J. Microbiol. Biotechnol. (2015), 25(5), 723–728
http://dx.doi.org/10.4014/jmb.1406.06049

jmb

In Silico Metagenomes Mining to Discover Novel Esterases with Industrial Application by Sequential Search Strategies S

Jorge Barriuso* and María Jesús Martínez*

Centro de Investigaciones Biológicas, Consejo Superior de Investigaciones Científicas, 28040 Madrid, Spain

Received: June 17, 2014
Revised: October 8, 2014
Accepted: December 4, 2014

First published online
December 12, 2014

**Corresponding authors*
J.B.
Phone: +34-918373112;
Fax: +34-915360432;
E-mail: jbarriuso@cib.csic.es
M.J.M.
Phone: +34-918373112;
Fax: +34-915360432;
E-mail: mjmartinez@cib.csic.es

S upplementary data for this paper are available on-line only at http://jmb.or.kr.

pISSN 1017-7825, eISSN 1738-8872

We present here an *in silico* search of fungal sterol-esterase/lipase and bacterial depolymerase sequences from environmental metagenomes. Both enzyme types contain the α/β-hydrolase protein fold. Analysis of DNA conserved motifs, protein homology search, phylogenetic analysis, and protein 3D modeling have been used, and the efficiency of these screening strategies is discussed. The presence of bacterial genes in the metagenomes was higher than those from fungi, and the sequencing depth of the metagenomes seemed to be crucial to allow finding enough diversity of enzyme sequences. As a result, a novel putative PHA-depolymerase is described.

Keywords: Sterol esterase, lipase, PHA-depolymerase, genome diversity

Microorganisms constitute the major reserve for genetic diversity on Earth. The study of DNA directly sequenced from an environmental sample is known as metagenomics, and allows obtaining information about taxonomic variability as well as the sequence of genes with biotechnological potential. By means of these approaches, we are able to explore the fraction (>99%) of the microorganisms that are non-culturable in the laboratory [4, 22]. In the last years, metagenomics has experienced great development, thanks to massive DNA sequencing, contributing to revising our views on microbial biodiversity. Sequencing projects are deposited in several public databases, among which MG-RAST, IMG/M, and CAMERA are three prominent systems [17, 19, 23].

Bioinformatics methodologies allow analyzing these metagenomes to discover new enzymes. Usually, these strategies are based in searches of homologous sequences, using BLAST [1] or more complex algorithms such as HMMER [8], or in the search of conserved motifs previously reported, using tools such as MEME [2] or Fuzznuc [20]. After identification of the candidate DNA sequences, their characteristics can be analyzed by means of sequence alignments, phylogenetic analysis, and/or modeling of the 3D structure of the putative proteins, making possible the prediction of some functionalities [5].

Lipases and sterol esterases belong to the α/β-hydrolase superfamily and are widely used in industry. Among them, lipases from the *Candida rugosa*-like family have gained special interest owing to their stability and broad substrate specificity [10]. Candidates from this family contain a GGGF conserved sequence forming the oxyanionic hole, and a GESAG sequence around the catalytic Ser [15, 16]. Some representative enzymes from this family, encoded by genes usually without introns, are the esterases from *C. rugosa* (CRL), *Ophiostoma piceae* [7], and *Melanocarpus albomyces* [13].

In addition, polyhydroxyalkanoate (PHA) depolymerases are prokaryotic esterases that act on PHA biodegradable polymers that can be used as a substitute of plastic [14]. They are classified depending on whether they are intra- or extracellular, and on their activity on short (PHAscl), or medium chain (PHAmcl) polymers. The best-studied PHA depolymerases are the intracellular PhaZ from *Pseudomonas putida* KT2442, and the extracellular PhaZ from *Pseudomonas fluorescens* GK13, that contain the SWGGA and the GISSG conserved motif in their catalytic centers, respectively [11]. In addition, an extracellular depolymerase family from actinobacteria that contains the conserved motif GHSQGG has been recently described [9].

In the present work, we carried out a sequential bioinformatics screening of publicly available metagenomes (Table S1) to look for fungal sterol-esterases/lipases and bacterial PHA depolymerases. The candidates were selected by taking into account sequence similarity, conserved motifs, phylogenetic affiliation, and sequence characteristics. The three-dimensional structure of selected candidates was modeled to discuss their potential catalytic properties.

