Database, 2018, 1–15
doi: 10.1093/database/baz122
Database tool

Database tool

UPObase: an online database of unspecific peroxygenases

Muniba Faiza[1], Dongming Lan[1], Shengfeng Huang[2,3,*] and Yonghua Wang[1,*]

[1]School of Food Science and Engineering, South China University of Technology, Wushan road, Tianhe district, Guangzhou 510640, Guangdong province, China. [2]Laboratory for Marine Biology and Biotechnology, Pilot National Laboratory for Marine Science and Technology (Qingdao) 1 Wenhai road, Aoshanwei, Jimo, Qingdao, Shandong, 266237, China and [3]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, No., 135, Xingang Xi road, Guangzhou, 510275, China

*Corresponding author: School of Food Science and Engineering, South China University of Technology, Guangzhou 510640, Guangdong Province, China. Tel/Fax: 86-020-87113842; Email: yonghw@scut.edu.cn

Citation details: Faiza,M., Lan,D., Huang,S. *et al.* UPObase: an online database of unspecific peroxygenases. *Database* (2019) Vol. 2019: article ID baz122; doi:10.1093/database/baz122

*Correspondence may also be addressed to Shengfeng Huang. Email: hshengf2@mail.sysu.edu.cn

Received ; Revised 23 August 2019; Accepted 17 September 2019

Abstract

There are many unspecific peroxygenases (UPOs) or UPO-like extracellular enzymes secreted by fungal species. These enzymes are considered special in their ways of catalyzing a wide variety of reactions such as epoxidation, peroxygenation and electron oxidations. This enzyme family exhibits diverse functions with thousands of UPOs and UPO-like sequences. These sequences are difficult to analyze without proper management tool and therefore desperately calls for a unified platform that can aide with annotation, classification, navigation and easy sequence retrieval. This prompted us to create an online database called Unspecific Peroxygenase Database (UPObase) (upobase.bioinformaticsreview.com) which currently includes 1948 peroxygenase-encoding protein sequences mined from more than 800 available fungal genomes. It provides information such as classification and motifs about each sequence and has functions such as homology search against UPObase sequence analyses such as multiple sequence alignments and phylogenetic trees. It also provides a new sequence submission facility. The database has been made user-friendly facilitating systematic search and filters. UPObase allows users to search for the sequences by organism name, cluster ID and accession number. Notably, in our previous study, 113 UPOs were classified into five subfamilies (I, II, III, IV and V) and an undetermined group (Pog) which remain established. In this study, using 1948 UPOs in our database, we were able to further identify six novel sub-superfamilies (Pog-a, Pog-b, Pog-c, Pog-d, Pog-e and Pog-f) with signature motifs and two distinct groups in Subfamily I and III, Ia and Ib, IIIa and IIIb, respectively. With the novel UPO-like sequences and classification, UPObase may serve for researchers working in the area of enzyme engineering and related fields.

Database URL: upobase.bioinformaticsreview.com

Introduction

Unspecific peroxygenases (UPOs) (EC 1.11.2.1) represent the oxidoreductase sub-subclass of heme-thiolate proteins obtained from fungal species (1). Fungal UPOs catalyze a wide variety of reactions such as epoxidation, dealkylation, hydroxylation, one- and two-electron oxidations and oxidation of aromatic and heterocyclic compounds, inorganic halides and organic heteroatoms (2–4). Fungal UPOs are considered as intriguing enzymes because of their various significant properties such as stability, specificity, catalytic activity, high specific activity, water-soluble nature and capability of catalyzing reactions using inexpensive peroxides and cofactors such as Mg^{2+}. Therefore, the UPOs are also termed as the 'ideal biocatalysts for (sub)-terminal hydroxylation of short-chain and medium-chain alkanes under mild conditions' (5).

Some UPOs which are known to date with experimental evidence include *Agrocybe aegerita* UPO (*Aae*UPO), *Marasmius rotula* (*Mro*UPO) and *Coprinellus radians* (*Cra*UPO), among which the protein crystal structure of *Aae*UPO (2YOR) and *Mro*UPO (5FUJ) is only available to date. UPOs are classified as heme-thiolate peroxidases (HTPs) due to their heme-ligation bond with *cysteine* and their similarity with other HTPs known as chloroperoxidases (CPOs). CPOs exhibit strong peroxidase activity but show less peroxygenase activity. There are existing known conserved motif patterns responsible for the catalytic activities of UPOs and CPOs (i.e. -PCP-EGD-R-E and -PCP-EHD-E, respectively) (6,7). However, in the preliminary publication, UPOs have been classified on the basis of phylogeny and sequence motifs, into five subfamilies and a superfamily which includes *Mro*UPO and some CPOs showing an intermediate behavior between the peroxygenases and peroxidases (8). But there are many other UPOs or UPO-like sequences which were not included in the previous analysis and thereby not classified under any known subfamilies and superfamily.

UPOs are considered intriguing enzymes, which could also possess some other necessary functions which may have not been discovered yet due to their limited information. There are many UPOs existing in the fungal kingdom with a wide range of activities but lack a proper classification and annotation for their systematic analysis. Therefore, in this study, core sequences of UPOs obtained in the previous study are used to search for more UPOs and organized into a proper classified system. Further, on the basis of new data obtained, the sequences are subclassified based on their phylogeny and sequence motifs, thereby constituting the Unspecific Peroxygenase Database (UPObase).

Sequence databases such as GenBank (9), Ensembl Fungi (10), MycoBank (11), EPPO-Q-Bank (12) archive information on nucleotide and protein sequences. Specialized databases use them as primary data, for instance, Pfam (13) which classifies sequences into families. Similarly, UPObase is a more specialized database consisting of protein sequences obtained by genome mining of all fungal genome sequences present in Ensembl Fungi (10). The sequences have been classified into new subfamilies and superfamilies based on their phylogenetic studies and motif patterns in their sequences. Some other enzyme-based databases exist such as Lipase Engineering Database (14) which provides information about lipases including their sequences and structures, PeroxiBase (15) which is a peroxidase database which is dedicated to peroxidases and other oxidoreductase enzymes and MEROPS (16) which is dedicated to peptidases. However, the comprehensive enzyme information system called BRENDA (17) is composed of multiple enzymes including their nomenclature and inhibitors but lacks information on UPOs. Any single database dedicated to the UPO enzyme is not available to date which can provide sequence details, submission portal and real-time sequence analyses. UPObase is the only all-UPO protein sequence database designed to perform a systematic analysis of sequence, function and phylogenetic relationships for these extracellular proteins found in fungi. Besides, this database provides more sequences along with detailed information which may help in discovering new potential functions of UPOs and study their physiological role in fungi. The sequences in UPObase are assigned to their corresponding subfamilies and superfamilies along with their signature motif patterns for their easy identification.

Methods

Genome sequence retrieval

A set of fungal genomes constituting 812 different species (or strains) were downloaded from the Ensembl Fungi genome database via FTP (ftp://ftp.ensemblgenomes.org/

pub/) (10). The genome sequences consist of a large number of peptide sequences. These sequences were used as primary data which were further subjected to mining composed of various filters.

Phylogenetic analysis

The phylogenetic analysis was carried out using MEGA7 software (18). A best-fit model to the data was selected using the PROTTEST3 (20) program. It recommended WAG+G+F, namely, WAG (19) amino acid substitution matrix, gamma distribution (under four rate categories) and empirical amino acid frequencies. Maximum likelihood trees were constructed with a bootstrap replicate of 300 using the same model.

Real-time sequence analyses

The multiple sequence alignments (MSAs), phylogenetic trees and their corresponding percent identity matrix (PIM) are generated in real time for each user query. A simple MSA was generated using Clustal Omega (21), and color-coded alignment was generated using MUSCLE (22). It uses Erik Sonnhammer's Belvu Editor (23) to color the alignment. The phylogenetic neighbor-joining (NJ) trees and their corresponding PIM are generated using ClustalW2 (24), and PhyD3 JavaScript (25) has been implemented for tree visualization in the form of phylogram or dendrogram.

