The Lipase Engineering Database – a navigation and analysis tool for protein families

Markus Fischer, Jürgen Pleiss

Corresponding author:

Jürgen Pleiss

Institute of Technical Biochemistry

University of Stuttgart

Allmandring 31

D-70569 Stuttgart, Germany

Phone: (+49) 711-6853191

Fax: (+49) 711-6853196

Email: Juergen.Pleiss@po.uni-stuttgart.de

Keywords:

Integrated database, sequence-structure-function relationship, relational database, sequence alignment, phylogenetic analysis

ABSTRACT

The Lipase Engineering Database (http://www.led.uni-stuttgart.de) integrates information on sequence, structure, and function of lipases, esterases, and related proteins. Sequence data on 806 protein entries are assigned to 38 homologous families, which are grouped into 16 superfamilies with no global sequence similarity between each other. For each family, multisequence alignments are provided with functionally relevant residues annotated. Pre-calculated phylogenetic trees allow navigation inside superfamilies. Experimental structures of 45 proteins are superposed and consistently annotated. The Lipase Engineering Database has been applied to systematically analyze sequence-structure-function relationships of this vast and diverse enzyme class. It is a useful tool to identify functionally relevant residues apart from the active site residues, and to design mutants with desired substrate specificity.

INTRODUCTION

Lipases (triacylglycerol hydrolases E.C. 3.1.1.3) are ubiquitous enzymes which play an important role in lipid metabolism, as they catalyze hydrolysis and synthesis of triglycerides and other water insoluble esters. Microbial lipases are widely used enzymes in biotransformation due to their high stability, activity, regio- and stereoselectivity (1,2). Although lipases belong to many different protein families without sequence similarity, they have the same architecture, the α/β-hydrolase fold (3) and a conserved active site signature, the GxSxG-motif (4). They all share the same catalytic machinery consisting of a catalytic triad (serine – histidine - aspartic or glutamic acid) and the oxyanion hole, formed by the backbone amides of two conserved residues. Experimental structure determination of free enzymes and complexes with substrate-analogous inhibitors (5-7) revealed insights into the structural determinants of enzymatic function, substrate specificity and selectivity (8,9). Other enzymes with the same fold and catalytic machinery are esterases (E.C. 3.1.1) like acetylcholinesterases (E.C. 3.1.1.7), cutinases (E.C. 3.1.1.74), carboxylesterases (E.C. 3.1.1.1), arylesterases (E.C. 3.1.1.2), phospholipases A_1 (E.C. 3.1.1.32), cholinesterases (E.C. 3.1.1.8), and juvenile-hormone esterases (E.C. 3.1.1.59), but also thioesterases (E.C. 3.1.2.14) and non-heme peroxidases (E.C. 1.11.1.10). The lipase family also includes proteins which are homologous to lipases or esterases, but have no enzymatic function like gliotactin, glutactin, neurotactin, neuroligin and thyroglobulin.

The Lipase Engineering Database (LED) has been designed to serve as a navigation tool for systematic analysis of the relationship of sequence, structure, and function of this rapidly growing, highly diverse protein class, and for the design of variants with optimized properties. The LED integrates information on sequence and structure. Proteins are assigned to homologous families and superfamilies based on sequence similarity. To study sequence variation inside and between families, sequence alignments are provided for each family. Functionally relevant residues which are involved in substrate binding or catalysis are annotated. To compare the binding sites, experimentally determined structures are superposed and annotated. A preliminary version of the database (10), although based on manually created HTML files and restricted to a limited number of entries, was useful for a deeper understanding of this protein class (10,11). The new version is based on an extensible data model and a relational database which integrates sequence, structure, and annotation information available in the public databases GenBank (12) and PDB (13).

CONSTRUCTION

Based on the classification of a preliminary release of the LED (10) representative sequences of 37 homologous families were selected. They were used to perform BLAST searches (14) in the non-redundant sequence database at GenBank with a cutoff of $E=10^{-10}$. An automated retrieval system extracted each resulting hit, parsed information on sequence, source organism, protein features, and descriptions, and loaded it into a relational database. From this information, a pre-classification into protein families was achieved. In a second round, protein entries were evaluated by multisequence alignments. All sequences with an overall sequence identity of more than 95 % were assigned to a single protein entry. For each protein entry the longest sequence was selected as a reference sequence for multisequence alignments. For GenBank entries referring to a PDB entry, monomers were extracted from the ExPDB database (15), superposed (11), and secondary structure information was created by DSSP (16).

After data extraction the sequences were analyzed. All sequences with high similarity were assigned to a single homologous family. Homologous families with low, but significant sequence similarity between each other were grouped into a single superfamily. Superfamilies have no significant sequence similarity between each other. Analysis occurred in three subsequent steps: 1) Sequences which were incorrectly assigned to protein entries were reassigned. 2) Based on multisequence alignments and phylogenetic trees of superfamilies, classification of homologous families and assignment of proteins to families was refined. If necessary, proteins were reclassified or new homologous families were created. 3) Finally, extracted annotation information was validated by analyzing conservation inside homologous families and superfamilies, and by comparing to structure information if available. Missing annotation was completed by transferring information among conserved residues in multisequence alignments. Contradictory annotation was removed.