Model Sequences and Datasets

Model sequences of representative proteins from the enzyme families of interest were selected. Lipases Lip1, Lip2, and Lip3 from *C. rugosa* [15, 16], and sterol-esterases from *O. piceae* (OPE) [7] and *M. albomyces* [13] were chosen among the members of the *Candida rugosa*-like family (accession numbers in Fig. 1). All of them are well characterized at the genetic and biochemical levels. Analysis using MEME software (http://meme.nbcr.net/meme/) corroborated the presence of the conserved amino acidic sequences GGGF and GESAG into the model proteins.

From the different PHA depolymerase families, mcl-PHAs from *P. putida* KT2440 and *P. fluorescens* GK13 [11] were selected as models of intra- and extracellular enzymes containing the conserved amino acidic sequences SWGGA and GISSG, respectively (accession numbers in Fig. 1). As members of the new family of depolymerases from actinomycetes, mcl-PHAs containing the conserved amino acidic sequence GHSQGG from *Streptomyces roseolus* SL3 and *Streptomyces venezuelae* SO1 were selected (accession numbers in Fig. 1). The genetic and biochemical traits of these enzymes are also well known [9].

Eighty-one nucleotide datasets from different metagenomes were downloaded from the servers of MG-RAST (http://metagenomics.anl.gov/), IMG/M (http://img.jgi.doe.gov), and CAMERA (http://camera.calit2.net/) databases. The metagenomes were selected from diverse environments to maximize genetic variability. Using these metagenomes, more than 7 million assembled contigs were analyzed (Table S1).

Metagenome Screening

We have used a two-step process to search for genes codifying for the proteins of interest in public metagenomes: the first step was to carry out a sequence similarity screening using two different strategies: The first similarity screening strategy consisted in comparing DNA sequences present in the 81 metagenomes translated in the six possible reading frames by means of BLASTx (e-value of 10^{-2}) against two custom databases, one for each target protein family. An esterase/lipase-specific database was built with a total of 4,805 protein sequences from the NCBI non-redundant (NR) database containing in their name the terms "esterase" or "lipase." In the case of the depolymerases, a custom database was made with a total of 1,275 sequences from the NCBI NR database containing the term "depolymerase," plus the sequences from mcl-PHA depolymerases present in the "PHA depolymerase Engineering Database" (DED) (http://www.ded.uni-stuttgart.de/).

The second strategy used to look for candidates was based on the translation in the six possible reading frames of the nucleotide sequences from the 81 metagenomes and their use as the reference database (40 Gb size). The model sequences described in the Model Sequences and Datasets section were compared against this database using BLASTp with an e-value of 10^{-40} to identify sequences highly related to the reference model proteins.

In all cases, the sequences giving a positive hit in the different BLAST comparisons were checked for the presence of introns, and nucleotide sequences were translated to protein in the correct reading frame and written in different Fasta files.

To carry out the second step in the screening for candidate sequences, each individual file with the selected translated sequences was filtered using the Bioedit 7.1.3 software to check for the conserved motifs from the model sequences. After that, the sequences shorter than 200 amino acids for PHA depolymerases and 500 amino acids for sterol-esterase/lipases were removed because they were too short to contain an ORF. In the case of the sterol-esterase/lipases, sequences with a distance longer than 100 amino acids between the motifs GGGF and GESAG were deleted.

Phylogenetic Analysis

A phylogenetic tree was built for each kind of enzyme using the representative model sequences and the putative sequences selected after metagenome screening. A bootstrap (1,000 repetitions) unrooted tree was built with the Mega 5.1 software (default parameters), using MUSCLE for

