REVIEW

Carboxylic ester hydrolases: Classification and database derived from their primary, secondary, and tertiary structures

Yingfei Chen,[1] Daniel S. Black,[2] and Peter J. Reilly[1]*

[1]Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011
[2]Information Technology Services, Iowa State University, Ames, Iowa 50011

Received 6 August 2016; Accepted 12 August 2016
DOI: 10.1002/pro.3016
Published online 17 August 2016 proteinscience.org

Abstract: We classified the carboxylic ester hydrolases (CEHs) into families and clans by use of multiple sequence alignments, secondary structure analysis, and tertiary structure superpositions. Our work for the first time has fully established their systematic structural classification. Family members have similar primary, secondary, and tertiary structures, and their active sites and reaction mechanisms are conserved. Families may be gathered into clans by their having similar secondary and tertiary structures, even though primary structures of members of different families are not similar. CEHs were gathered from public databases by use of Basic Local Alignment Search Tool (BLAST) and divided into 91 families, with 36 families being grouped into five clans. Members of one clan have standard α/β-hydrolase folds, while those of other two clans have similar folds but with different sequences of their β-strands. The other two clans have members with six-bladed β-propeller and three-α-helix bundle tertiary structures. Those families not in clans have a large variety of structures or have no members with known structures. At the time of writing, the 91 families contained 321,830 primary structures and 1378 tertiary structures. From these data, we constructed an accessible database: CASTLE (CArboxylic eSTer hydroLasEs, http://www.castle.cbe.iastate.edu).

Keywords: carboxylesterases; cholinesterases; cocaine esterases; cutinases; lysopholipases; phospholipases; triacylglycerol lipases

Introduction

Carboxylic ester hydrolases (CEHs) catalyze the hydrolysis of ester bonds into alcohols and carboxylic acids, and they are ubiquitous throughout life. Members of this enzyme group that attack different substrates, form different products, and have different names are listed as EC 3.1.1.1 to EC 3.1.1.98, with seven deleted entries.[1] Carboxylesterases (EC 3.1.1.1), triacylglycerol lipases (EC 3.1.1.3), phospholipase A2s (EC 3.1.1.4), lysophospholipases (EC 3.1.1.5), acetylcholinesterases (EC 3.1.1.7), butyrylcholinesterases (EC 3.1.1.8), aminoacyl-tRNA hydrolases (EC 3.1.1.29), and cocaine esterases (EC 3.1.1.84) are the most extensively researched CEHs.

According to the CATH database,[2] many CEHs have standard α/β hydrolase folds, which are composed of three α/β/α layers, with the second β-strand

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Peter J. Reilly, Department of Chemical and Biological Engineering, 2114 Sweeney Hall, Iowa State University, Ames, IA 50011-2230. E-mail: reilly@iastate.edu

being antiparallel to generally seven others in the β-sheet.[3,4] Other CEHs with other types of α/β hydrolase folds have different arrangements of their α-helices and β-strands. Some CEHs have six-propeller folds, which consist of a six-bladed β-sheet with a central axis. Others have four-layer β-sandwich folds, where several antiparallel β-strands are arranged in two β-sheets. Three-solenoid folds are also found in CEH structures; they consist of many parallel β-strands arranged into three β-sheets. The outer-membrane CEHs are commonly found in β-barrel folds.

To this point, there is no systematic structural classification of the CEHs. Such a classification is likely to yield very different results than found with Enzyme Commission (EC) numbering, because enzymes with different EC numbers and different names may have very similar amino acid sequences (primary structures) and three-dimensional (tertiary) structures. This has been demonstrated in the Carbohydrate-Active EnzYme (ThYme) databases and Thioester-Active EnzYme databases,[5,6] as well as elsewhere.