Database construction

A set of previously found 113 UPO encoding sequences belonging to different subfamilies and a superfamily were used to find more sequences using an improved pipeline which we created on the basis of our previous study. The database construction is based on an iterative process of searching for UPO encoding sequences for each and every new sequence that appeared in preliminary searches (Figure 1). In the first round, each of the core UPO sequences was used as a query for similarity search using PHMMER (http://hmmer.org; version 3.1b2) against the generated fungal genome database with an E-value and an inclusive E-value set to 10.0 and 0.01 respectively, providing ~1 false positive in every 100 searches. The output sequences were clustered using cd-hit software (26) at the 90% similarity cutoff and a word length of five residues. The resultant sequences were further clustered using graph-based clustering software MCL (27) at an inflation value set to 1.4. The obtained clustered sequences were then searched for sequence motifs corresponding to their subfamily type resulting in a large number of sequences. This step is repeated for each new sequence that appeared in the similarity search. In the second round, in order to reduce the redundant sequences, the resultant sequences were subjected to sequence-based clustering again at a 95% similarity cutoff providing a total of 1948 clusters. The representative sequences from each cluster which represent the operational taxonomical units (OTUs) were selected and then further analyzed which resulted in the reclassification of UPOs. Finally, we obtained 1948 total UPO encoding sequences (including *Aae*UPO, *Mro*UPO and *Lfu*CPO) constituting the database.

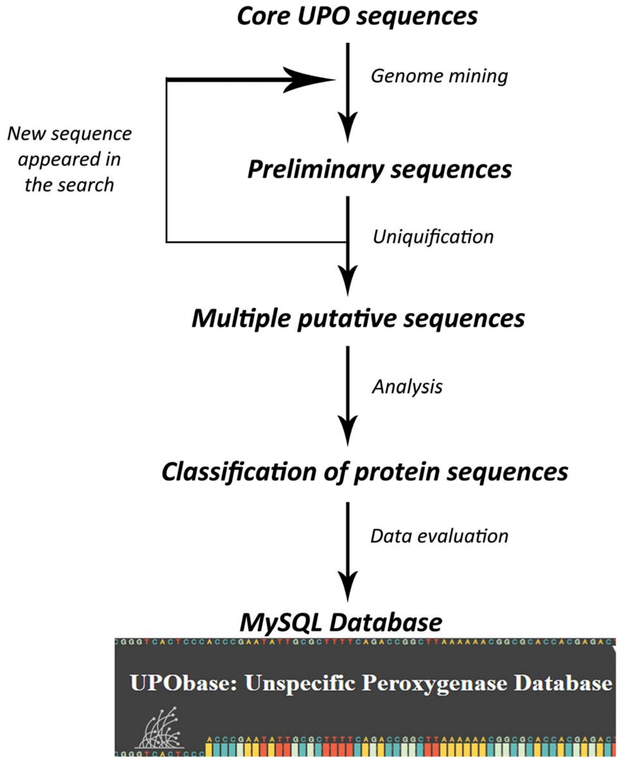
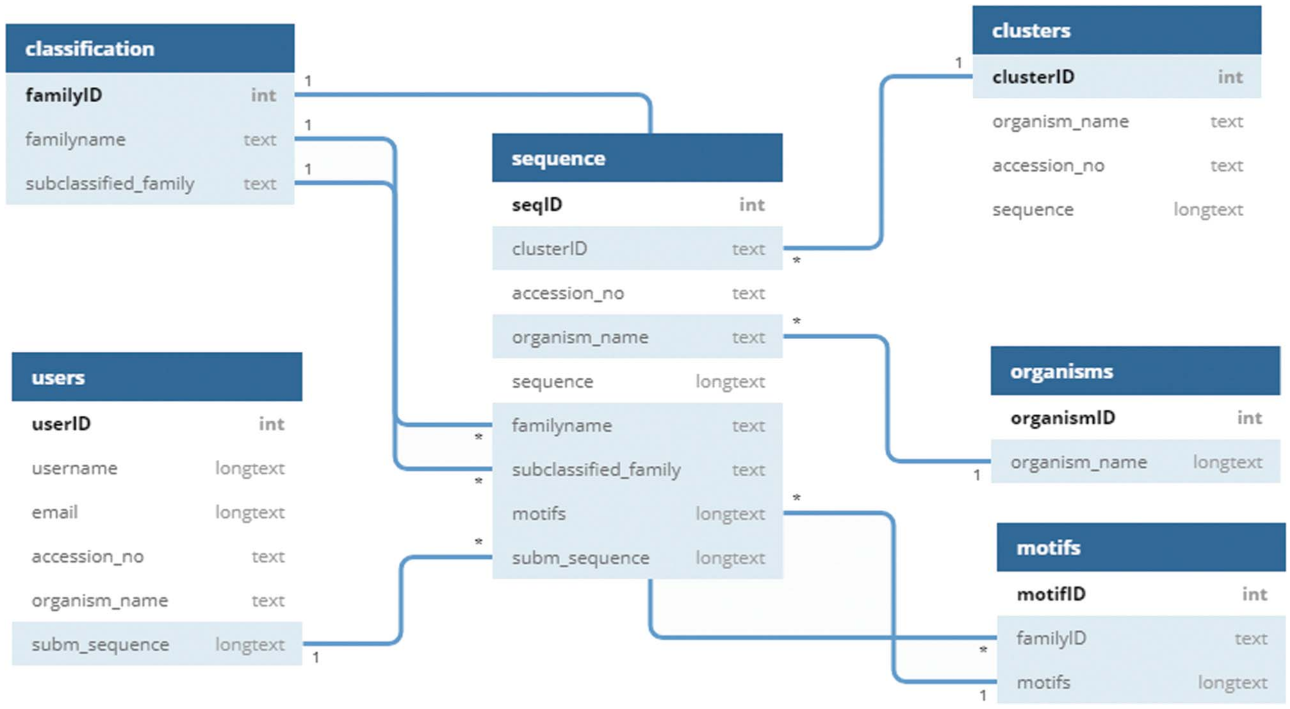

Core UPO sequences
Genome mining
New sequence appeared in the search
Preliminary sequences
Uniquification
Multiple putative sequences
Analysis
Classification of protein sequences
Data evaluation
MySQL Database
UPObase: Unspecific Peroxygenase Database

Figure 1. A scheme involved in the database development process.

Results

Sequence identification

To create the UPObase, we used a pipeline to search for UPOs and UPO-like sequences. This pipeline involves homology search refined with various filters such as blast, sequence-based and graph-based clustering and motif search. Additionally, the filtered sequences were again subjected to sequence-based clustering coupled with phylogenetic analysis in order to remove non-UPO sequences. Therefore, these sequences represent a complete and reliable set of UPOs or UPO-like protein sequences obtained from an *in silico* filtering including clustering, motif search and phylogenetic analysis. After a thorough sequence and phylogenetic analyses, these sequences were



classification	
familyID	int
familyname	text
subclassified_family	text

sequence	
seqID	int
clusterID	text
accession_no	text
organism_name	text
sequence	longtext
familyname	text
subclassified_family	text
motifs	longtext
subm_sequence	longtext

clusters	
clusterID	int
organism_name	text
accession_no	text
sequence	longtext

users	
userID	int
username	longtext
email	longtext
accession_no	text
organism_name	text
subm_sequence	longtext

organisms	
organismID	int
organism_name	longtext

motifs	
motifID	int
familyID	text
motifs	longtext

Figure 2. Schema of UPObase.

found to be exhibiting different motif patterns which led to their subclassification. The main purpose of UPObase is to provide a unified platform for systematic analysis of UPOs. Currently, the database consists of a complete set of 1948 protein sequences of UPOs or UPO-like extracted from 812 fungal genomes.

