environmental microbiology

Environmental Microbiology (2020) 22(8), 3561–3571 doi:10.1111/1462-2920.15138

Developing robust protein analysis profiles to identify bacterial acid phosphatases in genomes and metagenomic libraries

Zulema Udaondo,[1,2†] Estrella Duque,[1†] Abdelali Daddaoua,[3] Carlos Caselles,[1] Amalia Roca,[4] Paloma Pizarro-Tobias[4] and Juan L. Ramos [1*]

[1] *Estación Experimental del Zaidín, CSIC, Granada, E-18008, Spain.*

[2] *Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, 72205, USA.*

[3] *Department of Biochemistry and Molecular Biology II, Faculty of Pharmacy, University of Granada, Granada, Spain.*

[4] *Bio-Iliberis R&D, Peligros, Granada, Spain.*

Summary

Phylogenetic analysis of more than 4000 annotated bacterial acid phosphatases was carried out. Our analysis enabled us to sort these enzymes into the following three types: (1) class B acid phosphatases, which were distantly related to the other types, (2) class C acid phosphatases and (3) generic acid phosphatases (GAP). Although class B phosphatases are found in a limited number of bacterial families, which include known pathogens, class C acid phosphatases and GAP proteins are found in a variety of microbes that inhabit soil, fresh water and marine environments. As part of our analysis, we developed three profiles, named Pfr-B-Phos, Pfr-C-Phos and Pfr-GAP, to describe the three groups of acid phosphatases. These sequence-based profiles were then used to scan genomes and metagenomes to identify a large number of formerly unknown acid phosphatases. A number of proteins in databases annotated as hypothetical proteins were also identified by these profiles as putative acid phosphatases. To validate these *in silico* results, we cloned genes encoding candidate acid phosphatases from genomic DNA or recovered from metagenomic libraries or genes synthesized in vitro based on protein sequences recovered from metagenomic data. Expression of a number of these genes, followed by enzymatic analysis of the proteins, further confirmed that sequence similarity searches using our profiles could successfully identify previously unknown acid phosphatases.

Introduction

Phosphorous is a major component of cells in all living organisms and all prokaryotic and eukaryotic cells have developed mechanisms for the uptake of inorganic phosphate, which is used in the biosynthesis of phospholipids, sugar phosphates, nucleotides and other molecules (Barea and Richardson, 2015). Despite phosphorous being one of the most abundant non-metallic elements in the earth's crust, it is frequently found in forms that are not bioavailable—a reality that often leads to phosphorous nutrient limitation (Ågren *et al*., 2012; Sosa *et al*., 2019). Inorganic phosphorous forms are often solubilized by plants and microorganisms (bacteria and fungi) through the production of weak acids (Barea and Richardson, 2015). However, a number of common organic phosphorous compounds (i.e., phytic acid, sugar phosphates, nucleotides, phospholipids and others) must be first hydrolysed by phosphatases to yield inorganic phosphate, which can subsequently be taken up by microorganisms and plants to be used as a phosphorous source (Hayes *et al*., 2000; Alori *et al*., 2017; Thomashow *et al*., 2018). Evidence suggests that phosphatase activity in soils and aquatic environments is of ecological relevance and is a driver of the productivity of terrestrial ecosystems (Turner *et al*., 2013; Margalef *et al*., 2017) and influence primary and secondary production in fresh waters and marine environments (Martiny *et al*., 2019).

There are two types of phosphatases among the phosphoric ester hydrolases which are defined based on their optimal pH. Alkaline phosphatases are a broad group of well characterized enzymes that use different mechanisms and co-factors to carry out their function (Mullaney and Ullah, 2003; Ragot *et al*., 2015; Lidbury *et al*., 2017;

Received 16 June 2020; accepted 19 June, 2020. *For correspondence. E-mail juanluis.ramos@eez.csic.es; Tel. +34 958181600 ext. 289; Fax. +34 958572753. †These two co-authors contributed equally to this work.

Neal *et al*., 2018). Acid phosphatases are, in general, non-specific phosphatases with broad substrate specificity and are often secreted across the outer membrane or are located in the periplasmic space (Thaller *et al*., 1997). At least three different types of prokaryotic phosphatases that function at acidic pH have been distinguished mainly based on their sequences; they are known as types A, B and C (Thaller *et al*., 1997; Lidbury *et al*., 2017; Neal *et al*., 2018). It was noted that the B class phosphatases are generally associated with pathogenic microbes, whereas the other types are widely distributed in nature (Neal *et al*., 2018). Although the importance of acid phosphatases to the acquisition of phosphorous in soils, fresh waters and marine environments (Neal *et al*., 2018); Margalef *et al*. (2017) compiled phosphatase activities from a large number studies of natural ecosystems and made 329 observations for acid phosphatases versus 72 for alkaline phosphatases, highlighting the environmental importance of acid phosphatases.

The work described here aims to contribute further to the understanding of organic phosphorous mobilization in the environment by acid phosphatases. To this end, we developed three robust profiles that can unequivocally identify the different types of acid phosphatase types. We have empirically validated the profiles by cloning and expression of putative acid phosphatases rescued from genomes or recovered from functional metagenomic libraries or genes synthesized *in vitro* based on protein sequences recovered from metagenomic libraries (Fierer *et al*., 2013; Berini *et al*., 2017; Duque *et al*., 2018). This profiling methodology will serve as a valuable resource for the identification of these important enzymes within the preponderance of already sequenced genomes and widely available metagenomic data. Furthermore, this study provides a proof-of-concept for the successful use of profiles to characterize enzymes involved in biogenic cycles.

Results and discussion

As a first step towards the identification of bacterial acid phosphatases, we retrieved 4644 sequences annotated as bacterial acid phosphatases (either due to protein name or Pfam domain composition) from the Uniprot Database (UniProt: a worldwide hub of protein knowledge, 2019). A phylogenetic tree was constructed with a refined set of 3741 protein sequences (see Experimental procedures), and the results are shown in Fig. 1. The bacterial acid phosphatase tree has, as expected, three clear branches; one represented by the outer blue circle which corresponds to class B (Fig. 1), another represented by the outer purple circle that corresponds to class C and the other represented by the outer green circle that corresponds to Generic Acid Phosphatases class A (GAP) (Fig. 1). Supplementary Table 1 contains information collected from the Uniprot database for each of the refined datasets of acid phosphatases. The phylogenetic tree from Supplementary Fig. 1 shows that acid phosphatases from class GAP, B and C belong to three well-defined monophyletic groups. The unrooted tree also revealed that sequences from class A and C are closest relatives, and therefore, sequences from class B are from an evolutionary point of view more distant from the other two.

