Isolation and characterization of bacterial phosphorous metabolism genes from complex microbial communities

by

Adi Rolider

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2009

©Adi Rolider 2009

Author's declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Phosphorous (P) is an essential nutrient, playing a central role in the life of a bacterial cell. It is involved in cellular metabolic pathways, cell signaling and is a component of many of the cell's macromolecules. Since a majority of the biosphere's microorganisms have not yet been cultured, much more can be learned about the biochemical and genetic mechanisms that govern bacterial P metabolism. The function-driven approach to metagenomics was applied to study P metabolism in the bacterial communities present in pulp and municipal wastewater treatment plant activated sludge and soil, leading to the isolation and identification of three new phosphatases, genes involved in P transport, regulation of P related functions and additional genes which may be important for the bacterial cell's adaptation to the above communities.

The identification of two new nonspecific acid phosphatases (NSAPs) $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$ and an alkaline phosphatase, $phoA_{ACX6.71}$, belonging to the nucleotide pyrophosphatase phosphodiesterase (NPP) family is reported here. The genes for the three phosphatases were cloned, sequenced, and analysed for upstream regulatory sequences in addition to biochemical characterization of their protein products. PhoB-binding sites were found upstream to $phoA_{ACX6.71}$ and NSAP $phoN_{ACX6.13}$, suggesting these genes are governed by the mechanisms of the previously described "pho" regulon. The two NSAPs have pH optima in the acidic neutral range while the alkaline phosphatase has an optimal pH at 9.5. The three phosphatases appear to be distantly related to known bacterial phosphatase enzymes. Phylogenetic analysis shows the newly identified NSAPs appear on a separate clade from known bacterial NSAPs. Key amino acid residues involved in the catalytic site of

these NSAPs were identified in PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$.In PhoA$_{ACX6.71}$, key amino acid residues involved in catalysis and metal cofactor coordination were identified. The roles of these residues were confirmed based on the predicted molecular structure of these proteins. The structures indicate the three proteins are globular with folding patterns suitable for catalytic residues to bind and cleave the P substrate. This is the first report of functional characterization of phosphatases from uncultured bacteria.

In addition to exploring the hydrolysis of phosphate esters, the transport and metabolism of other P compounds was also investigated. By phenotypic complementation of phosphonate growth deficient mutants of the legume symbiont, *Sinorhizobium meliloti* and large scale sequencing of selected metagenomic clones, 92 ORFs were isolated. As expected, about 25% of these ORFs are P transport proteins and P related regulators. Genes involved in other regulatory functions made up about 12% of the total while genes related to Nitrogen metabolism and assimilation account for about 8% of the newly identified ORFs. About 30% of the ORFs encoded general cellular functions or hypothetical proteins of unknown function. The results of this investigation demonstrate the effectiveness of functional metagenomics in studying genetic diversity of bacteria inhabiting complex microbial communities and in identifying new proteins of interest.

Acknowledgements

Table of contents

List of figures

List of tables

List of abbreviations

2-AEP: 2-aminoethylphosphonate

4-HB: 4-hydroxybutyrate

ABC: ATP binding cassette

AMP: Adenosine monophosphate

ATP: Adenosine triphosphate

BAC: Bacterial artificial chromsome

BCIP: 5-bromo-4-chloro-3-indolyl phosphate

CAMERA: Community cyberinfrastructure for advanced marine microbial ecology research and

