C). During the same time, the Ser83 O–H bond length does not increase significantly, nor does the Ser83 hydrogen atom approach the thioester sulfur atom (Figure 6.4D).

The enzymatic reaction occurs near 3 ps of simulation time. The value of $CV_{nucl}$ first changes from ~0 to ~1, and quickly afterward $CV_{prot-ser}$ changes from ~1 to ~1. This shows that the nucleophilic attack occurs first, quickly followed by protonation of the thioester sulfur atom by the Ser83 hydrogen atom, completing hydrolysis of the thioester substrate. After reaction occurred, Asp65 became aspartic acid as it took a hydrogen from water during reaction (Figure 6.4B), the thioester bond broke and the fatty acid product was made (Figure 6.4C). This is irreversible, represented by the value of $CV_{nucl}$ that remains near 1 for the remainder of the simulation. However, the value of $CV_{prot-ser}$ changes after the reaction. After the thioester bond breaks, the CoASH product switches between its thiol and thiolate
forms, sharing a hydrogen atom with Ser83. A serine residue will not readily donate a proton when close to neutral pH, and for that to occur, it has to receive a proton as well. At the same time that Ser83 protonates the thioester, it receives the HIP134 hydrogen atom.

By summing all the Gaussian-like repulsive energy terms added during the metadynamics simulation in CV space, the free energy surface (FES) of the system as a function of CVnucl and CVprot–ser was reconstructed. Figure 6.5 shows the FES of the system, with its main reaction path states highlighted. From the FES, the energy barrier of the reaction is estimated to be between 20 and 25 kcal/mol. Based on the $k_{cat}$ values of hTHEM2 at two temperatures, the reaction activation energy was calculated to be 19.14 kcal/mol, showing good agreement between experimental and simulation estimates of the TE reaction energy barrier.

The reactant state (R), intermediate reactant state (R’), transition state (TS), intermediate product state (P’), and product state (P) can be distinguished in the FES. The simulation trajectory starts in R, passes to R’ with some recrossing between them, then explores TS, and falls into the product well after first passing through P’. The simulation trajectory oscillates between the P and P’ states (Figure 6.4A, CVprot–ser), but there is no recrossing the TS (Figure 6.4A, CVnucl). The product well is considerably deeper than the ~45 kcal/mol shown in Figure 6.5, as the simulation was stopped before reaching the bottom of the well, since the fatty acid and CoASH products would not reform back into the thioester substrate. The summation of Gaussian-like repulsive energy terms in the P state corresponds to ~45 kcal/mol.

The geometries of each reaction pathway state appear in Figures 6.6 and 6.7, relevant bond length values are found in Table 6.2, and the electrostatic potential charges of relevant atoms are in Table 6.3. The simulation first explored R and R’ where no reaction occurred. R’ differs from the reactants in that the water molecule comes closer to the thioester (Owat–Ccoh, Table 6.2) and an increase in thioester bond length occurs (Ccoh–Scoh, Table 6.2). Otherwise, the two states have roughly equal distances between the water molecule and Asp65, the thioester to Ser83, and Ser83 to HIP134 (Table 6.2). The electrostatic potential charges of relevant atoms do not differ much between R and R’ except for that of Scoh, which decreases as the thioester bond length increases.
In the TS (Figure 6.7), Asp65 takes a proton from the water molecule (Oasp–Hwat, Hwat–Owat, Table 6.2), the activated nucleophilic water molecule comes in close contact with the thioester carbon atom (Owat–Ccoh, Table 6.2), and the thioester bond length increases. A tetrahedral-like geometry is observed (Figure 6.7B) between Ccoh, Ocoh, Scoh, C2coh, and Owat. Gly57 helps to stabilize the tetrahedral-like intermediate, where Ocoh comes within hydrogen-bonding distance (2.32 Å) to the Gly57 hydrogen atom during the TS. Also in the TS, protonation is set to happen as the Ser83 hydrogen atom approaches Scoh, and the HIP134 hydrogen atom approaches the Ser83 oxygen atom. The electrostatic charges of relevant atoms change at the TS. The Oasp atom goes from negative to neutrally charged as it accepts a proton from the water molecule. The Owat negative charge decreases as it becomes hydroxide ion-like and approaches the thioester carbon atom, forming part of the tetrahedral-like intermediate seen in the TS. The charge on the thioester carbon atom, Ccoh, decreases slightly, and the thioester oxygen atom charge changes from close to zero to negative. The thioester sulfur atom charge decreases again as the thioester bond breaks. The Ser83 oxygen atom charge becomes less negative as its hydrogen atom approaches the thioester group and accepts a proton from HIP134.

In P and P’, Asp65 has accepted a proton from the water molecule, the thioester substrate has been hydrolyzed, and the fatty acid is formed. P and P’ differ in the protonation state of the CoASH product. In P’, Ser83 still holds its hydrogen atom while accepting the HIP134 hydrogen atom, and a thiolate product CoAS– appears. In P, the thiolate has accepted the Ser83 hydrogen and has become a thiol (CoASH), Ser83 has taken the HIP134 hydrogen atom, and HIP134 has lost a hydrogen atom, changing to its HID protonation state. The relevant atomic charges do not differ much between the P and P’ states except for Scoh, which is more negative in P’ as part of a thiolate group. The Ccoh charge has become more positive in P and P’, and the HIP134 charge has become negative, as it has lost its hydrogen atom to Ser83.

Non-reactive HID134 System

The QM/MM metadynamics simulation of the HID134 system ran for 12.18 ps (84,000 time steps). The values of \( CV_{nucl} \) and \( CV_{prot–asn} \) over the simulation time can be seen in
Figure 6S-C and the FES in Figure 6S-D. No reaction was observed in this metadynamics simulation. The simulation only explored the reactants, as the $CV_{nucl}$ and $CV_{prot-asn}$ values never became positive, and the FES was only constructed in the reactant state, where all repulsive Gaussian-like terms were added. The simulation was stopped when the energy barrier to escape the reactant state exceeded 100 kcal/mol.

The HID134 system did not include $CV_{prot-ser}$, as a serine residue will not readily give up a proton at near-neutral pHs without receiving an additional proton beforehand. The only residue in a position to protonate Ser83 is His134, and that cannot occur in its HID protonation state. Likewise, it is unlikely that Asn50 can give a proton at near-neutral pHs without receiving an additional proton. There are no residues in the enzyme that can donate a proton to the Asn50 C=O group, so that it can donate a proton from its NH$_2$ group to the thioester. Figure 6S-E shows residues that appear within proton-donating distance of Asn50 in the reactant state of the HID134 system: the C=O in the Asn50 side chain forms hydrogen bonds with backbone N-H groups from Ile52 and Thr54. Although not catalytic, Asn50 plays an important active-site role by positioning the substrate and stabilizing the TS tetrahedral-like intermediate (an Ocoh–Hasn distance of 3.27 Å).

**Conclusions**

Based on previous work$^{14}$ and the metadynamics QM/MM simulations performed in this study, a proposed mechanism of hTHEM2 can be seen in Figure 6.8. Asp65 acts as a base that activates a water molecule that performs a nucleophilic attack on the thioester carbon atom, hydrolyzing the substrate. HIP134 acts as an acid and donates a proton to Ser83, which in turn protonates the substrate thioester sulfur atom. In the TS, the thioester and nucleophilic water molecule form a tetrahedral-like intermediate while HIP134 starts to donate a proton to Ser83, which approaches the thioester sulfur atom. Asn50 and Gly57 play a role in positioning the substrate and stabilizing the TS tetrahedral-like intermediate. This study supports with first-principles calculations the basic character of Asp65, the hydrolysis of the thioester by a nucleophilic attack on the thioester carbon by an activated water molecule, and the proposed tetrahedral-like geometry of the TS. Additionally, new evidence is given that suggests that HIP134 protonates the thioester via Ser83.
Two main mechanisms have been proposed for HotDog-fold TEs: one where aspartate or glutamate residues act as a nucleophile, forming an enzyme-substrate intermediate that is hydrolyzed by a water molecule, and the second a one-step acid-base-like catalytic mechanism as proposed in this work for hTHEM2. Although only the first mechanism has been observed in Clan TE-A enzymes, both appear in Clan TE-B enzymes. It is likely that all HotDog-fold TEs employ either or both mechanisms for catalysis, unlike α/β hydrolase-fold TEs that appear to always use a serine protease-like mechanism.

Acknowledgments

The authors thank the Iowa State University High Performance Computation System for making its facilities available to us.

References

Tables

Table 6.1: RMSDs (Å) of Final Classical and QM/MM MD Frames Compared to Crystallographic Coordinates

<table>
<thead>
<tr>
<th>Residue(s)</th>
<th>HID classical MD</th>
<th>HID QM/MM MD</th>
<th>HIP classical MD</th>
<th>HIP QM/MM MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full enzyme</td>
<td>0.75</td>
<td>0.72</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>Asn50</td>
<td>0.68</td>
<td>0.75</td>
<td>0.80</td>
<td>0.65</td>
</tr>
<tr>
<td>Asp65</td>
<td>0.87</td>
<td>1.00</td>
<td>0.91</td>
<td>0.56</td>
</tr>
<tr>
<td>Ser83</td>
<td>0.91</td>
<td>1.04</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>His134</td>
<td>0.52</td>
<td>0.51</td>
<td>0.61</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 6.2: Bond Lengths (Å) along Reaction Pathway

<table>
<thead>
<tr>
<th>Bond</th>
<th>R</th>
<th>R’</th>
<th>TS</th>
<th>P’</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oasp–Hwat</td>
<td>1.61 ± 0.10</td>
<td>1.60 ± 0.11</td>
<td>1.10 ± 0.06</td>
<td>0.99 ± 0.01</td>
<td>0.99 ± 0.01</td>
</tr>
<tr>
<td>Hwat–Owat</td>
<td>1.03 ± 0.03</td>
<td>1.04 ± 0.03</td>
<td>1.41 ± 0.17</td>
<td>4.58 ± 0.52</td>
<td>4.48 ± 0.52</td>
</tr>
<tr>
<td>Owat–Ccoh</td>
<td>3.16 ± 0.17</td>
<td>2.82 ± 0.49</td>
<td>1.81 ± 0.09</td>
<td>1.35 ± 0.03</td>
<td>1.35 ± 0.03</td>
</tr>
<tr>
<td>Ccoh–Scoh</td>
<td>1.81 ± 0.10</td>
<td>2.11 ± 0.08</td>
<td>2.26 ± 0.04</td>
<td>4.62 ± 0.40</td>
<td>4.56 ± 0.33</td>
</tr>
<tr>
<td>Scoh–Hser</td>
<td>2.52 ± 0.16</td>
<td>2.33 ± 0.31</td>
<td>1.82 ± 0.07</td>
<td>1.58 ± 0.04</td>
<td>1.37 ± 0.07</td>
</tr>
<tr>
<td>Hser–Oser</td>
<td>1.03 ± 0.07</td>
<td>1.04 ± 0.07</td>
<td>1.09 ± 0.05</td>
<td>1.33 ± 0.05</td>
<td>2.04 ± 0.29</td>
</tr>
<tr>
<td>Oser–Hhip</td>
<td>1.78 ± 0.13</td>
<td>1.74 ± 0.14</td>
<td>1.34 ± 0.05</td>
<td>1.06 ± 0.04</td>
<td>1.02 ± 0.04</td>
</tr>
<tr>
<td>Hhip–Nhip</td>
<td>1.05 ± 0.03</td>
<td>1.06 ± 0.05</td>
<td>1.15 ± 0.05</td>
<td>1.60 ± 0.14</td>
<td>1.74 ± 0.16</td>
</tr>
</tbody>
</table>

Table 6.3: Electrostatic Potential Charges of Relevant Atoms along Reaction Pathway

<table>
<thead>
<tr>
<th>Atom</th>
<th>R</th>
<th>R’</th>
<th>TS</th>
<th>P’</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oasp</td>
<td>-0.15 ± 0.04</td>
<td>-0.13 ± 0.04</td>
<td>0.01 ± 0.02</td>
<td>0.00 ± 0.03</td>
<td>0.02 ± 0.03</td>
</tr>
<tr>
<td>Owat</td>
<td>-0.28 ± 0.04</td>
<td>-0.24 ± 0.05</td>
<td>-0.13 ± 0.04</td>
<td>0.00 ± 0.03</td>
<td>0.01 ± 0.04</td>
</tr>
<tr>
<td>Ccoh</td>
<td>0.07 ± 0.02</td>
<td>0.08 ± 0.02</td>
<td>0.04 ± 0.02</td>
<td>0.19 ± 0.03</td>
<td>0.19 ± 0.03</td>
</tr>
<tr>
<td>Ocoh</td>
<td>-0.01 ± 0.02</td>
<td>0.01 ± 0.04</td>
<td>-0.12 ± 0.02</td>
<td>-0.14 ± 0.12</td>
<td>-0.17 ± 0.10</td>
</tr>
<tr>
<td>Scoh</td>
<td>-0.05 ± 0.08</td>
<td>-0.26 ± 0.08</td>
<td>-0.43 ± 0.04</td>
<td>-0.65 ± 0.06</td>
<td>-0.45 ± 0.06</td>
</tr>
<tr>
<td>Oser</td>
<td>-0.16 ± 0.04</td>
<td>-0.14 ± 0.05</td>
<td>-0.05 ± 0.02</td>
<td>0.04 ± 0.03</td>
<td>-0.06 ± 0.06</td>
</tr>
<tr>
<td>Nhip</td>
<td>0.05 ± 0.02</td>
<td>0.06 ± 0.04</td>
<td>0.05 ± 0.01</td>
<td>-0.10 ± 0.04</td>
<td>-0.12 ± 0.05</td>
</tr>
</tbody>
</table>
Figures

Figure 6.1: A) Overall structure with inhibitor shown, and B) active-site residues of TE8 human THEM2 thioesterase (PDB 3F5O)

Figure 6.2: QM region. Active site residues in blue, and section of substrate in purple. Relevant atoms labeled in notation used throughout

Figure 6.3: Collective variables displayed in active site
Figure 6.4: Variation of CVs (A) and bond lengths (B–E) during TE8 human THEM2-catalyzed hydrolysis of the thioester substrate. X-axes represent simulation time in picoseconds. Y-axis in A represents value of CVs. Y-axis in B–E represents distance in Å.
Figure 6.5: Free energy surface of TE8 human THEM2-catalyzed hydrolysis of the thioester substrate. Contour lines are separated by 2.5-kcal/mol intervals.

Figure 6.6: Positions of catalytic residues, substrate, and water during the reaction pathway of TE8 human THEM2-catalyzed hydrolysis of the thioester substrate.
Figure 6.7: A) Transition-state positions of active-site residues, substrate, and water, and B) tetrahedral-like structure geometry of water and substrate during TE8 human THEM2-catalyzed hydrolysis of the thioester substrate.

Figure 6.8: Proposed mechanism of the TE8 human THEM2-catalyzed hydrolysis of the thioester bond in an acyl-CoA substrate.
Supplementary Information for Chapter 6
Mechanism of a HotDog-fold acyl-CoA thioesterase proposed by QM/MM metadynamics simulation

Figure 6S-A

Figure 6S-B

Figure 6S-C
Figure 6S-D

Figure 6S-E
CHAPTER 7
STRUCTURAL CLASSIFICATION AND PROPERTIES OF
KETOACYL REDUCTASES, HYDROXYACYL DEHYDRATASES,
AND ENOYL REDUCTASES

David C. Cantu, Tingsong Dai, Zachary S. Beversdorf, and Peter J. Reilly
Department of Chemical and Biological Engineering, Iowa State University

Modified from a paper published in Protein Engineering Design & Selection, 25, 803-811, with the same title and authors.

Abstract
Ketoacyl reductases (KRs), hydroxyacyl dehydratases (HDs), and enoyl reductases (ERs) are part of the fatty acid and polyketide synthesis cycles. Their reverse reactions, catalyzed by acyl dehydrogenases (equivalent to ERs), enoyl hydratases (equivalent to HDs), and hydroxyacyl dehydrogenases (equivalent to KRs), are part of fatty acid degradation by β-oxidation. These enzymes have been classified into families based on similarities in their primary and tertiary structures, and these families and structures are included in the ThYme (Thioester-active enzYmes) database. Members of each family have strong sequence similarity and have essentially the same tertiary structure, mechanism, and catalytic residues.

Introduction
Fatty acids and polyketides are a structurally and functionally diverse family of chemicals abundant throughout nature. In producing fatty acids, six biosynthetic steps, catalyzed by acyl-coenzyme A (CoA) synthases, acyl-CoA carboxylases, ketoacyl synthases, ketoacyl reductases (KRs), hydroxyacyl dehydratases (HDs), and enoyl reductases (ERs), elongate acyl chains by two carbon atoms per cycle until cleavage of CoA or acyl carrier protein (ACP) activating agents by thioesterases stops the process (Figure 3.1). This process can be catalyzed by a single multimodular enzyme, fatty acid synthase (FAS) in type I fatty acid synthesis, or by individual enzymes catalyzing each reaction in type II fatty acid synthesis.

In polyketide synthesis, the KR-, HD-, and ER-catalyzed steps do not have to occur at
every turn, thus leaving keto or hydroxy groups or double bonds in the acyl chain. Also, substrates other than malonyl-CoA can elongate the acyl chain, adding different functional groups to it. The polyketide synthesis cycle is terminated by hydrolysis of the thioester bond between the chain and ACP or CoA, either with water or with a hydroxy group in the chain, the latter cyclizing the molecule. These two cycles have been proposed as a potential source for industrial and specialty biorenewable chemicals.¹

Fatty acids are broken down by β-oxidation, where the KR-, HD-, and ER-catalyzed steps in the fatty acid synthesis cycle operate in reverse, with these enzymes given names characteristic of the reverse reactions. Oxidation of an acyl-CoA molecule to give 2-enoyl-CoA is catalyzed by acyl-CoA dehydrogenase (equivalent to ER) with a flavin adenine dinucleotide (FAD) prosthetic group. Hydration of 2-enoyl-CoA is catalyzed by 2-enoyl hydratase (equivalent to HD), yielding 3-hydroxyacyl-CoA. The third step is oxidation of 3-hydroxyacyl-CoA by catalysis with L-3-hydroxyacyl-CoA dehydrogenase (equivalent to KR) using nicotinamide adenine dinucleotide (NAD⁺) to give 3-ketoacyl-CoA. A final step, catalyzed by 3-ketoacyl-CoA thiolase and involving addition of CoA, removes an acetyl-CoA molecule and leaves an acyl-CoA molecule two carbon atoms shorter than before.

