PREDICTION REPORT

Protein Domain of Unknown Function DUF1023 Is an α/β Hydrolase

Mingzhu Zheng,1 Krzysztof Ginalsoki,2,3,4 Leszek Rychlewski,4 and Nick V. Grishin1,2
1Howard Hughes Medical Institute, University of Texas, Southwestern Medical Center, Dallas, Texas
2Department of Biochemistry, University of Texas, Southwestern Medical Center, Dallas, Texas
3Interdisciplinary Centre for Mathematical and Computational Modeling, Warsaw University, Warsaw, Poland
4BioInfoBank Institute, Poznań, Poland

ABSTRACT Pfam family DUF1023 consists entirely of uncharacterized proteins generated by sequencing the genomes of Actinobacteria (Bateman A., et al., Nucleic Acids Res. 2004;32 Database issue: D138–141.) Utilizing sequence similarity detection methods, we infer homology between DUF1023 and α/β hydrolases. DUF1023 proteins conserve the core secondary structures in α/β hydrolase fold, and share similar catalytic machinery as that of α/β hydrolases. We predict DUF1023 spatial structure and deduce that they function as hydrolases utilizing catalytic Ser-His-Asp triad with the serine as a nucleophile. Proteins 2005;59:1–6.

Key words: structure prediction; Actinobacteria; carboxylesterase; fungal lipase; triad; genomic context

INTRODUCTION

Pfam (version 12.0)1 family DUF1023 (domain of unknown function, accession number: PF06259) consists of 17 uncharacterized proteins derived from several genome sequences of Actinobacteria.2–5 Among the sequenced Actinobacteria, Streptomyces coelicolor and Streptomyces avermitilis are representative soil-dwelling bacteria that produce most of the natural antibiotics important in human and veterinary medicine, and Mycobacterium tuberculosis and Corynebacterium diphtheriae are pathogens that cause tuberculosis and diphtheria. Functional and structural prediction of the uncharacterized gene products is necessary in advancing our understanding of these important bacteria. Sequence similarity detection methods have been used extensively to assign potential functions and structures to new protein sequences from genomic sequencing efforts. Profile-based sequence similarity search tools such as PSI-BLAST6 and HMMER7,8 have increased the sensitivity of homology inference, which allows genome-wide, automated assignments of potential protein functions. In addition, fold recognition methods such as those available through the Meta Server (http://bioinfo.pl/meta) coupled with the 3D-Jury system9 greatly facilitate structural prediction. With the sequencing of complete genomes, the “genomic context” method emerges as a complementary functional prediction tool to homology-based methods. Genomic context method is based on the observation that genes encoding functionally associated proteins share similar selection pressures and tend to be maintained and regulated together. The major genomic context signal comes from gene fusion, conservation of gene-order or co-occurrence of genes in potential operons, and the co-occurrence of genes in genomes.10,11 Nevertheless, human judgment based on biological knowledge is often required in combining sequence and structure analysis methods, as well as genomic context information to study remote protein family relationships.

In this study we infer homologous relationship between DUF1023 and carboxylesterase/fungal lipase using PCMA12 alignment-seeded PSI-BLAST search. Results from sequence study and structure comparison on the basis of the multiple sequence alignment indicate that DUF1023 proteins adopt α/β hydrolase fold and conserve catalytic properties of α/β hydrolases, suggesting that DUF1023 proteins belong to the α/β hydrolase superfamily in SCOP (Structure Classification of Proteins) database.13 The superfamily of α/β hydrolase was defined by a similar α/β/α sandwich-like core structure and common location of the catalytic triad (nucleophile-His-acid).14 The core structure consists of a central β-sheet (five or more mostly parallel β-strands) flanked on both sides by α-helices.14–16 The loop regions connecting C-termini of β-strands and N-termini of α-helices constitute the substrate-binding site. The variability in the loop regions contributes versatile functionalities of α/β hydrolases, such as protease, lipase, esterase, dehalogenase, peroxidase, and...
epoxide hydrolase. Genomic context analysis suggests that two DUF1023 proteins from *Mycobacterium* may be involved in lipid metabolism pathway where they probably function as lipases.

**MATERIAL AND METHODS**

**Sequence Similarity Search and Multiple Sequence Alignment**

The PSI-BLAST program was used to search for homologs of DUF1023 proteins in the NCBI nonredundant database (nr, January 22, 2004; 1,594,288 sequences; 522,190,286 total letters). The e-value cutoff was 0.005 for inclusion of sequences into the profile, and the other parameters were default. To obtain better coverage, found homologs were used to construct global multiple sequence alignment with PCMA program, and the alignment was used to generate a position-specific scoring profile for a new round of BLAST searches. Each sequence from the alignment was used as a query sequence in a single iteration BLAST search against the nr database using the generated profile (e-value cutoff 0.01).