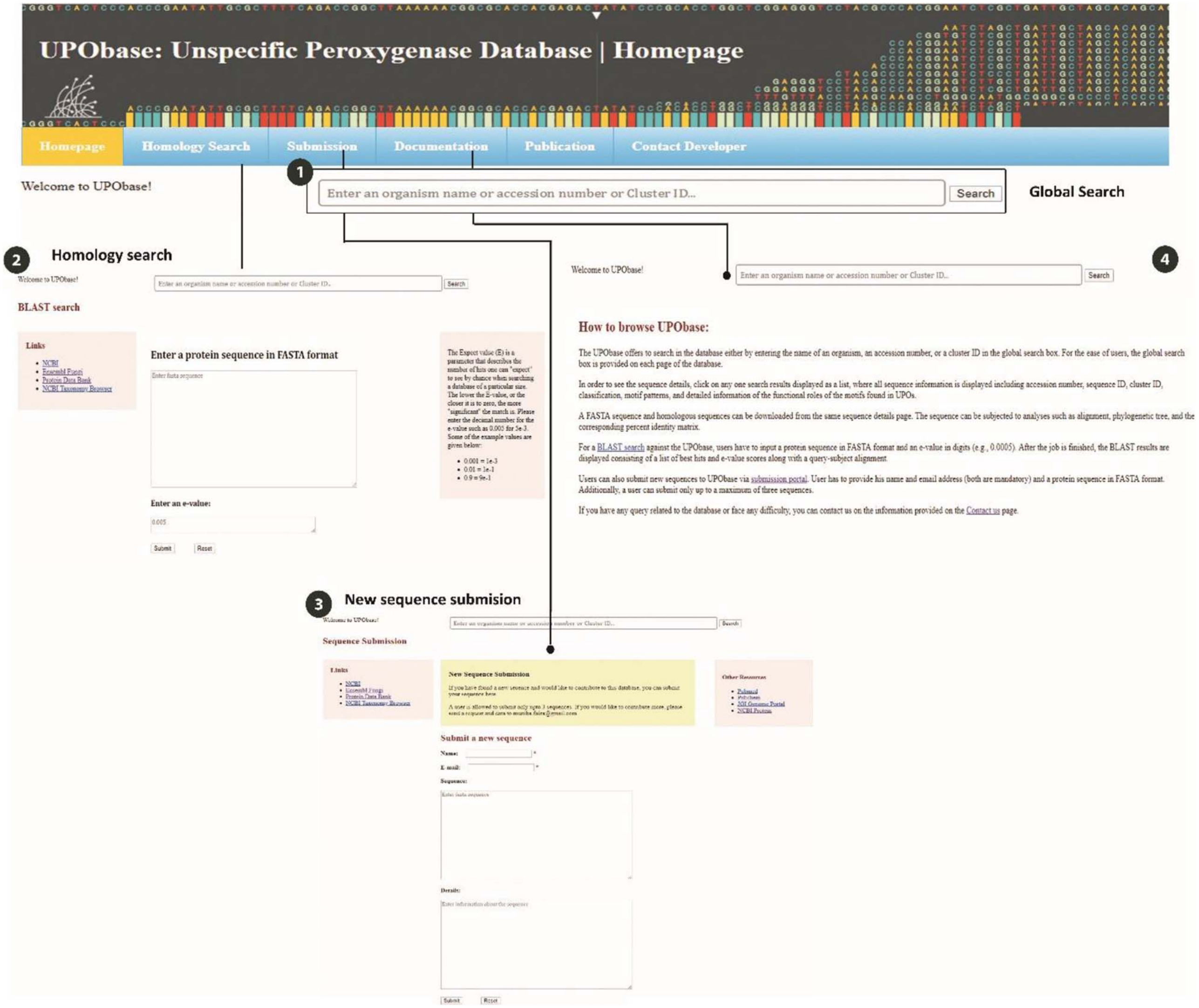
Database architecture

UPObase is a relational MySQL database, and its complete architecture is explained in Figure 2. It involves two different layers of sequences: UPOs and UPO-like sequences (thousands of sequences) > clustered highly similar sequences (1948 sequences with 95% sequence similarity). This helped to remove the redundant and insignificant sequences from the database. These two layers are linked together with the cluster IDs. Each cluster consists of various sequences sharing 95% and above similarity (layer 1), and a representative sequence from each cluster is selected for the next layer of sequences (layer 2).

The information regarding classification, motifs, organisms, and sequences is stored in separate tables linked to each other. The clusters with a specific ID are stored in a table, and only the representative sequence (OTU) is added in the sequence table with a linked cluster ID. The motifs are linked with the family and sequence tables where a motif pattern is assigned for each sequence depending upon its classification. The user-submitted sequences and other related information is stored in a separate table which will be added into the sequences after the validation and classification.

Web interface

The UPObase is available online at upobase.bioinformaticsreview.com, and its complete web interface is explained in Figure 3. The webpages can be easily accessed on any PHP and JavaScript supporting web browsers. A global search bar is given on each page to allow users to browse the database by any organism name, accession number or cluster ID which provides a list of entries in the database along with its sequence length and a direct link to download its FASTA sequence (Figure 4). A user can easily get all the information about any sequence by clicking the link. The details for each sequence include sequence ID, cluster ID, accession number, organism name, database source (from where the genome was downloaded), the sequence and the sequence features including sequence length, family, sub-subfamily, motif pattern and the tables which describe the functional role of motifs in detail. The sequence FASTA and corresponding homologous FASTA can be downloaded from the section provided in the right (Figure 4 (2)). In order to study the relationship among the other UPO-encoding sequences, real-time generated alignments and phylogenetic trees of each sequence are provided. The similarity among the homologous sequences can be seen in the real-time generated PIM corresponding to the alignment and the tree. Documentation provides information on browsing



The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The lower the E-value, or the closer it is to zero, the more "significant" the match is. Please enter the decimal number for the e-value such as 0.005 for 5e-3. Some of the example values are given below:

- 0.001 = 1e-3
- 0.01 = 1e-1
- 0.9 = 9e-1

New Sequence Submission

If you have found a new seuence and would like to contribute to this database, you can submit your sequence here.

A user is allowed to submit only upto 3 sequences. If you would like to contribute more, please send a request and data to [illegible]

The UPObase offers to search in the database either by entering the name of an organism, an accession number, or a cluster ID in the global search box. For the ease of users, the global search box is provided on each page of the database.

In order to see the sequence details, click on any one search results displayed as a list, where all sequence information is displayed including accession number, sequence ID, cluster ID, classification, motif patterns, and detailed information of the functional roles of the motifs found in UPOs.

A FASTA sequence and homologous sequences can be downloaded from the same sequence details page. The sequence can be subjected to analyses such as alignment, phylogenetic tree, and the corresponding percent identity matrix.

For a BLAST search against the UPObase, users have to input a protein sequence in FASTA format and an e-value in digits (e.g., 0.0005). After the job is finished, the BLAST results are displayed consisting of a list of best hits and e-value scores along with a query-subject alignment.

Users can also submit new sequences to UPObase via submission portal. User has to provide his name and email address (both are mandatory) and a protein sequence in FASTA format. Additionally, a user can submit only up to a maximum of three sequences.

If you have any query related to the database or face any difficulty, you can contact us on the information provided on the Contact us page.

Figure 3. An overview of the utilities of UPObase. (1) A global search box displayed at every page of the database to allow browsing convenient; (2) BLAST search feature where a user can enter any sequence and find homologous sequences corresponding to the input; (3) a new sequence submission portal; and (4) documentation page for help.

the database. In case of any difficulty, users can contact by sending an email provided at the contact information page.

Database utility

Sequence retrieval The sequences from UPObase can be easily retrieved either by entering an organism name, or accession number or a cluster ID. As shown in Figure 4, if a user searches for a term, for example, '*Sphaerobolus*', as a result, it will provide a complete list of the given entries in the database along with their sequence ID, sequence length and a direct link to download its FASTA sequence (Figure 4 (2)). The FASTA sequence and corresponding homologous FASTA sequences can also be downloaded from the sequence details page via the links given in the top right corner (Figure 4 (4)). If a user searches UPObase by providing a cluster ID, sequences belonging to that cluster will be displayed as a list, and it may include different organisms. If a user browses by an accession number, only a sequence linked to this accession number will be displayed. In case of any difficulty, users can refer to the examples for browsing UPObase that are provided on the documentation page with screenshots. The corresponding homologous sequences in FASTA format can be downloaded by exploring these entries displayed as a result of the search.

