Subcellular localization of marine bacterial alkaline phosphatases

Haiwei Luo[a,1], Ronald Benner[a,b], Richard A. Long[a,b], and Jianjun Hu[c]

[a]Department of Biological Sciences, [b]Marine Science Program, and [c]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved October 14, 2009 (received for review July 8, 2009)

Bacterial alkaline phosphatases (APases) are important enzymes in organophosphate utilization in the ocean. The subcellular localization of APases has significant ecological implications for marine biota but is largely unknown. The extensive metagenomic sequence databases from the Global Ocean Sampling Expedition provide an opportunity to address this question. A bioinformatics pipeline was developed to identify marine bacterial APases from the metagenomic databases, and a consensus classification algorithm was designed to predict their subcellular localizations. We identified 3,733 bacterial APase sequences (including PhoA, PhoD, and PhoX) and found that cytoplasmic (41%) and extracellular (30%) APases exceed their periplasmic (17%), outer membrane (12%), and inner membrane (0.9%) counterparts. The unexpectedly high abundance of cytoplasmic APases suggests that the transport and intracellular hydrolysis of small organophosphate molecules is an important mechanism for bacterial acquisition of phosphorus (P) in the surface ocean. On average, each marine bacterium possessed at least one suite of uptake of glycerol phosphate (ugp) genes (e.g., ugpA, ugpB, ugpC, ugpE) for dissolved organic phosphorus (DOP) transport, but only half of them had ugpQ, which hydrolyzes transported DOP, indicating that cytoplasmic APases play a role in hydrolyzing transported DOP. The most abundant heterotrophic marine bacteria, α- and γ-Proteobacteria, might hydrolyze DOP outside the cytoplasmic membrane, but the former could also transport and hydrolyze DOP in the cytoplasm. The abundant extracellular APases could provide bioavailable P for organisms that cannot directly access organophosphates, and thereby increase marine biological productivity and diversity.

PhoD | PhoX | PhoA | uptake of glycerol phosphate | protein localization

Phosphorus (P) is an essential element for life, and P cycling is intimately linked to carbon and nitrogen dynamics in the ocean (1). Although inorganic phosphate (P_i) is the preferred P source for microbial growth (2), it frequently becomes depleted in surface waters of many oceanic regions (2, 3). Dissolved organic phosphorus (DOP) dominates the total dissolved P pool in the surface ocean (4). Thus, the ability to use the DOP pool would be ecologically advantageous for marine microorganisms.

Alkaline phosphatases (APases) occur in a broad diversity of microorganisms and are important in the utilization of phosphoesters, one of the most abundant groups of DOP compounds in the ocean (5). To date, at least 3 prokaryotic APase gene families have been recognized (i.e., PhoA, PhoD, PhoX). They differ in substrate specificity and requirements of specific metal ions for their activities [supporting information (SI) Table S1]. Many marine microorganisms use APase enzymes to release P_i from phosphoesters, and thereby fulfill their P requirement for growth and reproduction (6).

APases have been reported primarily to be periplasmic in Gram-negative bacteria (7–10), but they also occur on the cell surface and extracellularly (11, 12). Because of their critical ecological role in organic P processing, efforts have been made to distinguish periplasmic and cell surface APase activities in marine bacteria (7). Quantification of the subcellular localizations of APases would provide valuable insights about the ecology of marine bacteria and enhance our understanding of the marine P cycle.

The Global Ocean Sampling (GOS) metagenomic database (13) provides an opportunity to investigate the distribution patterns of subcellular localizations of APases in the marine bacterial community. In this study, we developed a bioinformatics pipeline to identify APase peptide sequences and designed an algorithm to predict their subcellular localizations. The ecological and biogeochemical significance of our findings is discussed.

Results and Discussion

APase Distribution in World Ocean. A total of 935 PhoA, 887 PhoX, and 1,911 PhoD homologs were identified in this survey of GOS sampling sites. Normalizing APases to recA, a single-copy gene (14), indicated 0.50 and 0.32 APase per genome in the open ocean and coastal waters. This is consistent with previous observations in isolates in which 63% of open ocean bacterial isolates expressed APase compared with only 11–53% of coastal isolates (15, 16).

MEtaGenome Analyzer (MEGAN) (17) analysis binned the majority of APases as uncharacterized taxonomic groups (Fig. 1). This is in contrast to species distribution data, in which only 16.3% of the 16S rRNA was attributed to unclassified Proteobacteria and other bacteria (13). The substantial amount of unclassified APases indicated that APases in marine bacteria were highly diverged from their counterparts in characterized bacteria. Although α-Proteobacteria are more abundant than γ-Proteobacteria in the GOS samples (13), the latter made a greater contribution to the APase gene pool (Fig. 1). Planctomycetes accounted for 0.1% of the bacteria in the GOS samples, but they appeared to be an important source of APases (Fig. 1). In contrast, few characterized APases were affiliated with Gram-positive bacteria (Fig. 1), which represent 12% of bacteria in the GOS samples (13). *Alteromonadales*, *Burkholderiales*, and *Rhodobacterales* were consistently represented in all three APase families, whereas Planctomycetes, Bacteroidetes, and Cyanobacteria were overrepresented in PhoD, PhoA, and PhoX, respectively (Fig. 1). All six phylotypes were well represented in cytoplasmic, periplasmic, and extracellular APases (Fig. 1).

Examining the distribution of the individual APase families in the 46 open ocean and coastal water samples (Dataset S1), the mean numbers of PhoA and PhoX per genome were not significantly different (t test, $P = 0.09$), whereas the mean number of PhoD per genome was significantly greater than the mean number of either PhoA or PhoX per genome (t test, $P < 10^{-6}$ in either case). The mean number of each APase type per genome was greater in the open ocean than in coastal waters, but these differences were not significant (t test, $P > 0.05$). Strong

Author contributions: H.L. designed research; H.L. performed research; H.L. and J.H. contributed new reagents/analytic tools; H.L., R.B., R.A.L., and J.H. analyzed data; and H.L., R.B., R.A.L., and J.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: hluo2006@gmail.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907586106/DCSupplemental.