                     analysis

cAMP: cyclic adenosine monophosphate

cGMP: cyclic guanosine monophosphate

CAPSO: 3-[cyclohexylamino]-2-hydrosyl-1-propane-sulfonic acid

CP-lyase: Carbon-phosphorous lyase

DAPG: 2,4-diacetylphloroglucinol

DMSO: Dimethylsulfoxide

DTT: Dithiothreitol

EBNSAP: *Escherichia blattae* nonspecific acid phosphatase

EDTA: Ethylenediaminetetraacetic acid

EPSP: 3-enol-pyruvylshikimate-5-phosphate

G3P: Glucose-3-phosphate

G6P: Glucose-6-phosphate

GMP: Guanosine monophosphate

GTP: Guanosine triphosphate

LB:	Luria broth
MOPS:	Morpholino propane sulfonic acid
NAD:	Nicotinamide adenine dinucleotide
NADH:	Nicotinamide adenine hydrogen dinucleotide
NPP:	Nucleotide pyrophosphatase/phosphodiesterase
NSAP:	Nonspecific acid phosphatase
ORF:	Open reading frame
P:	Phosphorous
PAS:	Period circadian protein, Ah receptor translocator protein, single-minded protein
PAC:	C-terminal domain of PAC proteins
PDB:	Protein data bank
PHB:	Polyhydroxybutyrate
Pi:	Inorganic phosphate
pNPP:	p-nitrophenolphosphate
poly(Pi):	Polyphosphate
PPi:	Pyrophosphate
PPK:	Polyphosphate kinase
RBS:	Ribosome binding site
SBR:	Sequencing batch reactor
SDS-PAGE:	Sodium dodecylsulfate polyacrylamide gel electrophoresis
TAE:	Tris acetate EDTA
TEN:	Tris EDTA NaCl
TY:	Tryptone yeast extract
X-gal:	5-bromo-4-chloro-3-indolyl-β-galactopyranoside

Claims of contributions to scientific knowledge

1. Identified three unique metagenomic cosmid clones that confer BCIP utilization upon *E. coli* DH5α.

2. Subcloned and sequenced the relevant BCIP utilizing phosphatase genes and flanking regions. Identified two new class A nonspecific acid phosphatases *phoN*$_{ACX6.13}$ and *phoN*$_{BCX4.10}$ and one alkaline phosphatase, *phoA*$_{ACX6.71}$ belonging to the nucleotide pyrophosphatase / phosphodiesterase (NPP) superfamily.

3. Identified ribosome binding sites, signature motifs, regulatory promoter regions and subcellular location of the protein products of *phoN*$_{ACX6.13}$, *phoN*$_{BCX4.10}$ and *phoA*$_{ACX6.71}$. PhoB binding sites were predicted upstream to *phoN*$_{ACX6.13}$ and *phoA*$_{ACX6.71}$

4. Showed that PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ are phylogenetically distant from putative and experimentally determined phosphatases as they formed a new branch on a neighbor-joining phylogenetic tree. PhoA$_{ACX6.71}$ appeared in a clade with putative alkaline phosphatases of members of *γ-Proteobacterial* class. PhoN$_{ACX6.13}$ PhoN$_{BCX4.10}$ and PhoA$_{ACX6.71}$ have low amino acid identity to phosphatases from known bacterial isolates.

5. Showed pH profile of PhoN$_{ACX6.13}$, PhoN$_{BCX4.10}$ and PhoA$_{ACX6.71}$. A broad acidic-neutral optimal pH range was shown for NSAP PhoN$_{BCX4.10}$ while NSAP PhoN$_{ACX6.13}$ had a sharp optimal pH of 4.4. Alkaline phosphatase PhoA$_{ACX6.71}$ showed optimal activity at pH 9.

6. Predicted a globular molecular structure of phosphatase proteins and identified key amino acid residues involved in metal binding and catalysis as well as general protein folding pattern.

7. Provided the first report of expression and partial purification of phosphatases from uncultured bacteria. Detected expression and showed activity of PhoN$_{ACX6.13}$ and PhoA$_{ACX6.71}$ expressed in *E. coli*

8. Identified eight unique cosmids that confer upon existing *Sinorhizobium meliloti* mutants, the ability to grow using the herbicide glyphosate as the sole P source. Four of these cosmids complemented two mutants that were deficient in loci required for both the transport and degradation of glyphosate (*gly*).

9. A total of 92,896 bp of sequence, consisting of 92 protein coding sequences (CDS) 83 of which are complete open reading frames (ORFs) were assembled from 5 cosmids including the full sequence of cosmids pCX4-10F, pCX6-13F and pCX9-45F and the sequence of the *gly* loci in cosmids pCX4-3G and pCX6-8G.

10. Approximately 70% of the ORFs identified are predicted to be involved in specific cellular functions while only a general function was predicted for 16% of the ORFs. The function associated with the remaining 14% of the ORFs is unknown.