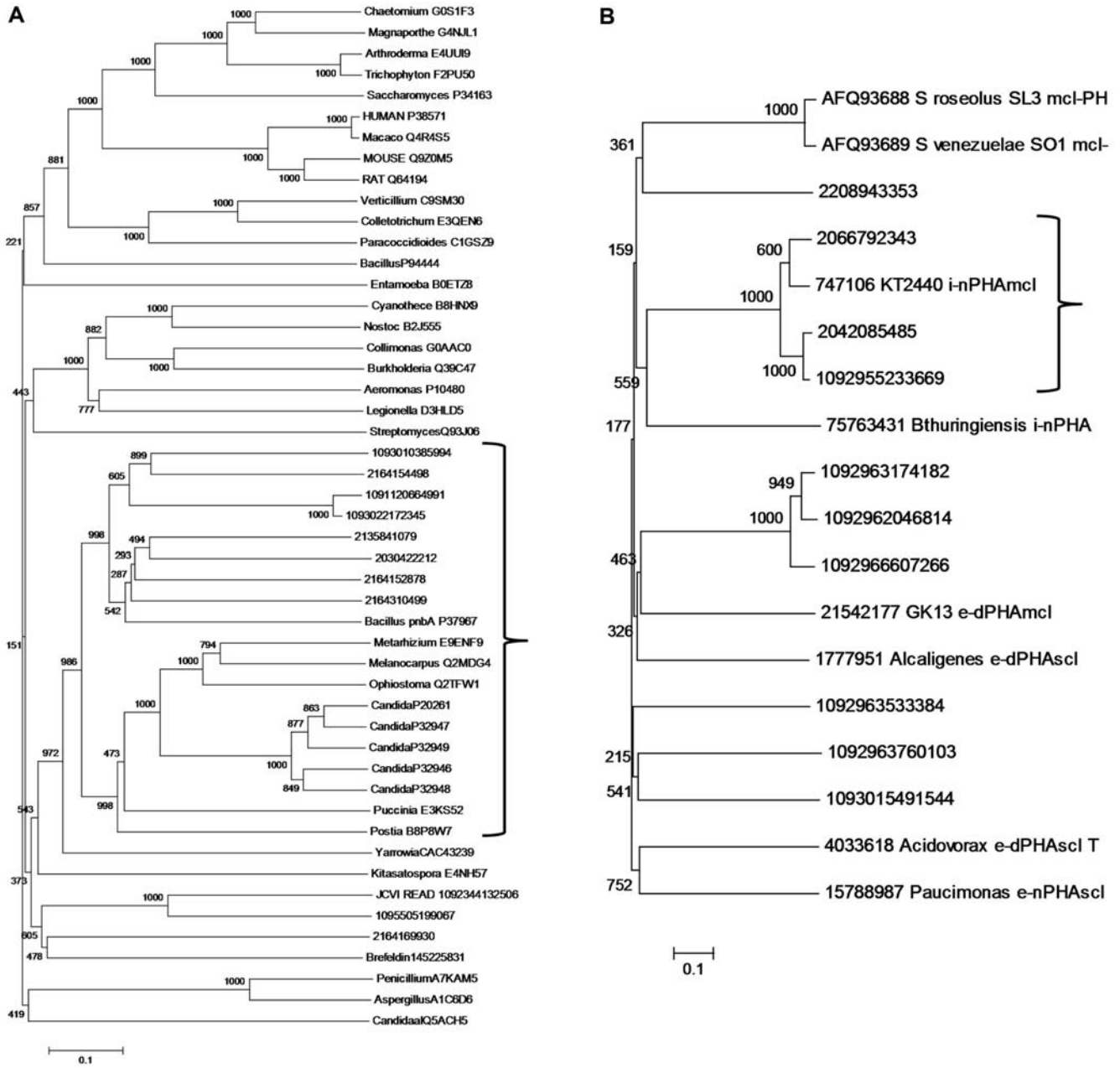

Fig. 1. Phylogenetic analysis.

(**A**) Phylogenetic tree of the putative sterol-esterases/lipases from the different metagenomes analyzed, as well as the model sequences from the *C. rugosa*-like family and representatives from other families. The selected samples are indicated in brackets, and came from soil metagenomes (Permafrost Bonanza Creek and Soil Miscanthus Kellogg), insect-associated habitats (Termite Fungus Garden and Atta Cephalotes Fungus Garden), and sea water (GS017 Coastal Caribbean and GS022 Open Ocean). (**B**) Phylogenetic tree of the putative PHA depolymerases, as well as the bacterial model sequences and sequences from other families. The selected samples are indicated in brackets, candidate 2066792343 was from sea water metagenome (Coastal North James Bay; MG-RAST ID 4441596) and 1092955233669 from insect-associated habitat (Termite Fungus Garden; JGI Project ID 401007). The trees were built using MUSCLE for multiple sequence alignment, and the Maximum-Likelihood to calculate distances. Bootstraping was made with 1000 repetitions.

multiple sequence alignment and the Maximum-Likelihood method to calculate distances. Sequences grouping with the versatile lipases and sterol-esterases from the *C. rugosa*-like family, the *P. putida* KT2442, *P. fluorescens* GK13, or actinomycetes-like mcl-PHA depolymerases, were selected for further analyses.

Sequence Analysis and Structural Models

The sequences from the sterol-esterase/lipase candidates selected in the previous section were compared against the NCBI NR database using BLASTp. Sequences from the PHA depolymerase candidates selected in the previous section were compared against the NCBI NR and "DED" databases (http://www.ded.uni-stuttgart.de/) using BLASTp.