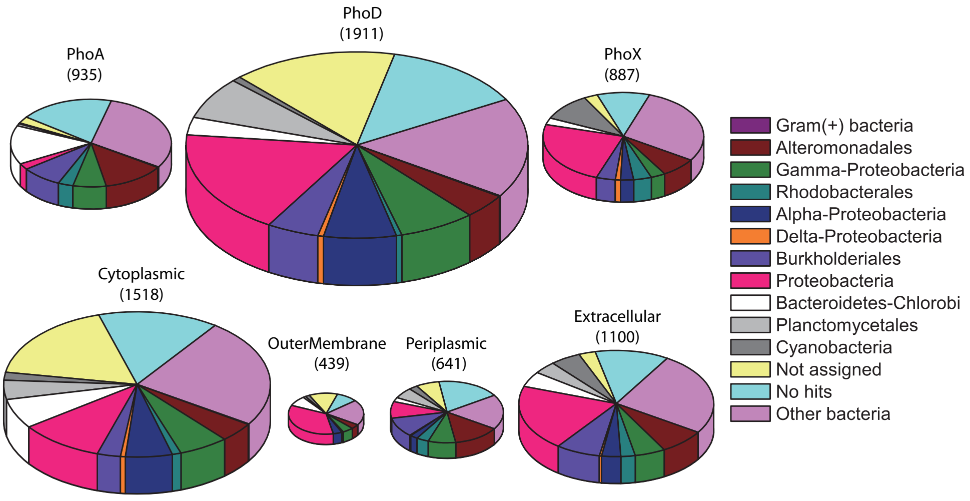

Fig. 1. Taxonomic distribution of APase genes recovered from the GOS metagenomic database. APases were sorted by family type and by subcellular localizations. The chart size indicates the relative abundance of APases, and the number of genes is shown. The γ-Proteobacteria category does not include Alteromonadales, which is shown separately. The α-Proteobacteria category does not include Rhodobacterales, which is shown separately. Proteobacteria only include those APase genes that cannot be assigned to any known phylotypes within Proteobacteria.

correlations were found between the number of PhoA, PhoD, and PhoX, respectively, and the total number of sequences sampled (r = 0.94, 0.91, and 0.92, respectively) (Dataset S1), indicating that each APase type is uniformly distributed across a variety of open ocean and coastal waters.

A recent study showed that PhoX is more abundant in the ocean than previously considered (18). We observed that PhoD is even more abundant than PhoX across a wide variety of marine habitats. Moreover, 4 pairs of PhoX and PhoD peptides were mapped to four paired reads and one pair of PhoX and PhoD was mapped to a single read (Dataset S2), indicating that PhoX and PhoD can co-occur in marine bacteria. Both PhoX and PhoD are activated by Ca^{2+}, an abundant ion in the ocean, whereas PhoA requires Zn^{2+} (Table S1), which often occurs at subnanomolar concentrations (18, 19). The replacement of Zn^{2+} with Ca^{2+} could be an important factor in the selection of PhoX and PhoD over PhoA in the ocean. We applied a loose but reliable searching criterion and identified more PhoA and PhoX homologs than previously shown (18). The abundant nature of PhoD genes suggests that they may play an important role in organophosphate hydrolysis in the surface oceans.

We then examined the subcellular localization of the hydrolysis of organophosphates. The MetaP algorithm was applied to the GOS datasets to sort APase sequences by their subcellular localizations. In total, there were 1,518 cytoplasmic, 641 periplasmic, 1,100 extracellular, 439 outer membrane, and 35 inner membrane APases. The mean number of cytoplasmic APases per cell across the open ocean and coastal waters was significantly greater than that of any other localization type per cell (t test, $P < 0.001$ in each case). The mean number of extracellular APases per cell was slightly greater than the mean number of the ectoenzymatic APases per cell, but this difference was not significant (t test, $P = 0.50$). Our finding of more periplasmic than outer membrane APases per cell is consistent with experimental evidence showing that more APase activity was in the periplasm than in the cell surface (7), but this difference was not significant (t test, $P = 0.28$). No significant differences were found between the mean number of each localization type per genome in the open ocean and in coastal waters (t test, $P > 0.05$). There was a significant correlation between the number of cytoplasmic APases and the total number of sequences sampled from the open ocean and coastal waters (r = 0.81). The same pattern was found for periplasmic and extracellular APases (r = 0.93 and 0.94, respectively). Membrane proteins were not included in the correlation analysis because of their low abundance. This indicates that the subcellular distribution of APases is generally homogeneous between the open ocean and coastal waters.

Examining individual APase families revealed different patterns in their subcellular localization. In both PhoA and PhoD, cytoplasmic proteins dominate over other proteins (one-sample proportion test, $P < 0.05$), whereas extracellular proteins comprised the majority of PhoX (one-sample proportion test, $P < 0.05$) (Fig. 2). The numbers of periplasmic proteins in both PhoA and PhoX were greater than in outer membrane proteins, although the inverse was observed for PhoD (Fig. 2). These patterns indicate PhoA, PhoD, and PhoX could be employed in different ecological strategies for using organophosphate in P-depleted surface waters.

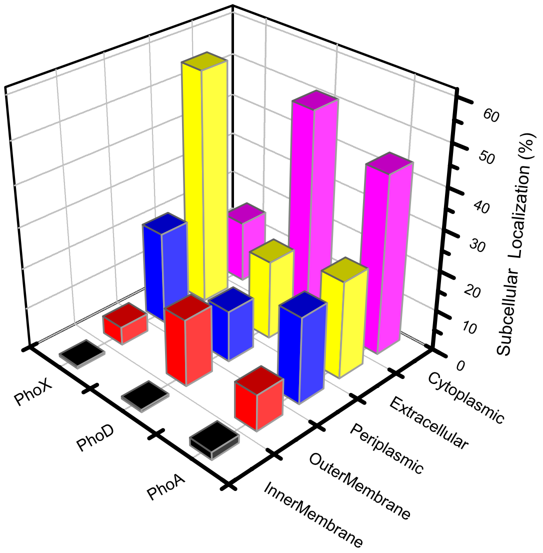

Fig. 2. Subcellular localization distributions of APases recovered from the GOS metagenomic database.