Information about enzymes and their catalytic domains has been organized into several databases such as GenBank,² UniProt,³ the Protein Data Bank⁴ (PDB), and the International Union of Biochemistry and Molecular Biology enzyme nomenclature system, which uses standardized Enzyme Commission (EC) numbers.⁵ GenBank and UniProt are enzyme sequence (primary structure) databases, which can expedite searches to group enzymes into families based on their similar primary structures. The PDB is a database for three-dimensional (tertiary) protein structures, useful for further family classification. On the other hand, the EC system specifies enzymes by their functions rather than by their structures. This is problematical when enzymes with similar sequences have multiple EC numbers. Furthermore, many enzymes with similar substrate specificities and products and therefore similar EC numbers have very different primary structures and therefore belong to non-related families.

We have established the ThYme (Thioester-Active EnzYmes) database⁶ in which fatty acid and polyketide synthesis enzymes are classified by their primary and tertiary structures.
In this article we report the classification of KR, HD, and ER enzyme groups. Each enzyme group consists of more than one family, where members of different families have essentially different sequences, implying that in general they may be descended from different protein ancestors. Enzymes in a family must have strongly similar primary structures, must have nearly identical tertiary structures (if known), and must share general mechanisms as well as catalytic residues located in the same tertiary structure positions.

**Computational methods**

Methods for identifying families were developed in our previous work\(^7\) and here are applied to KRs, HDs, and ERs. The process can be summarized in five major steps:

1) Experimentally verified enzyme sequences were used as queries. They were gathered from UniProt, using only reviewed entries noted as having “evidence at protein level”.

2) Successive Basic Local Alignment Search Tool\(^8\) (BLAST) searches and comparison among results reduced query sequences to a few representative ones.

3) The catalytic domains of representative query sequences were subjected to BLAST to populate the families, using an \(E\)-value of 0.001 to differentiate among them. These domains were identified by referring to Pfam-A,\(^9\) or they were found in the query sequence by constructing a hidden Markov model profile\(^10\) from a multiple sequence alignment (MSA) based on the initial BLAST result.

4) Experimentally confirmed enzymes were surveyed to search for missing enzyme families.

5) Families were confirmed by MSAs using MUSCLE,\(^11\) tertiary structure superpositions and comparisons with MultiProt,\(^12\) and catalytic residue positions.

Values of root mean square deviation (RMSD) between two structures, RMSD\(_{\text{ave}}\) (among three or more structures), and \(P_{\text{ave}}\) (the average percentage of \(\alpha\)-carbon atoms of the amino acid residues used to calculate the RMSD between two compared structures) were calculated, as explained in detail in the Supporting Information in reference 7.

**Results and discussion**

*Family Identification*

Seventeen putative KR, HD, and ER families were identified. They include four KR
families (KR1–KR4), eight HD families (HD1–HD8), and five ER families (ER2–ER6, ER1 having been deleted) (Table 7.1).

A representative tertiary structure and the characteristic structural fold of each family appear in Table 7.2 NAD(P)-binding Rossmann folds are found in families KR1–KR4, ER2, ER4, and ER5. KR2 enzymes have an additional 6-phosphogluconate dehydrogenase C-terminal fold, and ER5 enzymes have an additional GroES-like fold. Families HD1 and HD3–HD6 have HotDog folds. HD2 enzymes have ClpP/crotonase folds, while ER3 and ER6 members have TIM barrel folds. No tertiary structures are known for families HD7 and HD8.

Superimposing one structure per species within each family gave low RMSD values, ranging from 0.62 to 2.00 Å, and high $P_{ave}$ values, ranging from 77.9 to 99.7% (Table 7.2), indicating three-dimensional structural similarity among sequences in each family.

Furthermore, when catalytic residues in different members in the same family have been identified, they are nearly always the same, appear in the same relative positions, and can be superimposed (Table 7.3). Finally, as seen below, the same mechanism is found throughout a family.

**Family KR1**

Family KR1, with over 55,000 primary structures at the time of writing, is the largest family in any enzyme group in ThYme. They come from bacteria, eukaryota, and archaea, and they display many different functionalities. KR1 includes many reductases and dehydrogenases, most not part of fatty acid and polyketide synthesis or of β-oxidation. A sample of the functions of enzymes in KR1 appears in Table 7.1 along with a representative UniProt accession code.

The β-oxoacyl-ACP reductases (EC 1.1.1.100) are found in KR1. They reduce β-ketoacyl-ACP to β-hydroxyacyl-ACP in type II fatty acid biosynthesis, where the individual enzymes of the fatty acid synthesis cycle are separate proteins. They are present in bacteria as FabG enzymes and are also found in plant chloroplasts. These enzymes are also present in eukaryotic mitochondria, where type II fatty acid synthesis also occurs. 3-Oxoacyl-CoA reductases, the enzymes that catalyze the KR step in fatty acid elongation, are also found in
KR1.

Also in KR1 are ER enzymes that reduce enoyl groups in acyl chains to single bonds, 2,4-dienoyl-CoA reductases (DECR), which are auxiliary enzymes in fatty acid β-oxidation, and trans-2-enoyl-CoA reductases, present during fatty acid elongation.

Most enzymes in KR1 belong to the short-chain dehydrogenase/reductase (SDR) superfamily. These enzymes have been extensively reviewed and classified. Not all SDR enzymes are present in KR1. The present ones include the 3-hydroxyacyl-CoA dehydrogenases (EC 1.1.1.35) that catalyze the reverse (from a hydroxyacyl to a ketoacyl) reaction on various acyl-CoA substrates in fatty acid β-oxidation, as well as many short-chain dehydrogenases, carbohydrate dehydrogenases, and steroid dehydrogenases.

KR1 tertiary structures are conserved NAD(P)-binding Rossmann folds. Their catalytic sites include a highly conserved triad formed by Tyr, Lys, and Ser residues. The mechanism in KR1 enzymes is exhibited by *Streptomyces coelicolor* actinorhodin polyketide ketoreductase. The Ser residue forms a hydrogen bond with the substrate keto group. The Lys residue orients the ribosyl ring of NAD(P)H by forming a hydrogen bond with its hydroxyl groups. The protonated Tyr residue hydrogen-bonds with a hydroxyl group of the ribosyl ring and adds a proton to the substrate keto group. A hydride ion is transferred to the substrate carbon atom of the keto group from the nicotinamide ring of NAD(P)H, leaving the latter positively charged. A water-mediated proton relay then takes place to reprotonate the Tyr residue.

Most KR1 (and SDR) enzymes employ the above catalytic mechanism. The catalytic Tyr, Lys, and Ser residues of *Brassica napus* 3-oxoacyl-ACP reductase (FabG), *S. coelicolor* polyketide-ACP reductase, *Candida tropicalis* 3-hydroxyacyl-CoA dehydrogenase, *Bacillus megaterium* glucose dehydrogenase, and *Comamonas testosteroni* steroid dehydrogenase are in the same position when superimposed (Figure 7.1A). However, even though their structures are nearly identical to those of other KR1 members (Figure 7.1B), the catalytic residues of dienoyl-CoA reductases appear in a different conformation (Figure 7.1C). This has been noted, and the enzymes with ER function in KR1 have been described as "divergent SDRs".
Family KR2

KR2 mainly contains acyl-CoA dehydrogenases sequences involved with fatty acid β-oxidation. Their most common function is dehydrogenation of 3-hydroxyacyl-CoA substrates of various chain lengths. Included are the aerobic or anaerobic bacterial fatty acid oxidation complex subunit dehydrogenase domains22 (FadB or FadJ), plant peroxisomal fatty acid β-oxidation multifunctional protein (MFP) dehydrogenase domains23 expressed during seed germination, and eukaryotic peroxisomal and mitochondrial β-oxidation dehydrogenase domains, some active on long-chain fatty acids.24

Human 3-hydroxyacyl-CoA dehydrogenase (HADH) displays the KR2 enzyme mechanism. A nitrogen atom from a His ring abstracts a proton from the substrate hydroxyl group while its position is stabilized by a hydrogen bond between the other His-ring nitrogen atom and an adjacent Glu residue. Meanwhile, a hydride ion is transferred from the substrate carbon atom bonded to the dehydrogenated hydroxyl atom to a carbon atom in the NAD(H) pyrimidine ring.25 The His and Glu residues appear in the same conformation in a FabB enzyme (Table 7.3). Even though KR1 and KR2 share the same NAD(P)-binding Rossmann fold, they employ different mechanisms and contain different catalytic residues.

Family KR3

In KR3 are the β-oxoacyl-ACP reductase (EC 1.1.1.100) domains of fungal fatty acid synthases (FASs), which are part of type I fatty acid synthesis. Fungal FASs are composed of two subunits, with the KR domain residing in the subunit along with the ketoacyl synthase domain.26 Bacterial sequences are also present, all produced by actinobacteria.

Apparently no mechanism has been proposed for any enzyme in this family. Three residues in the active site of the KR3 domain of Saccharomyces cerevisiae fatty acid synthase are Ser827, Tyr839, and Lys843,27 the same residues found in essentially the same positions relative to each other as in KR1 enzymes (Table 7.3). This and the fact that the same NAD(P)-binding Rossmann fold is also found in KR3 suggest that the KR3 domain in the α subunit of fungal FASs employs the same mechanism as KR1, or SDR superfamily, enzymes.
Family KR4

The KR4 family is composed of $\beta$-oxoacyl-ACP reductases (EC 1.1.1.100) that form the KR4 domains of animal FASs, as well as bacterial and fungal multimodular polyketide synthases (PKSs). The animal FASs contain all catalytic domains, KR among them, in a single unit needed for fatty acid synthesis. PKSs vary according to their product molecule, yet most require a keto-reduction step and have KR domains.

KR4 enzymes also have the NAD(P)-Rossmann binding fold. Their catalytic sites, like those in KR1 and KR3 members, also include Tyr, Ser, Lys triads, and they employ a similar catalytic mechanism. However, in KR4 enzymes the Lys residue appears in a different conformation. Animal FASs and bacterial erythromycin PKSs have Tyr, Ser, and Lys residues in the same conformation as each other (Table 7.3).

Family HD1

The HD1 enzymes found in bacteria have enoyl-CoA hydratase activity and catalyze the reverse HD reaction, where an enoyl group is hydrated to a hydroxy group. The enoyl-CoA hydratases involved in polyhydroxyalkanoate biosynthesis are found in HD1; however, the biochemical in vivo function is not completely understood for most of the sequences. Hisano and coworkers proposed that an Asp residue activates a water molecule and a His residue donates a proton to the substrate. Johansson and coworkers also found that Asp and His residues are essential for catalysis in another HD1 bacterial enoyl-CoA hydratase. HD1 enzymes also include plant and animal sequences; among them is human 3-hydroxyacyl-ACP dehydratase, which catalyzes the HD step in mitochondrial type II fatty acid synthesis.

Family HD2

HD2 enzymes also have enoyl-CoA hydratase activity, and they are active on many substrates in several pathways. Included are the enoyl-CoA hydratases (ECHs) active in mitochondrial fatty acid $\beta$-oxidation. Also present are plant peroxisomal fatty acid $\beta$-oxidation MFP enoyl-CoA hydratase domains expressed during seed germination. Other proteins found in HD2 are eukaryotic peroxisomal and mitochondrial $\beta$-oxidation enoyl-CoA hydratase domains, some active on long-chain fatty acids. Some of these multimodular enzymes are also present in KR2, as their dehydrogenase domain belongs to that family.
The mechanism of HD2 enzymes is exhibited by enoyl-CoA hydratases. It was proposed that a Glu residue acts as an acid by donating a proton to the substrate while another Glu residue activates a water molecule for nucleophilic attack of the hydroxyl group.\textsuperscript{35}

**Family HD3**

HD3 enzymes are produced by bacteria and eukaryotes and are either hydratases or dehydratases. HD3 contains the HD domains of fungal multimodular FASs. Fungal FASs are composed of two subunits, with their HD domains residing in the subunit along with ER, acyltransferase, and thioesterase domains.\textsuperscript{26} HD3 enzymes also appear with KR3 and ER3 enzymes in the same fungal FAS sequences, although each family is found in a different catalytic domain. The enoyl-CoA hydratase domains of peroxisomal fatty acid β-oxidation bifunctional proteins are present in HD3.\textsuperscript{36}

The hydration mechanism is an attack by a water molecule coordinated by two catalytic residues, Asp808 and His813 in *Candida tropicalis* 2-enoyl-CoA hydratase 2, on the substrate double bond.\textsuperscript{37} These residues correspond to Asp1551 and His1564 in *Sacccharomyces cerevisae* FAS.\textsuperscript{27}

**Family HD4**

The HD4 family is composed mainly of 3-hydroxyacyl-ACP dehydratases (EC 4.2.1.61) that form the HD domains of animal multimodular FASs. Several HD domains from bacterial multimodular PKSs are also present. HD4 enzymes also appear with KR4 and ER4 enzymes in the same animal FAS sequences, although again each family is found in a different catalytic domain.

The proposed mechanism of HD4 dehydratases, by analogy with HD5 dehydratases, is deprotonation of the substrate by a His residue and protonation of the substrate’s hydroxyl group by an Asp residue, followed by detachment of a water molecule and formation of a double bond in the substrate.\textsuperscript{38}

**Family HD5**

Most sequences in HD5 are 3-hydroxyacyl-ACP dehydratases that catalyze the HD reaction in type II fatty acid synthesis and are known as FabA enzymes. They are almost
exclusively present in proteobacteria. These enzymes act on medium-length acyl chains, especially on ten-carbon ones.\textsuperscript{39}

In \textit{E. coli} β-hydroxydecanoyl-ACP dehydratase, the uncharged Nτ atom on a His residue abstracts a proton from the substrate’s C2 atom, leaving the former positively charged, while the protonated carboxyl group on an Asp residue of the adjacent enzyme subunit abstracts the hydroxyl group on the C3 atom, producing a water molecule and leaving the C3 atom negatively charged.\textsuperscript{40} Leesong and coworkers also identified the HotDog fold in FabA enzymes, which enzymes in HD1, HD3, HD4, and HD6 also contain.

\textit{Family HD6}

Most HD6 sequences are FabZ enzymes that catalyze the HD reaction in type II fatty acid synthesis. They are present in bacteria and in a few plants and other eukaryota. FabZ enzymes are very similar to FabA ones, but they have broad substrate specificities, acting both on short- and on long-chain acyl chains.\textsuperscript{39} Kostrewa and coworkers proposed a mechanism with \textit{Plasmodium falciparum} HD FabZ\textsuperscript{41} identical to the mechanism of HD5 FabA.\textsuperscript{40}

\textit{Family HD7}

Family HD7 contains mainly fungal sequences, among them 3-hydroxyacyl-ACP dehydratases, which catalyze the HD step in mitochondrial type II fatty acid synthesis.\textsuperscript{42} Autio and coworkers by phylogenetic analysis separated fungal (HD7) from animal (HD1) mitochondrial type II fatty acid synthesis HD sequences.\textsuperscript{33} No tertiary structures have been resolved for HD7, although it likely has a HotDog-like fold due to slight sequence similarity with family HD1.

\textit{Family HD8}

All HD8 sequences are eukaryotic, with fungal, plant, and animal sequences present; there is no known tertiary structure for this family. Sequences include the 3-hydroxyacyl-CoA dehydratases, which catalyze the HD step in fatty acid chain elongation in the endoplasmic reticulum.\textsuperscript{43}
Family ER1

This family has been merged into KR1 in ThYme, as all its entries are present in KR1.

Family ER2

ER2 enzymes are present mainly in bacteria, but they also are found in plants, green algae, and acomplexans. Most enzymes are enoyl-ACP reductases (EC 1.3.1.9), also known as FabI enzymes, that catalyze the ER step in type II fatty acid synthesis. ER2 enzymes also have a NAD(P)-Rossmann binding fold. They have also been labeled as “divergent SDRs”, as they possess the same fold as SDR enzymes, but their catalytic residues and mechanisms are different. They are known target enzymes for antimicrobial drugs, and their tertiary structures complexed with ACP have been resolved.

The mechanism of ER2 enzymes is exhibited by InhA, the Mycobacterium tuberculosis homolog of FabI enzymes. A Lys residue holds the NADH cofactor in place by a hydrogen bond between its amine group and a hydroxyl group of the NADH ribosyl group. The Tyr residue’s hydroxyl hydrogen atom forms a hydrogen bond with the substrate thioester carbonyl oxygen atom. A hydride ion from NADH is transferred to the C3 atom of the substrate, followed by a proton, presumably from the solvent, transferring to the C2 atom.

Family ER3

ER3 enzymes include the enoyl-ACP reductases that form the ER domains of multimodular fungal FASs, composed of two subunits, with the ER domain residing in the subunit along with HD, acyltransferase, and thioesterase domains. ER3 enzymes in fungal FASs have TIM barrel folds, and they are flavin mononucleotide (FMN)-dependent. Jenni and coworkers have proposed a two-step mechanism without specifying individual residues. Many actinobacterial sequences are also present in ER3, but their functions have not been experimentally verified.

Family ER4

The ER4 family includes the enoyl-ACP reductase domains of multimodular animal FASs, as well as bacterial and fungal multimodular PKSs. The animal FASs are homodimers containing all catalytic domains, ERs among them, needed for fatty acid synthesis.
vary according to their product molecule, yet most require an enoyl reduction step and have ER domains. ER4 enzymes have NAD(P)-binding Rossmann folds, but they are not part of the SDR superfamily.

Little is known about the catalytic mechanism of ER4 enzymes. Candidate proton donors to the substrate double bond after hydride transfer from NADPH are Lys1771 and Asp1797 (S. scrofa ER numbering).

**Family ER5**

Members of ER5 appear in eukaryota and bacteria. The largest number are trans-2-enoyl-CoA reductases (EC 1.3.1.38). They are usually mitochondrial in eukaryotes, and they have been characterized in fungi as ETR1’s (2-enoyl thioester reductase-1’s) and in mammals as MECRs (mitochondrial trans-2-enoyl-CoA reductases). They are thought to be involved with mitochondrial fatty acid synthesis.

By analogy with other ERs, Airenne and coworkers proposed that a Tyr residue of *Candida tropicalis* enoyl thioester reductase forms a hydrogen bond to the substrate carbonyl oxygen atom, stabilizing the transition state and allowing hydride ion transfer from NADPH to the substrate double bond. Later it was proposed that a Trp residue is also key for catalysis.

**Family ER6**

ER6 members include 2,4-dienoyl-CoA reductases (EC 1.3.1.34), known as FadHs in bacteria, that use NADPH to remove double bonds at even-numbered positions along the acyl chain used in fatty acid metabolism. ER6 also includes many other reductases and NADPH dehydrogenases such as Old Yellow Enzyme.

ER6 enzymes have TIM barrel folds. They are iron-sulfur enzymes that contain FMN and FAD, and they need NADPH to provide electrons for catalysis. A mechanism was proposed where in the final step Tyr and His residues protonate the substrate.