A tridimensional model of each selected putative protein was generated using the programs implemented in the protein homology-modeling server SWISS-MODEL (Swiss Institute of Bioinformatics) [12]. The template with higher similarity, automatically chosen by the SWISS-MODEL server for sequence 1092955233669, was the esterase 3IA2 from *P. fluorescens* (Qmean Z-Score 3.3112), for 2066792343 the hydrolase 3OM8 from *P. aeruginosa* PA01 (Qmean Z-score 3.033), and for PHA depolymerase from *P. putida* KT2440 was the hydrolytic enzyme PA3053 from *P. aeruginosa* PAO1 (4F0J) (Qmean Z-score 5.214). Since the crystal structure of the PHA depolymerase from *P. putida* KT2440 is still unknown, all proteins were modeled using the same template, the hydrolytic enzyme PA3053 from *P. aeruginosa* PAO1 (4F0J). The models were exhaustively analyzed using PyMol 1.1 (http://pymol.org/) and putative intramolecular tunnels were modeled from the catalytic serine using Caver 2.0 ver. 0.003 [18].

Search in Public Microbial Genomes

The search of putative enzymes was carried out by using different strategies. A comparison of the sequences from metagenomes with those in the custom esterase/lipase database rendered 14,237 candidates, but only 38 of them contained the two conserved motifs. After filtering the sequences, only 11 candidates remained. These candidates were translated in the correct reading frame, and the presence of introns was not detected (Table 1). In the case of the depolymerases, the comparison rendered 110,792 candidates. After filtering, 10 sequences containing the SWGGA conserved motif remained, and among them only six were not redundant. The GISSG motif was not present in any sequence, and the GHSQGG motif was found in four sequences (Table 1).

BLAST comparison against the combined metagenome database with the *C. rugosa*-like family model sequences did not render any hit. Using the model PHA depolymerase sequences provided four hits containing the conserved motif from *P. putida* KT2442, which were redundant with the above-mentioned (Table 1).

Differences in the results obtained with the different strategies are due to the stringency level used, the second approach being more restrictive and the first approach more prone to render false positives.

The differences found in the number of positive matches for the different groups of enzymes can be attributed to the scarce abundance of DNA from eukaryotes in the environment [6]. In addition, the presence of introns in some eukaryotic genes can be a drawback in the similarity search strategies that we have used.

Moreover, *Pseudomonas* members are among the most ubiquitous bacteria in the environment [24]. The fact that several sequences from *Pseudomonas* were repeated in different metagenomes, and the low abundance of sequences from actinomycetes and fungi, reflect the low sequencing depth of the different metagenomes. A much higher number of sequences seems to be crucial to allow finding less abundant or rare sequences like enzymes from the secondary metabolism.

Phylogenetic Analysis

In the case of fungal sterol-esterases/lipases (Fig. 1A), 8 out of 11 candidates grouped with the carboxyl esterases from *Bacillus*. This is a group of esterases (a/bH1.05) closely related with the lipases from *C. rugosa*, and also

Table 1. Number of candidates selected remaining at each step of the screening.

Strategy	Homology search	Conserved motif	Short sequences	Redundant sequences	Phylogenetic analysis
Custom databases	(Lipase) 14237	(GESAG) 38	11	11	0
	(Depol.) 110792	(SWGGA) 210	10	6	2
		(GISSG) 910	0	0	0
		(GHSQGG) 12	4	4	0
Metagenome database	(Lipase) 0	(GESAG) 0	0	0	0
	(KT2440) 16	(SWGGA) 4	4	4	2
	(GK113) 16	(GISSG) 0	0	0	0
	(Actinobacteria) 0	(GHSQGG) 0	0	0	0

The search criteria used are indicated in parentheses.