Earlier research has partially covered this topic, but with many fewer primary structures than amassed in this project. The ESTerases and α/β-Hydrolase Enzymes and Relatives (ESTHERs) database[7,8] covers some CEHs, focusing on α/β hydrolase structures. It is not limited to CEHs, but includes other enzymes such as peptidases and thioesterases that have this fold. ESTHER classifies sequences into three levels: blocks, rank 1 families, and rank 2 families, where each block is based on conserved and characteristic parts of sequences. At the time of writing, it had over 40,000 primary structures classified into four blocks, 94 rank 1 families, and 93 rank 2 families further divided from 11 rank 1 families.

The CAZy database has classified 15 homologous families of carbohydrate esterases, which catalyze the de-O- or de-N-acylation of substituted saccharides, by their primary structures. Twelve of these families contain CEHs, mainly acetyl xylan esterases, and mainly with α/β hydrolase folds. At the time of writing, about 32,000 primary structures of carbohydrate esterases were included, of which almost 21,000 were CEHs.

The lipase engineering database (LED)[9] classified three classes and 38 superfamilies with α/β hydrolase folds, of which 16 included lipases and 10 covered other CEHs, by their functions, sequences, and crystal structures. Its founders employed much smaller *E*-values in their use of Basic Local Alignment Search Tool (BLAST)[10] to gather primary structures than used in this work, implying that the family members in LED were more similar to each other than here, perhaps leading to more families. It encompassed 112 homologous families and almost 25,000 primary and over 1100 tertiary structures. However, it has not been maintained since 2009.

The MELDB database sorted microbial carboxylesterases and triacylglycerol lipases by their primary structure similarities.[11] It corresponded to parts of the LED database, but it appears to be no longer available.

The research reported in this article systematically classifies the CEHs by their primary, secondary, and tertiary structure similarities, as opposed to classifying them by their EC numbers. This will cast light on the various ways that CEHs with different structures catalyze the hydrolysis of ester bonds to yield carboxylic acids and alcohols.

Potential Family Identification

CEH families were generally identified by the techniques used for classifying fatty acid synthesis enzymes.[6,12,13] All the primary structures of CEHs, chosen by their EC numbers, with evidence at protein level in the UniProt database[14] were collected. These totaled 752 sequences. The criterion of evidence at protein level is to ensure that wet laboratory experiments had been conducted on these proteins to verify their functions as CEHs. This criterion ruled out most available CEH sequences, mainly those obtained from whole-genome projects, whose functions are putative because their sequences have been compared only with those of known CEHs rather than being verified experimentally.

The collected primary structures were checked on the Pfam database[15] to obtain their catalytic domains only. BLAST was used consecutively to find primary structures similar to these catalytic domains (query sequences) from the National Center for Biotechnology Information's up-to-date nr database,[16] which gathers nonredundant protein sequences from various sources such as the GenBank,[17] Protein Data Bank (PDB),[18] Protein Information Resource,[19] Protein Research Foundation,[20] RefSeq,[16] and Swiss-Prot[14] databases. The threshold *E*-value in BLAST was set to 0.001. Protein sequences with *E*-values lower than 0.001 were regarded as similar enough to the query sequence to be included in one potential family.[10] In-house Python and shell scripts were implemented to automate the process of obtaining catalytic domains of query sequences in Pfam, to conduct BLAST consecutively, and to further analyze the results of structure comparison. All the scripts were run on the Google cloud platform with Linux Cent OS7 installed.

Family Verification

Multiple sequence alignment (MSA), secondary structure comparison, and tertiary structure superposition are the three main techniques to verify membership in the potential families, possibly

merging or splitting them. It is assumed that all members of a family have the same protein ancestor.

A random sample of sequences in each potential family was used to perform MSA with ClustalX 2.1.[21] The alignment is to ensure that these sequences are similar enough, with several positions of amino acid residues conserved along the entire sample. Different potential families gathered by BLAST were subjected to joint MSA to ascertain whether they could be merged into one family. Conversely, if no amino acid residue is conserved and if clear differences are observed in the MSA result, then the potential family was split into two or more families. Occasionally no residue is conserved in what clearly is a family because of a sequence error in one or a few of its members.