A global multiple sequence alignment (Fig. 1) was constructed using PCMA program for DUF1023 and the two α/β hydrolase families (carboxylesterase and fungal lipase) found by alignment-seeded PSI-BLAST search. The multiple sequence alignment was manually adjusted according to secondary structure predictions by PSI-PRED, sequence-structure mappings by Meta Server (http://bioinfo.pl/meta), local pairwise alignments by PSI-BLAST, Dali structural alignment and Pfam seeds alignments. The full alignment is available at ftp://iole.swmed.edu/pub/hydrolases/.

**Genomic Context and Transmembrane Regions**

Databases MBGD (microbial genome database) and STRING (version 5.1) were used to analyze the conservation of gene neighborhood and potential operons. PSI-BLAST and STRING database were used for gene fusion prediction number: Q9X8R0; NCBI gene identification (gi) number: gi 2122205 from DUF1023 family, PSI-BLAST searches against the NCBI nonredundant protein database (nr, January 22, 2004; 1,594,288 sequences; 522,190,286 total letters; e-value cutoff 0.005) converged on the third iteration to 20 homologs, among which complete and nonredundant sequences are listed in Figure 1. These close homologs are all annotated as hypothetical proteins, and comprise different sets of paralogous members from different *Actinomycetales* species, which suggests that they may utilize similar biochemical properties in different biological pathways. Since PSI-BLAST searches with different query sequences can result in different hits, we carried out an extensive search to ensure completeness in coverage. The 20 found homologs were used to build global sequence alignment by the PCMA program, then each sequence of the 20 homologs was used as a query in a single iteration BLAST search of the nr database using the profile generated from the built alignment and an e-value cutoff of 0.01. We found statistically supported evidence that DUF1023 proteins are homologous to α/β hydrolases. For example, when we query with the sequence gi 25029180, one homolog annotated as phospholipase/carboxylesterase from *Brucella suis* 1330 (gi 23502167) was found with e-value 0.006. When this sequence was used as a query, it found structurally characterized carboxylesterase from *Pseudomonas fluorescens* (PIC, gi 3023719, EC3.1.1.1, PDB: 1aur) in the fourth iteration with e-value 4e-28. The same strategy also directly found a lipase sequence (gi 17564540, e-value 0.005) from *Caenorhabditis elegans* with the query sequence (gi 15842335). PSI-BLAST initiated with this lipase sequence found a triacylglycerol lipase with known structure from *Rhizomucor miehei* (RmL, gi 494918, EC 3.1.1.3, PDB: 5gbl) in the first iteration with e-value 2e-15. PIC and RmL have been classified respectively into carboxylesterase and fungal lipase, two families of α/β hydrolase superfamily in SCOP database. PIC represents the canonical α/β hydrolase fold. RmL has similar core structure to PIC, but differs at peripheral structural elements, namely the β-strand β4 and α-helix αF are missing, an antiparallel β-strand precedes the catalytic histidine, and the N-terminal extension in RmL contributes an α-helix H1 that occupies the place of αF in PIC (Fig. 2). Therefore the indicated homology between DUF1023 and carboxylesterase/fungal lipase helps to assign the fold and the biological functions to proteins in DUF1023.

To further explore the inferred homologous relationship between DUF1023 and α/β hydrolases, we utilized the Meta Server to predict fold for each sequence in the DUF1023 family. Most of the queries found α/β hydrolase fold among their top 3D-Jury hits. For example, when querying with the first sequence in DUF1023 (gi 21222058, residues 1–329), 18 out of the top 20 3D-Jury hits belong to α/β hydrolase superfamily (SCOP c.69.1), and their consensus scores range from 79.6 to 101.6. Importantly, 3D-Jury scores above 50 correspond to essentially correct predictions, meaning that in over 90% of the cases the overall fold of the model is similar to the experimental structure, although certain exceptions with α-helical domains exist. Obtained 3D-Jury results thus strongly support that DUF1023 members are homologous to α/β hydrolases and adopt an α/β hydrolase fold.