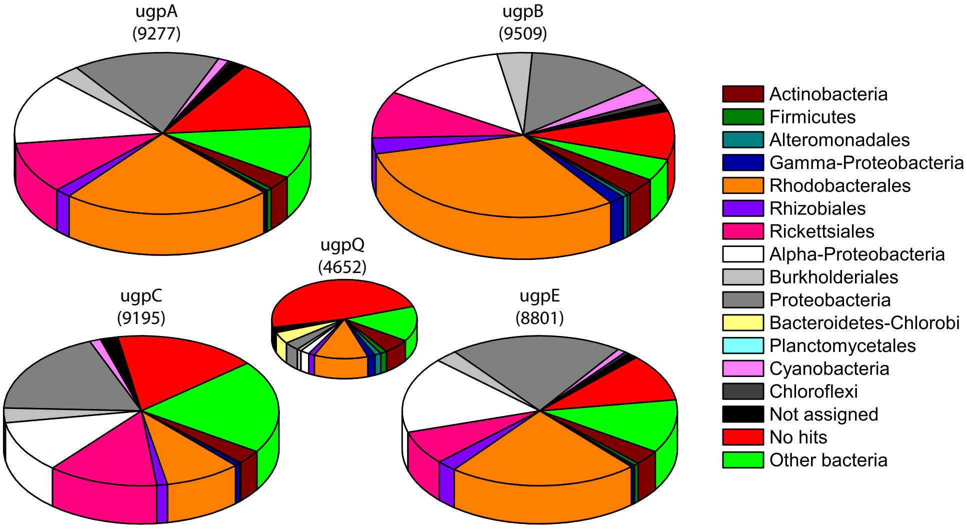

Fig. 3. Taxonomic distribution of ugp genes recovered from the GOS metagenomic database. The chart size indicates the relative abundance of ugp genes, and the number of genes is shown. The γ-Proteobacteria and Proteobacteria categories are the same as described for Fig. 1. The α-Proteobacteria category does not include Rhodobacterales, Rhizobiales, and Rickettsiales, which are shown separately.

Cytoplasmic APases and Uptake of Gllycerol Phosphate System for DOP Uptake. A substantial fraction (41%) of APases from the GOS database was located in the cytoplasm. Although no experimental evidence has shown the existence of cytoplasmic APases in isolated bacteria, it is possible that a greater diversity of APase subcellular localizations and metabolic pathways occurs in the ocean, because the GOS study found that over half of the 16S rRNA ribotypes at the species level (<97% identical) were not reported previously (13). A substantial number of highly diverged APases in the GOS database indicate that APases in marine bacteria could have evolved other functions (e.g., subcellular localizations).

Many bacteria possess the ability to take up monoesters and diesters of glycerol phosphate directly via the uptake of glycerol phosphate (ugp) system in P-limiting environments (20). Glycerol phosphate is the diacylation product of phospholipids (20), which are ubiquitous in bacteria and eukarya, accounting for ≈15–20% of total cellular P (1). In *Escherichia coli*, the ugp system consists of ugpB, ugpA, ugpE, ugpC, and ugpQ (20). UgpB is a specific binding protein, and ugpA and ugpE are membrane-imbedded proteins. UgpC is an ATP-binding protein that has a strong homology with the functionally exchangeable protein (MalK) for maltose transport (21). UgpQ is a glycerol phosphoryl phosphodiesterase that only hydrolyzes the assimilated diesters at the inner surface of the cytoplasmic membrane, and thus is not required for glycerol phosphate transport (20).

A substantial number of ugp transporter genes were identified in the GOS database (Table S2). Normalizing to the number of genomes represented by the recA gene, we found that each marine bacterial genome contained, on average, at least one suite of the ugp transport genes (Dataset S1), indicating that the ugp system may be prevalent in the ocean. In addition, the number of ugp systems per genome was significantly greater than the number of APases (sum of PhoA, PhoX, and PhoD) per genome (t test, $P < 10^{-16}$) across the GOS open ocean and coastal water samples, suggesting that the ugp system is more widespread than APases and that direct transport of small molecular DOP could be substantial in the ocean. Moreover, the mean number of ugp transporter genes (sum of ugpA, ugpB, ugpC, and ugpE) per genome in the open ocean was significantly greater than that in coastal waters (t test, $P < 0.05$), suggesting that transport of glycerol phosphate could be more substantial in the open ocean than in coastal waters. MEGAN analysis revealed that α-Proteobacteria was the dominant phylotype in the ugp gene pool, producing approximately half of the ugp genes, whereas γ-Proteobacteria made limited contributions (Fig. 3). This pattern was in sharp contrast to APase taxonomic distribution (Fig. 1), suggesting that these two most abundant marine bacterial phylotypes (13) mediate DOP utilization by distinct biological mechanisms. Likewise, Bacteroidetes and Planctomycetes were important sources for APases, but they produced few ugp transporter genes.

The function of the ugp system implies that the subcellular localizations of its component proteins are associated with the cytoplasmic membrane. We confirmed this hypothesis by examining ugp proteins from the GOS metagenome. Most metagenomic ugpA (92%) and ugpE (94%) were predicted to be inner membrane proteins. Most ugpC (94%) and ugpQ (97%) were cytoplasmic proteins, and periplasmic proteins comprised the majority (72%) of ugpB (Dataset S3), suggesting that the ugp system is linked to a transport and subsequent cytoplasmic hydrolysis strategy. An intriguing observation was that the number of ugpQ per genome was only about half that of the ugp transporter gene per genome, indicating that many marine bacteria possess ugp transporter genes but lack ugp hydrolyzer genes. Some marine bacteria might substitute cytoplasmic APases for ugpQ. Taxonomic binning of ugpQ and cytoplasmic APases showed that these 2 cytoplasmic phosphoester hydrolyzers dominate in different taxonomic groups (Figs. 1 and 3), which lends strong support for the occurrence of cytoplasmic APases in marine bacteria and strengthens the hypothesis that cytoplasmic APases play a similar role as ugpQ in hydrolyzing transported DOP in marine bacteria. Moreover, the mean sum of cytoplasmic APase and ugpQ per genome is significantly smaller than the mean number of the ugp systems per genome (t test, $P < 10^{-12}$), indicating that the identified number of cytoplasmic APases could be a conservative estimate. Cytoplasmic APases are likely associated with the inner surface of the cytoplasmic membrane and contribute to intracellular hydrolysis of transported phosphoesters, a scenario observed for ugpQ (20). This avoids hydrolysis of the cytoplasmic phosphoesters that are essential for metabolism.

Ecological Implications for the Ugp System and Cytoplasmic APases. Utilization of single or multiple phosphoester substrates by a marine microorganism depends on 2 independent mechanisms:

the occurrence of periplasmic or cell surface-bound phosphoesterases or the presence of phosphoester transporters embedded in the cytoplasmic membrane (1). However, the relative importance of these pathways remains unknown. Few studies have quantified the direct uptake of phosphoesters by marine bacteria. Our finding of the presence of a large number of phosphoester transporter genes in the GOS metagenome indicates that direct transport of dissolved phosphoesters by bacteria could be an important mechanism of P acquisition in the ocean. A large number of cytoplasmic APases and ugpQ suggests that intracellular hydrolysis of the exogenous DOP might be significant. However, considering the complex cellular metabolism and regulation by internal P compounds in the cytoplasm, these cytoplasmic APases could play a role in internal organophosphate hydrolysis. To support this shift in our changing perception of how marine bacteria acquire P requires additional field and laboratory studies regarding the intracellular fate of small phosphoester molecules in marine bacteria.