Sequence information Each sequence in UPObase is stored with its complete information including its classification, motif pattern, sequence ID and cluster ID. All the information is displayed for each entry in the database along with the tables illustrating the functional roles of motif

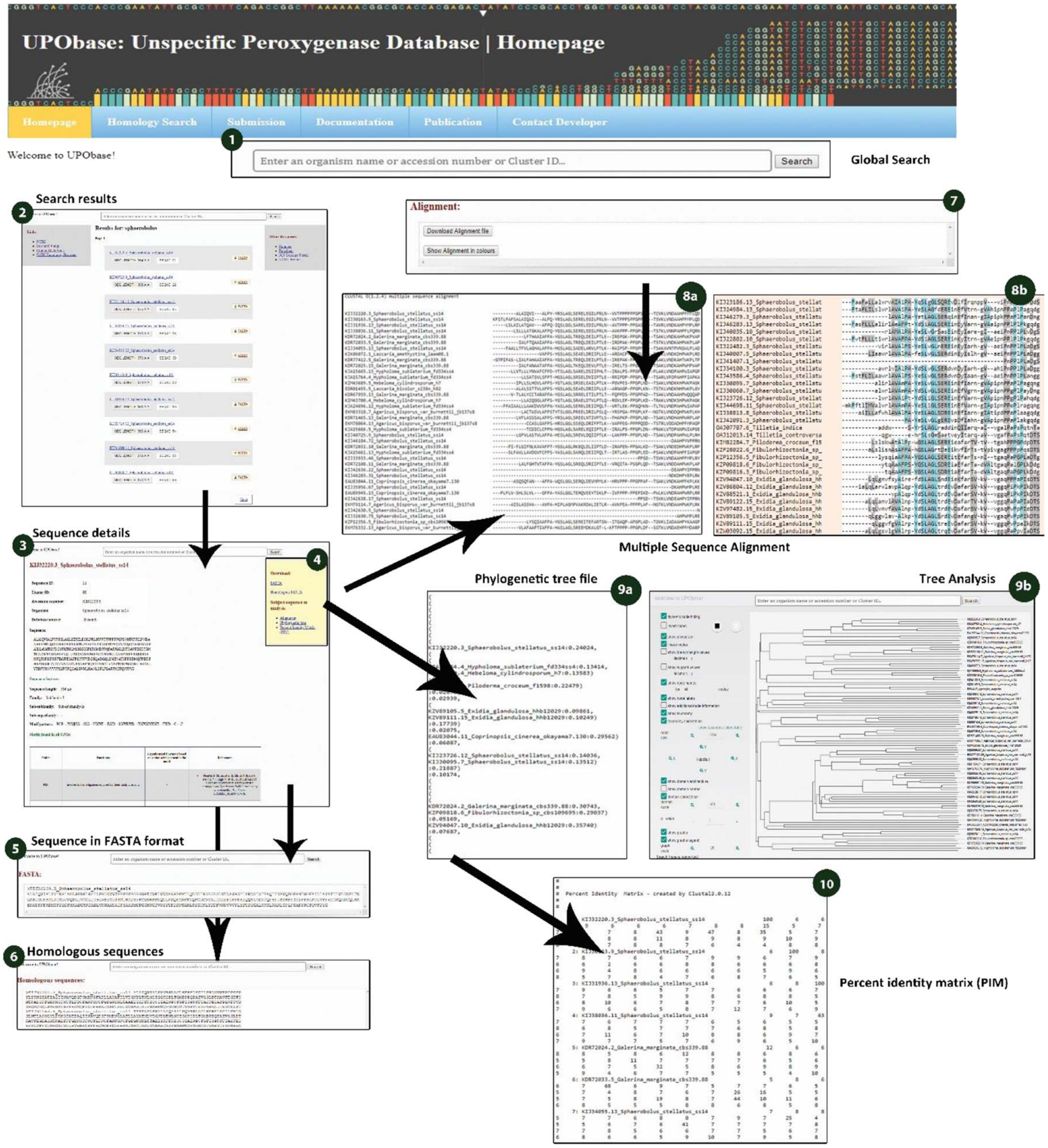
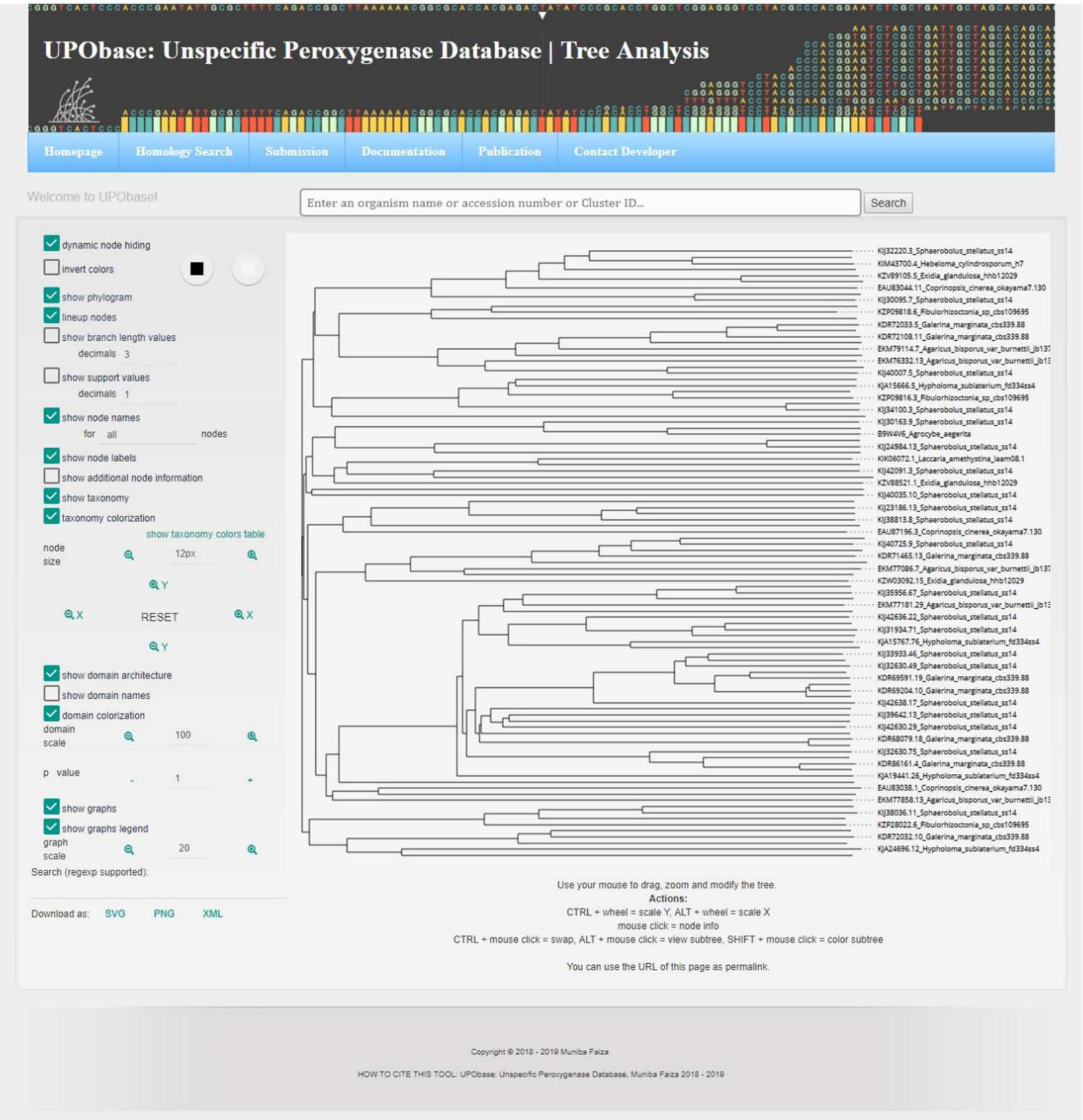

UPObase: Unspecific Peroxygenase Database | Homepage
Homepage
Homology Search
Submission
Documentation
Publication
Contact Developer
Welcome to UPObase!
1
Enter an organism name or accession number or Cluster ID...
Search
Global Search
2
Search results
Results for: sphaerobolus
3
Sequence details
KIJ32220.3_Sphaerobolus_stellatus_ss14
4
5
Sequence in FASTA format
FASTA:
6
Homologous sequences
Homologous sequences:
7
Alignment:
Download Alignment file
Show Alignment in colours
8a
8b
Multiple Sequence Alignment
Phylogenetic tree file
9a
Tree Analysis
9b
10
Percent Identity Matrix - created by Clustal2.0.12
Percent identity matrix (PIM)

Figure 4. Sequence details displayed for each and every sequence searched within UPObase. (1) the global search box; (2) search results displayed as a list to each search term; (3) sequence details; (4) download and subjecting sequence to analyses options; (5) sequence displayed in FASTA format; (6) FASTA sequences of the homologs corresponding to the sequence; (7) download files for alignment, tree and PIM; (8), (9) and (10) real-time created MSA, phylogenetic tree and PIM, respectively.

patterns found in all UPOs (Figure 4 (3)). This helps to identify the functional roles (either proved or hypothesized) of sequences belonging to different sub-subfamilies and sub-superfamilies. The conserved sequence patterns may also help in designing family-specific primers for screening new enzymes. The properly classified sequence information makes easy to further study their functional roles and to describe reasons behind their intriguing behavior.