**Families in fatty acid synthesis and elongation**

How KR, HD, and ER families appear in the different types of fatty acid synthesis is summarized in Table 7.4 and Figure 7.2. In fungal type I fatty acid synthesis, KR3, HD3, and
ER3 domains are present in large multidomain FASs made up of two subunits. For animal type I fatty acid synthesis, KR4, HD4, and ER4 domains are found in single large multidomain FASs. In type II fatty acid synthesis, present in bacteria and plants, with freestanding independent enzymes, the KR (FabG) enzymes are present in KR1, the HD enzymes are present in HD5 (FabA) and HD6 (FabZ), and the ER enzymes (FabI) are found in ER2. For mitochondrial type II fatty acid synthesis in eukaryotes, the KR enzymes appear in KR1, the animal HD enzymes are found in HD1, the fungal HD enzymes are in HD7, and the ER enzymes appear in ER5. Eukaryotic fatty acid elongation in the endoplasmic reticulum also includes KR, HD, and ER steps; their enzymes are found in KR1 and HD8. Both KR and ER steps appear in KR1, into which ER1 has been merged.

The NAD(P)-binding Rossmann fold in KRs is maintained in all types of fatty acid synthesis. The HotDog fold prevails in HDs, except for HD2 enzymes, which have ClpP/crotonase folds, and for HD7 and HD8, for which no tertiary structures have been resolved. ER folds vary by family. Fungal ER domains have TIM barrel folds, while both animal and bacterial ER enzymes contain NAD(P)-binding Rossmann folds, and mitochondrial ER enzymes contain a GroES-like fold in addition to the NAD(P)-binding Rossmann fold. The catalytic residues and mechanisms are conserved among individual KR, HD, and ER families.

Conclusions

KR, HDs, and ERs are essential components of the fatty acid synthesis and elongation, polyketide synthesis, and β-oxidation degradation cycles. They have been classified into different families by their primary and tertiary structures. This has led to four KR families, eight HD families, and five ER families. Their catalytic domains, functions, and mechanisms characterize each family.

References

### Tables

Table 7.1: Enzymes present in KR, HD, and ER families

<table>
<thead>
<tr>
<th>Family</th>
<th>General function</th>
<th>Specific function or enzyme name</th>
<th>Representative UniProt designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR1</td>
<td>Acyl-ACP reductase</td>
<td>3-Oxoacyl-ACP reductase (FabG)</td>
<td>P0AEK2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-Oxoacyl-ACP reductase (OAR1)</td>
<td>P35731</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-Oxoacyl-ACP reductase (MabA)</td>
<td>P0A5Y4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actinorhodin polyketide synthesis</td>
<td>P16544</td>
</tr>
<tr>
<td></td>
<td>Acyl-CoA reductase</td>
<td>3-Oxoacyl-CoA reductase</td>
<td>P38286</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,4-Dienoyl-CoA reductase (DECR)</td>
<td>Q16698</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trans-2-Enoyl-CoA reductase</td>
<td>Q9BY49</td>
</tr>
<tr>
<td></td>
<td>Other reductases</td>
<td>Carbonyl reductase</td>
<td>P16152</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Naphtalene reductase</td>
<td>Q12634</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xylulose reductase</td>
<td>Q7Z4W1</td>
</tr>
<tr>
<td></td>
<td>Acyl-CoA dehydrogenase</td>
<td>3-Hydroxyacyl-CoA dehydrogenase (HADH II)</td>
<td>Q99714</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peroxisomal multifunctional enzyme</td>
<td>P97852</td>
</tr>
<tr>
<td></td>
<td>Carbohydrate dehydrogenase</td>
<td>Glucose dehydrogenase</td>
<td>P40288</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mannitol 2-dehydrogenase</td>
<td>O93868</td>
</tr>
<tr>
<td></td>
<td>Steroid dehydrogenase</td>
<td>11-β-Hydroxysteroid dehydrogenase</td>
<td>P28845</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-β,20-α-Hydroxysteroid dehydrogenase</td>
<td>P19992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-β,17-β-Hydroxysteroid dehydrogenase</td>
<td>P19871</td>
</tr>
<tr>
<td></td>
<td>Other dehydrogenases</td>
<td>Alcohol dehydrogenase</td>
<td>P10807</td>
</tr>
<tr>
<td>KR2</td>
<td>Acyl-CoA dehydrogenase</td>
<td>3-Hydroxyacyl-CoA dehydrogenase (FadB)</td>
<td>P28793</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peroxisomal multifunctional protein</td>
<td>P07896</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydroxyacyl-CoA dehydrogenase (HADH)</td>
<td>Q16836</td>
</tr>
<tr>
<td>KR3</td>
<td>Acyl-ACP reductase</td>
<td>3-Oxoacyl-ACP reductase domain of FASs (subunit α)</td>
<td>P19097</td>
</tr>
<tr>
<td>KR4</td>
<td>Acyl-ACP reductase</td>
<td>3-Oxoacyl-ACP reductase domain of FASs</td>
<td>P49327</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-Oxoacyl-ACP reductase domain of PKSs</td>
<td>Q03131</td>
</tr>
<tr>
<td>H1</td>
<td>Acyl-CoA hydratase</td>
<td>Enoyl-CoA hydratase</td>
<td>O32472</td>
</tr>
<tr>
<td>-----</td>
<td>---------------------</td>
<td>-------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>Acyl-ACP dehydratase</td>
<td>Hydroxyacyl-ACP dehydratase</td>
<td>P86397</td>
</tr>
<tr>
<td>HD2</td>
<td>Acyl-CoA hydratase</td>
<td>Enoyl-CoA hydratase</td>
<td>P14604</td>
</tr>
<tr>
<td>HD3</td>
<td>Acyl-ACP dehydratase</td>
<td>3-Hydroxyacyl-ACP dehydratase domain of FASs (subunit β)</td>
<td>P07149</td>
</tr>
<tr>
<td></td>
<td>Acyl-CoA hydratase</td>
<td>Multifunctional enzyme type 2</td>
<td>P51659</td>
</tr>
<tr>
<td>HD4</td>
<td>Acyl-ACP dehydratase</td>
<td>3-Hydroxyacyl-ACP dehydratase domain of FASs</td>
<td>P49327</td>
</tr>
<tr>
<td>HD5</td>
<td>Acyl-ACP dehydratase</td>
<td>3-Hydroxyacyl-ACP dehydratase (FabA)</td>
<td>P0A6Q3</td>
</tr>
<tr>
<td>HD6</td>
<td>Acyl-ACP dehydratase</td>
<td>3-Hydroxyacyl-ACP dehydratase (FabZ)</td>
<td>Q5G940</td>
</tr>
<tr>
<td>HD7</td>
<td>Acyl-ACP dehydratase</td>
<td>Hydroxyacyl-ACP dehydratase</td>
<td>P38790</td>
</tr>
<tr>
<td>HD8</td>
<td>Acyl-CoA dehydratase</td>
<td>3-Hydroxyacyl-CoA dehydratase (HACD)</td>
<td>BOYJ81</td>
</tr>
<tr>
<td></td>
<td>Acyl-ACP reductase</td>
<td>Enoyl-ACP reductase (ENR) (FabI)</td>
<td>P0AEK4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Long-chain enoyl-ACP reductase (InhA)</td>
<td>P0A5Y6</td>
</tr>
<tr>
<td>ER2</td>
<td>Acyl-ACP reductase</td>
<td>Enoyl-ACP reductase domain of FASs (subunit β)</td>
<td>P07149</td>
</tr>
<tr>
<td>ER3</td>
<td>Acyl-ACP reductase</td>
<td>Enoyl-ACP reductase domain of FASs</td>
<td>P49327</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enoyl-ACP reductase domain of PKSs</td>
<td>Q03132</td>
</tr>
<tr>
<td>ER5</td>
<td>Acyl-ACP reductase</td>
<td>Enoyl-ACP reductase</td>
<td>Q8WZM3</td>
</tr>
<tr>
<td></td>
<td>Acyl-CoA reductase</td>
<td>Enoyl-CoA reductase</td>
<td>Q9BV79</td>
</tr>
<tr>
<td>ER6</td>
<td>Acyl-CoA reductase</td>
<td>2,4-Dienoyl-CoA reductase (FadH)</td>
<td>P42593</td>
</tr>
<tr>
<td></td>
<td>Other reductases</td>
<td>Old Yellow Enzyme</td>
<td>Q02899</td>
</tr>
</tbody>
</table>

a: non-exhaustive
Table 7.2: Representative tertiary structures and folds in KR, HD, and ER families

<table>
<thead>
<tr>
<th>Family</th>
<th>Representative tertiary structure</th>
<th>Fold</th>
<th>Total number of tertiary structures</th>
<th>RMSD, Å</th>
<th>$P_{ave}$, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR1</td>
<td>1EDO</td>
<td>NAD(P)-binding Rossmann</td>
<td>324</td>
<td>2.00</td>
<td>80.1</td>
</tr>
<tr>
<td>KR2</td>
<td>1WDK</td>
<td>6-Phosphogluconate dehydrogenase C-terminal and NAD(P)-binding Rossmann</td>
<td>20</td>
<td>1.20</td>
<td>91.3</td>
</tr>
<tr>
<td>KR3</td>
<td>2PFF (KR domain)</td>
<td>NAD(P)-binding Rossmann</td>
<td>6</td>
<td>0.93</td>
<td>98.5</td>
</tr>
<tr>
<td>KR4</td>
<td>2VZ8 (KR domain)</td>
<td>NAD(P)-binding Rossmann</td>
<td>10</td>
<td>1.28</td>
<td>98.1</td>
</tr>
<tr>
<td>HD1</td>
<td>2C2I</td>
<td>HotDog</td>
<td>6</td>
<td>1.47</td>
<td>78.1</td>
</tr>
<tr>
<td>HD2</td>
<td>1DUB</td>
<td>ClpP/crotonase</td>
<td>43</td>
<td>1.37</td>
<td>85.5</td>
</tr>
<tr>
<td>HD3</td>
<td>2PFF (HD domain)</td>
<td>HotDog</td>
<td>13</td>
<td>1.20</td>
<td>79.7</td>
</tr>
<tr>
<td>HD4</td>
<td>2VZ8 (HD domain)</td>
<td>HotDog</td>
<td>2</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>HD5</td>
<td>1MKA</td>
<td>HotDog</td>
<td>3</td>
<td>0.62</td>
<td>91.5</td>
</tr>
<tr>
<td>HD6</td>
<td>2GLL</td>
<td>HotDog</td>
<td>21</td>
<td>1.04</td>
<td>89.9</td>
</tr>
<tr>
<td>HD7</td>
<td>—</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>HD8</td>
<td>—</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ER2</td>
<td>1C14</td>
<td>NAD(P)-binding Rossmann</td>
<td>107</td>
<td>1.45</td>
<td>77.9</td>
</tr>
<tr>
<td>ER3</td>
<td>2PFF (ER domain)</td>
<td>TIM barrel</td>
<td>6</td>
<td>0.76</td>
<td>99.7</td>
</tr>
<tr>
<td>ER4</td>
<td>2VZ8 (ER domain)</td>
<td>NAD(P)-binding Rossmann</td>
<td>3</td>
<td>1.45</td>
<td>85.1</td>
</tr>
<tr>
<td>ER5</td>
<td>2VCY</td>
<td>GroES-like and NAD(P)-binding Rossmann</td>
<td>7</td>
<td>1.43</td>
<td>90.9</td>
</tr>
<tr>
<td>ER6</td>
<td>1PS9</td>
<td>TIM barrel</td>
<td>90</td>
<td>1.31</td>
<td>84.5</td>
</tr>
</tbody>
</table>

a: as of February 2012
Table 7.3: Catalytic residues in KR, HD, and ER families for representative structures

<table>
<thead>
<tr>
<th>Family</th>
<th>Catalytic residues</th>
<th>PDB file</th>
<th>Function</th>
<th>Producing organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR1</td>
<td>Ser154, Tyr167, Lys171</td>
<td>1EDO</td>
<td>3-Oxoacyl-ACP reductase (FabG)</td>
<td><em>Brassica napus</em></td>
</tr>
<tr>
<td>KR1</td>
<td>Ser144, Tyr157, Lys161</td>
<td>1X7G</td>
<td>Polyketide-ACP reductase</td>
<td><em>Streptomyces coelicolor</em></td>
</tr>
<tr>
<td>KR1</td>
<td>Ser150, Tyr163, Lys167</td>
<td>2ET6</td>
<td>3-hydroxyacyl-CoA dehydrogenase</td>
<td><em>Candida tropicalis</em></td>
</tr>
<tr>
<td>KR1</td>
<td>Ser145, Tyr158, Lys162</td>
<td>1GCO</td>
<td>Glucose dehydrogenase</td>
<td><em>Bacillus megaterium</em></td>
</tr>
<tr>
<td>KR1</td>
<td>Ser138, Tyr151, Lys155</td>
<td>1HXH</td>
<td>Steroid dehydrogenase</td>
<td><em>Comamonas testosteroni</em></td>
</tr>
<tr>
<td>KR2</td>
<td>His451, Glu463</td>
<td>1WDK</td>
<td>3-Hydroxyacyl-CoA dehydrogenase (FabB)</td>
<td><em>Pseudomonas fragi</em></td>
</tr>
<tr>
<td>KR2</td>
<td>His158, Glu170</td>
<td>3HAD</td>
<td>Hydroxyacyl-CoA dehydrogenase (HADH)</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>KR3</td>
<td>Ser827, Tyr839, Lys843</td>
<td>2PFF</td>
<td>3-Oxoacyl-ACP reductase domain of FASs</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>KR4</td>
<td>Lys1995, Ser2021, Tyr2034</td>
<td>2VZ8</td>
<td>3-Oxoacyl-ACP reductase domain of FASs</td>
<td><em>Sus scrofa</em></td>
</tr>
<tr>
<td>KR4</td>
<td>Lys1776, Ser1800, Tyr1813</td>
<td>2FR8</td>
<td>3-Oxoacyl-ACP reductase domain of PKSs</td>
<td><em>Saccharopolyspora eryth.</em></td>
</tr>
<tr>
<td>HD1</td>
<td>Asp31, His36</td>
<td>1IQ6</td>
<td>Enoyl-CoA hydratase</td>
<td><em>Aeromonas caviae</em></td>
</tr>
<tr>
<td>HD2</td>
<td>Glu144, Glu164</td>
<td>1DUB</td>
<td>Enoyl-CoA hydratase</td>
<td><em>Rattus norvegicus</em></td>
</tr>
<tr>
<td>HD3</td>
<td>Asp1559, His1564</td>
<td>2PFF</td>
<td>3-Hydroxyacyl-CoA dehydratase domain of FASs</td>
<td><em>S. cerevisiae</em></td>
</tr>
<tr>
<td>HD3</td>
<td>Asp808, His813</td>
<td>1PN2</td>
<td>Enoyl-CoA hydratase domain of multifunctional enzyme type 2</td>
<td><em>C. tropicalis</em></td>
</tr>
<tr>
<td>HD4</td>
<td>His878, Asp1033</td>
<td>2VZ8</td>
<td>3-Hydroxyacyl-ACP dehydratase domain of FASs</td>
<td><em>S. scrofa</em></td>
</tr>
<tr>
<td>HD5</td>
<td>His70, Asp84a</td>
<td>1MKA</td>
<td>3-Hydroxyacyl-ACP dehydratase (FabA)</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>HD6</td>
<td>His58, Glu72a</td>
<td>2GLL</td>
<td>3-Hydroxyacyl-ACP dehydratase (FabZ)</td>
<td><em>Helicobacter pylori</em></td>
</tr>
<tr>
<td>ER2</td>
<td>Tyr156, Lys163</td>
<td>1C14</td>
<td>Enoyl-ACP reductase (ENR) (FabI)</td>
<td><em>E. coli</em></td>
</tr>
<tr>
<td>ER2</td>
<td>Tyr158, Lys165</td>
<td>1BYR</td>
<td>Long chain enoyl-ACP reductase (InhA)</td>
<td><em>Mycobacterium tuberculosis</em></td>
</tr>
<tr>
<td>ER3</td>
<td>—</td>
<td>2PFF</td>
<td>Enoyl-ACP reductase domain of FASs</td>
<td><em>S. cerevisiae</em></td>
</tr>
<tr>
<td>ER4</td>
<td>Lys1771, Asp1797</td>
<td>2VZ8</td>
<td>Enoyl-ACP reductase domain of FASs</td>
<td><em>S. scrofa</em></td>
</tr>
<tr>
<td>ER5</td>
<td>Tyr94, Trp311</td>
<td>2VCY</td>
<td>Enoyl-CoA reductase</td>
<td><em>H. sapiens</em></td>
</tr>
<tr>
<td>ER6</td>
<td>Tyr166, His252</td>
<td>1PS9</td>
<td>2,4-Dienoyl-CoA reductase (FadH)</td>
<td><em>E. coli</em></td>
</tr>
</tbody>
</table>

a Separate subunit
Table 7.4: KR, HD, and ER families in fatty acid synthesis and elongation

<table>
<thead>
<tr>
<th>Fungal</th>
<th>Animal</th>
<th>Bacterial and plant</th>
<th>Eukaryotic mitochondrial</th>
<th>Eukaryotic endoplasmic reticulum</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD3</td>
<td>HD4</td>
<td>HD5/HD6</td>
<td>HD1/HD7</td>
<td>HD8</td>
</tr>
<tr>
<td>ER3</td>
<td>ER4</td>
<td>ER2</td>
<td>ER5</td>
<td>KR1</td>
</tr>
</tbody>
</table>
Figures

Figure 7.1: Catalytic residues and tertiary structures of KR1 members. A: Catalytic residues of *Brassica napus* 3-oxoacyl-ACP reductase (1EDO, white), *Streptomyces coelicolor* polyketide-ACP reductase (1X7G, green), *Candida tropicalis* 3-hydroxyacyl-CoA dehydrogenase (2ET6, yellow), *Bacillus megaterium* glucose dehydrogenase (1GCO, purple), and *Comamonas testosteroni* steroid dehydrogenase (1HXH, blue) are in the same position when superimposed. B: Tertiary structures of the same enzymes, 1EDO (white), 1X7G (green), 2ET6 (yellow) 1GCO (purple), 1HXH (blue), along with *Homo sapiens* enoyl-CoA reductase (1W6U, brown), are similar when superimposed. C: The catalytic residues of two enoyl-CoA reductases (1W6U, brown) and (1EDO, white) appear in different conformations.
Figure 7.2: Main fatty acid synthesis enzymes. KR3: *Saccharomyces cerevisiae* 3-oxoacyl-ACP reductase domain of FAS (2UV9); KR4: *Sus scrofa* 3-oxoacyl-ACP reductase domain of FAS (2VZ9); KR1: *Brassica napus* 3-oxoacyl-ACP reductase (1EDO); HD3: *Homo sapiens* acyl-CoA hydratase domain of multifunctional enzyme type 2 (1S9C); HD4: *S. scrofa* 3-hydroxyacyl-ACP dehydratase domain of FAS (2VZ8); HD5: *Escherichia coli* 3-hydroxyacyl-ACP dehydratase (FabA) (1MKA); ER3: *S. cerevisiae* enoyl-ACP reductase domain of FAS (2UV8); ER4: *S. scrofa* enoyl-ACP reductase domain of FAS (2VZ8); ER2: *E. coli* enoyl-ACP reductase (FabI) (1C14).