11. Approximately 25% are putative genes involved in P transport, metabolism and regulation: Identified genes include a *pit* inorganic phosphate transport operon, several genes that are part of the *pst* phosphate specific transport system, two *phoU* transport regulation proteins and a *phoB-phoR* two component regulatory system for the global control of the phosphate (*pho*) regulon.

12. Cloned and sequenced *orfA*$_{pAR010}$-*pitA*$_{pAR010}$ from cosmid pCX4-3G and showed these genes can complement the Gly⁻ phenotype of *phoCDET* and *pit*-deficient mutants of *S. meliloti*. RBS was found upstream to *orfA*$_{pAR010.}$

13. Cloned and sequenced *pstS*$_{pAR011}$*pstC*$_{pAR011}$*pstA*$_{pAR011}$*pstB*$_{pAR011}$*phoU*$_{pAR011}$ from cosmid pCX6-8G and showed these genes can complement Gly⁻ phenotype of *phoCDET* deficient mutants. RBS and a PhoB binding site was found upstream to *pstS*$_{pAR011.}$

14. Showed that $PstS_{pAR011}$ and $PstS_{pCX4-10F}$ have evolved different residues than the ones shown, by means of structural evidence, to be important for catalysis in PstS1 of *Mycobacterium tuberculosis*.

15. 12% of the identified genes, originating from the metagenomes of pulp and municipal waste water treatment plant activated sludge and soil, are predicted to be involved in signal transduction and regulatory mechanisms. Genes include a two-component regulatory system for heavy metal efflux, adenylate cyclase diguanylate cyclase and a nitrogenase inactivator.

16. Identified a nitrogenase operon and two putative *glnD* regulators of nitrogen assimilation from the metagenomes of municipal waste water treatment plant activated sludge and soil.

17. 5% of the genes identified, originating from the metagenomes of pulp and municipal waste water treatment plant activated sludge, are predicted to be involved in cellular defense mechanisms. These include two genes containing glutathione-S-transferase domains, $gst1_{pCX4-10F}$ and $gst2_{pCX4-10F}$, a glutathione-dependent formaldehyde dehydrogenase, $adhC_{pCX4-10F}$ and an S-formylglutatione hydrolase, $fghA_{pCX4-10F}$.

18. About 6% of the identified ORFs, originating from the metagenomes of pulp waste water treatment plant activated sludge and soil, are predicted to encode housekeeping

functions such as an inactivator of the chromosomal replication initiator (hdapCX4-10F), an integrase (intpCX9-45F), a poly(A) polymerase (pcnBpCX4-10F), a ribosomal protein S2 (rpsBpCX9-45F), a translational elongation factor (tsfpCX9-45F) and a methionine aminopeptidase, mappCX9-45F.

Chapter 1: Introduction

1.1 Soil and sludge microbial communities

Microbial life arose approximately 4 billion years ago, at least 2 billion years before the emergence of the eukaryotes (Xu 2006). Consequently, microbes have a diverse habitat range and are found in virtually every ecological niche on earth, covering a wide range of temperature, elevation, moisture and many more different conditions. Due to their ability to catalyze processes sustaining all life on earth, bacteria and archaea essentially drive the elements of the biogeochemical cycles on Earth (Torsvik *et al.*, 2002). Microbial diversity is well illustrated in two important ecosystems, soil and activated sludge. Activated sludge is essentially a mixed community of microorganisms that perform three biological processes: 1) utilization of energy from organic matter for cell growth (synthesis), 2) conversion of organic compounds into lower energy compounds such as carbon dioxide and water (respiration) 3) conversion of ammonia into nitrate (nitrification). By way of the above processes, microorganisms are used to biologically treat wastewater effluents from various sources such as pulp and paper mills or municipal sewage. The term "activated" refers to the use of "live" or "active" bacteria to biologically treat waste water effluents. Waste water consists of mineral, animal and vegetable matter in suspension and a large number of bacteria. Activated sludge used for its treatment contains predominantly heterotrophic bacteria, some autotrophic bacteria, protozoans, fungi and rotifers.