APase activity has been widely used as an indicator of P limitation in bacterioplankton (16, 22). The analog substrate 4-methylumbelliferyl phosphate (MUF-P) is commonly used to determine potential APase activity (22). Because MUF-P is not transported across the cytoplasmic membrane, such measurements cannot quantify the activity of cytoplasmic APases. The limitation of current APase activity measurements and the possibility of direct phosphoester uptake could potentially explain the previous observations that phosphate concentrations and APase activity are not inversely related in some aquatic ecosystems (23).

Phosphoesters are probably selectively transported by their uptake systems, because membrane transporters have specificity in substrate transport. Laboratory studies have shown that many phosphoesters cannot be taken up without hydrolysis in the periplasm or cell surface (24). The selective uptake of phosphoesters and potential subsequent intracellular dephosphorylation indicate that cytoplasmic APases could differ in substrate specificity from secreted APases. The wide distribution of bacterial cytoplasmic APases and limitations with current APase activity measurements suggest that marine bacteria may have even a greater role in the phosphoester utilization and the P cycle.

Ectoenzymatic and Extracellular APases. Ectoenzymes function in the periplasmic space and on the cell surface, whereas extracellular enzymes are released from cells (22). Marine bacterial APases are primarily considered as ectoenzymes rather than extracellular enzymes (7, 22). However, a substantial fraction (30%) of marine bacterial APases appears to be extracellular, indicating that the function of extracellular APases in substrate processing is important. This corresponds well with the results of some laboratory and field studies showing that dissolved APase accounts for ≈30% of total APase activity (16, 25, 26). However, the fraction of extracellular APase activity is highly variable (16, 26), and extracellular APase activity has been considered to occur primarily as a result of grazing, viral lysis, or filtration artifacts. Hence, it is reasonable to propose that the observed cell-free APase activity in the ocean is a combination of signal-targeted APases and APases released as a result of the above factors.

Although we cannot account for the expression and kinetics of these gene products, the subcellular locations of the enzymes have important ecological ramifications. The hydrolytic activities of extracellular APases and subsequent release of P_i can benefit source and neighboring cells, whereas cell-associated APases mainly provide P_i for the source cells that synthesized them. Differences in subcellular localization can lead to a variety of ecological relations. However, it is important to note that such factors as diffusion and a cell's microenvironment may influence the coupling of DOP hydrolysis and P_i uptake regardless of whether the enzyme is cell-associated or cell-free. The occurrence of extracellular APase might be important for P_i utilization by marine microorganisms lacking the enzymatic capability for using DOP in the surface ocean.

Microorganisms are found in much greater abundance on organic aggregates in comparison to the surrounding seawater (27). Attached bacteria actively transform particles, gels, and colloids (28), and even though it has been proposed that "extracellular enzymes" are critical in the degradation of these matrices (28), the subcellular locations of these enzymes remain unclear. Most organic matrices are porous and can effectively retain secreted enzymes, which, in turn, readily decompose the organic matrices. Therefore, it would be ecologically advantageous for attached bacteria to secrete extracellular enzymes. In the GOS Expedition, attached bacteria were largely excluded from collection in most sampling sites. Further studies should examine the subcellular locations of APases in attached marine bacteria.

The computational prediction of APase subcellular localization indicates a potential mechanistic shift in our understanding of bacterial utilization of phosphoesters; however, it is the expression and regulation of the different APases and the DOP transport system that are important. These studies will enhance our understanding of oceanic P cycling and its interconnection with carbon cycling.

Materials and Methods

Recovery of APase and Ugp Peptide Sequences from GOS Metagenomic Databases. Seawater samples collected from the GOS sampling sites were filtered (>0.1 μm and <0.8 μm) to concentrate marine microorganisms (13). To recover as many homologous sequences as possible, we used 2 seed queries obtained from 2 distantly related bacteria for each APase family (PhoA, PhoD, and PhoX) and ugp genes (ugpA, ugpB, ugpC, ugpE, and ugpQ). The accessions of query sequences are summarized in Table S3. We applied a position-specific iterated BLAST (29) with an expectation value of 0.1 to recover homologous sequences from the GOS database (30). In this step, some similar but nonhomologous sequences were also retrieved.

The retrieved sequences were verified using protein domain databases. We applied a reversed position-specific BLAST (29) to search all the retrieved sequences against a conserved domain database (CDD) (31). For each sequence, only the top hit known as PhoA, PhoD, and PhoX was accepted as an APase. Another putative APase family, PhoV (32), was not used in the analysis because their conserved domain has not been represented in a CDD. The same rule was applied to ugp peptides. The accessions of all APases and ugp peptides in the CDD are listed in Table S4. To identify the duplicate sequences attributable to paired reads, each APase and ugp peptide's J. Craig Venter Institute (JCVI) PEP number was mapped to JCVI Read ID and Mate ID (Datasets S2 and S3). Duplicate sequences were then assembled. All relevant information was parsed by Perl scripts. Ten GOS samples were not included in the analysis (Dataset S1), because organisms other than prokaryotes were collected. Twenty-five open ocean samples and 21 coastal water samples were used for statistical analyses. The open ocean samples from the Tropical South Pacific were not included because of their small sample sizes (Dataset S1). All statistical analyses were performed using the R statistical software package (33). Apparent taxonomic distributions of APases were estimated by MEGAN with the recommended parameter setting (min-score: 100, top-percent: 10%, min-support: 2) (17).

Subcellular Localization Prediction of APases. Gram-negative bacteria dominate (≈90%) the marine prokaryotic community in the GOS samples (13). Archaea are nearly absent in the GOS samples, which are surface samples (13). Proteins in Gram-negative bacteria have five possible subcellular localizations: cytoplasm, inner membrane, periplasm, outer membrane, and extracellular space. Although a variety of methods are available, for the purpose of predicting fragmentary peptide sequences (e.g., GOS peptides) and discovering unrecognized localizations of APases, only algorithms using amino acid compositional bias are useful, such as CELLO, SUBLOC, and LOCTree.

In some cases, different algorithms make different predictions (Dataset S2). To reconcile this discrepancy, we developed a metaalgorithm, MetaP. It works as follows. Given a protein sequence, its localization predictions from all independent algorithms (CELLO, SUBLOC, and LOCTree) are collected and transformed into a standard format using Perl scripts. In a previous metaalgorithm (34), different locations were regarded as independent classes and performance-based weighted voting was used to summarize predictions. MetaP considers the common properties shared by sorting signals targeting neighboring subcellular locations. We used the following weighted voting to incorporate these neighborhood relations among the subcellular locations as well as suboptimal predictions by base algorithms.