<table>
<thead>
<tr>
<th>Type I Fatty Acid Synthesis</th>
<th>Type II Fatty Acid Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fungal</strong></td>
<td><strong>Animal</strong></td>
</tr>
<tr>
<td>KR3</td>
<td>KR4</td>
</tr>
<tr>
<td>HD3</td>
<td>HD4</td>
</tr>
<tr>
<td>ER3</td>
<td>ER4</td>
</tr>
</tbody>
</table>
CHAPTER 8
ACYL CARRIER PROTEIN STRUCTURAL CLASSIFICATION AND NORMAL MODE ANALYSIS
David C. Cantu, Michael J. Forrester, Katherine Charov, and Peter J. Reilly
Department of Chemical and Biological Engineering, Iowa State University

Modified from a paper published in Protein Science, 21, 655-666, with the same title and authors.

Abstract
All acyl carrier protein (ACP) primary and tertiary structures were gathered into the ThYme database. They are classified into 16 families by amino acid sequence similarity, with members of the different families having sequences with statistically highly significant differences. These classifications are supported by tertiary structure superposition analysis. Tertiary structures from a number of families are very similar, suggesting that these families may come from a single distant ancestor. Normal vibrational mode analysis was conducted on experimentally determined freestanding structures, showing greater fluctuations at chain termini and loops than in most helices. Their modes overlap more so within families than between different families. The tertiary structures of three ACP families that lacked any known structures were predicted as well.

Introduction
Acyl carrier proteins (ACPs) usually have 70 to 100, but occasionally more, amino acid residues, and they are usually linked through an interior serine residue to the terminal phosphate group of a 4’-phosphopantetheine prosthetic group. In turn, the latter binds fatty acids, polyketides, and other moieties by a thioester bond to its terminal thiol group, activating them for reactions that usually produce longer acyl chains, but also many other compounds (Table 8.1).

ACP molecules have many more anionic than cationic residues and rather few hydrophobic residues.¹ Their tertiary structures feature three generally parallel α-helices
helices I, II, and IV) with a shorter crosswise $\alpha$-helix (helix III). Helix II has many conserved anionic residues and plays an important role in ACP-enzyme interactions. Helix III is not present in all ACP structures, and it displays helix-loop equilibrium conformations.

ACPs are either independent, freestanding structures, or they are covalently bound as part of multimodular enzymes such as fatty acid synthases (FASs), polyketide synthases (PKSs), and non-ribosomal peptide synthases (NRPSs). Acyl chains attached to freestanding ACPs are held within the hydrophobic pocket formed by the $\alpha$-helices until they are subjected to reaction, when they are expelled into the active site of the enzyme catalyzing the reaction. The mechanism of delivering an acyl substrate to an enzyme active site from ACP’s cavity, which may be accompanied by flexing of the ACP, is not completely understood. Those ACPs that are part of enzymes may not hold acyl chains within their hydrophobic pockets, as these chains are less exposed to solvent and to cell membranes.

Freestanding ACPs, at least, appear to be quite flexible, being found in multiple conformers and having very flexible loops and $\alpha$-helices.

We have gathered all available ACP primary and tertiary structures into the ThYme database. There we have classified ACPs into families, following the same techniques that we have used earlier with thioesterases and ketoacyl synthases, which are described in Computational Methods. In general, members of a protein family have strong sequence similarity. They should also have tertiary structures that can be superimposed with small root mean square deviations (RMSDs) between corresponding amino acid residues. These similarities may imply that members are descended from a common protein ancestor. Members of different families have primary structures with statistically highly significant differences. However, slight similarities in amino acid residue alignments between ACP families may suggest that those without known tertiary structures are related to those with known structures. We describe this work with ACPs for the first time here.

This article is an account also of two further efforts: 1) a normal mode analysis of experimentally determined tertiary structures of freestanding apo-ACPs to describe their dynamic structures; and 2) the attempted computational prediction of tertiary structures of freestanding ACPs in three families with no known experimentally determined structures.
Results and discussion

ACP families

Following the protocol described in the Computational Methods section led to 16 ACP families being defined. A multiple sequence alignment of representative members of these families shows very limited sequence similarity (Figure 8.1). Only the serine residue at position 39 (ACP1 numbering), to which the prosthetic group is attached, is almost completely conserved (except for ACP15). In addition, there is substantial conservation at positions 35 (glycine), 38 (aspartate), and 67 (threonine). A number of positions have almost exclusively hydrophobic residues. In ACP15, the only residue that is completely conserved near the position otherwise occupied by a serine residue is Thr40 (ACP1 numbering).

The same representative sequences of ACP families were subjected to pairwise sequence alignments to identify common residues between two sequences in different families. Even though there is low sequence similarity over the 16 families, there is substantial similarity from one family to the next (Figure 8S-A, Supplementary Information). The relationship of the families to each other is shown by two versions of a phylogenetic tree (Figure 8.2 and Figure 8S-B). They show ACP15 and ACP16 peripheral to the other ACP families, as they are in Figure 8.1.

Table 8.1 summarizes the 16 ACP families, showing 1) approximate numbers of sequences in each family; 2) domains of life that produce each family; 3) whether ACP families are composed of freestanding proteins or are parts of multidomain enzymes; 4) the end products of enzymes with which ACP families interact; 5) representative UniProt accession codes of the ACPs; and 6) related literature.

Families ACP1 through ACP3 are involved with fatty acid synthesis. ACP1 members are freestanding ACPs, present in type II fatty acid synthesis, bacterial long-chain fatty acid synthesis, and mitochondrial fatty acid synthesis. ACP2 and ACP3 proteins are parts of multidomain FASs, involved with fungal and animal type I fatty acid synthesis, respectively. A substantial amount of research has been conducted on these ACPs; much of it is covered in the exhaustive review of Chan and Vogel.10

Proteins in ACP4 and ACP5 are freestanding ACPs involved with polyketide synthesis. In ACP4 is AcpK, part of the pkGX pathway of Bacillus subtilis in making bacillaene.24 ACP5
includes ACPs involved with the synthesis of the antibiotics actinorhodin, frenolicin, and oxytetracycline in various *Streptomyces* species.

ACP6 through ACP9 and also ACP11 include the ACP domains of large multidomain PKSs. Among ACP6 sequences are the ACP domains of larger enzymes involved in making complex lipids found in *Mycobacterium* cell walls, including mycocerosic acid, and other sulfolipids. ACP7 includes the ACP domain of fungal PKS 6-methylsalicylic acid synthase. ACP8 includes mainly fungal sequences; among them are the ACP domains of PKSs involved with aflatoxin production, and the ACP domains of naphthopyrone PKSs that make the yellow pigment in conidia. ACP9 includes the ACP domain of PKS or chalcone synthase *stlA*, which produces acylpyrones. ACP11 includes the ACP domains of lovastatin syntheses.

ACP10, the family with the most members, includes the ACP domains of many PKSs, the peptide carrier protein (PCP) domain of NRPSs, hybrid PKS/NRPS enzymes, and ferrichrome synthases. The enzymes in this family make a variety of natural products from secondary metabolism. A representative sample of them is shown in Table 8.1.

Families ACP12 and ACP13 include the ACP domains of enterobactin synthases, isochorismatases, and mycobactin synthases. ACP14 has been merged into ACP10 and no longer exists.

The prosthetic group in ACP15 and ACP16 is 2’-(5”-phosphoribosyl)-3’-dephospho-CoA, instead of 4’-phosphopantetheine, linked to an interior serine residue of apo-ACP through its 5”-phospho group and to the acyl molecule with a thioester bond through its terminal thiol group. ACP15 proteins include the ACPs active with malonate decarboxylases in bacteria that convert malonate to acetate and CO₂ as an energy source. ACP16 enzymes include the ACPs active with citrate lyases that convert citrate to oxaloacetate and acetate in bacteria.

Members of ACP17 do not carry acyl groups, but instead they are D-alanyl carrier proteins, as the moiety bound by 4’-phosphopantetheine is D-alanine, which is ligated using adenosine triphosphate to poly(ribitol phosphate). These enzymes are involved with the production of D-alanyl lipoteichoic acid.

The members of all but four families, ACP1, ACP8, ACP10, and ACP13, are produced
exclusively by either bacteria or eukaryota. More specifically, virtually all members of ACP2 and ACP8 and all members of ACP7 and AC11 are produced by fungi (the latter three families all by ascomycota), while all ACP3 members come from animals. AC9 members are from slime molds. Members of different bacterial phyla produce different members of ACP4, ACP12, ACP15, ACP16, and ACP17. ACP5 and ACP6 members are all from actinobacteria, with the latter being only from *Mycobacterium* species.

In summary, ACPs have diverged into different families based on primary structures that have statistically highly significant differences. They are either freestanding or are covalently bound to enzymes, they are specific to different substrates, they are produced by different classes of organisms, and they have sharply defined roles.

**Existing ACP tertiary structures**

Families ACP1, ACP2, ACP3, ACP5, ACP8, ACP10, ACP12, ACP13, and ACP17 contain members with known tertiary structures (Figure 8.3). All but ACP8, ACP12, and ACP13 have more than one known structure. Most known structures and their properties were reviewed by Chan and Vogel. All tertiary structures are tabulated, with links to the Protein Data Bank (PDB), in the ACP section of the ThYme database.

We superimposed one tertiary structure per species within each putative ACP family and calculated their RMSDs using the protocol in the Computational Methods section. ACP2 and ACP17 have multiple known structures of freestanding ACPs or of ACP domains, but they all come from the same species. Therefore their RMSDs were not calculated, as this would represent structure conservation among the same sequence and not sequences in a family.

ACP2 domains in yeast FAS structures contain four extra helices in the C-terminal region, not seen in other ACPs, making them about twice as long as the others (Figure 8.2). We found ACP3 and ACP10 domains in enzymes containing them by superimposing freestanding ACPs and extracting the former for RMSD calculations. In ACP10, we chose to superimpose the structures in the A/H conformer, very similar to other ACPs, one of three conformers in which ACP10 structures have been found. Structures within putative families should have low RMSD<sub>ave</sub> values and high $P_{ave}$ values (average percentage of $\alpha$-carbon atoms of the amino acid residues between two structures compared for calculations). Shown
in Table 8.2 are the structures used, RMSD$_\text{ave}$ values (ranging from 1.74 Å to 2.02 Å), and $P_{\text{ave}}$ values (ranging from 84.8% to 94.4%). These findings further indicate that the structures represent members in the same families.

If different families have members with tertiary structures that can be closely aligned, even though their primary structures are different, then they may have a more distant common protein ancestor and may be gathered into the same clan. To determine this, one tertiary structure each from ACP1 (PDB accession code 1X3O), ACP2 (2UV8), ACP3 (2CG5), ACP5 (1OR5), ACP8 (2KR5), ACP10 (2JGP), ACP12 (2FQ1), ACP13 (2ROQ), and ACP17 (1HQB) (Figure 8.3) were superimposed. Of these, ACP2, ACP3, ACP8, ACP10, and ACP12 were domains in larger enzymes. The first four helices of the ACP2 structure were well superimposed on the structures of the other families. The superposition of the structures of these nine families resulted in an RMSD$_\text{ave}$ value of 2.34 Å and a $P_{\text{ave}}$ value of 76.2%, indicating that at least those families may descend from a distant common ancestor and may be gathered into the same clan. The moderate percentages of common residues between different pairs of families (Figure 8S-A) suggest that other ACP families may also have the same common ancestor. However, this conclusion must await determination of tertiary structures of members of these families.

**Normal mode analysis of freestanding ACP structures**

We subjected seven tertiary structures of freestanding apo-ACPs to normal mode analysis, using the Anisotropic Network Model (ANM) web server. We did not attempt to subject ACPs that were part of larger proteins to normal mode analysis, since their structures and fluctuations are most likely affected by their proximity of other parts of the protein. ACP1 members were PDB accession codes 2EHS and 2EHT (two conformers of the same protein) from *Aquiflex aeolicus*, 1T8K from *Escherichia coli*, 1X3O from *Thermus thermophilus*, and 2QNW from *Toxoplasma gondii*. The one ACP5 member was 2AF8 from *Streptomyces coelicolor*, while 1HQB in ACP17 was from *Lactobacillus rhamnosus*. All are produced by bacteria but one; *T. gondii* is an apicomplexan protozoan.

The residue fluctuations of the five slowest modes of each structure are plotted in Figure 8S-C. Chain ends, especially the N-terminal one, and loops were predicted to fluctuate more
than the three long and parallel α-helices (helices I, II, and IV), which move very little. The region between helices II and IV, which in ACP1 structures comprises the loop between helices II and III and the crosswise helix III, strongly fluctuated in many structures. In ACP5 and ACP17 structures helix III is replaced by a loop. This region was proposed to act as a gatekeeper in the acyl chain delivery process, as the acyl chain is exposed through a fissure following a conformational rearrangement. The acyl chain delivery process is unknown; sword-unsheathed or switchblade-like mechanisms have been proposed. Lower fluctuations are seen in the loop between helices I and II. The N-terminal region of helix II, where the 4′-phosphopantetheine prosthetic group is attached, shows relatively higher fluctuations than the rest of helix II. In general, the vibrations of the first five slowest modes in any structure are moderately to strongly correlated in location, although much less so in magnitude.

We compared the normal modes of two conformers (2EHS and 2EHT) of the same structure in ACP1, of two bacterial structures (1T8K and 1X3O) in ACP1, of bacterial (1T8K) and eukaryotic (2QNW) structures in ACP1, of ACP1 (1T8K) and ACP5 (2AF8) structures, of ACP1 (1T8K) and ACP17 (1HQB) structures, and of ACP5 (2AF8) and ACP17 (1HQB) structures. They were analyzed as explained in the Computational Methods section. Results are summarized in Table 8.3, overlap charts are shown in Figure 8S-D, and residue fluctuation comparisons of the three most overlapped modes for each case are shown in Figure 8S-E.

The two conformers of the same structure showed a very low frame-averaged RMSD (RMSD_f-ave, as defined in the Computational Methods section), showing that their structural similarity was conserved throughout the motion of the slowest normal mode. Several of their slowest normal modes showed some overlap (Figure 8S-D). Residue fluctuations (Figure 8S-E) show that 2EHT has larger amplitudes than 2EHS in the N-terminal region, while 2EHS has larger amplitudes in the central and C-terminal regions. Their fluctuations appear in very similar residue locations.

Two different structures within ACP1 display lower overlaps than two conformers of the same structure (Figure 8S-D). Values of RMSD_f-ave are higher than between two conformers of the same structure, but still are < 2 Å, as would be expected of two structures within a family. In both cases, fluctuations appear conserved in residue location but not so in
amplitude (Figure 8S-E), although they appear more correlated between the bacterial and eukaryotic structures.

Comparing the normal modes between structures in different families showed that the slowest mode of each structure was the most highly overlapped (Figure 8S-D), especially between ACP1 and ACP5 and between ACP5 and ACP17. The slowest modes in each structure display an overall twisting rather than stretching of discrete loops, and overlap highly due to their low flexibility. Their RMSD$_{f_{ave}}$ values are higher than in previous cases, showing structural differences between two families. However, within the most overlapped normal mode, structural differences are not great. Residue fluctuation for modes other than the most overlapped (Figure 8S-E) are poorly correlated in both amplitude and location, especially between ACP1 and ACP5 and between ACP5 and ACP17, but less so between ACP1 and ACP17.

Even though overlap charts (Figure 8S-D) do not differ much for comparisons within and between families, residue fluctuations (Figure 8S-E) are more conserved between two structures within a family than between two structures in different families. However, fluctuation amplitudes can be quite different in different structures produced by organisms in the same genus and, especially noteworthy, in different conformers of the same structure. Differences in fluctuation amplitudes may be due to slight crystal packing effects that could be present in loops.

**Determination of tertiary structures of freestanding ACPs by computation**

To determine whether the homology modeling and molecular dynamics (MD) protocol described in the Computational Methods section is a trustworthy method of determining tertiary structures of freestanding ACPs whose structures were previously unknown, we used experimentally determined freestanding structures of two families from ACP1 (2EHS), ACP5 (1OR5), and ACP17 (1HQB) as templates to predict the structure of the third family. We did not use ACP domains as templates or predict unknown ACP domains of large multidomain enzymes, as their folded states may be influenced by nearby sections of the larger protein.

Results are shown in Table 8.4. The amino acid identities, based on pairwise alignments
between individual family members, between template and target structures vary between 20.0% and 23.4%. Two RMSD values between predicted and crystal tertiary structures (RMSD$_{p-c}$) are reported: first the value of the predicted structure by homology modeling to its crystal structure, and second the predicted structure after MD refinement to its crystal structure. RMSD$_{p-c}$ values of homology modeling predictions vary between 1.17 Å and 2.10 Å, and RMSD$_{p-c}$ values after MD refinement vary between 1.42 Å and 1.96 Å. The predicted structures superimposed to their crystal structures can be seen in Figure 8.4. As indicated by low RMSD$_{p-c}$ values (< 2 Å) and by visual inspection, the protocol described can predict freestanding ACP structures when the template used for homology modeling is at least 20% identical in sequence to the target.

We then attempted to predict tertiary structures for sequences in ACP4, ACP15, and ACP16, which contain freestanding members with no known structures. The structure of Bacillus subtilis AcpK (UniProt ID Q7PC63) in ACP4 was predicted using as a template 2EHS from ACP1, with which it has an amino acid identity of 25.6%. The predicted structure of ACP4 (Figure 8.5) is very similar to structures in other ACP families, and it clearly shows a cavity and three long helices (helices I, II, and IV). When the predicted ACP4 structure was superimposed to structures in ACP1 (2EHS), ACP5 (1OR5), and ACP17 (1HQB), the resulting RMSDs were 1.92 Å, 2.12 Å, and 2.21 Å, respectively.

Subjecting Malonomonas rubra malonate decarboxylase (UniProt ID: O06925) from ACP15 and E. coli citrate lyase (UniProt ID: P69330) from ACP16 to the homology modeling and MD protocol resulted in unfolded structures that lacked the standard three parallel helices and cavity seen in other ACPs. However, they are most likely not intrinsically unstructured. Ten models were then predicted for each sequence using threading programs. Four of the ten ACP15 models (1, 5, 6, and 8) displayed two main α-helices and four β-strands (Figure 8S-F), while seven of the ten ACP16 models (1, 2, 4–8) showed two main α-helices with three β-strands. Representative models of the four and seven similar structures of ACP15 and ACP16, respectively, are shown in Figure 8.5. When superimposed, the four similar models in ACP15 have RMSD$_{ave}$ and $P_{ave}$ values of 0.73 Å and 88.7%, and the seven similar models of ACP16 show values of 0.99 Å and 91.3%.