This protocol resulted in a three-level hierarchy (superfamily, homologous family, sequence) with a most complete and consistent annotation of functionally relevant residues.

PROTEIN SEQUENCES

The LED contains 1367 sequences for 806 protein entries with 29 % being putative proteins. For 45 protein entries 150 structure data sets from PDB are available with a total number of 198 protein chains. Proteins were assigned to either of two classes, the GX and the GGGX class, according to sequence and structure of the oxyanion hole (10). The GX class consists of 11 superfamilies and 22 homologous families, which contain 376 protein entries with 600 sequences and 125 chains of known structure. This class comprises mainly bacterial and fungal lipases, eukaryotic lipases (hepatic, lipoprotein, pancreatic, gastric, and lysosomal acid lipases), cutinases, phospholipases and non-heme peroxidases. The GGGX class consists of 5 superfamilies and 16 homologous families including 430 protein entries with 767 sequences and 73 chains of known structure. It comprises bacterial esterases, α-esterases, eukaryotic carboxylesterases, bile-salt activated lipases, juvenile hormone esterases, hormone sensitive lipases, acetylcholinesterases, and thioesterases, as well as gliotactin, glutactin, neurotactin, neuroligin, and thyroglobulin.

ANNOTATED MULTISEQUENCE ALIGNMENTS

For each superfamily and each homologous family, multisequence alignments were generated using CLUSTALW (17) for one representative sequence per protein. For protein structures, all sequence entries were included and displayed with aligned secondary structure information. Proteins were labeled by the accession code of the source database; each label was linked to the original GenBank entry. Annotation information of individual residues (catalytic triad, oxyanion hole, anchor residue, disulfide bridges, scissile fatty acid binding site, alcohol binding site, mutations, glycosylation sites, signal, propeptide, ER retention signal, lid) are visualized by color-coding of the multisequence alignment. Upon selection of a colored residue, information on its position and all annotation information concerning this residue is displayed. Secondary structure information was calculated using DSSP and aligned to the corresponding protein sequence.

PHYLOGENETIC ANALYSIS

Phylogenetic trees for superfamilies and homologous families were calculated from the multisequence alignments using TREE-PUZZLE (18), visualized using PHYLODENDRON, and manually edited. The nodes representing the reference sequences are labeled and linked to the corresponding GenBank entry. A statistical significance is given for each branching by a bootstrap analysis.

The phylogenetic tree organizes the proteins by sequence similarity. However, this does not necessarily coincide with relationships between the source organisms. On one hand, sequences of a homologous family frequently come from very different organisms, like the homologous family for lipase 2 from *Moraxella* which comprises proteins from bacteria, fungi and eukaryotes. On the other hand, lipase and esterase sequences from the same genome are widely spread over the phylogenetic tree, like for *Drosophila melanogaster,* where 85 lipases and lipase-related proteins have been assigned to 5 different superfamilies and 14 homologous families.

ACCESSIBILITY

Information on superfamilies and homologous families, and tables of the protein entries are accessible at *http://www.led.uni-stuttgart.de*. The HTML pages can be accessed by any JavaScript capable WWW browser. Protein tables provide information on the protein name, the source organism, accession codes and links to the GenBank entries, short descriptions for the sequence entries, and links to fully annotated multisequence alignments, phylogenetic trees, and superposed protein structures if available. In addition, BLAST searches can be performed against all sequences represented by the database. The most similar hits and their corresponding homologous family and superfamily are listed in a HTML document and linked to the respective LED protein entries.

APPLICATIONS

A systematic analysis of sequence and structure of lipases has led to a deeper insight into the relation between sequence, structure, and function. From sequence alignment and structure superposition it was observed that the oxyanion hole is highly conserved inside the GX and the GGGX class (10). It could be shown recently that this classification has important consequences for activity. Numerous esterases and lipases of the GGGX class were screened for activity towards esters of tertiary alcohols, and most of them showed activity and even enantioselectivity (19). In contrast, none of the GX class enzymes was active towards this class of bulky substrates. This local pattern has a high predictive value and can be used for distinguishing active from inactive enzymes.

While the residue X in the GGGX class is mostly a hydrophobic residue, it is either hydrophobic or hydrophilic in the GX type (10). However, it is highly conserved inside homologous families and superfamilies. By comparing the orientation of the side chain it was observed that it interacts with a highly conserved "anchor residue" outside the active site. Although it does not contact the substrate, mutation studies (20,21) demonstrated that this residue is part of a functionally relevant network which stabilizes the local geometry of the oxyanion hole. The residues of this network can be identified by systematic comparisons inside protein families and across family borders.