Up to 50 tertiary structures from each potential family, if available in the PDB, were superimposed by MultiProt.[22] The monomer of each tertiary structure was extracted and compared. The root mean square deviation (RMSD) of the α-carbon atoms of the different tertiary structures was calculated, together with the P_{avg}, a measure of the percentage of these atoms that are close enough (<4.0 Å) to be compared.[12,13]

A further criterion to verify family membership is that active sites of potential members should remain in similar positions within each family. Also, secondary structure elements, based on the DSSP database[23,24] embedded in the PDB, were compared and analyzed to ensure that potential members of each family have almost the same elements.

Finally, memberships of potential families were manually inspected to confirm that they held a significant number of entries with names and EC numbers specific to CEHs.

Clan Identification and Verification

Clans are composed of two or more different families, where their active sites, reaction mechanisms, and secondary and tertiary structures are conserved from family to family, although their primary structures may not be significantly similar from one family to the next. It is assumed that family members of different clans are more distantly derived from the same protein ancestor than are members of the same family.

We used CATH-defined folds to first divide available tertiary structures in different families into separate groups. We then used two separate procedures to determine membership of families in clans. In the first, tertiary structure representatives from different families with similar tertiary structures were superimposed by MultiProt. Varying from previous methods to calculate RMSD and P_{avg} values of members of all potential families in a clan, pairwise RMSD and P_{avg} values after overlapping representative tertiary structures from pairs of families were calculated. This variation was caused by the large number of families with similar folds, making it difficult to visually distinguish them by PyMOL.[25] The MultiProt superposition, along with RMSD and P_{avg} calculations, were implemented by Python scripts, and RMSD and P_{avg} values were recorded in matrices. To cluster similar structures into potential clans, the pairwise RMSD matrices were imported into MEGA 6.06,[26] and neighbor-joining trees were produced as curved and circular trees. Although MEGA was intended to produce phylogenetic trees for the study of molecular evolution, the pairwise distance matrix used in MEGA is similar enough to be used for RMSD matrices. Potential clans were proposed according to the trees. Then the structures of potential clan members were superimposed and inspected in PyMOL.

In the second procedure, similar secondary and tertiary structures from different families were grouped roughly into potential clans, and then their structures were superimposed by MultiProt and visually inspected by PyMOL. If the superposition were satisfactory, RMSD and P_{avg} values for single representatives of all the potential family members of a potential clan were calculated. The proposed classification was tuned until the structures superimposed in PyMOL were in good alignment and RMSD and P_{avg} values were minimized. Active sites were checked, if available, to see whether the catalytic residues are in similar positions to act on the substrates, and whether they share the same mechanism in each clan.

Interestingly, the initial pairwise alignment procedure did not perform as well as the initial visual inspection procedure. Therefore, the latter technique was chosen to assign families to clans.

Results

BLAST searches using the 752 query sequences yielded 480,148 primary structures of CEHs and other enzymes. In addition, 2101 tertiary structures were gathered from the PDB. The primary structures were classified into 130 potential families.

The membership of each potential family was verified by three methods: MSA of primary structures, secondary structure analysis, and tertiary structure superpositions. The potential 130 families obtained by BLAST became 91 families after MSA using ClustalX, secondary structure analysis, and tertiary structure superposition by MultiProt and PyMOL, and after noting that some potential families had no or very few CEH members (Table I). After these operations, 321,830 primary structures, 1490 sequences with evidence at protein level, and 1378 tertiary structures remained.