ACP15 and ACP16 carrier proteins are used in reactions that break substrates into
smaller units, unlike other ACPs, which are involved in building large molecules from small units. Also, ACP15 and ACP16 do not use the same 4’-phosphopantetheine prosthetic group present in other ACPs. These facts further suggest that ACP15 and ACP16 may be structurally different from other ACP families, and if this is correct, ACP15 and ACP16 members may descend from different ancestors than most or all of the other ACPs. The multiple sequence alignment (Figure 8.1), phylogenetic tree (Figures 8.2, 8S-B), and threading predictions (Figures 8S-F) support this. However, only experimentally determined structures will show if ACP15 and ACP16 are indeed different from other ACP families.

**Computational methods**

*Family identification and phylogeny*

Families were identified following the procedures outlined earlier. In brief, families were based on sequences with evidence at protein level from the UniProt database or the literature, and populated by use of BLAST (downloadable version 2.2.19) to query the non-redundant (nr) protein sequence database. A cutoff $E$-value of 0.001 (likelihood that similarity between query and compared sequence is due to chance) was used as the exclusion criterion. Families were confirmed with multiple sequence alignments of the retrieved sequences, using MUSCLE 3.6, and with tertiary structure superimpositions.

Phylogenetic trees were built using Molecular Evolutionary Genetics Analysis 5 (MEGA5), based on a multiple sequence alignment containing three sequences per family made with MUSCLE 3.6. The minimum evolution algorithm was used, gaps were treated with pairwise deletion, and an amino acid Jones-Taylor-Thorton (JTT) matrix was chosen as the model. A bootstrap test with 1000 replicates was performed to further verify the results.

*Tertiary structure superposition and RMSD calculations*

In this study, all tertiary structures were superimposed with MultiProt. All RMSD values were calculated between $\alpha$-carbon atoms using MATLAB, to consider the most possible $\alpha$-carbon atoms in the calculation; MultiProt reports the RMSD for only aligned residues. Values of RMSD (between two structures), $\text{RMSD}_{\text{ave}}$ (between three or more
structures), and $P_{\text{ave}}$ (the average percentage of $\alpha$-carbon atoms of the amino acid residues used to calculate the RMSD between two compared structures) were calculated, as explained in detail in the Supporting Information in Cantu et al. (2010). When a NMR-resolved structure was superimposed, only the lowest-energy frame was used, chosen following single-point energy calculations with the ff99SB AMBER force field.

**Normal Mode Analysis**

Normal modes of individual ACPs were calculated using the ANM web server, by assuming that a molecule’s natural vibrations can be predicted by attaching a spring of uniform force constant to each $\alpha$-carbon atom of the molecule and allowing the system to oscillate. The cutoff distance and distance weights were changed until the best correlation between theoretical and experimental B-factors was achieved. Structures were submitted individually to the ANM server, which computed the vibrational normal modes of each molecule and returned an animated PDB file for the 20 slowest vibrational modes, their associated eigenvectors, and residue fluctuations. No significant crystal packing effects were found when viewing the symmetric molecules of the PDBs submitted.

To compare the normal modes between two structures, overlap charts showing the 20 slowest vibrational modes of two structures were made by taking the dot product of the eigenvectors from both structures, yielding an 20 x 20 overlap chart where each cell ranges in value from 0 (no overlap) to 1 (complete overlap). From this, the highest overlapped mode was chosen, and the animated structure PDB files of such mode were separated into individual PDB files for each frame. Then the frame from one structure was superimposed with the corresponding frame in the other, and RMSD and $P$ values was calculated. This was done for each frame, and the averages among all frames were taken, resulting in values of $\text{RMSD}_{f_{\text{ave}}}$ and $P_{f_{\text{ave}}}$. This notation is used here to differentiate the RMSD between two structures averaged over different frames in a normal mode (Table 8.3), from $\text{RMSD}_{\text{ave}}$, which refers to the average RMSD of superimposing three or more structures (Table 8.2).

**Tertiary Structure Determination**

A homology modeling and MD protocol was used for structure predictions. Homology
modeling was done with the I-TASSER server\textsuperscript{62,63} using other ACP structures as templates without alignment or restraints.

The resulting predicted structures were loaded in AmberTools1.4 using the ff99SB AMBER force field. The structures were placed in a 12-Å TIP3P water box,\textsuperscript{64} and Na\textsuperscript{+} ions were added to neutralize system charge.

The system was then simulated with the \textit{sander} module of AMBER.\textsuperscript{65} Solvent and ions were minimized first with 1000 steps while restraining the protein with 500 kcal/Å-mol weights. An unrestrained 10,000-step minimization followed. The system was then heated from 0 K to 300 K at constant volume for 25 to 100 ps while weakly restraining the protein with 10 kcal/Å-mol weights. Finally, the system was equilibrated at constant pressure and temperature and run unrestrained for 1.5 to 2 ns.

Sequences were predicted to not be intrinsically unfolded using the IUPred\textsuperscript{66} server that bases its predictions by calculating inter-residue energy interactions.

Threading predictions were done using the LOMETS\textsuperscript{67} server, which generates tertiary structures from sequences using eight different known threading programs. Unlike homology modeling, no templates are specified. The ten best threading models from the output were taken.

RMSDs between predicted structures to experimental crystal structures in the validation runs (Table 8.4) are labeled as \textit{RMSD}_{p-c}, to differentiate it from the other RMSDs used here.

\textbf{Concluding remarks}

Considering that ACPs, except those of ACP2, are all of roughly the same length and in nearly all cases they bind and activate acyl chains prior to their being subjected to enzymatic catalysis, it is noteworthy that they can be separated into 16 families by their highly significant differences in primary structure. It is perhaps equally noteworthy that many of these families have very similar tertiary structures, signifying that they may be descended from a common distant ancestor. We predicted the normal vibrational modes of freestanding ACPs, finding them more conserved within families. Also, we extended knowledge of three families possessing freestanding ACPs, one family by predicting its characteristic tertiary structure to be like those of many other ACP families, and two families by apparently
showing that they do not have similar structures.

Acknowledgments

The authors thank Michael T. Zimmermann for helpful discussions about normal mode analysis, and Christopher D. Warner for discussions about crystal packing effects.

References


44. Gehring AM, Bradley KA, Walsh CT (1997) Enterobactin biosynthesis in *Escherichia coli*: isochorismate lyase (EntB) is a bifunctional enzyme that is phosphopantetheinylated by EntD and then acylated by EntE using ATP and 2,3-dihydroxybenzoate. Biochemistry 36: 8495–8503.


bin/anm/anm1.cgi.


### Table 8.1: Acyl Carrier Protein Families

<table>
<thead>
<tr>
<th>Family</th>
<th>Approximate number of sequences&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Producing domain</th>
<th>Role</th>
<th>End product (non-exhaustive list)</th>
<th>UniProt accession code</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACP1</td>
<td>2,870</td>
<td>B, E</td>
<td>Freestanding ACPs:</td>
<td>Fatty acids</td>
<td>P0A6A8</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>in type II fatty acid synthesis</td>
<td>Long-chain fatty acids</td>
<td>P0A2W5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>in bacterial long-chain fatty acid synthesis</td>
<td>Fatty acids</td>
<td>P52505</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>in mitochondria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACP2</td>
<td>180</td>
<td>E</td>
<td>ACP domains of fungal FASs</td>
<td>Fatty acids</td>
<td>P19097</td>
<td>22</td>
</tr>
<tr>
<td>ACP3</td>
<td>190</td>
<td>E</td>
<td>ACP domains of animal FASs</td>
<td>Fatty acids</td>
<td>P49327</td>
<td>23</td>
</tr>
<tr>
<td>ACP4</td>
<td>50</td>
<td>B</td>
<td>Freestanding ACPs in polyketide synthesis</td>
<td>Bacillaene</td>
<td>Q7PC63</td>
<td>24</td>
</tr>
<tr>
<td>ACP5</td>
<td>100</td>
<td>B</td>
<td>Freestanding ACPs in polyketide synthesis</td>
<td>Actinorhodin</td>
<td>Q02054</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Frenolicin</td>
<td>Q54996</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Oxtetraacycline</td>
<td>P43677</td>
<td>27</td>
</tr>
<tr>
<td>ACP6</td>
<td>100</td>
<td>B</td>
<td>ACP domains in PKSs</td>
<td>Mycocerosic acid</td>
<td>Q02251</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Phthioceranic acids</td>
<td>O07798</td>
<td>29</td>
</tr>
<tr>
<td>ACP7</td>
<td>30</td>
<td>E</td>
<td>ACP domains in PKSs</td>
<td>6-Methylsalicylic acid</td>
<td>P22367</td>
<td>30</td>
</tr>
<tr>
<td>ACP8</td>
<td>240</td>
<td>B, E</td>
<td>ACP domains in PKSs</td>
<td>Aflatoxin</td>
<td>Q12053</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Conidial yellow pigment</td>
<td>Q03149</td>
<td>32</td>
</tr>
<tr>
<td>ACP9</td>
<td>80</td>
<td>E</td>
<td>ACP domain in PKSs/chalcone synthases</td>
<td>Acylpyrones</td>
<td>Q55E72</td>
<td>33</td>
</tr>
<tr>
<td>ACP10</td>
<td>7,000</td>
<td>B, E, A</td>
<td>ACP/PCP domain in PKSs/NRPSs and ACP domain in ferrichrome synthases</td>
<td>Surfactin</td>
<td>P27206</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Erythronolide</td>
<td>Q03131</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gramicidin</td>
<td>P0C061</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Phthiocerol</td>
<td>Q7TXL8</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tyrocidine</td>
<td>O30409</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pipastatin</td>
<td>O31827</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Enniatin</td>
<td>Q00869</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ferrichrome</td>
<td>Q9P7T1</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bacillaene</td>
<td>Q05470</td>
<td>42</td>
</tr>
<tr>
<td>ACP11</td>
<td>120</td>
<td>E</td>
<td>ACP domains in PKSs</td>
<td>Lovastatin</td>
<td>Q9Y8A5</td>
<td>43</td>
</tr>
<tr>
<td>ACP12</td>
<td>280</td>
<td>B</td>
<td>ACP domains in enterobactin synthases</td>
<td>Enterobactin</td>
<td>P0ADI4</td>
<td>44</td>
</tr>
<tr>
<td>ACP</td>
<td></td>
<td></td>
<td>ACP domains in isochorismatases</td>
<td>Mycobactins</td>
<td>P71717</td>
<td>45</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
<td>---</td>
<td>---------------------------------</td>
<td>-------------</td>
<td>--------</td>
<td>----</td>
</tr>
<tr>
<td>ACP13</td>
<td>220</td>
<td>B, E</td>
<td>ACP domains in enterobactin synthases</td>
<td>Enterobactin</td>
<td>P11454</td>
<td>46</td>
</tr>
<tr>
<td>ACP15</td>
<td>200</td>
<td>B</td>
<td>Freestanding ACPs active with malonate decarboxylases</td>
<td>Acetate</td>
<td>O06925</td>
<td>47</td>
</tr>
<tr>
<td>ACP16</td>
<td>300</td>
<td>B</td>
<td>Freestanding ACPs active with citrate lyases</td>
<td>Acetyl-CoA</td>
<td>P02903</td>
<td>48</td>
</tr>
<tr>
<td>ACP17</td>
<td>250</td>
<td>B</td>
<td>Freestanding D-alanyl carrier proteins</td>
<td>D-Alanyl lipoteichoic acid</td>
<td>P55153</td>
<td>49</td>
</tr>
</tbody>
</table>

*Present in the ThYme database at the time of writing this article.*

*b A, archaea; B, bacteria; E, eukaryota. Most prevalent producers bolded.*
Table 8.2. Acyl Carrier Protein Tertiary Structures

<table>
<thead>
<tr>
<th>Family</th>
<th>RMSD&lt;sub&gt;ave&lt;/sub&gt; (Å)</th>
<th>P&lt;sub&gt;ave&lt;/sub&gt; (%)</th>
<th>Tertiary structures&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Total number of PDB accessions in family</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACP1</td>
<td>2.01</td>
<td>86.4</td>
<td>1HY8, 1KLP, 1L0H, 1X3O, 2DNW, 2EHS, 2FQ0, 2FVE, 2KOO, 2KW2, 2KWL, 2L0Q, 2L3V, 2L4B, 2QNW</td>
<td>45</td>
</tr>
<tr>
<td>ACP2</td>
<td>—</td>
<td>—</td>
<td>2PFF</td>
<td>6</td>
</tr>
<tr>
<td>ACP3</td>
<td>2.02</td>
<td>94.4</td>
<td>2CG5&lt;sup&gt;c&lt;/sup&gt;, 2PNG</td>
<td>4</td>
</tr>
<tr>
<td>ACP5</td>
<td>1.75</td>
<td>84.8</td>
<td>1NQ4, 1OR5, 2AF8</td>
<td>14</td>
</tr>
<tr>
<td>ACP8</td>
<td>—</td>
<td>—</td>
<td>2KR5</td>
<td>1</td>
</tr>
<tr>
<td>ACP10</td>
<td>1.74</td>
<td>85.9</td>
<td>1DNY, 2GDW, 2JGP&lt;sup&gt;c&lt;/sup&gt;, 2JU2, 2VSQ&lt;sup&gt;c&lt;/sup&gt;</td>
<td>9</td>
</tr>
<tr>
<td>ACP12</td>
<td>—</td>
<td>—</td>
<td>2FQ1</td>
<td>1</td>
</tr>
<tr>
<td>ACP13</td>
<td>—</td>
<td>—</td>
<td>2ROQ</td>
<td>1</td>
</tr>
<tr>
<td>ACP17</td>
<td>—</td>
<td>—</td>
<td>1DV5</td>
<td>2&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> For ACP1, ACP3, ACP5, and ACP10 all structures listed were superimposed to yield reported RMSD<sub>ave</sub> and P<sub>ave</sub> values; for others a representative sequence is shown. Only one structure per species within a family was superimposed.

<sup>b</sup> RMSD not calculated, as all resolved ACP domains in existing tertiary structures come from the same species.

<sup>c</sup> Only ACP domain of a larger structure.
Table 8.3: Normal Mode Analysis Comparisons

<table>
<thead>
<tr>
<th>Structural comparison</th>
<th>PDB designations</th>
<th>Highest overlap between two modes</th>
<th>RMSD_{f-ave} P_{f-ave}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same structure, two conformers</td>
<td>2EHS, 2EHT</td>
<td>0.47</td>
<td>0.90 99.1</td>
</tr>
<tr>
<td>Same family, two bacterial structures</td>
<td>1T8K, 1X3O</td>
<td>0.42</td>
<td>1.51 80.9</td>
</tr>
<tr>
<td>Same family, bacterial and eukaryotic</td>
<td>1T8K, 2QNW</td>
<td>0.47</td>
<td>1.05 95.7</td>
</tr>
<tr>
<td>ACP1 to ACP5</td>
<td>1T8K, 2AF8</td>
<td>0.65</td>
<td>2.70 61.2</td>
</tr>
<tr>
<td>ACP1 to ACP17</td>
<td>1T8K, 1HQB</td>
<td>0.58</td>
<td>2.08 85.9</td>
</tr>
<tr>
<td>ACP5 to ACP17</td>
<td>2AF8, 1HQB</td>
<td>0.64</td>
<td>2.67 71.1</td>
</tr>
</tbody>
</table>

Table 8.4: Tertiary Structure Prediction Validation

<table>
<thead>
<tr>
<th>Target structure</th>
<th>Template structure</th>
<th>Amino acid identity (%)(^a)</th>
<th>RMSD_{p-c} (homology) (Å)</th>
<th>RMSD_{p-c} (MD) (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACP1</td>
<td>ACP5</td>
<td>23.4</td>
<td>1.17</td>
<td>1.42</td>
</tr>
<tr>
<td>ACP1</td>
<td>ACP17</td>
<td>20.8</td>
<td>1.44</td>
<td>1.48</td>
</tr>
<tr>
<td>ACP5</td>
<td>ACP1</td>
<td>21.7</td>
<td>1.97</td>
<td>1.87</td>
</tr>
<tr>
<td>ACP5</td>
<td>ACP17</td>
<td>20.5</td>
<td>2.10</td>
<td>1.96</td>
</tr>
<tr>
<td>ACP17</td>
<td>ACP1</td>
<td>20.0</td>
<td>1.73</td>
<td>1.93</td>
</tr>
<tr>
<td>ACP17</td>
<td>ACP5</td>
<td>21.3</td>
<td>1.78</td>
<td>1.88</td>
</tr>
</tbody>
</table>

\(^a\) Percentage of same residues in same position between template and target structures divided by the number of residues in the template
Figures

Figure 8.1: Multiple sequence alignment of single members of each of the 16 ACP families, roughly arranged in order of similarity to each other. ACP1: *Borrelia burgdorferi*; ACP2: *Schizosaccharomyces pombe*; ACP3: *Gallus gallus*; ACP4: *Bacillus subtilis*; ACP5: *Streptomyces rimosus*; ACP6: *Mycobacterium tuberculosis*; ACP7: *Penicillium patulum*; ACP8: *Aspergillus nidulans*; ACP9: *Dictyostelium discoideum*; ACP10: *Saccharopolyspora erythraea*; ACP11: *Aspergillus terreus*; ACP12: *Escherichia coli*; ACP13: *E. coli*; ACP15: *Klebsiella pneumoniae*; ACP16: *K. pneumoniae*; ACP17: *Lactobacillus rhamnosus*.

Figure 8.2: Phylogenetic tree of ACP families.
Figure 8.3: Tertiary structures of single members of ACP families. ACP1: *Thermus thermophilus* (PDB accession code 1X3O); ACP2: *Saccharomyces cerevisiae* (2UV8); ACP3: *Homo sapiens* (2CG5); ACP5: *Streptomyces roseofulvus* (1OR5); ACP8: *Aspergillus parasiticus* (2KR5); ACP10: *Brevibacillus parabrevis* (2JGP); ACP12: *E. coli* (2FQ1); ACP13: *E. coli* (2ROQ); and ACP17: *Lactobacillus casei* (1HQB).