Furthermore, a global multiple sequence alignment was constructed between DUF1023 and the two homologous

---

* PSI-BLAST searches that query with the sequence gi 2122205 against the NCBI non-redundant protein database (nr, March 29, 2004; 2,724,447 sequences; 759,694,065 total letters; e-value cutoff 0.01) converged to 21 DUF1023 family homologues; all of them are uncharacterized proteins.
Fig. 1. Multiple sequence alignment of DUF1023 proteins (Pfam database accession number: PF06259) and representative α/β hydrolases (carboxylesterase and fungal lipase families of α/β hydrolases in SCOP database). Each sequence is identified by an NCBI gene identification (gi) number followed by the abbreviation of the species name. The abbreviations for bacteria and eukaryotes are marked in black and red, respectively. The gi numbers of sequences with known structure are underlined. The first and last residue numbers are shown before and after each sequence, and the total sequence length is indicated in parenthesis at the end. Long insertions in loop regions are not displayed; instead the number of omitted residues is indicated in parenthesis. Uncharged residues at dominantly hydrophobic positions are marked in yellow and small residues (G, A, S, T, C, P) at sites containing mostly small residues are shadowed in gray. The conserved catalytic residues S, D and H are highlighted with bold white letter in black background. The residues probably forming the oxyanion hole are marked in bold red letter with numerical letters ("1" and "2") above the corresponding residues for easy referring. The diagrams of the secondary structure elements for the two α/β hydrolase families are shown below the corresponding sequences. The diagrams are drawn according to the structures of Pseudomonas fluorescens carboxylesterase (PIC, PDB: 1aur, chain A, gi:3023719) and Rhizomucor miehei fungal lipase (Rml, PDB: 5tgl, gi:494918): blue cylinders and yellow arrows represent α-helices and β-strands, respectively, a black triangle (▲) denotes specific insertions; the β-hairpins S1–S4 in PIC or the short α-helix H2 functioning as a "lid" in Rml. The PSI-PRED secondary structure predictions (E: strands; H: helix) and reliability scores (9: highest; 0: lowest) for a DUF1023 protein (gi:2122271) are shown below its sequence. The abbreviations of species names: Mbo, Mycobacterium bovis subsp. bovis; MtuC, Mycobacterium tuberculosis CDC1551; MtuH, Mycobacterium tuberculosis H37Rv; Cef, Corynebacterium efficiens YS-314; Cgl, Corynebacterium glutamicum ATCC 13032; Cdi, Corynebacterium diphtheriae; Sav, Streptomyces avermiltis MA-4680; Sco, Streptomyces coelicolor A3(2); Xca, Xanthomonas campestris pv. campestris str. ATCC 33913; Hsa, Homo sapiens; Spo, Schizosaccharomyces pombe; Cel, Caenorhabditis elegans; Cpn, Chlamydomphila pneumoniae CWL029; Bce, Bacillus cereus ATCC 10987; Rmi, Rhizomucor miehei.
families (carboxylesterase and fungal lipase) to validate the suggested relationship (see Materials and Methods). The multiple sequence alignment in Figure 1 excludes residues before β3 and after αF since they are too divergent to generate a reliable alignment. Figure 1 shows that the predicted secondary structure elements of DUF1023 match well the experimentally determined PfC structure; the conserved hydrophobicity pattern and the small residue positions (colored in yellow and gray, respectively, Fig. 1) are consistent with the secondary structure prediction. In the core structure of α/β hydrolase fold, depicted with a typical example (PIC, Fig. 2(a)), β-strands β3 to β7 are shielded from solvent by the α-helices αA to αF; therefore these central β-strands comprise mostly hydrophobic residues (Fig. 1, yellow blocks). The characteristic hydrophobicity pattern of amphipathic α-helices is also mostly conserved among the aligned homologous (Fig. 1). Our prediction indicates that DUF1023 proteins conserve all the fundamental secondary structure elements of α/β hydrolase fold, β3-β8 and αA-αF.

N-Terminal Regions in Some DUF1023 Proteins

DUF1023 proteins have sequence extension ranging from 50 to 340 amino acids N-terminally to DUF1023 domain. Using PSI-BLAST we were not able to link any of these regions to any characterized protein domain, however, five of the N-terminal regions (gi29833238, gi21220842, gi21222171, gi29831122, and gi31793262) show local similarity to each other. We noticed a conserved hydrophobic segment covering ∼22 amino acids at the N-terminus of some of these sequences and the protein gi29831075, annotated as putative membrane protein in Pfam database. After the N-terminal regions of these sequences were submitted to TMHMM Server (version 2.0)21,22 and SignalP Server (version 3.0)23 the hydrophobic segments were all predicted as transmembrane α-helices and possible N-terminal signal peptides.

In conclusion, the N-terminal regions of DUF1023 proteins vary in length and possibly functions; some DUF1023 proteins may have an N-terminal signal peptide causing secretion from the cell, while some may be bound to the membrane through a putative transmembrane α-helix.

Carboxylesterases, Fungal Lipases, and α/β Hydrolases

Carboxylesterases and fungal lipases are close relatives that both hydrolyze carboxylic ester bonds. Carboxylesterases hydrolyze short chain aliphatic and aromatic esters with broad substrate specificity.27 Fungal lipases break down long-chain acyl-triglycerides to di- and monoglycerides, glycerol, and free fatty acids at a water/oil interface.25 Carboxylesterase, fungal lipase and other hydrolytic enzymes with known structures from widely diverse species and with versatile functionalities comprise the α/β hydrolase superfamily in SCOP database.13 The core of canonical α/β hydrolases14 forms a three-layer sandwich. The central β-sheet consists of eight mostly parallel β-strands (order 12435678) with the α-helix.