The ClustalX sequence alignment files, using 50 representative sequences of each family (or all that were available, if <50), are in Supporting Information

Table I. *Clans and Families of Carboxylic Ester Hydrolases*

Family	Number of sequences	Number of sequences with evidence at protein level	Number of known tertiary structures (representative PDB structures)	Producing organisms[a]	Dominant enzyme names	Dominant EC numbers
Clan A (α/β-hydrolase, three-layer α/β/α sandwich, Rossmann fold, second β-strand antiparallel with sequence 1, 2, 4, 3, 5, 6, 7, and 8)						
1	1216	5	1 (3QIT)	**B**, E[a]	α/β-Hydrolase, esterase, thioesterase	3.1.1.–
2	31,202	41	131 (5ALM)	A, **B**, E	α/β-Hydrolase, 3-oxoadipate enol-lactonase	3.1.1.24
3	26,277	69	53 (3L1J)	A, **B**, E	α/β-Hydrolase, acetylesterase, esterase/lipase	3.1.1.–
4	497	5	10 (4CG1)	**B**	Lipase, triacylglycerol lipase	3.1.1.3
5	2191	5	14 (3FYU)	**B**	Acetyl xylan esterase	3.1.1.72
6	1181	10	2 (3C5V)	B, **E**	Protein phosphatase methylesterase, peroxidase	3.1.1.89, 1.11.1.–
7	1437	8	6 (3D59)	B, **E**	Carboxylic ester hydrolase, platelet-activating factor acetylhydrolase	3.1.1.–, 3.1.1.47
8	730	2	4 (4G4G)	**B**, E	Acetyl xylan esterase	3.1.1.72
9	2432	13	2 (1K8Q)	**E**	Lysomal acid lipase, lipase member M	3.1.1.–
10	3896	3	2 (3HXK)	**B**, E	Xylanase, pectin acetylesterase, esterase	3.2.1.8, 3.1.1.–
11	1359	4	14 (4UYU)	B, **E**	Pectin acetylesterase, protein notum homolog	3.1.1.–
12	24,560	169	280 (2JGJ)	A, B, **E**	Carboxylesterase, carboxylic ester hydrolase, acetylcholinesterase, cholinesterase, cocaine esterase	3.1.1.8, 3.1.1.84
13	4538	6	19 (3ZI7)	**B**, E	Esterase	3.1.1.–
Clan B (α/β-hydrolase, three-layer α/β/α sandwich, Rossmann fold, all β-strands parallel with sequence 2, 1, 3, 4, and 5)						
14	410	5	2 (1YQE)	**A**, E	D-Aminoacyl-tRNA deacylase	3.1.1.96
15	5264	4	7 (1U8U)	**B**	GDSL-like lipase, arylesterase, acyl-CoA thioesterase	3.1.1.–, 3.1.1.2, 3.1.2.2
16	1869	2	13 (1R50)	**B**, E	Lipase	3.1.1.–
17	2960	12	60 (1XZG)	**B**, E	Cutinase, acetyl xylan esterase	3.1.1.74, 3.1.1.72
18	1463	14	10 (1BWQ)	**B**, E	GDSL-like lipase, acetylhydrolase	3.1.1.–
19	2262	6	6 (1DEO)	A, **B**, E	Rhamnogalacturonan acetylesterase, GDSL family lipase, carbohydrate esterase family 12 protein	3.1.1.86, 3.1.1.–
20	1717	15	43 (1EB8)	B, **E**	α/β-Hydrolase, esterase	3.1.1.–
21	1985	5	4 (1ESD)	**B**, E	GDSL family lipase, triacylglycerol lipase	3.1.1.–, 3.1.1.3
22	2374	9	18 (1CVL)	**B**, E	Lipase, lactonizing lipase	3.1.1.3
23	498	8	7 (4X92)	B, **E**	Phospholipase A2, lecithin-cholesterol acyltransferase	3.1.1.4, 2.3.1.43
24	1537	17	15 (4X71)	**B**, E	Lipase, triacylglycerol lipase	3.1.1.–, 3.1.1.3
Clan C (α/β-hydrolase, three-layer α/β/α sandwich, Rossmann fold, first β-strand antiparallel with sequence 1, 3, 2, 4, 5, 6, and 7)						
25	6115	6	19 (1ZJ5)	A, **B**, E	Carboxymethylenebutenolidase, dienelactone hydrolase	3.1.1.45