Figure 8.4: Predicted tertiary structure of an ACP1 member using ACP5 (green) and ACP17 (blue) tertiary structures as templates, compared to a known ACP1 (red) tertiary structure; of an ACP5 member using ACP1 (yellow) and ACP17 (blue) tertiary structures as templates, compared to a known ACP5 (red) tertiary structure; of an ACP17 member using ACP1 (yellow) and ACP5 (green) tertiary structures as templates, compared to a known ACP17 (red) tertiary structure. Known structure templates: ACP1: 2EHS; ACP5: 1OR5; ACP17: 1HQB.
Figure 8.5: Tertiary structure of ACP4 (*Bacillus subtilis* AcpK) predicted by homology modeling and MD. Representative models of tertiary structures of ACP15 (*Malonomonas rubra* malonate decarboxylase ACP) and ACP16 (*E. coli* citrate lyase ACP) predicted by threading.
**Supplementary Information for Chapter 8**

**Acyl carrier protein structural classification and normal mode analysis**

Figure 8S-A: Color map representation of amino acid identities between randomly chosen sequences in two families based on pairwise alignments. Values are shown as a percentage; those above the diagonal are based on the shorter sequence compared, while those below the diagonal are based on the longer sequence compared.

Figure 8S-B: Detailed phylogenetic tree of the 16 ACP families.

Figure 8S-C: Individual residue fluctuations of the five slowest vibrational normal modes for Protein Data Bank (PDB) ACP1 structures 1T8K, 1X3O, 2EHS, 2EHT, 2QNW, ACP5 structure 2AF8, and ACP17 structure 1HQB. Fluctuation scale normalized to largest residue fluctuation within each structure. Residues corresponding with α-helices are highlighted.

Figure 8S-D: Overlap charts between the 20 slowest vibrational normal modes of two structures, 0 being no overlap and 1 complete overlap. Shown for two conformers (2EHS and 2EHT) of a single structure in ACP1, two bacterial structures (1T8K and 1X3O) in ACP1, a bacterial (1T8K) and a eukaryotic (2QNW) structure in ACP1, an ACP1 (1T8K) and ACP5 (2AF8), an ACP1 (1T8K) and ACP17 (1HQB), and an ACP5 (2AF8) and ACP17 (1HQB).

Figure 8S-E: Comparative residue fluctuations of the three most overlapped vibrational normal modes between compared cases. Fluctuation scale normalized to largest residue fluctuation between the two compared structures. See table for corresponding modes in each comparison and color key:

<table>
<thead>
<tr>
<th>Comparison</th>
<th>1st most overlapped modes</th>
<th>2nd most overlapped modes</th>
<th>3rd most overlapped modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two conformers</td>
<td>6 (red) and 6 (red lined)</td>
<td>2 (green) and 1 (green lined)</td>
<td>7 (blue) and 5 (blue lined)</td>
</tr>
<tr>
<td>Two Bacterial</td>
<td>1 (red) and 4 (red lined)</td>
<td>2 (green) and 3 (green lined)</td>
<td>2 (green) and 1 (green dotted)</td>
</tr>
<tr>
<td>Bacterial and Eukaryotic</td>
<td>1 (red) and 1 (red lined)</td>
<td>1 (red) and 3 (red dotted)</td>
<td>6 (green) and 5 (green lined)</td>
</tr>
<tr>
<td>ACP1 and ACP5</td>
<td>1 (red) and 1 (red lined)</td>
<td>2 (green) and 4 (green lined)</td>
<td>2 (green) and 3 (green dotted)</td>
</tr>
<tr>
<td>ACP1 and ACP17</td>
<td>4 (red) and 2 (green lined)</td>
<td>1 (green) and 1 (green lined)</td>
<td>2 (blue) and 3 (blue lined)</td>
</tr>
<tr>
<td>ACP5 and ACP17</td>
<td>1 (red) and 1 (green lined)</td>
<td>2 (green) and 5 (green lined)</td>
<td>2 (green) and 9 (green dotted)</td>
</tr>
</tbody>
</table>

Figure 8S-F: Tertiary structure models for ACP15 and 16 predicted by threading.
Figure 8S-A
Figure 8S-B

ACP12_Escherichia_coli
ACP12_Mycobacterium_tuberculosis
ACP12_Bacillus_subtilis
ACP7_Aspergillus_ochraceus
ACP7_Penicillium_wardii
ACP7_Aspergillus_terreus
ACP8_Aspergillus_ nidulans
ACP8_Xanthoria_elegans
ACP13_Escherichia_coli
ACP13_Salmonella_typhi
ACP13_Mycobacterium_smegmatis
ACP10_Chondromyces_crocatus
ACP10_Bacillus_subtilis
ACP10_Brevibacillus_parabrevis
ACP4_Bacillus_subtilis
ACP4_Burkholderia_pseudomallei
ACP4_Pseudomonas_fluorescens
ACP11_Cochliobolus_heterostrophus
ACP11_Aspergillus_terreus
ACP11_Monascus_pilosus
ACP1_Borrelia_burgdorferi
ACP1_Escherichia_coli
ACP1_Aquifex_aeolicus
ACP17_Clostridium_difficile
ACP17_Bacillus_subtilis
ACP17_Lactobacillus_rhamnosus
ACP6_Mycobacterium_bovis
ACP6_Mycobacterium_leprae
ACP6_Mycobacterium_tuberculosis
ACP5_Streptomyces_roseofulvus
ACP5_Streptomyces_rimosus
ACP5_Streptomyces_coelicolor
ACP9_Dictyostelium_discoideum_1
ACP9_Dictyostelium_discoideum_2
ACP9_Dictyostelium_discoideum_3
ACP3_Gallus_gallus
ACP3_Sus_scrofa
ACP3_Homo_sapiens
ACP2_Saccharomyces_cerevisiae
ACP2_Penicillium_wardii
ACP15_Pseudomonas_aeruginosa
ACP15_Klebsiella_pneumoniae
ACP15_Malonomonas_rubra
ACP16_Escherichia_coli
ACP16_Klebsiella_pneumoniae
ACP16_Malonomonas_rubra
ACP16_Streptococcus_pyogenes
Figure 8S-D
Figure 8S-E

Residue Fluctuations of Two Conformers of the Same Structure

Residue Fluctuations of Two Bacterial Sequences in ACP1
Residue Fluctuations of a Bacterial and Eukaryotic Structure in ACP1

Residue Fluctuations of ACP1 and ACP5
Figure 8S-F

<table>
<thead>
<tr>
<th>ACP15, Model 1</th>
<th>ACP15, Model 2</th>
<th>ACP15, Model 3</th>
<th>ACP15, Model 4</th>
<th>ACP15, Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACP15, Model 6</td>
<td>ACP15, Model 7</td>
<td>ACP15, Model 8</td>
<td>ACP15, Model 9</td>
<td>ACP15, Model 10</td>
</tr>
<tr>
<td>ACP16, Model 1</td>
<td>ACP16, Model 2</td>
<td>ACP16, Model 3</td>
<td>ACP16, Model 4</td>
<td>ACP16, Model 5</td>
</tr>
<tr>
<td>ACP16, Model 6</td>
<td>ACP16, Model 7</td>
<td>ACP16, Model 8</td>
<td>ACP16, Model 9</td>
<td>ACP16, Model 10</td>
</tr>
</tbody>
</table>
CHAPTER 9
CONCLUSIONS

The ThYme database provides a useful source of information on fatty acid and polyketide synthesis enzymes. These enzymes have been classified and organized into families based on amino acid sequence similarity, which can help to predict active sites, catalytic residues, and mechanisms of individual sequences, as well as providing a standardized nomenclature. Family membership was confirmed by multiple sequence alignments, structure superimpositions, and positions of the catalytic residues. Detailed and timely reviews on thioesterases (TEs), ketoacyl reductases, hydroxyacyl dehydratases, enoyl reductases, and acyl carrier proteins (ACP), as well as their products and pathways, were also published.

Normal vibrational mode analysis showed that the dynamic behavior of ACP structures was conserved within families. Also, the tertiary structures of the members of three ACP families were predicted: those of one family were like those of many other ACPs, while those of the other two appear not to be similar.

A TE family whose members determine fatty acid chain length was statistically divided into subfamilies based on primary structure. This guided the choice of sequences for experimental characterization, to explore their substrate specificities and enzyme activity spaces. From the experimental work that followed, sequences that provide previously unknown functional diversities relative to chain length specificity as well as acyl chains that contain additional chemical functionalities were revealed. However, subfamily membership and substrate specificity were not correlated.

TEs had not been studied before with first-principles methods. From quantum mechanics/molecular mechanics metadynamics simulations, the mechanism of a TE was confirmed, the electronic and atomic details of its transition state were revealed, its associated energy barrier was estimated, and new evidence was found suggesting proton-donating catalytic residues previously not considered. The proposed mechanism can be extended to all members of this TE family, and it further confirms that TEs of the HotDog fold perform catalysis differently than other TEs.
APPENDIX

THEORY AND COMPUTATION SHOW THAT ASP463 IS THE CATALYTIC PROTON DONOR IN A MANNOSIDASE

David C. Cantu¹, Wim Nerinckx², and, Peter J. Reilly¹

Department of Chemical and Biological Engineering¹, Iowa State University, and
Department of Biochemistry, Physiology, and Microbiology², Ghent University

Modified from a manuscript published in Carbohydrate Research, 343, 2235-2242, with the same title and authors.

Abstract

It has been difficult to identify the proton donor and nucleophilic assistant/base of endoplasmic reticulum α-(1→2)-mannosidase I, a member of glycoside hydrolase Family 47, which cleaves the glycosidic bond between two α-(1→2)-linked mannosyl residues by the inverting mechanism, trimming Man₉GlcNAc₂ to Man₈GlcNAc₂ isomer B. Part of the difficulty is caused by the enzyme’s use of a water molecule to transmit the proton that attacks the glycosidic oxygen atom. We earlier used automated docking to conclusively determine that Glu435 in the yeast enzyme (Glu599 in the corresponding human enzyme) is the nucleophilic assistant. The commonly accepted proton donor has been Glu330 in the human enzyme (Glu132 in the yeast enzyme). However, for theoretical reasons this conclusion is untenable. Theory, automated docking of α-D-³S₁-mannopyranosyl-(1→2)-α-D-⁴C₁-mannopyranose and water molecules associated with candidate proton donors, and estimation of dissociation constants of the latter have shown that the true proton donor is Asp463 in the human enzyme (Asp275 in the yeast enzyme).

Introduction

Endoplasmic reticulum α-(1→2)-mannosidase I (ERManI, EC 3.2.1.113) belongs to glycoside hydrolase family 47 (GH47). It cleaves the glycosidic bond between two α-(1→2)-linked mannosyl residues by the inverting mechanism and trims Man₉GlcNAc₂ to
Man$_8$GlcNAc$_2$ isomer B. Other GH47 $\alpha$-(1→2)-mannosidases cleave other mannosyl residues to yield other products.

Members of GH47 have an ($\alpha,\alpha$)$_7$-barrel fold (Figure A1) and an active-site calcium ion that is necessary for high enzyme activity and thermostability. Saccharomyces cerevisiae and human ERManI enzymes are 35% similar in primary sequence, but the amino acids involved in catalysis are practically the same. Both enzymes have essentially the same crystal structure, and inhibitor binding causes little conformational change.

In the inverting mechanism, an amino acid residue acting as a nucleophilic assistant/base helps a water molecule to perform a nucleophilic attack on the anomeric carbon. A second amino acid residue aids glycosidic bond cleavage by donating a proton to the glycosidic oxygen atom. However, in ERManI the proton must be relayed by a water molecule, as no active-site carboxyl group is close enough to the glycosidic oxygen for direct proton donation. In addition, all three potential proton-donating groups coordinate water molecules (Figure A2). This is very unusual.

The complexed substrate conformation changes during cleavage. Human ERManI binds the inhibitors kifunensine and 1-deoxymannojirimycin in its subsite −1 in the unusual $^1C_4$ conformation. The glycon of methyl 2-S-(α-D-mannopyranosyl)-2-thio-α-D-mannopyranoside (S-Man$_2$) is bound in the $^3S_1$ conformation, leading to the suggestion that its transition state is a $^3H_4$ conformer, intermediate between $^1C_4$ and $^3S_1$ conformers. An automated docking study indicated that the substrate glycon in yeast ERManI must be in the $^1C_4$ conformation to enter the active site. It then passes through $^3H_2$, $^0S_2$, $^3O_B$, and $^3S_1$ conformations to reach the putative $^3E$ transition-state conformer, structurally adjacent to the $^3H_4$ conformer. After hydrolysis, the β-mannose molecule that had been the glycon finds itself successively in the $^1C_4$, $^1H_2$, and $B_{2.5}$ conformations before being expelled from the enzyme active site.

It has been difficult to identify the ERManI catalytic proton donor and nucleophilic assistant/base. A crystal structure of the yeast enzyme with glycerol in the active site led to two hypotheses: 1) that Glu132 is the nucleophilic assistant to the water nucleophile, and that Asp275 or Glu435, probably the former, is the proton donor; and 2) alternatively and less likely, that Glu435 is the nucleophilic assistant, with Glu132 being the proton donor.
companion study with human ERManI again led to two hypotheses similar to those above: 1) that Glu599 (Glu435 in the corresponding yeast enzyme) is the nucleophilic assistant to Water5, with Glu330 (Glu132) as the proton donor, transmitting a proton through Water8; and 2) that Asp463 (Asp275) is the nucleophilic assistant, with Water17 being the actual nucleophile, with Glu330 as the proton donor (Figure A2). A later study on human ERManI adopted the first hypothesis, suggesting that Arg334 (Arg136) contributed to the general acid function. Work on *Hypocrea jecorina* and mouse GH47 enzymes with more capacious active sites yielded the same conclusions about the catalytic residues. An automated docking study of yeast ERManI did not challenge Glu132 as the proton donor, acting through Water195, and identified Glu435 rather than Asp275 as the nucleophilic assistant to Water5.

The nine invariant yeast ERManI acidic residues were mutated before any crystal structure was available. E214Q, D275N, E279Q, E435Q, and E503Q were not active, whereas D86N, E132Q, E438Q, and E526Q had <2% of the activity of wild-type ERManI. A similar study on a GH47 enzyme from *Aspergillus saitoi* found the activity of E124Q (E132Q in yeast ERManI), E124D, D269N (D275N), D269E, E411Q (E435Q), and E411D as 0.02%, 0.2%, 0%, 1.9%, 0%, and 0.74% of the wild-type enzyme. A third mutagenesis project conducted on human ERManI gave \( k_{\text{cat}}/K_M \) values of 3.5%, 0.1%, 0.0005%, 0.006%, and 0.0003% of the wild-type value for E330Q (E132Q in yeast ERManI), D463N (D275N), E599Q (E435Q), E330Q/E599Q, and D463N/E599Q, respectively.

In summary, identification of the ERManI catalytic proton donor and its associated water molecule is uncertain, because all three potential catalytic carboxyl groups coordinate water molecules, and because mutating each of these groups in yeast, *Aspergillus*, and human ERManI causes loss of all or nearly all activity. Therefore we have in this article considered the theory of GH catalysis and then the relative merits of putative proton donor/water systems, and have followed this with extensive use of computation, both by automated docking to determine orientations of substrate and water molecules, and by estimating pK\(_a\)’s of these groups.
Theory

*Electrostatic transition state stabilization in relation to syn- versus anti-protonation*

Enzyme-catalyzed reactions are mediated by preferential stabilization of the transition state,\(^\text{17}\) and electrostatic factors contribute the most to this stabilization.\(^\text{18-20}\) At the glycoside substitution transition state, the local charge distribution of the glycon ring oxygen atom differs most substantially from that of the ground state or any local minimum conformation. In the latter cases, the ring oxygen atom always bears two fully occupied \(sp^3\) lone pairs, whereas that in a glycoside transition state bears a fully occupied \(sp^2\)-hybrid and an electron-deficient \(2p_z\) orbital that overlaps with the anomeric carbon atom’s antibonding orbital from the partially leaving or incoming groups. Since the ring oxygen atom is sterically relatively accessible, one expects that GHs will strategically position at least one electron-rich functional group (e.g. with a correctly oriented free electron pair) to intercept and stabilize this transient change in the local charge. A search for such a strategically positioned enzyme residue at the syn-A and/or syn-B space quadrants, axially above and/or underneath the ring oxygen atom of the glycon complexed in subsite \(-1\) of GHs from different families, has indeed confirmed that syn-protonators, with their proton-donating carboxyl residues residing in the syn-half-space and close to the glycon’s ring oxygen atom, invariably use the conjugate base of the proton donor for electrostatic transition-state stabilization.\(^\text{21}\) On the other hand, anti-protonators, their proton donors being in the anti-half-space and thus inherently far away from the ring oxygen, contain at least one electrostatic transition-state-stabilizing residue within the syn-half-space.

The Glu330/Water8 system as the putative proton donor

An indication that Glu330 in human ERManI (PDB code 1X9D) may not be the proton donor, with Water8 as the transferer of its proton to the glycosidic oxygen atom, is that this residue is in contact distance (3.05 Å) with Arg334 (Figure A2). The Glu330/Arg334 system is expected to be zwitterionic, with Glu330 deprotonated and Arg334 protonated. Another counter-indication is that the possible proton-transferring Water8 is not semilaterally positioned versus the average ring plane of the S-Man\(_2\) glycon occupying subsite \(-1\) of the 1X9D complex structure, but it is instead near-orthogonally positioned. Indeed, a proton
donor is expected to reside near-laterally, within the *anti*- or the *syn*-half-space (Figure A3), to the glycon occupying subsite –1, as observed in many GH families.22,23 Instead, it is the observed oxygen atom of Water17, associated with the former putative nucleophilic assistant Asp463, that is clearly lateral and *anti*-positioned, at 3.60 Å from S-Man2’s glycosidic sulfur atom, which fits with the following observations.

*The occurrence of the exo-anomeric effect in relation to syn- versus anti-protonation*

In the 1X9D complex, the S-Man2 O5’–C1’–S2–C2 dihedral angle is 73°, and therefore the thioglycosidic bond displays the *exo*-anomeric effect.24 The same is observed for the O5’–C1’–O2–C2 dihedral angles in the best-docked Man2 ligands of this study. This effect stabilizes the glycosidic bond by about 4 kcal/mol,25 which is mainly attributed to a hyperconjugative overlap of the O5’–C1’ antibonding orbital with an antiperiplanar-oriented lone-pair orbital lobe of the glycosidic heteroatom. This lobe is semilaterally positioned versus the glycon and in 1X9D is directed toward Water17, which resides in the *anti*-half-space (Figure A3). It indicates that GH47 enzymes are *anti*-protonators, since protonation of the lone pair that is involved in the *exo*-anomeric effect automatically removes this stabilizing effect en route to the transition state. Oligosaccharides that span subsites –1 and +1 of *anti*-protonating GHs appear to always show their glycosidic bonds in conformations dictated by the *exo*-anomeric effect, with protonation of the lone pair that is involved in it, whereas those complexed in *syn*-protonating enzymes consistently do not show an *exo*-anomeric effect conformation at this subsite junction.21

In the case of *anti*-protonators, where the proton donor is situated semi-laterally versus the sugar’s average ring plane and rather near to the C2 atom of the sugar entity in subsite –1, the protonation necessarily has to occur on a lone pair of the glycosidic oxygen that is also semi-laterally positioned, and this is automatically so when the glycosidic bond resides in a conformation that conforms to the *exo*-anomeric stabilizing effect. In the case of *syn*-protonators, where the proton donor is again situated semi-laterally versus the sugar’s average ring plane but is now rather near to the ring oxygen atom, the β- or α-scissile glycosidic bond must turn clockwise or counterclockwise, respectively, out of the *exo*-anomeric effect conformation, thereby bringing a lone pair of the glycosidic oxygen atom
into a semi-lateral position that is within reach of the syn-proton donor. This is a consequence of the original syn- versus anti-protonation insight by Heightman and Vasella. It is also derived from the correlation of the non-exo-anomeric effect conformation of the glycosidic bond versus syn-protonation that can be repeatedly observed in crystal structures with Michaelis complexes spanning subsites –1 and +1 of syn-protonating glycoside hydrolases, such as in PDB structures 1QJW, 1OVW, 2QZ3, 1W2U, and 1ITC. On the other hand, the exo-anomeric effect is preserved with Michaelis complexes of anti-protonators, such as in PDB structures 1VO3, 1JYW, 1IEX, 4A3H, 1KWF, and 1CKX.