The blue branch of the tree grouped 512 sequences that corresponded with annotated acid phosphatases of class B, the purple branch included 1701 sequences of annotated class C acid phosphatases; whereas the other set, which we named GAP, comprised 1528 non-specific class A acid phosphatases. The analysis of the sequences at the family level showed that class C and GAP proteins were found widely distributed among microbes that inhabit soils, fresh water and marine environments. In contrast, class B acid phosphatases are present in a limited number of microbial families which include Enterobacteriaceae, Pasteurellaceae, Morganellaceae, Aeromonadaceae and Vibrionaceae (Figs 1 and 2), of which some are pathogens (Supplementary Tables 2A–C). Conversely, it should be noted that class C and GAP acid phosphatases were also present in some Enterobacteriaceae. For example, in *Salmonella* and *Klebsiella* genomes, GAP proteins were identified and in a number of *Enterobacter* species (mainly *cloacae*) class C proteins were found. In contrast in the *Escherichia coli* species, despite being a broad taxonomic group (Abram *et al*., 2020), only class B acid phosphatases were identified (Supplementary Tables 2A–C).

Bacterial acid phosphatases have previously been identified through a number of signatures; for example, the database of families and domain proteins PROSITE (Sigrist *et al*., 2002) identified bacterial acid phosphatase sequences based on short sequence pattern motifs defined by the signature PS01157 (pattern G-S-Y-P-S-G-H-T). The compendium of protein fingerprints *PRINTS* database (Attwood *et al*., 2000) contains the signature PR00483 which corresponds to a five-element fingerprint from bacterial acid phosphatases derived from an initial alignment of a limited number of sequences. Four profiles were available from TIGRFAM database that were constructed using a limited set of acid phosphatase sequences (TIGR03397, 01675, 01672 and 01668); however, these profiles were found to have no discriminatory power. Other databases, such as Pfam domain protein database (El-Gebali *et al*., 2019) and Simple Modular Architecture Research Tool (SMART) (Letunic and Bork, 2018), contain a number of entries related to identification and classification of bacterial acid phosphatases. Nonetheless, none of the above motifs and classifications

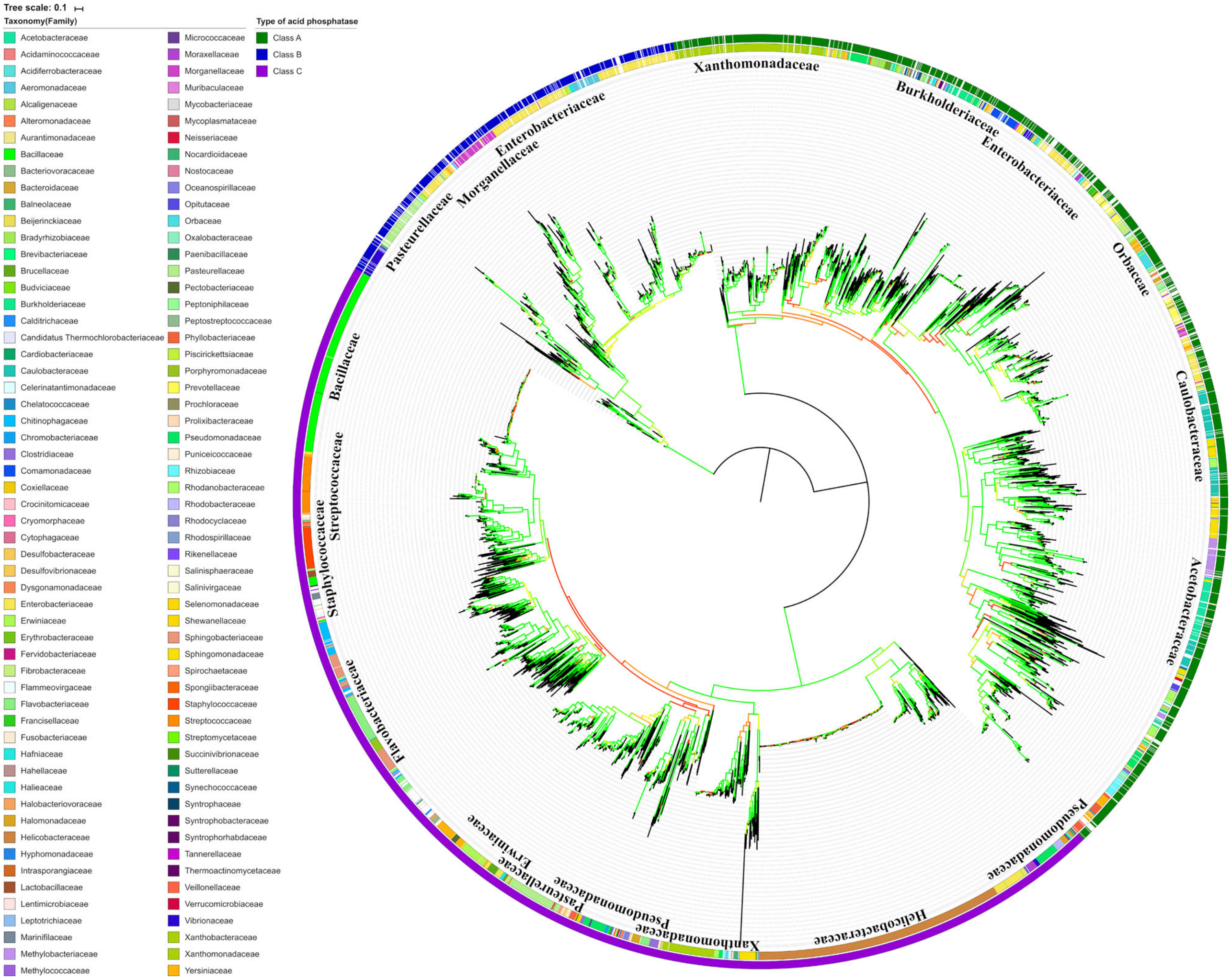

Fig. 1. Maximum likelihood phylogenetic tree of bacterial acid phosphatases. The maximum likelihood tree was inferred from a simultaneous comparison of 3741 protein sequences of bacterial acid phosphatases. Tree topology and branch lengths were calculated by maximum likelihood using the WAG+F+R10 model of evolution for amino acid sequences in IQ-TREE software Nguyen *et al*., 2015. The tree was rooted by using clade B as an outgroup that shows a clear separation between the three clades of acid phosphatase proteins. Colours of the branches represent levels of significance obtained in the bootstrapping analysis using 1000 bootstrap replications. Green indicates percentages close to 100% of confidence in the bootstraping analysis. The unrooted tree obtained using the same sample set it is shown in Supplementary Fig. 1.

distinguishes unequivocally between the three classes of bacterial acid phosphatases.