Homology search The database sequences can be searched and compared with any other enzyme using the homology search which may help in the prediction of possible functions of unknown proteins. Users can adjust the e-value for the blast search against the database according to their requirements as shown in Figure 3 (2). The most relevant BLAST hits are displayed as output which consists of the subject query alignment along with e-value, bit score, percent identity and length of the subject sequence.

Sequence submission New sequences can be submitted via the submission portal where a user has to provide his


UPObase: Unspecific Peroxygenase Database | Tree Analysis
Homepage
Homology Search
Submission
Documentation
Publication
Contact Developer
Welcome to UPObase!
Enter an organism name or accession number or Cluster ID...
Search
show taxonomy colors table
RESET
Download as: SVG PNG XML
KIJ32220.3_Sphaerobolus_stellatus_ss14
KIM43700.4_Hebeloma_cylindrosporum_h7
KZV89105.5_Exidia_glandulosa_hhb12029
EAU83044.11_Coprinopsis_cinerea_okayama7.130
KIJ30095.7_Sphaerobolus_stellatus_ss14
KZP09818.6_Fibulorhizoctonia_sp_cbs109695
KDR72033.5_Galerina_marginata_cbs339.88
KDR72108.11_Galerina_marginata_cbs339.88
EKM79114.7_Agaricus_bisporus_var_burnettii_jb137
EKM76332.13_Agaricus_bisporus_var_burnettii_jb13
KIJ40007.5_Sphaerobolus_stellatus_ss14
KJA15666.5_Hypholoma_sublaterium_fd334ss4
KZP09816.3_Fibulorhizoctonia_sp_cbs109695
KIJ34100.3_Sphaerobolus_stellatus_ss14
KIJ30163.9_Sphaerobolus_stellatus_ss14
B9W4V6_Agrocybe_aegerita
KIJ24984.13_Sphaerobolus_stellatus_ss14
KIK06072.1_Laccaria_amethystina_laam08.1
KIJ42091.3_Sphaerobolus_stellatus_ss14
KZV88521.1_Exidia_glandulosa_hhb12029
KIJ40035.10_Sphaerobolus_stellatus_ss14
KIJ23186.13_Sphaerobolus_stellatus_ss14
KIJ38813.8_Sphaerobolus_stellatus_ss14
EAU87196.3_Coprinopsis_cinerea_okayama7.130
KIJ40725.9_Sphaerobolus_stellatus_ss14
KDR71465.13_Galerina_marginata_cbs339.88
EKM77086.7_Agaricus_bisporus_var_burnettii_jb137
KZW03092.15_Exidia_glandulosa_hhb12029
KIJ35956.67_Sphaerobolus_stellatus_ss14
EKM77181.29_Agaricus_bisporus_var_burnettii_jb13
KIJ42636.22_Sphaerobolus_stellatus_ss14
KIJ31934.71_Sphaerobolus_stellatus_ss14
KJA15767.76_Hypholoma_sublaterium_fd334ss4
KIJ33933.46_Sphaerobolus_stellatus_ss14
KIJ32630.49_Sphaerobolus_stellatus_ss14
KDR69591.19_Galerina_marginata_cbs339.88
KDR69204.10_Galerina_marginata_cbs339.88
KIJ42638.17_Sphaerobolus_stellatus_ss14
KIJ39642.13_Sphaerobolus_stellatus_ss14
KIJ42630.29_Sphaerobolus_stellatus_ss14
KDR68079.18_Galerina_marginata_cbs339.88
KIJ32630.75_Sphaerobolus_stellatus_ss14
KDR86161.4_Galerina_marginata_cbs339.88
KJA19441.26_Hypholoma_sublaterium_fd334ss4
EAU83038.1_Coprinopsis_cinerea_okayama7.130
EKM77858.13_Agaricus_bisporus_var_burnettii_jb13
KIJ38036.11_Sphaerobolus_stellatus_ss14
KZP28022.6_Fibulorhizoctonia_sp_cbs109695
KDR72032.10_Galerina_marginata_cbs339.88
KJA24696.12_Hypholoma_sublaterium_fd334ss4

☑ dynamic node hiding
□ invert colors
☑ show phylogram
☑ lineup nodes
□ show branch length values
decimals 3
□ show support values
decimals 1
☑ show node names
for all nodes
☑ show node labels
□ show additional node information
☑ show taxonomy
☑ taxonomy colorization
node size 12px
☑ show domain architecture
□ show domain names
☑ domain colorization
domain scale 100
p value - 1 +
☑ show graphs
☑ show graphs legend
graph scale 20
Search (regexp supported):

Use your mouse to drag, zoom and modify the tree.
Actions:
CTRL + wheel = scale Y, ALT + wheel = scale X
mouse click = node info
CTRL + mouse click = swap, ALT + mouse click = view subtree, SHIFT + mouse click = color subtree

You can use the URL of this page as permalink.

Copyright © 2018 - 2019 Muniba Faiza

HOW TO CITE THIS TOOL: UPObase: Unspecific Peroxygenase Database, Muniba Faiza 2018 - 2019

Figure 5. Tree analysis showing various key features.

name, email address and details about the new sequence including the source and type of the sequence whether hypothesized or expressed (Figure 3 (3)). A single user can submit a maximum of three sequences to the database. If users wish to submit more sequences, then they can send a request and data to the email address mentioned at the contact information page on the website. The criterion to submit any sequence in UPObase includes the following: sequence can be hypothetical or expressed, must belong to the fungal kingdom and must be longer than 100 amino acid residues. The user-submitted sequences will be added into the database after manual curation and validation. The curation involves motif pattern search to identify the subfamily/superfamily and classification of the organism. This new sequence submission portal allows the UPObase to grow and helps in making available all the new sequences discovered so far.

Sequence analyses A sequence in the database can be easily subjected to analysis by creating MSA with the other corresponding homologous sequences present in UPObase. Phylogenetic analyses can also be carried out, and in order to identify the similarity amongst these sequences, a PIM is also generated. In order to include the new and updated sequences in the analysis, the generation of MSA, phylogenetic tree and PIM is completely automated. In addition,