The predicted location of MetaP for a sequence s is the one that has the maximum sum of weighted voting for that subcellular location. The prediction can be denoted formally as $P_s = \arg\max_i \sum_{j=1}^{N} P(i, j)$, where N is the total number of base predictors and i is the index of a predicted subcellular compartment: cytoplasmic ($i = 1$), cytoplasmic membrane ($i = 2$), periplasmic ($i = 3$), outer membrane ($i = 4$), and extracellular ($i = 5$). $P(i, j)$ denotes the voting weight of the prediction(s) of the jth element predictor for compartment i. It is defined as $P(i, j) = \sum_{k=0}^{M_j} 2^{-|c_k - i|} \cdot w_k$, where M_j is the number of predictions of the jth predictor. It means that the voting weight of a prediction by the jth predictor for compartment i depends on the offset of the index c_k of its predicted class with regard to the index i as well as its normalized score w_k.

The performance of MetaP to predict fragmentary protein sequences was evaluated using sets of testing sequences whose localizations were verified by experiments. We showed that MetaP is an accurate method in predicting cytoplasmic, periplasmic, and extracellular proteins (Table S5).

Among the three base algorithms (CELLO, SUBLOC, and LOCTree), only CELLO can predict inner membrane and outer membrane proteins, whereas membrane proteins must be removed from the datasets before analysis by SUBLOC and LOCTree (35). We showed that CELLO makes accurate predictions for membrane proteins (Table S6). We initially applied CELLO to identify membrane APase peptides, and the remaining nonmembrane sequences were predicted by all base algorithms and MetaP. For ugp proteins, CELLO was used to predict ugpA and ugpE, which are inner membrane proteins, whereas all three base algorithms and MetaP predicted ugpB, ugpC, and ugpQ, which are nonmembrane proteins. Duplicate sequences attributable to paired reads were assembled before prediction. All predictions were collected (Dataset S2). More details are provided in *SI Methods*.

ACKNOWLEDGMENTS. We are grateful to Dr. Mary Ann Moran and Shalabh Sharma (Moran laboratory, University of Georgia), Dr. Winfried Boos (University of Konstanz), Karl Kaiser (Benner laboratory, University of South Carolina), Dr. Hongmei Zhang (University of South Carolina), Dr. Jijun Tang and Yiwei Zhang (Tang laboratory, University of South Carolina), Dr. Tao Tao (National Center for Biotechnology Information), Dr. Bangqin Huang (Xiamen University), Dr. Tian Xiao (Chinese Academy of Sciences), Michael Chiu (Community Cyberinfrastructure for Advanced Marie Microbial Ecology Research and Analysis), Dr. Lukas Käll (Stockholm University), Dr. Aron Marchler-Bauer (National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health), Dr. Henrik Nielsen (Technical University of Denmark), Dr. Chin-Sheng Yu (Feng Chia University), Dr. Jenn-Kang Hwang (National Chiao Tung University) and 2 anonymous reviewers' constructive comments. We especially thank Dr. Shruti Rastogi and Guy Yachdav (Rost laboratory, Columbia University) for their packaging and distributing of LOCTree software for us. J.H. was supported by the National Science Foundation (Grant BIO-DBI-0845381). R.A.L was supported by the National Science Foundation (Grant MCB-0729594). All computations were performed on a 128-core shared memory computer of the High Performance Computing Group at the University of South Carolina.

1. Karl DM, Bjorkman KM (2002) Dynamics of DOP. *Biogeochemistry of Marine Dissolved Organic Matter*, eds Hansell DA, Carlson CA (Academic) pp 249–366.
2. Karl DM (2000) Phosphorus, the staff of life. *Nature* 406:31–32.
3. Wu J, Sunda W, Boyle EA, Karl DM (2000) Phosphate depletion in the western North Atlantic Ocean. *Science* 289:759–762.
4. Karl DM, et al. (2001) Ecological nitrogen-to-phosphorus stoichiometry at station ALOHA. *Deep-Sea Res* 48:1529–1566.
5. Clark LL, Ingall ED, Benner R (1998) Marine phosphorus is selectively remineralized. *Nature* 393:426.
6. Martinez J, Smith DC, Steward GF, Azam F (1996) Variability in ectohydrolytic enzyme activities of pelagic marine bacteria and its significance for substrate processing in the sea. *Aquatic Microbial Ecology* 10:223–230.
7. Martinez J, Azam F (1993) Periplasmic aminopeptidase and alkaline phosphatase activities in a marine bacterium: Implications for substrate processing in the sea. *Mar Ecol Prog Ser* 92:89–97.
8. Thompson LMM, MacLeod RA (1974) Biochemical localization of alkaline phosphatase in the cell wall of a marine pseudomonad. *J Bacteriol* 117:819–825.
9. Hassan HM, Pratt D (1977) Biochemical and physiological properties of alkaline phosphatases in five isolates of marine bacteria. *J Bacteriol* 129:1607–1612.
10. Doonan BB, Jensen TE (1977) Ultrastructural localization of alkaline phosphatase in the blue-green bacterium Plectonema boryanum. *J Bacteriol* 132:967–973.
11. von-Tigerstrom RG (1984) Production of two phosphatases by Lysobacter enzymogenes and purification and characterization of the extracellular enzyme. *Appl Environ Microbiol* 47:693–698.
12. Poole K, Hancock REW (1983) Secretion of alkaline phosphatase and phospholipase C in Pseudomonas aeruginosa is specific and does not involve an increase in outer membrane permeability. *FEMS Microbiol Lett* 16:25–29.
13. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* 5:e77.
14. Karl DM, et al. (2008) Aerobic production of methane in the sea. *Nat Geosci* 1:473–478.
15. Kobori H, Taga N, Simidu U (1979) Properties and generic composition of phosphatase-producing bacteria in coastal and oceanic seawater. *Bulletin of the Japanese Society of Scientific Fisheries* 45:1429–1433.
16. Hoppe H-G (2003) Phosphatase activity in the sea. *Hydrobiologia* 493:187–200.
17. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
18. Sebastian M, Ammerman JW (2009) The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. *ISME Journal* 3:563–572.
19. Lohan MC, Statham PJ, Crawford DW (2002) Total dissolved zinc in the upper water column of the subarctic North East Pacific. *Deep-Sea Res* 49:5793–5808.
20. Brzoska P, Rimmele M, Brzostek K, Boos W (1994) The Ugp paradox: The phenomenon that glycerol 3-phosphate, exclusively transported by the Escherichia coli Ugp system, can serve as a sole source of phosphate but not as a sole source of carbon is due to trans inhibition of Ugp-mediated transport by phosphate. *Phosphate in Microorganisms Cellular and Molecular Biology*, eds Torriani-Gorini A, Yagil E, Silver S (ASM, Washington, DC) pp 30–36.
21. Hekstra D, Tommassen J (1993) Functional exchangeability of the ABC proteins of the periplasmic binding protein-dependent transport systems Ugp and Mal of Escherichia coli. *J Bacteriol* 175:6546–6552.
22. Gaas BM, Ammerman JW (2007) Automated high resolution ectoenzyme measurements: Instrument development and deployment in three trophic regimes. *Limnol Oceanogr* 5:463–473.
23. Neddermann K, Nausch M (2005) Effects of organic and inorganic nitrogen compounds on the activity of bacterial alkaline phosphatase. *Aquatic Ecology* 38:475–484.
24. Bengis-Garber C, Kushner DJ (1982) Role of membrane-bound 5′-nucleotidase in nucleotide uptake by the moderate Halophile Vibrio costicola. *J Bacteriol* 149:808–815.
25. Hoch MP, Bronk DA (2007) Bacterioplankton nutrient metabolism in the Eastern Tropical North Pacific. *J Exp Mar Bio Ecol* 349:390–404.
26. Labry C, Delmas D, Herbland A (2005) Phytoplankton and bacterial alkaline phosphatase activities in relation to phosphate and DOP availability within the Gironde plume waters (Bay of Biscay). *J Exp Mar Bio Ecol* 318:213–225.
27. Alldredge AL, Cole JJ, Caron DA (1986) Production of heterotrophic bacteria inhabiting macroscopic organic aggregates (marine snow) from surface waters. *Limnol Oceanogr* 31:68–78.
28. Smith DC, Simon M, Alldredge AL, Azam F (1992) Intense hydrolytic enzyme activity on marine aggregates and implications for rapid particle dissolution. *Nature* 359:139–142.
29. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
30. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A Community Resource for Metagenomics. *PLoS Biol* 5:e75.
31. Marchler-Bauer A, et al. (2009) CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* 37:D205–D210.
32. Wagner KU, Masepohl B, Pistorius EK (1995) The cyanobacterium Synecchococcus sp. strain PCC 7942 contains a second alkaline phosphatase encoded by phoV. *Microbiology* 141:3049–3058.
33. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria).
34. Liu J, Kang S, Tang C, Ellis LBM, Li T (2007) Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Res* 35:e96.
35. Gardy JL, Brinkman FSL (2006) Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol* 4:741–751.