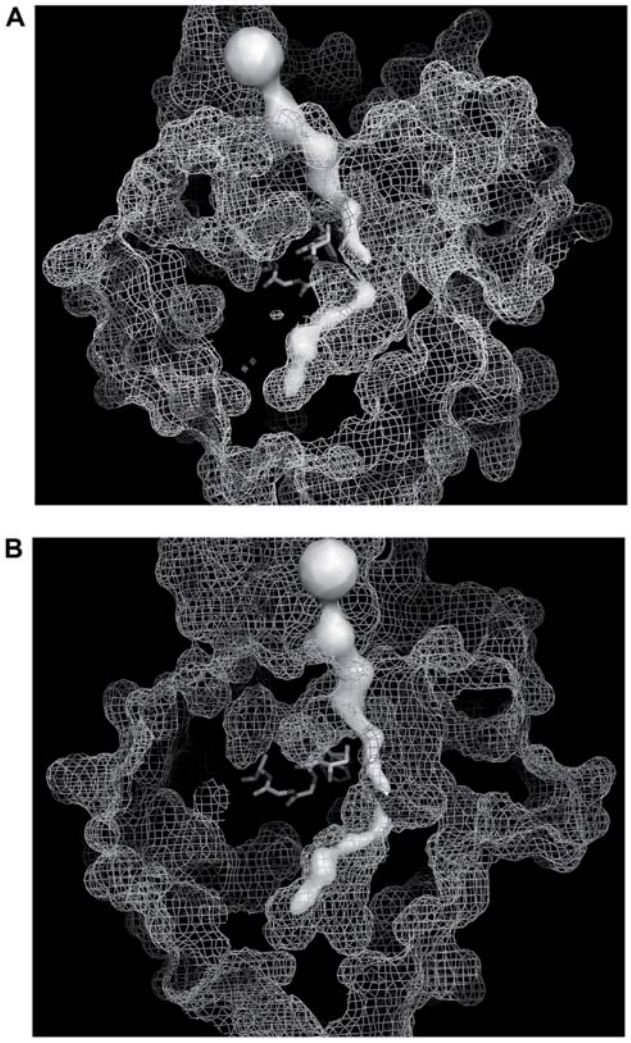

Fig. 2. Tridimensional structure models of the proteins. (**A**) 747106_KT2440_i-nPHAmcl. (**B**) Depolymerase 1092955233669, generated using the SWISS-MODEL server. Internal tunnels in each structure were modeled using Caver 2.0.

contains the motifs GGGF and GESAG [21]. *Bacillus* is a very common genus in many habitats [3]. Three of the candidates grouped with the lipases from the *B. subtilis* Brefeldin A family [25]; however, these enzymes have not been reported to hydrolyze sterol esters.

Fig. 1B shows how candidates 2066792343, 2042085485, and 1092955233669, containing the SWGAA motif, presented high similarity with the PHA depolymerase from *P. putida* KT2440. Sequence 2208943353, which grouped with the sequences from actinomycetes, did not present homology with any PHA depolymerase after BLAST comparison, but it presented similarity with α/β hydrolases from *Dietzia*, *Rhodococcus*, *Kribella*, and *Nocardia*. Three sequences grouped with GK13, but they did not contain the GISSG motif, and three more did not group with any model sequence.

The candidates 2066792343 and 1092955233669 gave positive matches in the two strategies employed and were selected for further analyses (Fig. 1B).

As a general consideration, our sequence-based screening methods rely on similarity to known sequences or to conserved motifs. Therefore, this approach is unable to detect genes with novel sequences. Nevertheless, unlike function-based methods, this approach is useful in screening genes without need of heterologous gene expression and protein folding in the selected host.

Molecular Modeling Analysis of the Selected Candidates

The two putative PHA depolymerase sequences were compared against the NCBI database using BLASTp. Sequence 1092955233669 had 99% identity with a predicted poly(3-hydroxyalkanoate) depolymerase from the genome of *P. fluorescens* WH6, presenting three amino acid substitutions. Sequence 2066792343 displayed 98% identity with a predicted poly(3-hydroxyalkanoate) depolymerase from the genome of *P. syringae* pv. *aesculi*, with 16 amino acid substitutions.

Molecular 3D models of the candidates and the model sequence were generated using the same template (Fig. 2). The model from *P. putida* KT2440 showed a typical esterase structure, with the catalytic Ser close to the mouth of the catalytic pocket. The prediction of the internal tunnels showed a big channel coincident with the substrate binding pocket (Fig. 2A). Candidate 2066792343 displayed a very similar model structure (not shown), but in the case of 1092955233669 the internal tunnel looked wider (Fig. 2B). Small differences in the substrate binding pocket could affect the enzyme specificity, admitting bigger substrates, or catalytic efficiency [5, 16].

In conclusion, we present here a useful sequential strategy to identify enzymes with potential biotechnological interest by means of metagenome mining. Search of genes from eukaryotes may be penalized mainly by the presence of introns and the predominance of DNA from prokaryotes in the environmental samples. After screening more than 7,000,000 sequences, two putative PHA depolymerases from *Pseudomonas* were selected. One of them presented specific features that may confer different properties.