The Asp463/Water17 system as the putative proton donor

The Oδ₁ atom of the carboxyl group of Asp463 in human ERManI (1X9D) is 3.70 Å from Water17 and 2.72 Å from the C4–O atom of the D-mannosyl moiety in subsite +1 (Figure A2). Furthermore, Water17 is in contact distance to the lone pair involved in the exo-anomeric effect. The Oδ₂ atom of Asp463 is 4.24 Å from Water17 and 2.51 Å from the C3–O atom of this D-mannosyl residue. It is thus possible that Asp463 has a double role: 1) as an indirect proton donor, through Water17, to the glycosidic oxygen atom; and 2) as an interactor with the D-mannosyl residue occupying subsite +1.

If Water17 is indeed the transmitter of the proton from Asp463 to the substrate glycosidic oxygen atom, then one of its hydrogen atoms should be pointing toward this glycosidic oxygen atom while one of its lone pairs should be directed toward Asp463. One purpose of this work is to investigate the specific orientation of Water17 by means of automated docking experiments.

The role of the Glu330/Arg334/Water8 system

Within the syn-A space quadrant of the glycon in the S-Man₂ complex with human ERManI, Oε₁ of the nucleophilic assistant Glu599 is 3.54 Å from, and nearly axial to, the ring oxygen atom, a situation analogous to many other anti-protonating α-GHs (Figures A2 and A3). At syn-B the Water8 oxygen atom is 4.32 Å from and axial to the ring oxygen atom. It should be noted that the glycon in the observed complex resides in a ⁸S₁ conformation, whereas in passing to a ⁴H₄- or ⁵E-type transition state this ring oxygen atom will
move even closer to Water8. For Water8 to be an electrostatic transition-state stabilizer rather than the transmitter of a donated proton, one of its free electron pairs should be directed toward the ring oxygen atom, so that it can intercept and stabilize the transient electron-deficient $2p_z$ orbital. This may very well be so, since the nearest neighbors of the Water8 oxygen atom are 1) the ligand’s glycosidic sulfur atom at 3.23 Å; 2) $\text{Oe}_2$ of the likely deprotonated Glu330 at 2.63 Å, each to where the hydrogen atoms of Water8 may very well be pointing; as well as 3) $\text{No}_2$ of the likely protonated Arg334 at 2.73 Å, to where the other lone pair of Water8 may be directed. This suggests that the role of the conserved Glu330/Arg334 system is to specifically orient Water8 for electrostatic transition-state stabilization of the glycon's ring oxygen atom. If this rationale is correct, then it should be possible to reproduce this specific orientation of the hydrogen atoms of Water8 by automated docking, which is the other purpose of this work.

**Computational methods**

*Plan of work*

This project uses automated docking, which we have used previously with yeast ERManI$^{13–15}$ and with several other GHs and lectins, to more surely identify the catalytic proton donor in ERManI. Here we docked substrates and water molecules into the human ERManI crystal structure to determine whether Glu330 paired with Arg334 is mediating proton donation through Water8, or whether it is Asp463 mediating proton donation through Water17. We have supplemented this work with computation to determine the probable protonation states of the putative catalytic residues.

*Automated docking*

We docked ligands using AutoDock 3.0$^{26}$ into the human ERManI crystal structure 1X9D.$^{10}$ The normal ligands used were $\alpha$-D-mannopyranosyl-(1→2)-$\alpha$-D-mannopyranose (Man$_2$), with its glycon and aglycon having $3S_1$ and $4C_1$ conformations, respectively, and a water molecule. Ligands were given the desired three-dimensional conformation and hydrogen atoms were added using PCModel (Serena Software, Bloomington, IN, http://www.serenasoft.com). Man$_2$ was pair-fitted with S-Man$_2$ in the enzyme active site with
PyMOL (DeLano Scientific, Palo Alto, CA, http://pymol.sourceforge.net) so that both had the same coordinates. Then charges were assigned to the ligands using GAMESS.\textsuperscript{27} AutoTors in the AutoDock suite was used to define the ligand torsions. The designations of nonpolar hydrogen atoms, those bonded to carbon atoms, were changed so that the program could differentiate them from the polar hydrogen atoms bonded to oxygen atoms. Hydrogen atoms were added to the enzyme using the WHAT IF\textsuperscript{28} webpage. Charges of each atom were added, as well as solvation parameters. Nonpolar hydrogen atoms were specified. Grid maps with 0.375 Å spacing were created using AutoGrid in the AutoDock package. To calculate the electron-affinity map, AutoGrid assumes that full and fractional charges on atoms are located at their centers. However, this is not so with heteroatoms containing a free electron pair. Oxygen atoms have two local negative charges, both somewhere near the centers of the lone pairs. It is remarkable that AutoDock/AutoGrid and other empirical modeling programs yield reliable dockings, even with such a serious deviation from the real situation.

AutoDock was used to find ligand docking positions in the enzyme active site and to calculate the total binding energy ($E_{\text{Total}}$), while holding the enzyme but not the ligands rigid. This was a two-step process, a global search using the Lamarkian genetic algorithm option followed by a Solis and Wets local search.\textsuperscript{26} Our global searches were stopped after 1000 runs, yielding 1000 possible ligand locations in the enzyme. These were gathered into clusters so that all members of each cluster were within a root mean squared deviation (RMSD) of 1 Å of all other members. After the global search, the best-fitted member of each of six clusters with significant numbers of members was chosen based on its conformation and location in the enzyme active site, and on its $E_{\text{Total}}$ value. Then, to enhance dockings, local minimizations were done on each of the six ligands found by the global search.\textsuperscript{13} The optimally-docked ligand was chosen from the six locally-minimized ones based on the criteria stated above, and final values of the intermolecular energy, $E_{\text{Inter}}$, between ligand and enzyme were determined. This rigorous docking procedure characteristically yields much more negative $E_{\text{Inter}}$ values than those normally attained with AutoDock.

To confirm that AutoDock places molecules in the correct location in the active site of ERManI, we docked S-Man\textsubscript{2} and measured the RMSD between docked and crystal-structure
ligands as 0.23 Å. Also, we docked a Water8 molecule and a Water17 molecule in ERManI with S-Man2 bound in the active site, yielding distances between the oxygen atoms of the docked and crystal water structures of 0.44 Å and 0.53 Å, respectively.

Following this, each regular docking set had the same protocol: S-Man2 in the human ERManI structure was removed and Man2 was docked, followed by a water molecule, either Water8, adjacent to Glu330 and Arg334, or Water17, adjacent to Asp463, into the enzyme active site. This was followed by redocking Man2 and then the water molecule.

In the first docking set, Man2 and Water8 were docked in ERManI while varying the protonation states of Glu330, a putative proton donor, and Arg334, adjacent to it. The enzyme’s putative nucleophilic assistant, Glu599, was deprotonated in all dockings, as was Asp463, the other putative proton donor. All the water molecules in the enzyme were removed except for Water5, which is coordinated by Glu599 and is maintained in its crystal-structure position. There are fifteen possible ways in which Glu330 and Arg334 can be protonated or deprotonated: four in which both are deprotonated (Glu−/Arg0a, Glu−/Arg0b, Glu−/Arg0c, and Glu−/Arg0d), where a proton has been abstracted from each of four positions of the two Arg334 amino groups, one in which Glu330 is deprotonated and Arg334 is protonated (Glu−/Arg+), eight in which Glu330 is protonated and Arg334 is deprotonated (Glu0e/Arg0a, Glu0e/Arg0b, Glu0e/Arg0c, Glu0e/Arg0d, Glu0f/Arg0a, Glu0f/Arg0b, Glu0f/Arg0c, and Glu0f/Arg0d), and two in which both are protonated (Glu0e/Arg+ and Glu0f/Arg+), where a proton is found on each of two Glu330 oxygen atoms.

In the second docking set, Man2 and Water17 were docked in ERManI with varying protonation states of Asp463, the other putative proton donor. Glu330 and Glu599 were deprotonated and Arg334 was protonated. Water5 and Water8 were located in their crystal-structure positions. There are three possible ways in which Asp463 can be deprotonated or protonated (Asp−, Asp0a, and Asp0b), and all three were investigated.

*Computational determination of pKₐ values of potential catalytic residues*

The web-accessible program H++ automatically computes pKₐ values of dissociable groups in macromolecules.²⁹,³⁰ It was used here to estimate the pKₐ’s of Glu330, Arg334, Asp463, and Glu599 in the unliganded human ERManI crystal structures 1FMI⁶ and 1X9D¹⁰.
and in the latter crystal structure when it was complexed with S-Man$_2$. H++ requires a contiguous amino acid sequence; however, 1X9D lacks the coordinates of Pro676, which is near the surface on the opposite side of the enzyme from the active site. These coordinates in 1X9D were restored by taking those from an automatic overlap with 1FMI using the Swiss–PDB viewer.\textsuperscript{31} Conversely, 1FMI lacks the coordinates of Trp389 and Thr390, located on the enzyme surface 15 Å from its active site. The unbroken sequence of the structure was restored by using the residues’ coordinates from an automatic overlap with 1X9D.

The H++ program can process only one ligand within a macromolecule, so all water molecules were removed (solvation effects are implicitly accounted for by the program’s methodology), as were SO$_4^{2-}$, Ca$^{2+}$, and 1,4-butandiol (when it was present). Removal of Ca$^{2+}$, even though it is essential for ERManI recognition of the glycon through the latter’s C2’–OH and C3’–OH groups, should not drastically influence the pK$_a$ values of the residues (the putative proton donors Glu330 and Asp463, Arg334, and the putative nucleophilic assistant Glu599), since Ca$^{2+}$ is sufficiently far away from them. The ligand atom names were indicated as LIGAND to be recognized as such by H++. Default physical conditions were used: a salinity of 0.15, internal and external dielectric constants of 6 and 80, respectively, and a pH of 6.5.

**Results and discussion**

*Docking of Man$_2$ and Water8 with different Glu330 and Arg334 protonation states*

The first docking set (Tables A1 and A2) was designed to study the effect of the different protonation states of Glu330 and Arg334 while docking Man$_2$ and Water8, the latter located between these two amino acid residues and Man$_2$. Many of these protonation states, as when Glu330 is unprotonated, requiring Arg334 to be the proton donor, are unlikely in practice but serve as controls to validate ligand dockings that could indicate successful proton donation to the glycosidic oxygen atom. Exploring all possible protonation states also tests the ability of AutoDock to differentiate between viable and non-viable ones.

A number of criteria can be employed to choose successful protonation states. Among them are 1) significantly negative sums of $E_{\text{inter}}$ for Man$_2$ and Water8 docking; 2) low RMSDs of docked Man$_2$ to crystal-structure Man$_2$; 3) low distances between the docked
oxygen atom in Water8 and the crystal-structure Water8 oxygen atom (no protons appear in crystal-structure Water8); 4) intermediate distances between the nearer proton in docked Water8 and the docked glycosidic oxygen atom (O2 of Man2), between the proton in Glu330 and the docked Water8 oxygen atom, between the oxygen atom of docked Water8 and the ring oxygen atom of the glycon (O5’ of docked Man2), and between a proton in Arg334 and the docked Water8 oxygen atom; and 5) correct docked Water8 orientation. Proton donation by the Glu330/Arg334/Water8 system is suggested if one of Water8’s protons is aimed at the docked Man2 glycosidic oxygen atom and its oxygen atom is facing Glu330 and Arg334.

Ability to electrostatically stabilize the transient and electron-deficient 2pz orbital of the O5’ atom at the transition state is suggested if a lone pair of the oxygen atom in Water8 is aimed at it.

Sums of $E_{\text{Inter}}$ values for Man2 and Water8 docking range from $–174.3$ to $–227.9$ kcal/mol (Table A1). Sums less negative than $–190$ kcal/mol are unlikely to indicate successful proton donation. These are generated by the protonation states Glu$^-$/Arg$^{0b}$, Glu$^-$/Arg$^+$, Glu$^{\alpha}$/Arg$^{0c}$, Glu$^{\alpha}$/Arg$^{0a}$, Glu$^{\alpha}$/Arg$^{0b}$, and Glu$^{\alpha}$/Arg$^{0c}$.

RMSD values of docked Man2 to crystal-structure S-Man2 range from 0.80 to 1.29 Å (Table A1). Relatively large RMSD values (greater than ~1.1 Å) when ERManI is in the Glu$^-$/Arg$^{0a}$, Glu$^-$/Arg$^{0c}$, Glu$^{\alpha}$/Arg$^{0a}$, and Glu$^{\alpha}$/Arg$^+$ protonation states lessen their likelihood of proton donation compared to those with lower RMSD values.

Docking of Water8 yields two different ranges of distances (0.27–0.54 Å and 1.32–2.55 Å) between their docked and crystal-structure oxygen atoms (Table A1). Protonation states, such as Glu$^-$/Arg$^{0a}$, Glu$^-$/Arg$^{0b}$, Glu$^{\alpha}$/Arg$^{0a}$, Glu$^{\alpha}$/Arg$^{0b}$, Glu$^{\alpha}$/Arg$^{0d}$, Glu$^{\alpha}$/Arg$^{0b}$, and Glu$^{\alpha}$/Arg$^+$, having distances in the second range are unlikely to successfully donate a proton through Water8 to the glycosidic oxygen atom.

The distances 1) between the nearer proton of docked Water8 and the O2 atom of docked Man2 range between 1.68 and 4.03 Å; 2) those between the docked Water8 oxygen atom and the proton associated with the Oε2 atom in crystal-structure Glu330 are from 1.73 and 4.78 Å; 3) those between the oxygen atom of docked Water8 and the O5’ atom in docked Man2 range from 3.90 to 5.21 Å; and 4) those between the docked Water8 oxygen atom and the nearer proton associated with the Nω2 atom in crystal-structure Arg334 range from 1.65 and
4.61 Å (Table A1). Values of >4 Å in the first two cases indicate a lesser probability of successful proton donation from Glu330 to Water8 to the glycosidic oxygen atom, while values of >5 Å in the third case suggest that a lone pair of Water8 would be unlikely to electrostatically stabilize the glycon ring oxygen during the transition state, and values of >4 Å in the fourth case suggest a lessened ability of Arg334 to stabilize and orient Water8 or to donate a proton through it to the glycosidic oxygen atom. These criteria suggest that the protonation states Glu–/Arg0a, Glu0e/Arg0a, Glu0e/Arg0b, Glu0f/Arg0a, Glu0f/Arg0b, and Glu0f/Arg+ are less likely candidates for successful proton donation.

Also measured were the distances between the oxygen atom of crystal-structure Water5, the putative nucleophile, and the C1' atom of docked Man2, which it attacks, for the different protonation states. In all cases these values are in an acceptable range between 2.89 and 3.38 Å.

The ligand dihedral angle (O5′–C1′–O2–C2) should be about 70°, indicating the presence of the exo-anomeric effect. Furthermore, the orientation of Water8 is extremely important for proton donation. Its oxygen atom should face Glu330 and Arg334, one of its lone pairs needs to face the ring oxygen atom of Man2, and one of its protons should face the glycosidic oxygen atom. Only two protonation states, Glu–/Arg0d and Glu–/Arg+, satisfy these criteria (Table A2).

Use of the criteria listed above suggests that only one of these fifteen protonation states, Glu–/Arg0d, is a good proton donor candidate. However, in that state Glu330 has no proton to donate and Arg334 is not positively charged, leaving no readily available proton for donation.

Chemical reasoning suggests that the Glu330/Arg334 system can donate a proton only from double-protonated systems such as Glu0eArg+ and Glu0fArg+, with the proton being donated by a protonated Glu330. However, its protonation is not predicted by H+++, as will be noted below. With Glu0eArg+, Water8 is oriented correctly for proton transfer, but it is misoriented for transition-state stabilization. Glu0fArg+ fails many criteria necessary for successful proton donation.
Docking of Man$_2$ and Water17 with different Asp463 protonation states

Man$_2$ and Water17 were docked in the second set, with Asp463 in different protonation states (Tables A2 and A3). Five criteria are important here: 1) the sums of $E_{\text{inter}}$ values for Man$_2$ and Water17 docking should be more negative than $\sim$190 kcal/mol, as before; 2) RMSD values between crystal-structure and docked Man$_2$ should be less than $\sim$1.1 Å, as before; 3) distances between oxygen atoms of crystal-structure and docked Water17 should be less than $\sim$1.1 Å; 4) distances between a proton in docked Water17 and the O2 atom of docked Man$_2$ should be $<4$ Å, between a proton in Asp463 and the oxygen atom of docked Water17 should be $<4$ Å, and between the oxygen atom of crystal-structure Water8 and the O5' atom of docked Man$_2$ should be should be $< 5$ Å; and 5) docked Water17 should be oriented so that a proton is aimed at the O2 atom of docked Man$_2$ and a lone pair is facing Asp463.

Sums of $E_{\text{inter}}$ values in Asp$^-$, Asp$^{0a}$, and Asp$^{0b}$ are $-186.7$, $-193.9$, and $-193.4$ kcal/mol, respectively (Table A3), less negative than in many cases when the protonation states of Glu330 and Asp334 are varied. The least negative energy sum occurs with Asp$^-$, the only state of the three in which proton donation is not possible.

In all three cases, the RMSDs of docked Man$_2$ relative to crystal-structure S-Man$_2$ are under 1.0 Å (Table A3). The distances between the oxygen atoms of crystal-structure and docked Water17 are $\leq 1.1$ Å. Distances between the nearer proton in docked Water17 and the O2 atom of docked Man$_2$ are $\sim 3$ Å, while those between the nearer proton in Asp463, when present, and the Water17 oxygen atom are $\leq 3$ Å, allowing proton donation from Asp463 through Water17 to the Man$_2$ glycosidic bond. Furthermore, distances between Water8’s oxygen atom and the O5' atom range between 4.41 and 4.83 Å, suggesting that Water8 can electrostatically stabilize Man$_2$ during the transition state. Distances between the oxygen atom of crystal-structure Water5 and the C1' atom of docked Man$_2$ are between 3.06 and 3.30 Å in all four cases, an acceptable range.