To establish a new criterion defining the three kinds of acid phosphatases represented in the phylogenetic tree, we decided to explore the construction of PROSITE generalized profiles, which are not available in the PROSITE database (https://prosite.expasy.org). Profiles are weight matrices that are useful for grouping proteins into families (Gromiha, 2010) and use quantitative motif descriptors which are given as linear sequences that comprise weighted match or mismatch residues and insert sequences in a profile position (Sigrist *et al*., 2002). Given that the phylogenetic tree defined three branches, according to differences in their amino acid sequences, we expected that a Profile for each of the branches would result in a net gain in specificity for identification and assignation of the entire collection of bacterial acid phosphatases.

To construct the three new profiles, we proceeded as suggested by PROSITE (https://prosite.expasy.org/prosuser.html#meth_prf). To create these profiles, we used the three sets of proteins identified in each of the branches of the tree, the profile for class B phosphatases (Prf-B-Phos) was constructed using a set of 512 seed sequences, for the profile for class C we used 1701 sequences, whereas for the profile for GAP (Pfr-GAP), due to the high variability in the sequence similarity and sequence length from members of this class, we used a filtered set of 948 of the 1528 sequences from the previous analysis (Supplementary Tables 3A–C). The three

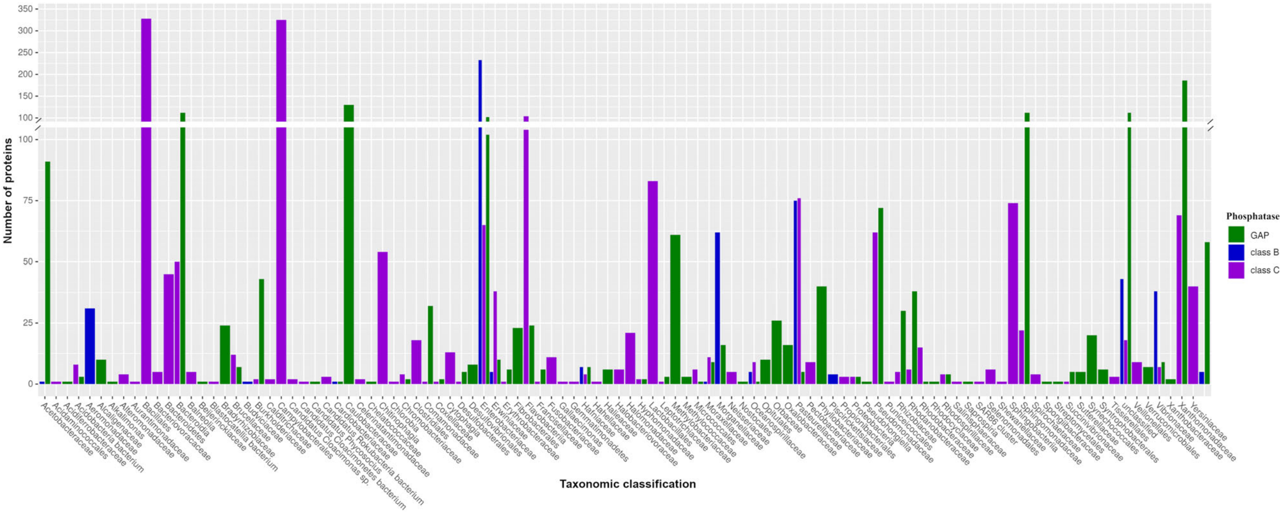

Fig. 2. Taxonomic distribution of the number of sequences used to construct the three acid phosphatase profiles (Prf-GAP, Prf-B and Prf-C). Sequences were downloaded from the Uniprot database according to their functional annotation. The number of proteins per taxonomic group was plotted using ggplot2 library in R (Wickham, 2016).

profiles obtained in this study are publicly available in Supplementary Table 4. The generation of a profile requires a multiple-alignment of the seed sequences as input, which was performed using Muscle (Edgar, 2004). The consensus sequences derived from the multiple-alignments (Supplementary Tables 3A–C) showed conserved regions with high-sequence identity scattered throughout the full sequence of the proteins. This reflects the existence of several functional constraint regions, with lower site-specific substitution ratio distributed along the protein sequences belonging to each of the classes. This is in contrast with most sequence patterns where high-sequence identity regions are restricted to active sites, cofactor binding domains or specific DNA binding regions (Fuglebakk *et al*., 2012).

The multiple-alignment revealed that the short patterns used previously to define these acid phosphatases were in a wider sequence identity context, and this warranted the construction of profiles to encompass the full gamut of acid phosphatase sequences belonging to these families. We used the script *pfmake* to translate the multiple-alignment into a matrix table of positions and convert frequency distributions into positive specific amino acid weights and gaps according to the original algorithms of Sibbald and Argos (Sibbald and Argos, 1990) and Gribskov *et al*. (1987). Once the profiles were constructed, we proceeded to calibrate and validate the profiles as recommended by PROSITE (described in Experimental procedures); for this, the profiles were run against a database to produce a list of sorted scores. It has been previously empirically determined that cut-off values of *Z*-scores equal or greater than 8.5 are biologically significant and warrant the correct assignment of a protein to a family (Gallegos *et al*., 1997; Sigrist *et al*., 2002; Godoy *et al*., 2010).

As a proof of concept, the three profiles were used as input for *pfsearch v2.3* from the PTOOLS suite to scan the complete set of Uniref100 proteins (downloaded from the UniProt database on May 24, 2019). As a result, 6000 proteins were matched with Pfr-GAP (Fig. 3 and Supplementary Fig. 2), 2132 protein sequences were matched by the Pfr-B-Phos (Fig. 3 and Supplementary Fig. 3) and 10,494 with Pfr-C-Phos (Fig. 3 and Supplementary Fig. 4).

We found that Pfr-B-Phos identified acid phosphatases preferentially from enterobacteria, vibrios and other microorganisms mainly from orders Pasteurellales and Bacillales (see Supplementary Fig. 3) whose life style indicated a close relationship with eukaryotes, as mentioned above, and confirming previous studies (Gandhi and Chandra, 2012; Neal *et al*., 2018). Conversely, we found that Pfr-C-Phos and Pfr-GAP identified acid phosphatases from a variety of different sources in a highly-specific and sensitive manner, including Acidobacteria, Actinobacteria, alpha, beta, gamma and epsilon proteobacteria, Firmicutes, Verrumicrobia and Bacterioidetes among many others (see Supplementary Figs 2 and 4). The results obtained with the three profiles against Uniref100 database (Suzek *et al*., 2015) demonstrated the ability of Pfr-GAP, Pfr-C-Phos and Pfr-B-Phos to discriminate between all classes of acid phosphatases displayed in the phylogenetic tree and within a wide taxonomic range. It is worth noting that although the three profiles were developed using only bacterial sequences, presumed eukaryotic acid phosphatases were also found in all cases. The complete set of raw hits sorted by output