Supporting Information

Luo et al. 10.1073/pnas.0907586106

SI Methods

Database Download. Bioinformatics databases were downloaded from the web site (ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/) of the National Center for Biotechnology Information (NCBI) CDD in April 2009 (1). The CDD is a collection of protein sequence models that represent protein domains conserved in molecular evolution (1). The CDD imports various well-curated domain collections, including Pfam-A seed alignments from the protein families database of alignments and Hidden Markov Models (HMMs) (Pfam), clusters of orthologous groups (COGs), the Simple Modular Architecture Research Tool (SMART), and Protein Clusters, and domain databases are also established internally by the CDD research group at the NCBI (1).

GOS metagenomic databases at Community Cyberinfrastructure for Advanced Marine Microbial Ecology (CAMERA) were also downloaded from the web site (http://camera.calit2.net/) (2).

Algorithm Design for Predicting Subcellular Localization of Metagenomic APases. Many computational algorithms have been developed to predict the subcellular localization of proteins in Gram-negative bacteria. These algorithms employ a variety of supervised machine learning techniques and different information sources to make predictions. They can be generally classified into four types (3). One type is based on signal peptide prediction, such as SignalP and Phobius. It cannot be applied to incomplete peptides, such as metagenomic sequences. The second type, such as Proteome Analyst, uses localization information from well-annotated homologous sequences identified by BLAST. It is not suitable to make discoveries of APases with different subcellular localizations. The third type (e.g., CELLO, SUBLOC, LOCTree) usually predicts protein localization using features, such as amino acid/dipeptide compositional bias, physicochemical properties of amino acids, and others derived from whole protein sequences. Yet another type is a hybrid of the above methods, such as PSORT-B. It is also not applicable because it utilizes inappropriate modules from the initial two types. Only the third approach is useful for the purpose of this study.

Signal peptide-based methods, such as SignalP (4, 5), Phobius (6), and TatP (7), have been widely used to predict protein localizations. However, they have important limitations. They can only predict a protein as a secretory or nonsecretory protein. They cannot provide further information regarding the finer locations, such as periplasm and extracellular space. Next, the accuracy of these methods requires that the signal peptides actually exist and are complete at the N-terminal part. Because there is no guarantee that the N-terminal part is correctly assembled in the GOS metagenomic sequences, the reliability of signal peptide-based algorithms is questionable. Finally, some secretory proteins have their signal peptide locating in the middle part or C-terminal part or even do not possess a signal peptide. In this case, the signal peptide-dependent methods cannot make correct predictions. A previous study demonstrated that *Synechococcus elongatus* PCC 7942 possesses an atypical APase that does not have a cleavable signal peptide at its N-terminal even though it is transported across the cytoplasmic membrane and into the periplasmic space (8). We found that both SignalP and Phobius erroneously predict it as a cytoplasmic protein. Proteome Analyst (9) is another category of method, which initially applies a BLAST search against the SwissProt database to obtain homologs with manually curated annotation (3). This approach is restricted to find known localizations for a specific protein, and it cannot make other predictions.

Recently, a number of algorithms using a support vector machine [CELLO (10, 11), SUBLOC (12), LOCTree (13), SLP (14), PSLpred (15), and P-CLASSIFIER (16)] have been developed to predict the subcellular localizations of proteins in Gram-negative bacteria (3). They use information on amino acid/dipeptide and other compositional biases at the whole sequence level (3). Hence, most algorithms in this category are minimally affected by the incompleteness of peptide sequences. Among these algorithms, PSLpred requires a single sequence submission at a time (3), which is not appropriate for high-throughput analysis. P-CLASSIFIER is currently not available. SLP performs undesirably on the incomplete sequences (Table S5) because it still relies heavily on the presence of N-terminal protein sequences (14). Therefore, PSLpred, P-CLASSIFIER, and SLP were not applied, whereas CELLO, SUBLOC, and LOCTree were useful in this study. Because all algorithms have their own bias, the predictions from CELLO, SUBLOC, and LOCTree were frequently inconsistent (Dataset S1).

To address this issue, we developed a metaalgorithm (MetaP), which combines predictions from CELLO, SUBLOC, and LOCTree to get weighted consensus predictions. A previous simple metaprediction algorithm reported improvement of prediction accuracy and is superior to all base algorithms (17). The MetaP algorithm proposed here considers neighborhood relations among subcellular localizations and also suboptimal predictions, and thus has the benefit of resolving conflicting predictions by the base algorithms and achieves higher precision and accuracy of prediction.