The orientation in which Water17 docks indicates its probability of proton donation. With Asp$^-$ and Asp$^{0b}$, proton donation cannot occur because Water17 is not correctly oriented so that one of its hydrogen atoms is pointing toward the glycosidic oxygen atom while one of its lone pairs is directed toward Asp463 (Table A2). However, Water17 docks in the correct
position and with the right orientation to donate a proton from Asp\textsuperscript{0a}.

These docking results suggest that the only protonation state of Asp463 eligible to donate a proton to the glycosidic oxygen atom of Man\textsubscript{2} is Asp\textsuperscript{0a}. Since the Glu330/Arg334 system has no protonation states that appear likely to be proton donors, Asp463 appears by docking analysis to be the actual proton donor in ERManI.

**Determination of $pK_a$ values with H++**

Finally, we used H++ to predict the $pK_a$ values and protonation states of the potential catalytic residues in human ERManI (Table A4). Glu330 and Arg334 have predicted $pK_a$'s of near zero and much above 14, respectively, so at physiological pH's the former is deprotonated and the latter is protonated, to the extent that its protons are so strongly bound that they cannot be donated. The putative proton donor Asp463, with a predicted $pK_a$ in the basic range, appears to be mainly or completely protonated, while the putative nucleophilic assistant Glu599, with a predicted $pK_a$ greatly below zero, indeed appears to be completely deprotonated.

H++ uses single-structure continuum solvent methodology, giving an average $pK_a$ error of about one unit, with potentially larger errors at both very negative and very positive $pK_a$ values.\textsuperscript{29} However, it should give reasonable estimates of whether a dissociable group is protonated or not at these extremes. It is clear that movement of amino acid side-chains during substrate binding and catalysis can change their $pK_a$ values. However, Glu330, Asp334, Asp463, and Glu599 are all part of $\alpha$-helices (Figure A2) and therefore have low potential for movement. Furthermore, Arg334 is located one turn on an $\alpha$-helix from Glu330 and will always face in the same direction as the latter despite any movement. These results further confirm, as predicted by theoretical considerations and confirmed by docking analysis, that Asp463 acts as the proton donor in human ERManI.

**Non-suitability of Glu330 as the nucleophilic assistant/base**

With Glu330 eliminated as the putative proton donor, the question arises as whether it could be the nucleophilic assistant instead of Glu599. Glu330 is on the opposite side of the ligand from Water5, the natural nucleophile, preventing contact between them. An alternative
possibility is for Water8 to be the nucleophile with Glu330 as the nucleophilic assistant. This would require that GH47 members hydrolyze substrates through an internal SN$_i$ substitution, with the leaving group departing and the nucleophile replacing it from the same side. Although such a mechanism is known,$^{32,33}$ the leaving group would need to be much better than the mannosyl residue here. Furthermore, a classical base–assistant system rather than the Glu330–Arg334 zwitterionic system would need to compensate for a highly oxocarbenium-type SN$_i$ transition state.

**Conclusions**

We have conducted this project to clearly identify the proton donor in ERManI, and by extension in all GH47 α-1,2-mannosidases. This question has remained open because no carboxyl group is close enough to the C1’ atom of the glycon for direct proton transfer, and because all three carboxyl groups near the substrate’s glycosidic bond coordinate water molecules. In summary, our theoretical, docking, and pK$_a$ prediction studies show that Asp463 is the proton donor in human ERManI. Theoretical considerations based on the ERManI crystal structure$^{10}$ indicate that GH47 enzymes are anti-protonators, and that Asp463 is the only potential proton donor located in the anti-quadrant of the ERManI active site. Only protonated Asp463 allows a water molecule to be positioned to donate a proton to the Man$_2$ glycosidic oxygen atom and to allow another water molecule to electrostatically stabilize the ring oxygen atom of Man$_2$. Glu330 paired with Arg334 is not the proton donor because Glu330 is deprotonated, due to its proximity to Arg334. Arg334 cannot be the proton donor because even when positively charged, it binds protons too tightly for them to be released.

**Acknowledgments**

The authors thank Blake Mertz and Anthony Hill for their help in the initial docking runs, ISU for furnishing its computational facilities for this research, and Professor Arthur Olson of the Scripps Research Institute for supplying AutoDock.
References


### Table A1: Energies and distances from docking Man$_2$ and Water8 into the human ERManI active site

<table>
<thead>
<tr>
<th>Charge status</th>
<th>Man$_2$ docking</th>
<th>Water8 docking</th>
<th>Distance, Å</th>
<th>H(Water8)$\cdot$H(Arg334)$^b$</th>
<th>H(Glu330)$\cdot$O$_2$(Man$_2$)</th>
<th>O(Water8)</th>
<th>O5'(Man$_2$)</th>
<th>O(Water8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E_{\text{inter}}$, kcal/mol</td>
<td>RMSD, Å</td>
<td>$E_{\text{inter}}$, kcal/mol</td>
<td>Distance, Å</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-$ 0$^a$</td>
<td>-151.2</td>
<td>1.18</td>
<td>-49.6</td>
<td>1.73</td>
<td>3.60</td>
<td>—</td>
<td>5.21</td>
<td>4.07</td>
</tr>
<tr>
<td>$-$ 0$^b$</td>
<td><strong>-151.3</strong></td>
<td>0.99</td>
<td><strong>-37.6</strong></td>
<td><strong>1.73</strong></td>
<td>1.98</td>
<td>—</td>
<td>4.19</td>
<td>3.57</td>
</tr>
<tr>
<td>$-$ 0$^c$</td>
<td>-161.7</td>
<td>1.29</td>
<td>-34.3</td>
<td>0.51</td>
<td>3.36</td>
<td>—</td>
<td>4.20</td>
<td>1.87</td>
</tr>
<tr>
<td>$-$ 0$^d$</td>
<td>-162.0</td>
<td>0.80</td>
<td>-33.1</td>
<td>0.40</td>
<td>3.31</td>
<td>—</td>
<td>4.57</td>
<td>2.18</td>
</tr>
<tr>
<td>$-$ 0$^e$</td>
<td>-141.6$^g$</td>
<td>0.87</td>
<td><strong>-33.4</strong></td>
<td>0.38</td>
<td>3.47</td>
<td>—</td>
<td>4.44</td>
<td>1.65</td>
</tr>
<tr>
<td>0$^f$</td>
<td>0$^g$</td>
<td>-182.6</td>
<td>0.94</td>
<td>-32.9</td>
<td>1.81</td>
<td>1.91</td>
<td>2.33</td>
<td>4.33</td>
</tr>
<tr>
<td>0$^f$</td>
<td>0$^h$</td>
<td>-186.4</td>
<td>0.94</td>
<td>-41.5</td>
<td><strong>2.55</strong></td>
<td>2.07</td>
<td>2.81</td>
<td>4.60</td>
</tr>
<tr>
<td>0$^i$</td>
<td>0$^i$</td>
<td>-149.0</td>
<td>1.09</td>
<td><strong>-32.6</strong></td>
<td>0.38</td>
<td>2.45</td>
<td>1.73</td>
<td>4.23</td>
</tr>
<tr>
<td>0$^i$</td>
<td>0$^i$</td>
<td>-152.4</td>
<td>1.06</td>
<td><strong>-39.2</strong></td>
<td><strong>1.32</strong></td>
<td>1.68</td>
<td>1.99</td>
<td>3.99</td>
</tr>
<tr>
<td>0$^i$</td>
<td>+</td>
<td>-169.3</td>
<td>1.10</td>
<td>-31.9</td>
<td>0.41</td>
<td>2.45</td>
<td>1.82</td>
<td>4.76</td>
</tr>
<tr>
<td>0$^j$</td>
<td>0$^j$</td>
<td>-148.2</td>
<td>1.24</td>
<td><strong>-30.6</strong></td>
<td>0.54</td>
<td><strong>4.03</strong></td>
<td>2.85</td>
<td>4.14</td>
</tr>
<tr>
<td>0$^j$</td>
<td>0$^j$</td>
<td>-150.6</td>
<td>1.08</td>
<td><strong>-36.2</strong></td>
<td><strong>1.68</strong></td>
<td>1.82</td>
<td><strong>4.21</strong></td>
<td>4.10</td>
</tr>
<tr>
<td>0$^j$</td>
<td>0$^j$</td>
<td>-140.6</td>
<td>0.99</td>
<td><strong>-33.7</strong></td>
<td>0.27</td>
<td>3.23</td>
<td>3.04</td>
<td>4.28</td>
</tr>
<tr>
<td>0$^j$</td>
<td>0$^j$</td>
<td>-162.6</td>
<td>0.87</td>
<td><strong>-35.1</strong></td>
<td>0.37</td>
<td>3.51</td>
<td>3.09</td>
<td>4.65</td>
</tr>
<tr>
<td>0$^j$</td>
<td>+</td>
<td>-167.8</td>
<td><strong>1.25</strong></td>
<td><strong>-35.0</strong></td>
<td><strong>2.13</strong></td>
<td>1.74</td>
<td><strong>4.78</strong></td>
<td>3.90</td>
</tr>
</tbody>
</table>

$^a$ Proton missing in more distant position of Arg334 No$_2$ atom  
$^b$ Proton missing in nearer position of Arg334 No$_2$ atom  
$^c$ Proton missing in nearer position of Arg334 No$_1$ atom  
$^d$ Proton missing in more distant position of Arg334 No$_1$ atom  
$^e$ Proton associated with Glu330 O$_{\varepsilon}$ atom missing  
$^f$ Proton associated with Glu330 O$_{\varepsilon1}$ atom missing  
$^g$ Bolded numerals signify values that lessen possibility of proton donation  
$^h$ Distance is that to the nearer available proton bound to Arg334 No$_2$ atom
<table>
<thead>
<tr>
<th>Protonation state</th>
<th>$\phi$, degrees$^a$</th>
<th>$\psi$, degrees$^b$</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu-/Arg$^{0a}$</td>
<td>122.9$^c$</td>
<td>-169.7</td>
<td>Water8 proton aimed at Man$_2$ O5'</td>
</tr>
<tr>
<td>Glu-/Arg$^{0b}$</td>
<td>115.9</td>
<td>-145.5</td>
<td>Water8 proton aimed at Man$_2$ O5'</td>
</tr>
<tr>
<td>Glu-/Arg$^{0c}$</td>
<td>124.2</td>
<td>-168.7</td>
<td>Water8 lone pair mainly oriented to Man$_2$ O5'; may improve with $^3H_4$ transition-state conformation</td>
</tr>
<tr>
<td>Glu-/Arg$^{0d}$</td>
<td>81.8</td>
<td>-146.1</td>
<td>Water8 lone pair mainly oriented to Man$_2$ O5'; may improve with $^3H_4$ transition-state conformation</td>
</tr>
<tr>
<td>Glu-/Arg$^{0e}$</td>
<td>117.8</td>
<td>-158.5</td>
<td>Water8 proton aimed at Man$_2$ O5', but misaligned lone pairs of Water8 and Man$_2$ O2$^d$</td>
</tr>
<tr>
<td>Glu-/Arg$^{0f}$</td>
<td>107.6</td>
<td>-150.4</td>
<td>Misoriented Water8</td>
</tr>
<tr>
<td>Glu-/Arg$^{0g}$</td>
<td>108.3</td>
<td>-164.5</td>
<td>Water8 lone pair mainly oriented to Man$_2$ O5'; may improve with $^3H_4$ transition-state conformation</td>
</tr>
<tr>
<td>Glu-/Arg$^{0h}$</td>
<td>122.7</td>
<td>-166.8</td>
<td>Water8 lone pair mainly oriented to Man$_2$ O5'; may improve with $^3H_4$ transition-state conformation</td>
</tr>
<tr>
<td>Glu-/Arg$^{0i}$</td>
<td>66.9</td>
<td>-152.1</td>
<td>Water8 misoriented for transition-state stabilization; oriented correctly for protonation by Glu330 or Arg334</td>
</tr>
<tr>
<td>Glu$^{0j}$/Arg$^{0a}$</td>
<td>121.2</td>
<td>-162.7</td>
<td>Misoriented Water8</td>
</tr>
<tr>
<td>Glu$^{0j}$/Arg$^{0b}$</td>
<td>118.5</td>
<td>-169.5</td>
<td>Misoriented Water8</td>
</tr>
<tr>
<td>Glu$^{0j}$/Arg$^{0c}$</td>
<td>110.2</td>
<td>-165.7</td>
<td>Misoriented Water8</td>
</tr>
<tr>
<td>Glu$^{0j}$/Arg$^{0d}$</td>
<td>72.8</td>
<td>-142.1</td>
<td>Misoriented Water8</td>
</tr>
<tr>
<td>Glu$^{0j}$/Arg$^{0e}$</td>
<td>121.4</td>
<td>-164.5</td>
<td>Misoriented Water8</td>
</tr>
<tr>
<td>Asp$^{0a}$</td>
<td>72.7</td>
<td>-149.9</td>
<td>Water17 lone pair not oriented to Asp463</td>
</tr>
<tr>
<td>Asp$^{0b}$</td>
<td>79.8</td>
<td>-152.4</td>
<td>Favorable orientation for Asp463 proton transfer to Water17</td>
</tr>
<tr>
<td>Asp$^{0c}$</td>
<td>79.5</td>
<td>-154.3</td>
<td>Misoriented Water17</td>
</tr>
</tbody>
</table>

$^a$ $\phi = O5'-C1'-O2-C2$

$^b$ $\psi = C1'-O2-C2-C1$

$^c$ Bold in $\phi$ and $\psi$ columns signifies that ligand is not in exo-anomeric state, but instead is in a near-eclipsed conformation

$^d$ Bold in Comments column signifies traits that lessen possibility of proton donation
**Table A3: Results from docking Man$_2$ and Water17 into the ERManI active site**

<table>
<thead>
<tr>
<th>Charge status</th>
<th>Man$_2$ docking</th>
<th>Water17 docking</th>
<th>Distance, Å</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E_{\text{Inter}}$, kcal/mol</td>
<td>RMSD, Å</td>
<td>$E_{\text{Inter}}$, kcal/mol</td>
</tr>
<tr>
<td>Asp463</td>
<td>−155.2</td>
<td>0.89</td>
<td>−31.5</td>
</tr>
<tr>
<td></td>
<td>−161.4</td>
<td>0.91</td>
<td>−32.5</td>
</tr>
<tr>
<td></td>
<td>−164.0</td>
<td>0.94</td>
<td>−29.4</td>
</tr>
</tbody>
</table>

$^a$ Proton associated with Asp463 δ$_2$ atom missing  
$^b$ Proton associated with Asp463 δ$_1$ atom missing

**Table A4: Dissociation constants of active-site residues in human ERManI**

<table>
<thead>
<tr>
<th>Residue</th>
<th>1FMI</th>
<th>1X9D</th>
<th>1X9D–Man$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu330</td>
<td>−1.9</td>
<td>0.6</td>
<td>1.7</td>
</tr>
<tr>
<td>Arg334</td>
<td>26.9</td>
<td>25.8</td>
<td>23.4</td>
</tr>
<tr>
<td>Asp463</td>
<td>8.5</td>
<td>9.3</td>
<td>13.1</td>
</tr>
<tr>
<td>Glu599</td>
<td>−9.0</td>
<td>−8.5</td>
<td>−15.0</td>
</tr>
</tbody>
</table>
Figures

Figure A1: Human ERManI crystal-structure 1X9D showing complexed S-Man$_2$.

Figure A2: Crystal structure of human ERManI active site. Pink: Glu599 nucleophile; blue: possible proton donors, either Asp463 or instead Glu330 paired with Arg334; yellow: S-Man$_2$; red: oxygen atoms; green: calcium ion. Water5 is between Glu599 and S-Man$_2$, Water8 is between Glu330 and S-Man$_2$, and Water17 is between Asp463 and S-Man$_2$. 
Figure A3: Six-panel illustration of the proposed substitution mechanism conducted by GH47 enzymes. These are idealized two-dimensional projections based on the actual three-dimensional arrangement within the human ERManI 1X9D crystal structure, and they are shown in relation to the anti-A, anti-B, syn-A and syn-B space quadrants. The ring oxygen’s lone pair involved in the substitution process is indicated in black. When Asp463 is deprotonated, Oδ₁ and Oδ₂ of the carboxylate residue are equivalent because of resonance, as are Oε₁ and Oε₂ of Glu330 and Glu559 when the latter are deprotonated; Nω₁ and Nω₂ of Arg334’s guanidino cation are also equivalent.

**Panel 1:** Enzyme-occupied situation before the catalytic event. The D-mannosyl glycon resides in an ALPH-compliant 3S₁-like conformation, with the C1’–O2 glycosidic bond axial and in an antiperiplanar position versus the axial lone pair (in black) of the ring oxygen atom. A lone pair of the glycosidic oxygen atom is antiperiplanar to the C1’–O5’ bond, giving rise to the exo-anomeric effect. The proton on Oδ₁ of Asp463 is transferred to the nearby Water17.

**Panel 2:** The lone pair of O2 that is involved in the exo-anomeric effect acquires a proton from the protonated Water17, which removes this stabilizing effect. The proton donor Asp463 as well as the proton-shuttling Water17 reside in the anti-half-space. This enzyme is therefore an anti-protonator.

**Panel 3:** The ultimate substitution starts. Helped by Glu599 as nucleophilic assistant, Water5 axially attacks the glycosidic bond.

**Panel 4:** The transition state with the D-mannosyl glycon in a 3H₄ (or the adjacent 3E) conformation. The transient 2pₓ orbital (in black) of the ring oxygen atom, overlapping with the breaking and forming bonds, is electron-deficient and is locally accessible by the enzyme. It is electrostatically stabilized from syn-A by Oε₁ of Glu599, and from syn-B by a lone pair from Water8. This water molecule is oriented in the shown position by hydrogen bonding to the remaining lone pair of the glycosidic oxygen atom and to Oε₂ of Glu330, and by a hydrogen bond from Nω₂ of Arg334 to its other lone pair.

**Panel 5:** End of the anomeric substitution. The D-mannose product’s end conformation is the 1C₄ inverted chair. The hydroxyl group of the newly formed hemiacetal is still protonated, and the nearby Oε₁ of Glu599 finally acquires this proton.

**Panel 6:** End situation of the catalytic event. The products subsequently leave the enzyme, and the liberated D-mannose molecule will flip into the ground-state 4C₁ chair in an independent conformational process. Finally, a proton exchange between Glu599 and Asp463 will reset the enzyme for a next catalytic cycle.