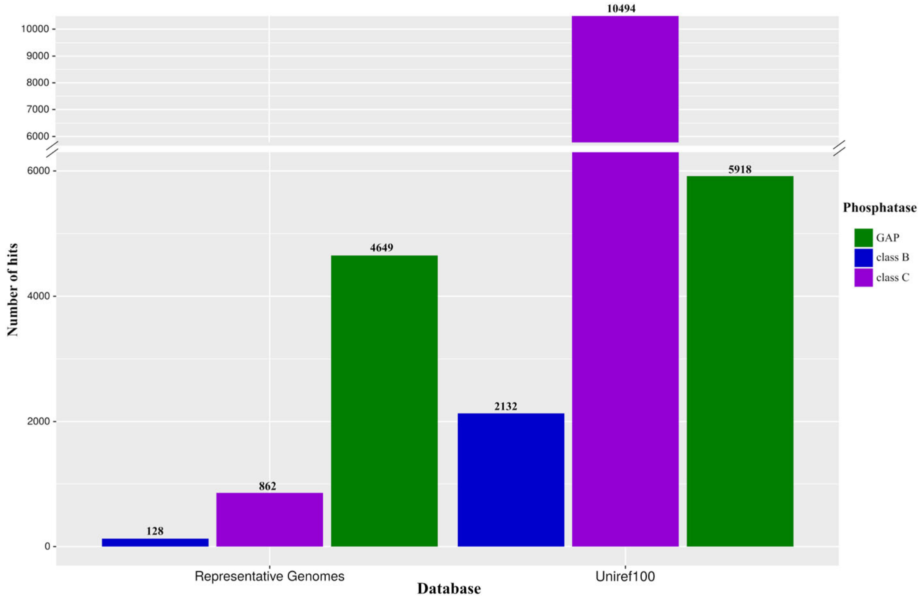

Fig. 3. Number of hits found by the constructed profiles of three classes of acid phosphatases (Prf-GAP, Prf-B and Prf-C). Using *pfscan* tool on protein sequences from the Uniref100 database (the last three columns on the right) and a local database of proteomes of 5639 representative bacterial and archaea genomes (the three most left columns) downloaded from NCBI.

score is shown in Supplementary Table 5. Remarkably, although these are non-filtered results, the high accuracy of the three profiles allowed the identification of proteins belonging to each of the classes at very low score numbers. The specificity of the profiles also identified a large number of putative acid phosphatase sequences which were annotated in the Uniref100 database as ‘uncharacterized protein’.

To further validate the new profiles, we decided to test if Pfr-GAP, Pfr-C-Phos and Pfr-B-Phos could identify acid phosphatases within available annotated whole genomes, in metagenomic libraries in which proteins are annotated as Hypothetical Proteins of unknown function, as well as proteins recovered from functional metagenomic libraries after screening for positive phosphatase activity. We found that the Pfr-GAP, Pfr-C-Phos and Pfr-B-Phos profiles could indeed identify a number of potential acid phosphatases in all of these screens. Specifically, we found that in the annotated reference, genomes collected from the NCBI database 4649 proteins were identified by the Prf-A GAP profile, 862 by the Pfr-C-Phos profile and 128 proteins by the Pfr-B-phos profile (Fig. 3). For most type of strains, the number of GAP acid phosphatases and class C was between 1 and 3, although we found 13 genomes with six GAP acid phosphatases and two genomes with up to five class C acid phosphatase. In those genomes, in which an acid phosphatase of class B was present, a single gene was always found, except in one case in which a duplication was identified, and another genome which bore four class B acid phosphatase genes. As validation of the proof of concept, we rescued acid phosphatases from the genomes of two microorganisms (i.e., *Pyrococcus*, and *Bacillus subtilis* strain 168). A search using the three profiles with *pfsearch* against the genomes of *Pyrococcus furiosus* DSM 3638 and *Bacillus subtilis* str. of note, 168 identified the protein sequences PF0040 and BSU_36530 as putative GAP acid phosphatases, encoded in each genome respectively. *Bacillus subtilis* BSU_36530 was previously annotated as undecaprenyl diphosphatase, whereas PF0040 from *P. furiosus* was annotated as an acidic acid phosphatase. To confirm these ‘hits’ empirically, we used whole chromosomal DNA from these microorganisms and cloned the amplified DNA into pET28 as described in Experimental procedures (Table 1).

As an initial step for confirmation of phosphatase activity, we spread the cells on LB medium supplemented with BCIP and found that colonies turned deep blue, suggesting that the cloned genes encoded, as expected, phosphatases. A single random clone bearing the gene from each of the two microorganisms was kept. Then, cells were grown in liquid LB and acid phosphatase activity determined over a wide pH range in permeabilized cells as described in Experimental procedures. The results revealed that the optimal pH was in the range of 5–6 (Table 2).

Our laboratory previously screened a functional metagenomic library from hydrocarbon-polluted soil after land farming and identified a clone, named FOS M2-62, that had robust phosphatase activity (see Experimental procedures). The fosmid of this clone was sequenced, and our profiles were used to identify it as a putative GAP acid phosphatase. We subsequently cloned it into pET28 to generate pET28_FOS M2-62. Phosphatase assays revealed that the AP-M2-62 protein had high activity

Table 1. Strains and plasmids used in this study.

Strains or plasmids	Genotype or relevant characteristics	Reference
Escherichia coli EPI 300	*recA1*, *endA1*, *araD139*, *rpsL*, *nupG*, *trfA*	Epicentre
Escherichia coli BL21 (DE3)	F′/*ompI*, *hsdS*, *gal*, *dam*, *met*	(Studier *et al*., 2009)
Plasmids		
pMBL	Vector for cloning PCR amplicons, Ap	Dominion
pET28a	Expression vector, 6xHis, Km	Novagen
pET28::FOS M2-62	pET28 containing the complete gene encoding acid phosphatase FOSM 2-62	This study
pET28:BSU	pET28 containing the complete gene encoding acid phosphatase from *Bacillus subtilis*	This study
pET28:PYR	pET28 containing the complete gene encoding acid phosphatase from *Pyrococcus furiosus*	This study
pET28:MET_A1	pET28 containing the complete gene encoding the MEAT_A1 GAP acid phosphatase deduced from environmental metagenomes	This study
pET28:MET_A2	pET28 containing the complete gene encoding the MEAT_A2 GAP acid phosphatase deduced from environmental metagenomes	This study
pET28:MET_B1	pET28 containing the complete gene encoding the MEAT_B1 class B acid phosphatase deduced from environmental metagenomes	This study
pET28:MET_B2	pET28 containing the complete gene encoding the MEAT_B2 class B acid phosphatase deduced from environmental metagenomes	This study
pET28:MET_C1	pET28 containing the complete gene encoding the MEAT_C1 class C acid phosphatase deduced from environmental metagenomes	This study
pET28:MET_C2	pET28 containing the complete gene encoding the MEAT_C2 class C acid phosphatase deduced from environmental metagenomes	This study

Ap and Km stand for resistance to ampicillin and kanamycin.

between pH 5 and pH 7 (Table 2), but lower activity at pH greater than 7 or lower than 5. This suggests that AP-M2-62 is indeed an acid phosphatase.