The numbers, types and activities of soil microorganisms are influenced by factors such as organic matter content, texture, pH, moisture and aeration. The most abundant microorganisms in the soil are bacteria (10^8-10^9 per gram soil) followed by fungi and algae

1

(10^4-10^5 per gram soil) (Chen *et al.*, 2003). Soil microorganisms are involved in cycling of basic elements such as carbon (Butler *et al.*, 2003) (Singh *et al.*, 2004), phosphorous (Gupta *et al.*, 2002a) and nitrogen (Paterson, 2003), and make an impact on plant nutrition, plant health, soil structure and soil fertility (Kirk *et al.*, 2004). Soil ecosystems are largely characterized by the types of interactions microorganisms have with plants. One large group of soil bacteria is the root colonizing *Rhizobia*. These organisms form root structures, called nodules, in which they differentiate into bacteroids. By this mechanism, *Rhizobia* have a symbiotic relationship with leguminous plants, providing fixed atmospheric nitrogen in exchange for plant photosynthate. Contrary to the example of this symbiotic relationship, plant-microbe interactions may also be antagonistic as plants are susceptible to bacterial or fungal soilborne plant pathogens which cause root and crown rots, wilts and damping off in a large number of crops (Weller *et al.*, 2002). Examples of such pathogens are the wilt-causing *Fusarium* fungi, mainly *Fusarium oxysporum*, the potato scab producing *Streptomyces scabies* and other *Streptomyces* species and the apple replant-causing fungi, *Cylindrocarpon destructans*, *Phytopthora cactorum, Pythium spp.* and *Rhizoctonia solani* (Weller *et al.*, 2002).

One of the significant roles of soil bacteria is nitrogen fixation, the conversion of atmospheric nitrogen gas into soluble nitrogenous compounds readily available to plants for growth (Chen *et al.*, 2003). Nitrogen compounds are further metabolized by nitrifying and denitrifying bacteria. Nitrifying bacteria such as *Nitrosomonas europaea, Nitrospira briensis* and *Nitrosococcus nitrosus* oxidize ammonia to nitrite while nitrifying bacteria such as *Nitrobacter winogradsky, Nitrospina gracilis* and *Nitrococcus mobilis* oxidize nitrite into

2

nitrate (Delgado & Follett, 2002). One of the ways soil microorganisms affect the carbon cycle is by acting as a biological sink to atmospheric CH_4 (Conrad, 1996). There are two known types of Methanotrophic bacteria (MB), types I and II, which are distinguished by their phylogenetic affiliation (γ-*Proteobacteria* vs. α-*Proteobacteria*) and by structural and biochemical features (Ricke *et al.*, 2005).

In wastewater treatment plant activated sludge, bacteria are active participants in the biological removal of nitrogen (N) and phosphorous (P). Nitrogen removal is facilitated by processes such as nitrification/denitrification (Lim *et al.*, 2005b; Neufeld *et al.*, 2001), and nitrogen fixation (Neufeld *et al.*, 2001) while P removal is mostly due to polyphosphate accumulation (Bond *et al.*, 1995; Garcia Martin *et al.*, 2006; Jeon *et al.*, 2003; Shoji *et al.*, 2006). One of the most dominant phylotypes in activated sludge is the *Proteobacteria* consisting of α- and β-*Proteobacteria* and green nonsulfur (GNS) bacteria (Smith *et al.*, 2003). *Proteobacteria* are believed to have a role in nitrogen conversion since many genera in this division are capable of this process (Smith *et al.*, 2003). Members of the genus *Paracraurococcus* were also detected and believed to have a role in the reduction of nitrate to nitrite as characterized members of this genus are aerobic or facultative photoheterotrophs. A similar denitrification process is carried out in activated sludge by members of the *Acidovorax* genus. (Smith *et al.*, 2003).