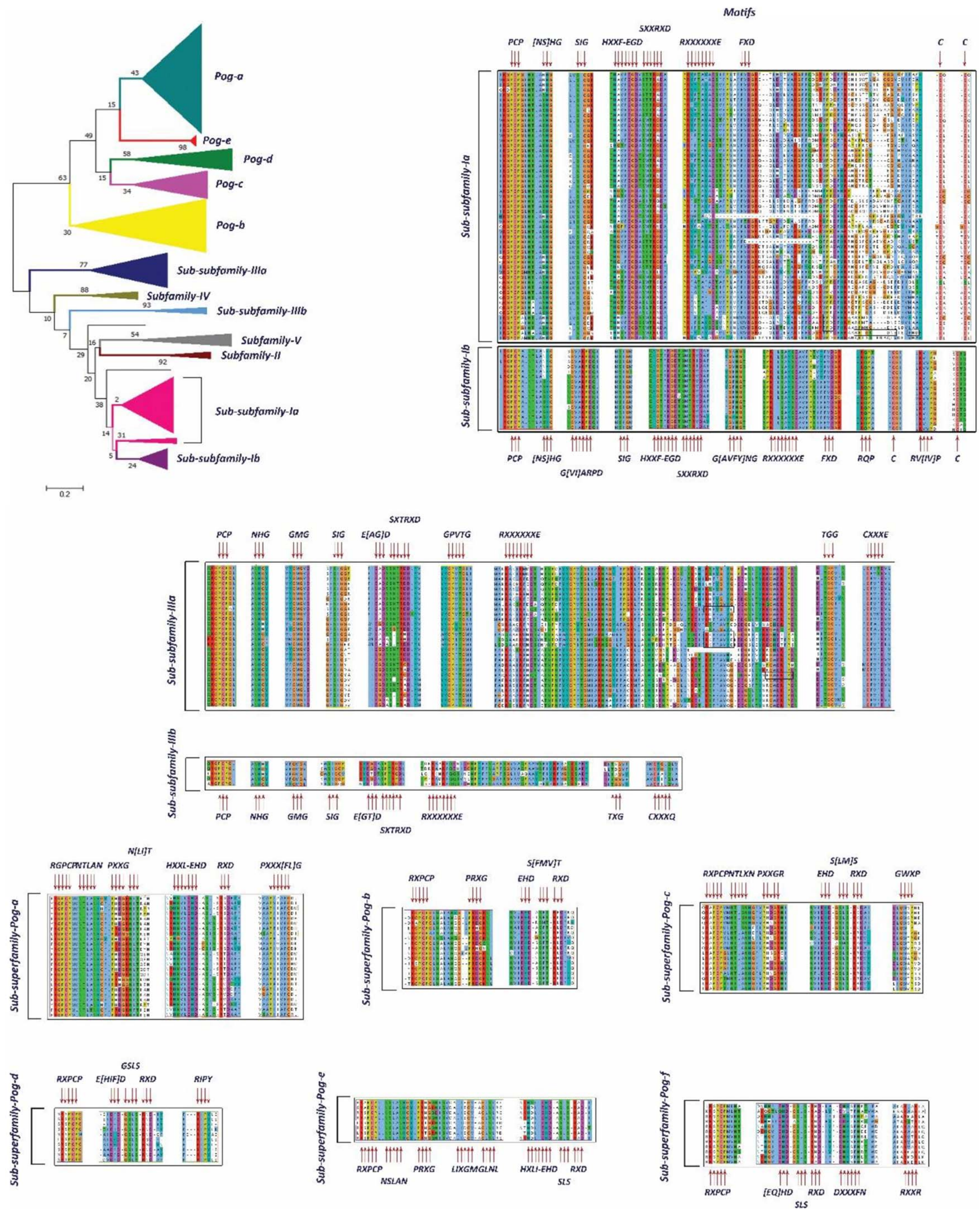


Figure 6. A phylogenetic tree and MSAs of UPO encoding sequences belonging to sub-subfamilies and sub-superfamilies which are reclassified. The motifs specific to each sub-subfamily are signified with a red arrow.

the MSA can be visualized in a color scheme showing the conserved residues (Figure 4 (8b)). The generated phylogenetic trees can be analyzed in the form of a phylogram or a dendrogram with various other visualization options (Figure 4 (9b)). Phyd3 offers various features to analyze Newick and XML tree files including information for each node in the tree, visualize branch lengths, support values, adjust the graph, see the graph in different background and foreground colors and display/hide node names and labels (25). The tree graph can be exported in SVG, PNG and XML format (Figure 5).

In summary, UPObase has been designed to study and analyze all fungal UPOs but it also works as a platform to perform similarity search and comparison of any other enzyme of interest with the UPOs. The conserved patterns and classification of UPObase can also be used for identify-

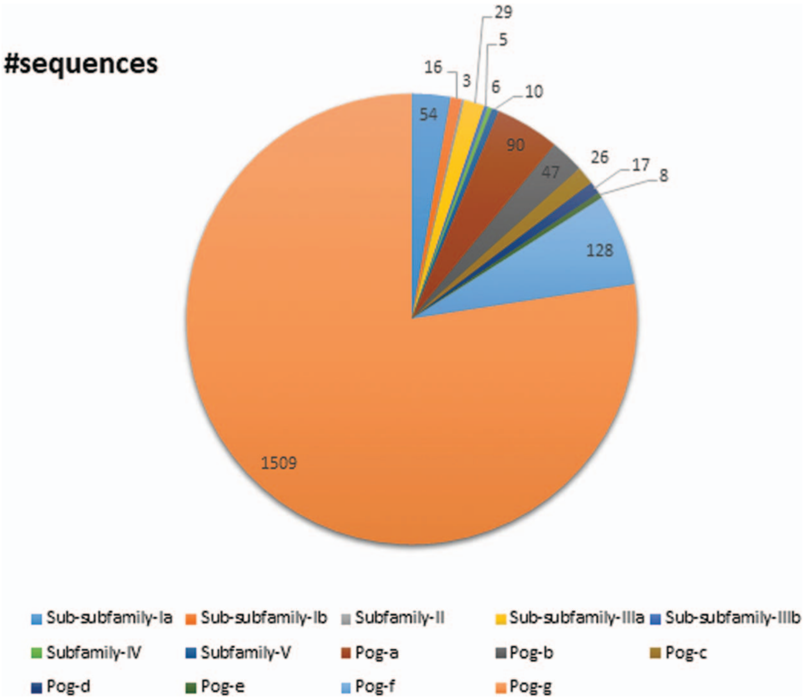


Figure 7. A pie chart showing the total number of sequences present in the database classified into subfamilies and superfamilies.

ing functions for the unknown proteins. Besides, these discovered new members in the families may reveal some novel characteristics in addition to those exhibited by the UPOs.

Classification of UPOs

In our preliminary work, we found 113 putative UPO sequences, which were classified into five different subfamilies (I, II, III, IV and V) and a superfamily (Pog) based on the motifs present in their sequences and the phylogenetic analysis. Here, in this study, we have found 20 times more UPO and UPO-like sequences at our disposal. Previously, Subfamily I was found to have a specific motif pattern (Table 1). Based on the current data of UPOs, a new slightly different motif pattern has been found to exist in this subfamily, and hence, it is subclassified into two sub-subfamilies: Sub-subfamily Ia having the former motif pattern and Sub-subfamily Ib with a newly found motif pattern (Figure 6). However, some motifs such as [NS] HG, SIG and SXXTRXD which were present in all UPOs are still present in the new sub-subfamilies. After re-clustering in the second step, Subfamily II remains with a very few numbers of OTUs and not further subclassified, which is found to have the same motif pattern as explained previously. According to the phylogenetic and sequence analyses of Subfamily III sequences, it has been classified into two new sub-subfamilies: Sub-subfamily IIIa and Sub-subfamily IIIb (Figure 5). These two sub-subfamilies were found to have some additional motifs in their sequences to the pattern explained previously (Tables 1 and 2). The Subfamily IV and Subfamily V UPO encoding sequences consisted of the same motif pattern as explained previously (Table 1). No new motif pattern was found to exist in these sequences. However, the Pog superfamily which was previously not found to be consisting of any signature motif, after finding more sequences belonging to this superfamily, led to its subclassification into seven sub-superfamilies based on the phylogenetic tree and the sequence motifs (Figure 5). The hypothesized functions of the newly found motifs are explained in Table 2 to allow users to identify the possible roles of subfamilies and superfamilies.

Database sequences

UPObase is composed of 1948 sequences of UPOs classified into five subfamilies and a superfamily which are subclassified into different sub-subfamilies and sub-superfamilies, respectively (Figure 7). Subfamily I consists of 70 sequences in total including *Aae*UPO categorized into two sub-subfamilies: Ia (54 sequences) and Ib (16 sequences). Subfamily II consists of three sequences. Subfamily III consists of 34 sequences in total categorized into two sub-subfamilies: IIIa (29 sequences) and IIIb (5 sequences). Subfamilies IV and V are not further categorized into sub-subfamilies and consists of 6 and 10 sequences, respectively. The Pog superfamily which consists of the maximum

Table 1. represents the motif patterns specific to sub-subfamilies and sub-superfamilies.