We then explored the ability of our constructed profiles to identify hypothetical proteins as putative acid phosphatase from metagenomic libraries. To this end, we screened 1,552,866 hypothetical proteins from soil metagenomes and 4,925,568 sequences from marine metagenomes (downloaded in June 2019 from the NCBI database), and we found that the search yielded a total of 539 hypothetical proteins from the soil metagenome and 351 hypothetical proteins from marine metagenomes using Pfr-GAP profile (Supplementary Table 5). The Pfr-C-Phos profile was able to find 242 proteins from marine metagenomes and 23 from terrestrial metagenomes. The Pfr-B-Phos profile was able to find only 11 proteins from marine metagenomes. These results are in line with the initial phylogenetic tree results in the sense that class B proteins are poorly represented in marine and terrestrial ecosystems.

This data confirmed that among non-characterized acid phosphatases, generic acid phosphatases and class C phosphatases were more abundant than class B, and that class C and GAP can be considered cosmopolitan proteins as they can be found in a wide range of niches. We found that among the set of non-characterized proteins, one acid phosphatase could be rescued per 3000

Table 2. Relative acid phosphatase activity of genes amplified from genomic DNA and recovered from metagenomic libraries at different pHs.

Enzyme source								
pH	MET_A1	MET_A2	MET_B1	MET_B2	MET_C1	MET_2	M2-62	Bacillus
2	1	5	30	5	3	2	2	2
3	9	15	30	59	23	8	8	15
4	41	23	41	56	77	21	21	22
5	79	90	50	55	106	30	85	90
5.5	100	100	100	100	100	100	100	100
6	97	16	73	43	61	80	98	97
7	93	7	59	35	16	10	93	81
8	71	1	39	17	7	9	47	30
9	33	2	32	5	5	6	16	9

The set of acid phosphatases were expressed in *Escherichia coli*, and the assays carried out as described in Materials and methods at different pH in Britton-Robisson poly-buffer. Activities are expressed as relative activity, the maximum activity for all of the enzymes was at pH 5.5, and the corresponding value is considered 100% in each case. Results shown are the average of at least three replicates with standard deviations below 20% of the given values. Supplementary Table 5 shows the activity for each enzyme at pH 5.5 in nanomoles of *p*-nitrophenol produced per minute per milligram of cell dry weight at 25°C.

sequences in soil metagenomes, whereas one acid phosphatase protein was found every 14,000 sequences in marine metagenomes. Because the quality of metagenomic sequences is non-homogeneous and because our data are raw hit counts, at present, we cannot make any conclusions regarding the biogeographic distribution of acid phosphatases based on metagenomic data.

Considering the apparent abundance of these sequences, we explored whether the identified sequences were indeed acid phosphatases. To this end, we choose two sequences with the highest Z-score from each acid phosphatase family (Supplementary Table 6) and synthesized the corresponding genes. We then cloned and expressed them in *Escherichia coli* and enzyme activity was determined in permeabilized whole cells using the Britton-Robinson poly-buffer. We found that the six metagenomic acid phosphatase had optimal activity at acidic pH (Table 2 and Supplementary Table 7). These results further validate the ability of the profiles to find acid phosphatase enzymes from metagenomes. It is worth mentioning that although the MET_A1 enzyme exhibited the highest activity at pH 5.5–6, it had significant activity pH in the pH range between 5 and 9 (Table 2).

To further characterize in more detail, the kinetics properties of the metagenomic acid phosphatases, we purified three proteins (see Experimental procedures) and the kinetics parameters determined using isothermal titration calorimetry (ITC) (Watt, 1990; Williams and Toone, 1993). The initial rate of reaction (V_o) with different concentrations of pNPP was determined from the slope of the linear portion of the curve of integrated heats versus time as described by Bianconi (2003). We found that values for V_o followed typical Michaelis–Menten kinetics and K_{cat} and K_M were calculated by fitting the curve to the Michaelis–Menten kinetics equation using non-linear regression (Ababou and Ladbury, 2006). For MET_A_1, M2-F62 and MET_C_1, values of K_M were 49.3 ± 2.6 μM, 29.7 ± 0.02 μM and 23.8 ± 6.9 μM respectively; and K_{cat} were $0.63\ s^{-1}$, $0.55\ s^{-1}$ and $0.26\ s^{-1}$ respectively. Our results revealed that the substrate affinities were in the low micromolar range with up to twofold differences; K_{cat} values differed by up to 2.5-fold. The K_M values we determined are lower than those measured for acid phosphatases from different sources using classical spectrophotometric assays (Reilly *et al*., 2009; Zhang *et al*., 2013; Wang *et al*., 2018).

Conclusions

In conclusion, we have constructed a phylogenetic tree for acid phosphatases that grouped them into three branches. For each of the branches, a Prosite profile was constructed and validated; the three profiles were shown to be effective in the differentiation of the three sets of acid phosphatases. These profiles were able to assign a set of proteins annotated as hypothetical proteins in databases as being acid phosphatases (Supplementary Table 4). We tested our 'hits' empirically and confirmed phosphatase activity at acidic pH. Use of these profiles and the underlying strategy could serve as a powerful approach to explore the role that acid phosphatases play in primary productivity in edaphic and aquatic environments.

Experimental procedures

Phylogenetic tree construction

Sequences were downloaded from the Uniprot database by filtering proteins that belong to the Domain = bacteria and the annotation = acid phosphatase and 5′ nucleotidase lipoprotein ep4 family; the later corresponds to class C acid phosphatases. Using these filters (on April 26, 2019), we retrieved 4644 protein sequences. Muscle v3.8.1551 (Edgar, 2004) alignment software with parameter—maxiters 1000—was used to align the set of 4644 protein sequences and construct the phylogenetic tree. Very divergent sequences were filtered and removed from the alignment until a final set of 3741 amino acid sequences were kept. The final set of sequences was aligned again using Muscle v3.8.1551 with the same parameters. Aligned sequences were used as input for the IQ-TREE software v1.6.10 (Nguyen *et al*., 2015) with parameters -nt AUTO, -bb 1000 -m TESTMERGE. The maximum likelihood tree was constructed following the model of evolution WAG with parameters F+R10 (IQ-TREE uses ModelFinder). Phylogenetic trees were plotted using the Interactive Tree of Life (iTOL) suite software v4 (Letunic and Bork, 2016).

Profile construction

To construct PROSITE 'generalized' profiles, first, we established the 'seed protein sequences' that would determine the sensitivity and average quality of the profiles.