Table I. *Continued*

Family	Number of sequences	Number of sequences with evidence at protein level	Number of known tertiary structures (representative PDB structures)	Producing organisms[a]	Dominant enzyme names	Dominant EC numbers
26	8649	9	14 (4ETW)	A, **B**, E	α/β-Hydrolase, hydrolase (biotin biosynthesis), carboxylesterase	3.1.1.1
27	4643	15	5 (4UUQ)	A, **B**, E	α/β-Hydrolase, lysophospholipase	3.1.1.5
28	5655	30	9 (3CN9)	**B**, E	Carboxylesterase, phospholipase	3.1.1.1, 3.1.1.–
29	4262	3	18 (1TQH)	A, **B**	Esterase, carboxylesterase	3.1.1.–, 3.1.1.1
30	1818	4	2 (3BF8)	A, **B**, E	α/β-Hydrolase, 3-oxoadipate enol lactonase, 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase	3.1.1.24, 4.2.99.20
31	314	1	14 (1LBS)	B, **E**	Lipase	3.1.1.–
Clan D (six-bladed β-propeller)						
32	8864	16	42 (4GN9)	A, **B**, E	Gluconolactonase, SMP-30/gluconolactone/LRE-like region	3.1.1.17
33	672	18	7 (3SRE)	B, **E**	Serum paraoxonase/arylesterase 2	3.1.1.2, 3.1.1.81
Clan E (three α-helix bundle)						
34	2791	325	288 (1TG1)	**E**	Phospholipase A2	3.1.1.4
35	738	14	1 (1POC)	B, **E**	Phospholipase A2	3.1.1.4
36	269	5	3 (2WG7)	B, **E**	Phospholipase A2	3.1.1.4
Not part of a clan (*α/β-Hydrolase, three-layer α/β/α sandwich, Rossmann fold, various β-strand arrangements*)						
37	1715	15	8 (3ERJ)	A, B, **E**	Peptidyl-tRNA hydrolase	3.1.1.29
38	7996	29	15 (3EB9)	**B**, E	6-Phosphogluconolactonase	3.1.1.31
39	4994	56	12 (1LPB)	B, **E**	Pancreatic glycerol lipase, phospholipase A1	3.1.1.3, 3.1.1.32
40	2593	32	35 (4GBG)	B, **E**	Lipase	3.1.1.–
41	1493	11	1 (2YIJ)	B, **E**	Phospholipase A1, lipase	3.1.1.32
42	1059	11	1 (1CJY)	**E**	Lysophospholipase	3.1.1.5
43	688	2	1 (3KVN)	**B**	Esterase	3.1.1.–
44	869	6	4 (1UWC)	**E**	Diacylglycerol lipase	3.1.1.–
45	11,783	17	43 (4HOY)	**B**, E	Peptidyl-tRNA hydrolase	3.1.1.29
46	43	1	1 (1TIA)	**E**	Lipase	3.1.1.–
47	12,511	8	4 (1CHD)	A, **B**	Chemotaxis-specific regulator protein, protein-glutamate methylesterase	3.1.1.61
Patatin-like fold						
48	11,787	40	4 (4PKB)	B, **E**	Patatin, patatin-like phospholipase family	3.1.1.–
α/β TIM barrel						
49	8906	10	15 (3CL6)	A, **B**, E	Polysaccharide deacetylase	3.1.1.58
50	3765	7	7 (4DI9)	**B**, E	2-Pyrone-4,6-dicarboxylate hydrolase, amidohydrolase	3.1.1.57, 3.5.1.–, 3.5.2.–
Seven-bladed β-propeller						
51	8410	8	5 (3FGB)	A, **B**, E	6-Phosphogluconolactonase, 3-carboxy-muconate cyclase	3.1.1.31, 5.5.1.5

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.