Subfamily/superfamily	Sub-subfamily/sub-superfamily	Motif pattern
Subfamily-I	Sub-subfamily-Ia	PCP—[NS]HG—SIG—HXXF—EGD—SXXRXD—RXXXXXXE—FXD—C—C
	Sub-subfamily-Ib	PCP—[NS]HG—GVARPD—SIG—HXXF—EGD—SXXRXD—G[AVFY]NG—RXXXXXXE—FXD—RQP—C—RV[IV]P—C
Subfamily-II	-	PCP—NHG—RGN—S[IL]G—VPPLPG—IDG—HGRF—EGD—SMTRXD—RXXXXXXE—TXXXXXXR
Subfamily-III	Sub-subfamily-IIIa	PCP—NH[NG]—G[ML]G—SIG—E[GA]D—SXTRXD—GPXTG—RXXXXXXE—TGG—CXXXE
	Sub-subfamily-IIIb	PCP—NH[NG]—G[ML]G—SIG—E[GT]D—SXTRXD—RXXXXXXE—TXG—CXXXQ
Subfamily-IV	-	PCP—N[HY][NG]—FXXXD—S[IL]G—CDA—HXXF—EGD—SLTRXD—RXXXXXXE—GAAXXXYE
Subfamily-V	-	EDXXH—PCP—NHG—SIG—GXG—EGD—SVTRXD—RXXXXXXE
Pog-superfamily	Pog-a	RGPCP—NTL[AT]N—PXXG—NXT—HXXL—EHD—RXD—PXXXFG
	Pog-b	RXPCP—PRXG—[EQ]HD—S[FMV]T—RXD
	Pog-c	RXPCP—NTLXN—PXXGR—EHD—S[ML]S—RXD—GWXP
	Pog-d	RXPCP—E[IHF]D—GSLS—RXD—RIPY
	Pog-e	RXPCP—NSLAN—PRXG—LIXGM—GLNL—HXLI—EHD—SLS—RXD
	Pog-f	RXPCP—[EQ]HD—S[LM]S—RXD—DXXXFN—RXXR
	Pog-g	No signature motif

number of sequences (1825 including *Leptoxyphium fumago* and *Marasmius rotula*) in the database is further subclassified into seven sub-superfamilies: Pog-a (90 sequences), Pog-b (47 sequences), Pog-c (26 sequences), Pog-d (17 sequences), Pog-e (8 sequences), Pog-f (128 sequences) and Pog-g (1509 sequences), where Pog-g sequences do not have any signature motif pattern of their own except the *Cys* ligation to the heme which is a characteristic of all HTPs.

Concluding Remarks and Future Enhancements

We provide a unified platform to analyze all fungal UPOs and UPO-like sequences systematically with easy retrieval and browsing, which can also be successfully used to compare with other enzymes. UPObase also provides a sequence submission portal for new sequences. Besides, it provides a complete classification of UPOs based on their phylogeny and sequence study, and conserved set of sequence motif patterns among these species. UPObase may work as a beneficial tool for the scientists working in the area of fungal UPOs, as it provides annotated data to work on and allows to explore insights to further advance in studying the main physiological role of fungal UPOs. Further developments to UPObase include the better display of homologous searches in the database, search for more UPO and UPO-like sequences and include the protein crystal structures which are currently limited as only two of the fungal UPO protein structures (*Aae*UPO and *Mro*UPO) have been experimentally resolved to date.

Author Contributions

The original idea of this study was conceived by Y.W., S.H. and D.L. M.F., and S.H. designed the experiments performed by M.F. and collected the data. All authors analyzed the data. The manuscript was drafted by M.F., S.H. and D.L. critically revised by all the co-authors. All authors read and approved the final manuscript.

Funding

National Outstanding Youth Science Foundation of China (31725022); Key Program of Natural Science Foundation of China (31930084); Molecular Enzyme and Engineering International Cooperation Base of South China University of Technology (2017A050503001); Special Program of Guangdong Province for Leader Project in Science and Technology Innovation: Development of New Partial Glycerin Lipase (2015TX01N207); Marine S&T Fund of Shandong Province (2018SDKJ0302-2); National Key R&D Program of China (2018YFD0900503). Funders had no role in the

Table 2. summarizes the hypothesized functions of the preliminary and newly found subfamilies and/ sub-subfamilies and sub-superfamilies.

Subfamily	Sub-subfamily/ superfamily	Motif	*Roles of amino acids present in the motif	Hypothesized functions of the subfamily/superfamily
I	Ia	FXD	*Phe* is basically involved in stacking interactions with other aromatic side-chains and the *Asp* is frequently involved in salt-bridges interacting with positively charged amino acids to create stabilizing H-bonds which can be important for proteins stability.	may actively involve in interacting with aromatic residues and in forming stable H-bonds imparting to the structural stability, and in substrate recognition.
		Cys-Cys	the disulfide bond is mostly involved in providing stability to protein structure.	
	Ib	GVARPD	*Gly* provides the conformational stability; *Val* may play a role in substrate recognition; *Ala* is involved in substrate recognition and specificity; *Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability; *Pro* plays an important role in molecular recognition; and Asp residues create a stable H-bonds.	
		G[AVFY]NG	Again *Gly* provides the conformational stability; *Tyr* and *Phe* make stacking interactions with the aromatic side chains; the *Asn* is involved as proteins active and binding sites.	
II	-	RGN	*Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability; the *Gly* provides the conformational stability, and the *Asn* is involved as proteins active and binding sites.	may potentially interact with the hydrophobic ligands such as lipids and may show specificity for some polar substrates.
		IDG	*Ile* in the IDG motif is involved in recognizing hydrophobic ligands; *Asp* forms stable H-bonds with positively charged amino acids required for proteins stability, and the *Gly* again may provide conformational stability.	
		TXXXXXXR	*Thr* is often found in protein centers and capable of forming H-bonds with the polar substrates.	
III	IIIa	G[ML]G	the *Gly* provides the conformational stability; *Met* and *Leu* play a role in binding and recognition of hydrophobic ligands.	may play an important role in substrate specificity/recognition, specific to aromatic residues, and capable of forming strong H-bonds with the polar substrates.
		GPXTG	*Gly* provides the conformational stability; *Pro* plays an important role in molecular recognition; *Thr* is often	

(Continued)

Table 2. Continued

Subfamily	Sub-subfamily/ superfamily	Motif	*Roles of amino acids present in the motif	Hypothesized functions of the subfamily/superfamily
			found in protein centers and capable of forming H-bonds with the polar substrates.	
	CXXXE	*Cys* may act as a reactive center of an enzyme; *Glu* residues create a stable H-bonds.		
IIIb	CXXXQ	*Gln* is involved in protein active and binding sites.		
IV	-	CDA, FXXXDG, GAAXXXYE, and HXXF	*Ala* is involved in substrate recognition and specificity; *Tyr* makes stacking interactions with the aromatic side chains; *His* is involved in protein metal binding sites; and *Phe* also makes stacking interactions with aromatic side chains.	may show large interactions with the aromatic substrates and these motifs are perhaps involved in substrate recognition and binding.
V	-	EDXXH	*His* is most commonly involved in active and binding sites especially in metal binding sites and the *Asp* and *Glu* residues create the stable H-bonds.	may play an important role in reacting with positively charged amino acids.
		GXG	*Gly* provides the conformational stability	
	Pog-a	NTL[AT]N	*Asn* is involved as proteins active and binding sites; *Tyr* makes stacking interactions with the aromatic side chains; *Leu* plays a role in binding and recognition of hydrophobic ligands	may play an important role in reacting with hydrophobic ligands and polar substrates
		NXT	Again, *Asn* is involved as proteins active and binding sites; and *Thr* is often found in protein centers and capable of forming H-bonds with the polar substrates.	
Pog superfamily		HXXL	*His* is most commonly involved in active and binding sites especially in metal binding sites.	
		RXD	*Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability, and *Asp* residues create the stable H-bonds.	
		PXXXFG	*Pro* plays an important role in molecular recognition; and *Phe* is basically involved in stacking interactions with other aromatic side-chains.	
	Pog-b	PRXG	*Pro* plays an important role in molecular recognition; *Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability; *Gly* provides the conformational stability.	may be involved in the interaction with aromatic substrates and hydrophobic ligands.