Fig. 2. (a) Ribbon representation of the carboxylesterase (PIC) structure from Pseudomonas fluorescens (PDB: 1aur, chain A). β-strands and α-helices included in the multiple sequence alignment in Figure 1 are colored in yellow and blue, respectively, except the specific elements αA, αF and β4 are colored in pale green corresponding to the structural replacements in RmL. The β-hairpins S1–S4 insert is colored in orange. The remaining secondary structure elements are colored in white. Annotation for secondary structure follows the tradition of canonical α/β hydrolase for easy comparison. β-strands are labeled numerically, α-helices alphabetically. The residues forming catalytic triad (Ser114, His199 and Asp168 in PIC) are shown in ball-and-sticks. N- and C-termini are labeled. The PIC inhibitor, sulfonitouene, is represented by red sticks. (b) Ribbon representation of the fungal lipase (RmL) structure from Rhizomucor miehei (PDB: 5tgi) in open conformation complexed with an inhibitor, n-hexylphosphonate ethyl ester (red sticks). The catalytic triad includes Ser114, His257, and Asp203 (ball and sticks). The "lid," an α-helix insert H2, is colored in orange. The N-terminal α-helix H1 and β-strand β2 that spatially corresponding to α-helices αA, αF and β-strand β4 in PIC are colored in pale green. The color scheme, representation and annotation for the remaining secondary structure elements are the same as in (a). These diagrams were drawn with MOLSCRIPT program.31
The Active-Site Cleft, Catalytic Triad and Oxyanion Hole of DUF1023 Family Proteins

The conservation of secondary structure elements among DUF1023 proteins, carboxylesterases and fungal lipases indicates that DUF1023 proteins have active-site cleft configuration similar to that of PfC and RmL. In PfC six loops after the carboxyl ends of the parallel β-strands (β3–β8) form the cleft, whose bottom is formed by the three loops after β3, β5, and β6. The β-hairpins [S1–S4, Fig. 2(a)] after β4, and the loops around them, form a wall on one side of the cleft and two other loops after β7 and β8 form the other wall on the opposite side. The β-hairpins appear to block long-chain fatty acid esters from binding to PfC active site [Fig. 2(a)]. DUF1023 proteins have the corresponding six loops except that the loop after β4 is shorter and the β-hairpins are absent. In the RmL structure, a short α-helix insert H2 after β3 functions as a "lid" that covers the active site when the enzyme is in closed conformation and opens in the presence of lipid aggregates, a phenomenon called interfacial activation [Fig. 2(b)]. There is no corresponding secondary structure element in DUF1023 proteins that can function as the "lid." Thus compared to PfC and RmL, the DUF1023 protein structures may have a relatively open and more shallow active-site crevice.

The most conserved residues throughout the aligned homologs are the serine after β5, the aspartic acid after β7 and the histidine after β8 (highlighted with black background, Fig. 1). These three residues constitute the catalytic Ser-His-Asp triad in α/β hydrolases. As shown in PfC structure [Fig. 2(a)], the β-sheet is twisted and the loops nesting the catalytic triad are arranged to bring the catalytic residues near each other to make hydrogen bonds. In DUF1023 family, the "nucleophilic elbow" is constructed by [Cys/Gly][His/Tyr]SerTyrGly, which has been observed as the most conserved motif in PSI-BLAST based on the proximity of their open reading frames (ORF). For example, a DUF1023 protein encoded by ORF Mb2105 (gi:31793262) and a putative lipoprotein LppJ (ORF: Mb2106) in Mycobacterium bovis subsp. bovis are transcribed as a unit; and in Mycobacterium tuberculosis, a DUF1023 protein encoded by ORF Rv2797C (gi:15842335) and a putative lipoprotein LppV (ORF: Rv2796C) are in one transcription unit.

Lipoproteins are characterized by a lipidated N-terminus that is presumed to anchor lipoproteins into the outer leaflet of the cell membrane. The common lipoprotein biogenesis pathway in bacteria uses phospholipids. Phospholipids are degraded through specific lipases. Based on the gene proximity of DUF1023 and some lipoproteins, we infer that DUF1023 proteins may be lipases that mediate the concentration of their substrates (e.g. phospholipids) that are acyl donors in the lipoprotein biosynthesis pathway.

ACKNOWLEDGMENTS

We thank Dr. Lisa Kinch for critical reading of the manuscript and helpful discussions. This work was supported by NIH grant GM67165 to NVG.

REFERENCES