Similarly, results from mutation analysis of substrate binding can be interpreted by annotating all fatty acid binding residues (11). Replacement of F95 which is part of the scissile fatty acid binding site of *Rhizopus* lipase by the more bulky and hydrophilic tyrosine decreases binding of long chain fatty acids (22). More drastically, the binding tunnel of *Candida rugosa* lipase can be completely blocked for fatty acids longer than C6 by replacing an appropriately positioned P246 by phenylalanine (23)

A careful and comprehensive analysis of sequence and structure similarities between lipases and esterases also allows identification of conserved sequence patterns which are specific for a superfamilies, and thus the design of family-specific primers for screening DNA for new members of a superfamily.

The LED system was developed and applied for the class of lipase-related enzymes. However, it has been designed in such a way that is can be easily adapted to investigate other protein families with an even larger number of members. To the expert scientist working with a large protein class, it may serve as a useful tool to organize information, analyze sequence-structure-function relationships, and navigate among highly diverse protein families.

REFERENCES

1. Schmid, R.D. and Verger, R. (1998) Lipases: interfacial enzymes with attractive applications. *Angew. Chem. Int. Ed. Engl.*, **37,** 1608-1633.
2. Kazlauskas, R.J. and Bornscheuer, U.T. (1998) *Biotransformations with lipases*. Wiley-VCH, Weinheim, New York.
3. Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J.D. *et al.* (1992) The α/β hydrolase fold. *Protein Eng.*, **5,** 197-211.

4. Chapus, C., Rovery, M., Sarda, L. and Verger, R. (1988) Minireview on pancreatic lipase and colipase. *Biochimie*, **70,** 1223-1234.

5. Derewenda, U., Brzozowski, A.M., Lawson, D.M. and Derewenda, Z.S. (1992) Catalysis at the interface: the anatomy of a conformational change in a triglyceride lipase. *Biochemistry*, **31,** 1532-1541.

6. Uppenberg, J., Patkar, S., Bergfors, T. and Jones, T.A. (1994) Crystallization and preliminary X-ray studies of lipase B from Candida antarctica. *J. Mol. Biol.*, **235,** 790-792.

7. Lang, D.A., Mannesse, M.L., de Haas, G.H., Verheij, H.M. and Dijkstra, B.W. (1998) Structural basis of the chiral selectivity of Pseudomonas cepacia lipase. *Eur J Biochem*, **254,** 333-340.

8. Scheib, H., Pleiss, J., Stadler, P., Kovac, A., Potthoff, A.P., Haalck, L., Spener, F., Paltauf, F. and Schmid, R.D. (1998) Rational design of Rhizopus oryzae lipase with modified stereoselectivity toward triradylglycerols. *Protein Eng.*, **11,** 675-682.

9. Schulz, T., Pleiss, J. and Schmid, R.D. (2000) Stereoselectivity of Pseudomonas cepacia lipase toward secondary alcohols: a quantitative model [In Process Citation]. *Protein Sci***,** 1053-1062.

10. Pleiss, J., Fischer, M., Peiker, M., Thiele, C. and Schmid, R.D. (2000) Lipase Engineering Database - Understanding and exploiting sequence-structure-function relationships. *J. Mol. Catal. B-Enzym.*, **10,** 491-508.

11. Pleiss, J., Fischer, M. and Schmid, R.D. (1998) Anatomy of lipase binding sites: the scissile fatty acid binding site. *Chem. Phys. Lipids*, **93,** 67-80.

12. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002) GenBank. *Nucleic Acids Res*, **30,** 17-20.

13. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S. *et al.* (2002) The Protein Data Bank. *Acta Cryst.*, **D58,** 899-907.

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

15. Schwede, T., Diemand, A., Guex, N. and Peitsch, M.C. (2000) Protein structure computing in the genomic era. *Res Microbiol*, **151,** 107-112.

16. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22,** 2577-2637.

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

18. Schmidt, H.A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18,** 502-504.

19. Henke, E., Pleiss, J. and Bornscheuer, U.T. Activity of lipases and esterases towards tertiary alcohols: New insights into structure-function relationships. *Angew. Chem.*, (in press).

20. Joerger, R.D. and Haas, M.J. (1994) Alteration of Chain Length Selectivity of a Rhizopus delemar Lipase Through Site-Directed Mutagenenesis. *Lipids*, **29,** 377-384.

21. Beer, H.D., Wohlfahrt, G., McCarthy, J.E.G., Schomburg, D. and Schmid, R.D. (1996) Analysis of the catalytic mechanism of a fungal lipase using computer-aided design and structural mutants. *Protein Eng.*, **9,** 507-517.

22. Atomi, H., Bornscheuer, U., Soumanou, M.M., Beer, H.D., Wohlfahrt, G. and Schmid, R.D. (1996), *Oils-Fats-Lipids, 21st World Congress Int. Soc. Fat. Res.* PJ Barnes & Associates, Bridgewater, pp. 49-50.

23. Schmitt, J., Brocca, S., Schmid, R.D. and Pleiss, J. Blocking the tunnel: engineering of Candida rugosa lipase mutants with short chain length specificity. *Protein Eng.*, (in press).