Once visualized, the phylogenetic tree branches annotated as class B phosphatases, class C and generic acid phosphatases were aligned separately and filtered according to observed divergences in the alignment. Then *pfw* and *pfmake* scripts from PFTOOLS v2.3 (Gribskov *et al*., 1987; Sigrist *et al*., 2002; Bucher *et al*., 2015) were used to compute new weights for each individual sequence from the multiple sequence alignment and to construct the profile respectively. The matrix BLOSUM 45 was selected for the construction of the profile.

Pfsearch and *pfscan* were used to calibrate each profile against a calibration database. The calibration database was made from the entire collection of Swiss-Prot protein sequences filtered by Taxonomy = bacteria. The database contained a total of 334,009 sequences that where shuffled randomly with a sliding window of 20 residues using the script fasta-shuffle-letters from MEME suite v5.0.2 (Bailey *et al*., 2015).

Searches with the three profiles using Uniref100 database, a local database of representative bacterial and archaea sequences and hypothetical protein databases from metagenomic samples, were all done using *pfscan* script from PFTOOLS v2.3 with parameters -z -f (Sigrist *et al*., 2002).

Sequences in databases

Uniref100 database was downloaded to be used locally in May, 2019. The set of protein FASTA sequences from representative strains was downloaded from the NCBI database in August, 2019. The set of representative strains was obtained via genome browse from NCBI https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/ and then filtered by 'archaea' AND 'bacteria' AND 'representative genome'. The two sets of hypothetical proteins used in these analyses were obtained from NCBI protein database using filters: 'soil metagenome' AND 'hypothetical protein' and 'marine metagenome' AND 'hypothetical protein'.

Construction of a functional soil metagenomic library

Soil samples were taken from hydrocarbon polluted soil after land farming. High-molecular-weight DNA extraction was performed from the soil using the commercial GNOME DNA kit (MP, Biomedicals) according to the manufacturer's instructions. DNA fragments of approximately 40 kb were recovered and ligated into the pCC1FOS vector (Epicentre®), and the product was transduced into *E. coli* EPI300 (Raleigh *et al*., 2002) according to the manufacturer's protocol. Screening for phosphatase activity was performed by replicating the metagenomic library onto agar LB plates with 40 mg per mL of 5-bromo-4-chloro-3-indolyl phosphate (BCIP Applichem, Darmstadt, Germany) as substrate, supplemented with 12.5 μg per mL *chloramphenicol and 0.01% L-arabinose*. Following replication, the colonies were incubated for 24 h at 37°C. A total of 64 clones with phosphatase activity were identified and detected as pale to dark blue colonies. A single clone, named M2-62, that turned deep blue on these plates was used for further analysis in this study.

Cloning of putative acid phosphatases in Escherichia coli

DNA from *Pyrococcus furiosus* DSM 3638 and *Bacillus subtilis* DSM 204 were obtained from the DSMZ culture collection. The *Bacillus subtilis* gene was PCR amplified with the following primers 5′-TTGAACTACGAAATTTTTAAAGCAATCC-3′ and 5′-TTCTTAGAAATTTTGATCGGTTGG-3′, whereas the *Pyrococcus* gene was amplified using the following pair of primers 5′-ATGCTGGCAATACTTACGGCAA-3′ and 5′-TCACTTATCCACTTTAAAAAAGATGCGC-3′; amplified DNA was subsequently cloned into pTOPO and further subcloned into pET28 after digestion with NdeI and EcoRI. Plasmids were transformed into *E. coli* BL21 (DE3) (Studier *et al*., 2009). For amplification of the open reading frame encoding the AP-M2-62 protein, fosmid DNA was prepared and the following primers: 5′-CATATGAAAAAAAATACCTGAACCCTTC-3′ (forward) and 5′-GGATCCTCAGTGCTGGGTCAG-3′ (reverse) were used. Following PCR amplification, under standard conditions, the fragment was cloned into the pMBL vector to yield pMBL_FOSM2-62. The plasmid was subsequently digested with NdeI/BamHI, and the 806 bp fragment bearing the ORF AP-M2-62 was cloned into pET28b (+) digested with the same enzymes (Table 1).

Cloning of putative metagenomic acid phosphatases in Escherichia coli

Protein sequences retrieved from metagenomic libraries with a high Z-score for GAP, class B and class C were manually curated. The protein sequences were then translated into DNA sequences with optimized codon usage for *E. coli*, synthesized in vitro by Genescript, cloned into pET28 and expressed from the P_{lac}.

Growth of Escherichia coli *and* in vivo *acid phosphatase activity*

Escherichia coli BL21 (DE3) transformed with the corresponding plasmid was grown in 100 ml conical flasks containing 25 ml of LB supplemented with 0.025 mg/ml kanamycin (pET28). Cultures were incubated at 37 °C with shaking until they reached a turbidity at 660 nm (OD_{660}) of 0.6, at which point 0.1 mM isopropyl-α-D-thiogalactopyranoside (IPTG) was added, to induce expression, incubation was continued overnight. After growth of *E. coli*, the turbidity of the cultures was adjusted to 1 in 600 μl of lysis buffer (100 mM acetate, pH 5.5, $CaCl_2$, 1 mM, and Tween 80, 0.01% or a drop of toluene) (Lassen *et al*., 2001). The assay was performed by combining 100 μl of permeabilized cells with 10 μl of a solution of 100 mM *p*-nitrophenyl phosphate (pNNP) dissolved in 0.1 M Na-acetate buffer, pH 5.5. The reaction mixture was incubated for 30 min at 25°C. Subsequently, 100 μl of 0.5 M sodium hydroxide in water was added to stop the reaction. The samples were then centrifuged in a bench centrifuge (5 min at 10000 rpm), and the absorbance at 405 nm was measured in a spectrophotometer. To determine the optimal pH range, the Britton-Robinson poly-buffer (40 mM boric acid, 40 mM acetic acid and 40 mM phosphoric acid) was adjusted with NaOH to a pH between 2 and 9 (Souri *et al*., 2013). Other conditions for the acid phosphatase assays are those mentioned above.

Protein purification

For protein purification, cells were suspended in 25 ml of buffer A (50 mM Hepes pH 6.9; 300 mM NaCl; 1 mM dithiothreitol) with EDTA-free protease inhibitor mixture. Cells were lysed by two passes through a French Press at a p.s.i. of 1000. The cell suspension was then centrifuged at 20,000*g* for 1 h. The pellet was discarded and the supernatant was filtered and loaded onto a 5 ml His-Trap chelating column (GE Healthcare, St. Gibes, UK). The proteins were eluted with a 10–500 mM gradient of imidazol in buffer A. The purity of the eluate was determined by running 12% SDS-PAGE gels. Homogenous protein preparations were dialyzed overnight against buffer A but supplemented with 10% [v/v] glycerol. Dialyzed protein was collected at a concentration of about 1 mg/ml and stored in 1 ml aliquots at −80°C.