Actually, sorting signals targeting different subcellular locations usually share some similarities. For example, sorting signals targeting the periplasm and outer membrane both have N-terminal positively charged regions. In this case, prediction algorithms usually have some ambiguity for distinguishing these neighboring compartments. When an algorithm predicts a protein as a periplasmic protein with the highest confidence, it also implies that the protein has a probability of being located in its neighboring compartments, including the cytoplasm, inner membrane, outer membrane, and extracellular space, with higher probability assigned to the locations closest to the periplasm. Indeed, neighboring compartments are usually reported as suboptimal predictions by the component algorithms (CELLO, SUBLOC, and LOCTree).

The voting weight w_k for kth prediction is defined on the basis of its relative score by comparison with all other predictions made by this algorithm. Because raw scores of predictions from different component base algorithms are not directly comparable, the raw score s_k is converted into a normalized probability $p(k) = p(s \leq s_k)$ by calculating the percentage of predictions with lower raw scores among all predictions for a given algorithm. w_k is then defined as $w_k = p(_k)$. In the case of the reported atypical APase (8), different base algorithms produced contradictory results. For instance, CELLO, SUBLOC, and LOCTree predicted it as an outer membrane protein, periplasmic protein, and cytoplasmic protein, respectively. However, MetaP correctly predicted it as a periplasmic protein.

The prediction accuracy of the proposed ensemble algorithm MetaP will be influenced by the prediction accuracy of the component base predictors. Although the predication accuracies of the three base prediction algorithms CELLO (10, 11), SUB-

LOC (12), and LOCTree (13) were reported on different datasets using different criteria (10, 11–14), they were mainly restricted to complete sequences. These performance data may not be directly applied to the GOS incomplete peptides occurring in the GOS metagenomic database. We downloaded testing protein sequences with known localizations from multiple literature sources (15, 18, 19), manually removed their N-terminal peptides with different lengths, and applied these base algorithms and MetaP to predict their localizations. We pooled testing sequences located at the periplasm and extracellular space as secreted proteins. Membrane proteins were not used because they cannot be predicted by SUBLOC and LOCTree (3). SLP was used as a comparison and was not included in MetaP. Table S5 shows their accuracy for predicting complete and incomplete protein sequences.

Statistical Test for One-Sample Proportion. To test whether PhoX localization patterns (mainly extracellular) are different from PhoD and PhoA (mainly cytoplasmic) localizations (Fig. 2), we designed a statistical approach. The basic method is the one-sample test on proportions. For each APase family, we define the sample space comprising cytoplasmic and extracellular APases and defined a statistic, $z = (p\text{–p0})/(\text{p0}\cdot(1\text{–p0})/\text{n})^{0.5}$, where n is the sample size, which is the number of genes for each APase family; p is the sample proportion, which is the proportion of the number of cytoplasmic APases (or other localizations) among the sample size n; and p0 is a designated true proportion (population proportion, p0 = 0.39 here). For large sample size n (as in our case), the statistic z fits approximate standard normal distribution. This property is used to test whether the sample proportion is significantly different from the designated true proportion. We found that, for PhoX, the sample proportion of extracellular proteins is significantly greater than 0.39, whereas the proportions of others are significantly smaller than 0.39 ($P < 0.05$ in all cases). In contrast, for both PhoD and PhoA, the sample proportion of cytoplasmic proteins is significantly larger than 0.39, whereas the proportions of others are significantly smaller than 0.39 ($P < 0.05$ in all cases). We concluded that PhoX is dominated by extracellular proteins, whereas PhoD and PhoA are dominated by cytoplasmic proteins. The difference of the distribution pattern is significant.

1. Marchler-Bauer A, et al. (2009) CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* 37:D205–D210.
2. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A Community Resource for Metagenomics. *PLoS Biol* 5:e75.
3. Gardy JL, Brinkman FSL (2006) Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol* 4:741–751.
4. Dyrlov Bendtsen J, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340:783–795.
5. Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* 6:122–130.
6. Käll L, Krogh A, Sonnhammer ELL (2007) Advantages of combined transmembrane topology and signal peptide prediction—The Phobius web server. *Nucleic Acids Res* 35:W429–W432.
7. Bendtsen J, Nielsen H, Widdick D, Palmer T, Brunak S (2005) Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 6:167.
8. Ray JM, Bhaya D, Block MA, Grossman AR (1991) Isolation, transcription, and inactivation of the gene for an atypical alkaline phosphatase of Synechococcus sp. strain PCC 7942. *J Bacteriol* 173:4297–4309.
9. Lu Z, et al. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 20:547–556.
10. Yu C-S, Lin C-J, Hwang J-K (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* 13:1402–1406.
11. Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. *PROTEINS*: *Structure*, *Function*, *and Bioinformatics* 64:643–651.
12. Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17:721–728.
13. Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol* 348:85–100.
14. Matsuda S, et al. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci* 14:2804–2813.
15. Bhasin M, Garg A, Raghava GPS (2005) PSLpred: Prediction of subcellular localization of bacterial proteins. *Bioinformatics* 21:2522–2524.
16. Wang J, Sung W-K, Krishnan A, Li K-B (2005) Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinformatics* 6:174.
17. Liu J, Kang S, Tang C, Ellis LBM, Li T (2007) Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Res* 35: e96.
18. Menne KML, Hermjakob H, Apweiler R (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* 16:741–742.
19. Gardy JL, et al. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* 31:3613–3617.

Table S1. APase properties

APase type	Substrate specificity	Metal cofactor
PhoA	Monoesters (1)	Zn, Mg (2)
PhoD	Monoesters, diesters (3–5)	Ca (4)
PhoX	Monoesters, diesters (6)	Ca (6)

1. Sone M, Kishigami S, Yoshihisa T, Ito K (1997) Roles of disulfide bonds in bacterial alkaline phosphatase. *J Biol Chem* 272:6174–6178.
2. Coleman JE (1992) Structure and mechanism of alkaline phosphatase. *Annu Rev Biophys Biomol Struct* 21:441–483.
3. Yamane K, Maruo B (1978) Alkaline phosphatase possessing alkaline phosphodiesterase activity and other phosphodiesterases in Bacillus subtilis. *J Bacteriol* 134:108–114.
4. Yamane K, Maruo B (1978) Purification and characterization of extracellular soluble and membrane-bound insoluble alkaline phosphatases possessing phosphodiesterase activities in Bacillus subtilis. *J Bacteriol* 134:100–107.
5. Eder S, Shi L, Jensen K, Yamane K, Hulett FM (1996) A Bacillus subtilis secreted phosphodiesterase/alkaline phosphatase is the product of a Pho regulon gene, phoD. *Microbiology* 142:2041–2047.
6. Wu J-R, et al. (2007) Cloning of the gene and characterization of the enzymatic properties of the monomeric alkaline phosphatase (PhoX) from Pasteurella multocida strain X-73. *FEMS Microbiol Lett* 267:113–120.