(Continued)

Table 2. Continued

Subfamily	Sub-subfamily/ superfamily	Motif	*Roles of amino acids present in the motif	Hypothesized functions of the subfamily/superfamily
		S[FMV]T	*Ser* is capable of H-bonds with polar substrates; *Met* plays a role in binding and recognition of hydrophobic ligands; and *Thr* is often found in protein centers and capable of forming H-bonds with the polar substrates.	
	Pog-c	NTLXN	*Asn* is involved as proteins active and binding sites; and *Thr* is often found in protein centers and capable of forming H-bonds with the polar substrates; *Leu* plays a role in binding and recognition of hydrophobic ligands.	may get involved in making interactions with polar substrates and non-protein ligands.
		PXXGR	*Pro* plays an important role in molecular recognition; *Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability; *Gly* provides the conformational stability.	
		S[ML]S	*Ser* is capable of H-bonds with polar substrates; *Met* and *Leu* play a role in binding and recognition of hydrophobic ligands.	
		GWXP	*Trp* may be involved in binding with non-protein ligands.	
	Pog-d	GSLS	*Gly* provides the conformational stability; *Ser* is capable of H-bonds with polar substrates; and *Leu* plays a role in binding and recognition of hydrophobic ligands.	may react with aromatic substrates and hydrophobic ligands.
		RIPY	*Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability; *Ile* plays a role in binding and recognition of hydrophobic ligands; and *Tyr* makes stacking interactions with the aromatic side chains.	
	Pog-e	NSLAN	*Asn* is involved as proteins active and binding sites; *Ala* may be involved in substrate recognition or specificity.	may show specificity for some hydrophobic ligands.
		LIXGM	*Ile* and *Leu* is involved in recognizing hydrophobic ligands; *Met* plays a role in binding and recognition of hydrophobic ligands.	
		GLNL	*Gly* provides the conformational stability; *Leu* is involved in recognizing hydrophobic ligands; *Asn* is involved as proteins active and binding sites.	
		HXLI	*His* is involved in protein metal binding sites; *Ile* and *Leu* are involved in recognizing hydrophobic ligands.	

(Continued)

Table 2. Continued

Subfamily	Sub-subfamily/ superfamily	Motif	*Roles of amino acids present in the motif	Hypothesized functions of the subfamily/superfamily
	Pog-f	DXXXFN	*Asp* forms stable H-bonds with positively charged amino acids required for proteins stability; *Phe* makes stacking interactions with the aromatic side chains; *Asn* is involved as proteins active and binding sites.	may show strong structural stability with substrate specificity.
		RXXR	*Arg* is frequently involved in making salt-bridges with the negatively charged amino acids creating stable H-bonds which may be crucial for the structure stability.	

design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of interest. None declared.

References

1. Ullrich,R., Nüske,J., Scheibner,F. *et al.*. Novel haloperoxidase from the agaric basidiomycete Agrocybe aegerita oxidizes aryl alcohols and aldehydes. *Appl Environ Microbiol [Internet]*. (2004 [cited 2019 Jun 3]; **70**(8):4575–81. Available from: http://aem.asm.org/.
2. Gutiérrez,A., Babot,E.D., Ullrich,R. *et al.*. Regioselective oxygenation of fatty acids, fatty alcohols and other aliphatic compounds by a basidiomycete heme-thiolate peroxidase. *Arch Biochem Biophys [Internet]*. 2011 Oct 1 [cited 2019 Jun 3]; **514**(1–2):33–43. Available from: https://www.sciencedirect.com/science/article/abs/pii/S000398611100289X.
3. Hofrichter,M. and Ullrich,R. (2014) Oxidations catalyzed by fungal peroxygenases. *Curr. Opin. Chem. Biol. [Internet]*, Apr 1 [cited 2019 Jun 3];**19**:116–25. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1367593114000106.
4. Peter,S., Kinne,M., Wang,X. *et al.* (2011) Selective hydroxylation of alkanes by an extracellular fungal peroxygenase. *FEBS J. [Internet]*, Oct 1 [cited 2019 Jun 3];**278**(19):3667–75. Available from: https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-4658.2011.08285.x%4010.1002/%28ISSN%291742-4658%28CAT%29VirtualIssues%28VI%29MolecularEnzymology2012.
5. Bordeaux,M., Galarneau,A. and Drone,J. (2012) Catalytic, mild, and selective oxyfunctionalization of linear alkanes: current challenges. *Angew. Chemie Int. Ed. [Internet]*, Oct 22 [cited 2018 Jul 27];**51**(43):10712–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22996726.
6. Hofrichter,M., Kellner,H., Pecyna,M.J. *et al.* (2015) Fungal unspecific peroxygenases: heme-thiolate proteins that combine peroxidase and cytochrome P450 properties. *Adv. Exp. Med. Biol. [Internet]*, [cited 2018 Jun 19]. p. 341–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26002742.
7. Pecyna,M.J., Ullrich,R., Bittner,B. *et al.* (2009) Molecular characterization of aromatic peroxygenase from Agrocybe aegerita. *Appl. Microbiol. Biotechnol. [Internet]*, Oct 12 [cited 2018 Jun 19];**84**(5):885–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19434406.
8. Faiza,M., Huang,S., Lan,D. *et al.* (2019). New insights on unspecific peroxygenases: Superfamily reclassification and evolution. *BMC Evolutionary Biology,* **19**(1), 1–19. Available from: https://doi.org/10.1186/s12862-019-1394-3.
9. Benson,D.A., Cavanaugh,M., Clark,K.*et al.* (2017). *GenBank. Nucleic Acids Research,* **45**(D1), D37–D42. Available from: https://doi.org/10.1093/nar/gkw1070.
10. Kersey,P.J., Allen,J.E., Allot,A. *et al.* (2018). Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research,* **46**(D1), D802–D808. Available from: https://doi.org/10.1093/nar/gkx1011.
11. Crous,P.W., Gams,W., Stalpers, J.A. *et al.*. (2004). MycoBank: An online initiative to launch mycology into the 21st century. Studies in Mycology (Vol. 50). Retrieved from www.indexfungorum.org.
12. Bonants,P., Edema,M., and Robert V. (2013). Q-bank, a database with information for identification of plant quarantine plant pest and diseases. *EPPO Bulletin*, **43**(2), 211–215. https://doi.org/10.1111/epp.12030.
13. Bateman,A. (2000) The Pfam protein families database. *Nucleic Acids Res. [Internet]*, Jan 1 [cited 2019 Aug 21];**28**(1):263–6. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.263.
14. Fischer,M. and Pleiss,J. (2003) The lipase engineering database: a navigation and analysis tool for protein families. *Nucleic Acids Res. [Internet]*, Jan 1 [cited 2019 Jul 9];**31**(1):319–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12520012.
15. Passardi,F., Theiler,G., Zamocky,M. *et al.*. (2007). PeroxiBase: The peroxidase database. Phytochemistry, **68**(12), 1605–1611. https://doi.org/10.1016/j.phytochem.2007.04.005.
16. Rawlings,N.D., Waller,M., Barrett,A.J. *et al.* (2014) The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res. [Internet]*, Jan 1 [cited 2019 Jul 10];**42**(D1):D503–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt953.
17. Schomburg,I., Chang, A., Hofmann,O. *et al.*. (2002). BRENDA: A resource for enzyme data and metabolic information. *Trends*

in Biochemical Sciences. Elsevier Ltd. https://doi.org/10.1016/S0968-0004(01)02027-8.

18. Kumar,S., Stecher,G. and Tamura,K. (2015) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol [Internet]*, [cited 2018 Jun 19];**33**(7):1870–4. Available from: https://www.megasoftware.net/pdfs/KumarStecher16.pdf.
19. Whelan,S. and Goldman,N.A. (2001) General empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol. [Internet]*, May 1 [cited 2019 Jun 4];**18**(5):691–9. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a003851.
20. Darriba,D., Taboada,G.L., Doallo,R. *et al.* (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics [Internet]*, Apr 15 [cited 2018 Dec 11];27(8):1164–5. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr088.
21. Sievers,F., Wilm,A., Dineen,D. *et al.* (2014) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol [Internet]*, Apr 16 [cited 2019 Jun 7];**7**(1):539–539. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21988835.
22. Edgar,R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* **32**(5), 1792–1797. https://doi.org/10.1093/nar/gkh340.
23. Sonnhammer,E.L.L. and Hollich,V. (2005). Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics*, **6**. https://doi.org/10.1186/1471-2105-6-108.
24. Larkin,M.A., Blackshields,G., Brown,N.P. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics [Internet]*, Nov 1 [cited 2019 Jun 7];**23**(21):2947–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm404.
25. Kreft,L., Botzki,A., Coppens,F. *et al.* (2017). PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, **33**(18), 2946–2947. https://doi.org/10.1093/bioinformatics/btx324.
26. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics [Internet]*, Jul 1 [cited 2018 Jun 19];**22**(13):1658–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16731699.
27. Schaeffer,S.E. (2000). Graph clustering by flow simulation. *Computer Science Review*. University of Utrecht. https://doi.org/10.1016/j.cosrev.2007.05.001.
28. Dongen,S.v. (2000) *Graph Clustering by Flow Simulation*. University of Utrecht