Acknowledgements

Work in Granada was supported by grant RTI2018-094370-B-I00, at CSIC and grant EC-H2020-685474 at Bio-Iliberis R&D. Part of the data analysed in this work was performed on the High-Performance Computing Facilities, in particular the Grace cluster, provided by the University of Arkansas for Medical Sciences (UAMS), managed by the Department of Biomedical Informatics. We thank Ben Pakuts for critical reading of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

Ababou, A., and Ladbury, J.E. (2006) Survey of the year 2004: literature on applications of isothermal titration calorimetry. *J Mol Recognit* **19**: 79–89.

Abram, K., Udaondo, Z., Bleker, C., Wanchai, V., Wassenaar, T.M., Robeson, M.S., and Ussery, D.W. (2020) What can we learn from over 100,000 *Escherichia coli* genomes? *bioRxiv*: 708131. https://doi.org/10.1101/708131.

Ågren, G.I., Wetterstedt, J.Å.M., and Billberger, M.F.K. (2012) Nutrient limitation on terrestrial plant growth – modeling the interaction between nitrogen and phosphorus. *New Phytol* **194**: 953–960.

Alori, E.T., Glick, B.R., and Babalola, O.O. (2017) Microbial phosphorus solubilization and its potential for use in sustainable agriculture. *Front Microbiol* **8**: 141–146.

Attwood, T.K., Croning, M.D.R., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., *et al*. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res* **28**: 225–227.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015) The MEME suite. *Nucleic Acids Res* **43**: W39–W49.

Barea, J.-M., and Richardson, A.E. (2015) Phosphate mobilisation by soil microorganisms. In *Principles of Plant-Microbe Interactions: Microbes for Sustainable Agriculture*, Lugtenberg, B. (ed). Cham: Springer International Publishing, pp. 225–234.

Berini, F., Casciello, C., Marcone, G.L., and Marinelli, F. (2017) Metagenomics: novel enzymes from non-culturable microbes. *FEMS Microbiol Lett* **364**: fnx211.

Bianconi, M.L. (2003) Calorimetric determination of thermodynamic parameters of reaction reveals different enthalpic compensations of the yeast hexokinase isozymes. *J Biol Chem* **278**: 18709–18713.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (2015) A flexible motif search technique based on generalized profiles. *Comput Chem* **20**: 3–24.

Duque, E., Daddaoua, A., Cordero, B.F., Udaondo, Z., Molina-Santiago, C., Roca, A., *et al*. (2018) Ruminal metagenomic libraries as a source of relevant hemicellulolytic enzymes for biofuel production. *J Microbial Biotechnol* **11**: 781–787.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., *et al*. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* **47**: D427–D432.

Fierer, N., Ladau, J., Clemente, J.C., Leff, J.W., Owens, S. M., Pollard, K.S., *et al*. (2013) Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342**: 621–624.

Fuglebakk, E., Echave, J., and Reuter, N. (2012) Measuring and comparing structural fluctuation patterns in large protein datasets. *Bioinformatics* **28**: 2431–2440.

Gallegos, M.T., Schleif, R., Bairoch, A., Hofmann, K., and Ramos, J.L. (1997) Arac/XylS family of transcriptional regulators. *Microbiol Mol Biol Rev* **61**: 393–410.

U. Gandhi, N. and B. Chandra, S. (2012) A comparative analysis of three classes of bacterial non-specific acid phosphatases and archaeal phosphoesterases: evolutionary perspective. *Acta Inform Med* **20**: 167–173.

Godoy, P., Molina-Henares, A.J., Torre, J.D.L., Duque, E., and Ramos, J.L. (2010) Characterization of the RND family of multidrug efflux pumps: in silico to in vivo confirmation of four functionally distinct subgroups. *J Microbial Biotechnol* **3**: 691–700.

Gribskov, M., McLachlan, A.D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *PNAS* **84**: 4355–4358.

Gromiha, M.M. (2010) *Protein Bioinformatics: From Sequence to Function*. New Delhi: Academic Press.

Hayes, J.E., Richardson, A.E., and Simpson, R.J. (2000) Components of organic phosphorus in soil extracts that are hydrolysed by phytase and acid phosphatase. *Biol Fertil Soils* **32**: 279–286.

Lassen, S.F., Breinholt, J., Østergaard, P.R., Brugger, R., Bischoff, A., Wyss, M., and Fuglsang, C.C. (2001) Expression, gene cloning, and characterization of five novel phytases from four Basidiomycete fungi: *Peniophora lycii*, *Agrocybe pediades*, *a Ceriporia* sp., and *Trametes pubescens*. *Appl Environ Microbiol* **67**: 4701–4707.

Letunic, I., and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**: W242–W245.

Letunic, I., and Bork, P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* **46**: D493–D496.

Lidbury, I.D.E.A., Fraser, T., Murphy, A.R.J., Scanlan, D.J., Bending, G.D., Jones, A.M.E., *et al*. (2017) The 'known' genetic potential for microbial communities to degrade organic phosphorus is reduced in low-pH soils. *MicrobiologyOpen* **6**: e00474.

Margalef, O., Sardans, J., Fernández-Martínez, M., Molowny-Horas, R., Janssens, I.A., Ciais, P., *et al*. (2017) Global patterns of phosphatase activity in natural soils. *Sci Rep* **7**: 1–13.

Martiny, A.C., Lomas, M.W., Fu, W., Boyd, P.W., Chen, Y.L., Cutter, G.A., *et al*. (2019) Biogeochemical controls of surface ocean phosphate. *Sci Adv* **5**: eaax0341.

Mullaney, E.J., and Ullah, A.H.J. (2003) The term phytase comprises several different classes of enzymes. *Biochem Biophys Res Commun* **312**: 179–184.

Neal, A.L., Blackwell, M., Akkari, E., Guyomar, C., Clark, I., and Hirsch, P.R. (2018) Phylogenetic distribution, biogeography and the effects of land management upon bacterial non-specific acid phosphatase gene diversity and abundance. *Plant and Soil* **427**: 175–189.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B. Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268–274.

Ragot, S.A., Kertesz, M.A., and Bünemann, E.K. (2015) *phoD* alkaline phosphatase gene diversity in soil. *Appl Environ Microbiol* **81**: 7281–7289.

Raleigh, E.A., Elbing, K., and Brent, R. (2002) Selected topics from classical bacterial genetics. *Curr Protoc Mol Biol* **59**: 1.4.1–1.4.14.

Reilly, T.J., Chance, D.L., Calcutt, M.J., Tanner, J.J., Felts, R.L., Waller, S.C., *et al*. (2009) Characterization of a unique class C acid phosphatase from *Clostridium perfringens*. *Appl Environ Microbiol* **75**: 3745–3754.