Acknowledgments

This work was supported by the Spanish projects BIO2009-0844, BIO2012-36372, and S-2009AMB-1480. J. Barriuso is thankful for the financial support from the JAE-

DOC CSIC program. We thank the CIB-CSIC bioinformatics facility personnel.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol*. **215:** 403-410.
2. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, *et al*. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. **37:** W202-W208.
3. Barriuso J, Ramos Solano B, Santamaría C, Daza A, Gutierrez Mañero FJ. 2008. Effect of inoculation with putative PGPR isolated from *Pinus* sp. on *Pinus pinea* growth, mycorrhization and rhizosphere microbial communities. *J. Appl. Microbiol*. **105:** 1298-1309.
4. Barriuso J, Valverde JR, Mellado RP. 2011. Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics* **14:** 473.
5. Barriuso J, Prieto A, Martínez MJ. 2013. Fungal genomes mining to discover novel sterol esterases and lipases as catalysts. *BMC Genomics* **18:** 712.
6. Bolduc B, Shaughnessy DP, Wolf YI, Koonin EV, Roberto FF, Young M. 2012. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J. Virol*. **6:** 5562-5573.
7. Calero-Rueda O, Plou FJ, Ballesteros A, Martínez AT, Martínez MJ. 2002. Production, isolation and characterization of a sterol esterase from *Ophiostoma piceae. BBA Proteins Proteomics* **1599:** 28-35.
8. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. **39:** W29-W37.
9. Gangoiti J, Santos M, Prieto MA, de la Mata I, Serra JL, Llama MJ. 2012. Characterization of a novel subgroup of extracellular medium-chain-length polyhydroxyalkanoate depolymerases from actinobacteria. *Appl. Environ. Microbiol*. **78:** 7229-7237.
10. Jaeger KE, Eggert T. 2002. Lipases for biotechnology. *Curr. Opin. Biotechnol*. **13:** 390-397.
11. Jiang Y, Ye J, Wu H, Zhang H. 2004. Cloning and expression of the polyhydroxyalkanoate depolymerase gene from *Pseudomonas putida*, and characterization of the gene product. *Biotechnol. Lett*. **26:** 1585-1588.
12. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. 2009. The SWISS-MODEL repository and associated resources. *Nucleic Acids Res*. **37:** D387-D392.
13. Kontkanen H, Tenkanen M, Reinikainen T. 2006. Purification and characterisation of a novel steryl esterase from *Melanocarpus albomyces*. *Enzyme Microb. Technol*. **39:** 265-273.
14. Lee SY. 1996. Bacterial polyhydroxyalkanoates. *Biotechnol. Bioeng*. **49:** 1-14.
15. Lotti M, Tramontano A, Longhi S, Fusetti F, Brocca S, Pizzi E, Alberghina L. 1994. Variability within the *Candida rugosa* lipases family. *Protein Eng*. **7:** 531-535.
16. Mancheño JM, Pernas MA, Martínez MJ, Ochoa B, Rua ML, Hermoso JA. 2003. Structural insights into the lipase/esterase behavior in the *Candida rugosa* lipases family: crystal structure of the lipase 2 isoenzyme at 1.97A resolution. *J. Mol. Biol*. **332:** 1059-1069.
17. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, *et al*. 2012. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*. **40:** D123-D129.
18. Medek P, Benes P, Sochor J. 2007. Computation of tunnels in protein molecules using Delaunay triangulation. *J. WSCG* **15:** 107-114.
19. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, *et al*. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **19:** 386.
20. Olson SA. 2002. EMBOSS opens up sequence analysis. European molecular biology open software suite. *Brief Bioinform*. **3:** 87-91.
21. Pleiss J, Fischer M, Peiker M, Thiele C, Schmid RD. 2000. Lipase engineering database: understanding and exploiting sequence-structure-function relationships. *J. Mol. Catal. B Enzym*. **10:** 491-508.
22. Schmeisser C, Steele H, Streit WR. 2007. Metagenomics, biotechnology with non-culturable microbes. *Appl. Microbiol. Biotechnol*. **75:** 955-962.
23. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, *et al*. 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res*. **39:** D546-D551.
24. Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, Streit WR. 2003. Prospecting for novel biocatalysts in a soil metagenomes. *Appl. Environ. Microbiol*. **69:** 6235-6242.
25. Wei Y, Contreras JA, Sheffield P, Osterlund T, Derewenda U, Kneusel RE, *et al*. 1999. Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase. *Nat. Struct. Biol*. **6:** 340-345.