Table S2. Number of ugp genes in GOS database

Gene name	No. genes
ugpA	9,277
ugpB	9,509
ugpC-MalK*	9,195
ugpE	8,801
ugpQ	4,652

*The ugp system is a multicomponent system driven by an ATP-hydrolyzing subunit, which is known as ugpC. UgpC is functionally exchangeable with MalK, the ATP-binding cassette of the maltose transport system (1, 2). UgpC and MalK sequences are very similar, making it difficult to distinguish one from the other using reversed position-specific BLAST searches against CDDs. Hence, we consider ugpC and MalK as a unit in comparison to other ugp component genes.

1. Hekstra D, Tommassen J (1993) Functional exchangeability of the ABC proteins of the periplasmic binding protein-dependent transport systems Ugp and Mal of Escherichia coli. *J Bacteriol* 175:6546–6552.
2. Brzoska P, Rimmele M, Brzostek K, Boos W (1994) The Ugp paradox: The phenomenon that glycerol 3-phosphate, exclusively transported by the Escherichia coli Ugp system, can serve as a sole source of phosphate but not as a sole source of carbon is due to trans inhibition of Ugp-mediated transport by phosphate. *Phosphate in Microorganisms Cellular and Molecular Biology*, eds Torriani-Gorini A, Yagil E, Silver S (ASM, Washington, DC) pp 30–36.

Table S3. Seed query information for APases

Gene name	Seed query NCBI accession no.	Source organism
PhoA	YP_002396458.1	*Escherichia coli*
	ABP37735.1	*Chlorobium phaeovibrioides*
PhoD	NP_388144.1	*Bacillus subtilis*
	YP_001092963.1	*Shewanella loihica*
PhoX	ABL09520.1	*Pasteurella multocida*
	YP_001225533.1	*Synechococcus* sp.
ugpA	P10905.1	*Escherichia coli*
	ABQ74637.1	*Mycobacterium tuberculosis*
ugpB	P0AG80.1	*Escherichia coli*
	ABQ74635.1	*Mycobacterium tuberculosis*
ugpC	P10907.3	*Escherichia coli*
	ABQ74634.1	*Mycobacterium tuberculosis*
ugpE	P10906.1	*Escherichia coli*
	ABQ74636.1	*Mycobacterium tuberculosis*
ugpQ	P10908.1	*Escherichia coli*
	ABQ75671.1	*Mycobacterium tuberculosis*

Table S4. Conserved domain accession numbers for APases

Source database	PhoA	PhoD	PhoX	ugpA	ugpB	ugpC	ugpE	ugpQ
COG	COG1785	COG3540	COG3211	COG1175	COG1653	COG3839	COG0395	COG0584
Pfam	Pfam00245	Pfam09423	NA	NA	NA	NA	NA	Pfam03009
PRK	PRK10518	NA	NA	PRK10561	PRK10974	PRK11650 PRK11000	PRK10973	PRK09454
SMART	smart00098	NA	NA	NA	NA	NA	NA	NA
Curated at NCBI*	cd00016	NA	NA	NA	NA	cd03301	NA	NA

*This domain database uses 3D structure information to define domain boundaries and aligned blocks explicitly and to amend alignment details. It is curated by the National Center for Biotechnology Information (NCBI) CDD group, which is in contrast to other external databases (COG, SMART, Pfam, and PRK) imported to the CDD (1). COG, cluster of orthologous groups of proteins (2); NA, not available; Pfam, Pfam-A seed alignments from the protein families database of alignments and Hidden Markov Models (HMMs) (3); PRK, PRotein K(c)lusters (4); SMART, the Simple Modular Architecture Research Tool (5, 6).

1. Marchler-Bauer A, et al. (2009) CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* 37:D205–D210.
2. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637.
3. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
4. Klimke W, et al. (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res* 37:D216–D223.
5. Letunic I, Doerks T, Bork P (2009) SMART 6: Recent updates and new developments. *Nucleic Acids Res* 37:D229–D232.
6. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc Natl Acad Sci USA* 95:5857–5864.

Table S5. Performance of MetaP in predicting extracellular and periplasmic proteins*

	Cytoplasmic			Periplasmic			Extracellular		
Database	Precision†	Recall‡	Accuracy§	Precision	Recall	Accuracy	Precision	Recall	Accuracy
db2	0.82	0.97	0.91	0.85	0.77	0.87	0.92	0.79	0.92
db3	0.82	0.98	0.92	0.84	0.77	0.87	0.92	0.79	0.92

*The first 200 amino acids at the N-terminal were removed in the testing sequences, and the sequences with no less than 30 amino acids were used in the analysis.

†Precison = TP/(TP + FP).

‡Recall = TP/(TP + FN).

§Accuracy = (TP + TN)/(TP + TN + FP + FN), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. For instance, in the case of extracellular proteins, TP is the number of proteins predicted to be extracellular that are indeed extracellular, TN is the number of proteins predicted to be nonextracellular that are indeed nonextracellular, FP is the number of proteins predicted to be extracellular that are indeed nonextracellular, and FN is the number of proteins predicted to be nonextracellular that are indeed extracellular.

Here, only extracellular and periplasmic proteins from database 2 (db2) (1) and database 3 (db3) (2) were used so as to balance the number of true-positive and true-negative cases. Database 1 (db1) (3) was not used, because many secretory sequences in db1 are not labeled with finer subcellular localizations (e.g., extracellular or periplasmic).

1. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, et al. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* 31:3613–3617.
2. Bhasin M, Garg A, Raghave GPS (2005) PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* 21:2522–2524.
3. Menne KML, Hermjakob H, Apweiler R (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* 16:741–742.

Table S6. Performance of MetaP in predicting inner and outer membrane proteins*

	Inner membrane			Outer membrane		
Database	Precision	Recall	Accuracy	Precision	Recall	Accuracy
db2	0.99	0.74	0.94	0.97	0.86	0.95
db3	1	0.76	0.94	0.97	0.90	0.96

*The first 200 amino acids at the N-terminal were removed in the testing sequences, and the sequences with no less than 30 amino acids were used in the analysis. The definitions of precision, recall, and accuracy are the same as described in Table S5.

Other Supporting Information Files

Dataset S1

Dataset S2

Dataset S3