Sibbald, P.R., and Argos, P. (1990) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J Mol Biol* **216**: 813–818.

Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., *et al*. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **3**: 265–274.

Sosa, O.A., Repeta, D.J., DeLong, E.F., Ashkezari, M.D., and Karl, D.M. (2019) Phosphate-limited ocean regions select for bacterial populations enriched in the carbon–phosphorus lyase pathway for phosphonate degradation. *Environ Microbiol* **21**: 2402–2414.

Souri, E., Kaboodari, A., Adib, N., and Amanlou, M. (2013) A new extractive spectrophotometric method for determination of rizatriptan dosage forms using bromocresol green. *DARU J Pharm Sci* **21**: 12.

Studier, F.W., Daegelen, P., Lenski, R.E., Maslov, S., and Kim, J.F. (2009) Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. *J Mol Biol* **394**: 653–680.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**: 926–932.

Thaller, M.C., Schippa, S., Bonci, A., Cresti, S., and Rossolini, G.M. (1997) Identification of the gene (aphA) encoding the class B acid phosphatase/phosphotransferase of *Escherichia coli* MG1655 and characterization of its product. *FEMS Microbiol Lett* **146**: 191–198.

Thomashow, L.S., LeTourneau, M.K., Kwak, Y.-S., and Weller, D.M. (2018) The soil-borne legacy in the age of the holobiont. *J Microbial Biotechnol* **12**: 51–54.

Turner, B.L., Lambers, H., Condron, L.M., Cramer, M.D., Leake, J.R., Richardson, A.E., and Smith, S.E. (2013) Soil microbial biomass and the fate of phosphorus during long-term ecosystem development. *Plant and Soil* **367**: 225–234.

UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**: D506–D515. https://doi.org/10.1093/nar/gky1049.

Wang, Z., Tan, X., Lu, G., Liu, Y., Naidu, R., and He, W. (2018) Soil properties influence kinetics of soil acid phosphatase in response to arsenic toxicity. *Ecotoxicol Environ Saf* **147**: 266–274.

Watt, G.D. (1990) A microcalorimetric procedure for evaluating the kinetic parameters of enzyme-catalyzed reactions: kinetic measurements of the nitrogenase system. *Anal Biochem* **187**: 141–146.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York. ISBN 978-3-319-24277-4: Springer-Verlag. https://ggplot2.tidyverse.org.

Williams, B.A., and Toone, E.J. (1993) Calorimetric evaluation of enzyme kinetic parameters. *J Org Chem* **58**: 3507–3510.

Zhang, G.-Q., Chen, Q.-J., Sun, J., Wang, H.-X., and Han, C.-H. (2013) Purification and characterization of a novel acid phosphatase from the split gill mushroom *Schizophyllum commune*. *J Basic Microbiol* **53**: 868–875.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Supplementary Table 1. Metadata obtained from Uniprot database for sequences used to construct acid phosphatase profiles. Sequences were obtained according to their functional annotation and were filtered as described in materials and methods.

Supplementary Table 2A. Identification of GAP acid phosphatases in enterobacteria. The basic information of the set of identified GAP phosphatases in enteric bacteria was determined using the complete set of sequences in Supplementary Table 1.

Supplementary Table 2B. Identification of class B acid phosphatases in enterobacteria. The basic information of the set of identified class B phosphatases in enteric bacteria was determined using the complete set of sequences in Supplementary Table 1.

Supplementary Table 2C. Identification of class C acid phosphatases in enterobacteria. The basic information of the set of identified class C acid phosphatases in enteric bacteria was determined using the complete subset of sequences in Supplementary Table 1.

Supplementary Table 3A. Summary of the sub-alignment utilized to construct the profile of 1528 sequences identified as Acid Phosphatase GAP. The consensus sequence derived from the multiple-alignment of the 948 members of the GAP Class is shown at the bottom of the table.

Supplementary Table 3B. Summary of the sub-alignment utilized to construct the profile of 512 sequences identified as Class B Acid Phosphatase. The consensus sequence derived from the multiple-alignment of the 512 members of the Class B is shown at the bottom of the table.

Supplementary Table 3C. Summary of the sub-alignment utilized to construct the Profile of 1701 sequences identified as Class C Acid Phosphatase. The consensus sequence derived from the multiple-alignment of the 1701 members of the Class C is shown at the bottom of the table.

Supplementary Table 4. Acid phosphatase profiles for classes GAP, B and C. The three generalized profiles were constructed using *pfsearch v2.3* from the PFTOOLS suit. They can be find concatenated in the same file in order (GAP, B, C). These profiles are ready to be used using PFTOOLS software (https://prosite.expasy.org/prosuser.html).

Supplementary Table 5. Summary of hits found using databases of soil and marine proteins from metagenomic samples annotated as hypothetical proteins. Databases were scanned using *pfscan* script from PFTOOLS v2.3 (20,21) and the three acid phosphatase profiles (Prf-GAP, Prf-B, Prf-C).

Supplementary Table 6. Amino acid sequences of acid phosphatases identified among 'uncharacterized proteins' in screening of metagenomic libraries. Two high Z-score proteins of each class were converted into DNA sequences using optimized codon usage for *Escherichia coli*. The synthesized genes were cloned into the pET28a plasmid and transformed into *E.coli* BL21 (DE3) for overexpression.

Supplementary Table 7. Acid phosphatase activity in permeabilized whole cells. Assays conditions are detailed in Materials and Methods; related data is also presented in Table 2. The enzymatic activity was determined at pH 5.5 and 25 °C for 15 to 30 min. Activity is expressed as nanomoles of *p*-nitrophenol produced per mg of cell dry wieigth per min.

Supplementary Fig. 1. Unrooted maximum likelihood phylogenetic tree of bacterial acid phosphatases. The tree was inferred from a simultaneous comparison of 3741 sequences of bacterial acid phosphatases. Tree topology and branch lengths were calculated using IQ-TREE software (18) and the WAG+F + R10 model of evolution for amino acid sequences. Colours of the branches represent levels of significance obtained in the bootstrapping analysis using 1000 bootstrap replications. Green indicates percentages close to 100% of confidence.

Supplementary Fig. 2. Number of hits found in Uniref100 database using *pfscan* script from PFTOOLS v2.3 and Prf-GAP profile. Bar plots were performed using ggplot2 library in R (Wickham *et al*., 2016).

Supplementary Fig. 3. Number of hits found in Uniref100 database using *pfscan* script from PFTOOLS v2.3 and Prf-B profile. Bar plots were performed using ggplot2 library in R (Wickham *et al*., 2016).

Supplementary Fig. 4. Number of hits found in Uniref100 database using *pfscan* script from PFTOOLS v2.3 and Prf-C profile. Bars plots were performed using ggplot2 library in R (Wickham *et al*., 2016).