Microbial populations in activated sludge vary in terms of size and relative abundance according to the nutritional environment. For example, in activated sludge in a sequencing batch reactor (SBR) with acetate supplying the only carbon source, a dominant population of

3

coccus-shaped microorganisms capable of polyphosphate accumulation was observed (Jeon *et al.*, 2003). These organisms were affiliated with the *Rhodocyclus* group of the β-*Proteobacteria* subclass and were found to be responsible for biological phosphorous removal in acetate-supplied SBR-activated sludge (Jeon *et al.*, 2003). On the other hand, the abundance of polyphosphate accumulating organisms was found to be significantly affected by the electron acceptor conditions illustrated by a decrease in the abundance of such organisms as the oxygen supply decreased (Shoji *et al.*, 2006). Members of the α-*Proteobacteria* class and the *Planctomycetes* group were also found to be abundant in SBR-activated sludge however their role in P removal is unclear due to the lack of significant differences in their numbers between phosphate and non-phosphate removing communities (Bond *et al.*, 1995). Finally, members affiliated with the Flexibacter-Cytophaga-Bacteroides group were also found to be part of the SBR-activated sludge microbial community but were more abundant in the non-phosphate removal SBR suggesting these organisms may inhibit phosphate removal or are out-competed by phosphate removing organisms (Bond *et al.*, 1995). So far, some progress has been made in the study of bacterial phosphorous metabolism but more can be learned about their functional and genetic diversity with culture independent approaches.

Effective nutrient cycling carried out by microorganisms determines the viability of soil and activated sludge habitats. It is thus of great interest to identify and understand how microorganisms carry out their respective functions in such communities. Therefore, effective methodology needs to be developed in order to account for the microbial populations occupying such complex communities and to elucidate their function.

1.2 Metagenomics

1.2.1 Overview

In order to successfully study bacterial populations in complex communities such as soil and activated sludge, there are two key approaches to be considered: (i) identification of the microorganisms inhabiting the community; (ii) determination of the various functions carried out by microorganisms in these communities. Traditionally, microbiologists first isolated pure cultures (or co-cultures) of microorganisms from the environment followed by an analysis of their physiological and biochemical traits. Although cultivation allows researchers to accurately and extensively describe particular organisms, the major drawback of this approach is the inability to describe the diversity of a microbial population due to the inability to obtain the majority of the microorganisms in pure culture. To quantify active cells in environmental samples, viable cell counts and most-probable-number techniques are frequently used (Amann *et al.*, 1995). However, these methods select for only a small percentage of dominant organisms and therefore greatly underestimate the number of microorganisms that make up these communities. In the case of soil and activated sludge, the culturability (defined as the percentage of culturable bacteria in comparison with total direct cell counts) is 0.3% and 1-15%, respectively (Amann *et al.*, 1995).

To characterize the diversity of microbial populations, a variety of tools are available for microbiologists. For example, a profile of the metabolic properties of a microbial community can be analyzed by assessing the presence of certain enzymes, phenotypic tests, sensitivity to phages and the API/BIOLOG biochemical identification systems (Morris *et al.*, 2002). Other approaches to studying microbial communities include the assessment of

pathogenicity (towards plants or animals), pathogenicity-like processes and toxin production. These include potential to infect and the presence of markers of virulence (Morris *et al.*, 2002). Patterns of resistance to antimicrobial compounds (antibiotics, bacteriocins, pesticides, heavy metals and polyaromatic hydrocarbons) can also provide information about microbial communities. Additional phenotypic properties can also be described by using markers such as protein profiles, lipopolysaccharides, antigens, fatty acids and isozymes (Morris *et al.*, 2002). With the advancement of molecular biology techniques, DNA-based approaches to study microbial communities have been developed targeting particular DNA sequences or analyzing total DNA. The latter approach may involve Restriction Length Fragment Polymorphism (RFLP) without probes, repetitive sequence-based PCR (rep-PCR), Amplification Fragment Length Polymorphism (AFLP), DNA amplification finger printing (DAF), Pulsed-Field Gel Electrophoresis (PFGE), DNA-DNA hybridization, Arbitrary Signature from Amplification Profiles (ASAP) and Randomly Amplified Polymorphic DNA (RAPD) (Morris *et al.*, 2002). Effective DNA-based analyses may be performed by targeting specific DNA sequences. These approaches include the assessment of plasmid profiles, Fluorescent in situ hybridization (FISH), RFLP with specific probes, analysis of sequences of ribosomal genes, PCR polymorphism, Internal transcribed spacer (ITS), Denaturing gradient gel electrophoresis (DGGE) and Amplifed R-DNA restriction analysis (ARDRA) (Morris *et al.*, 2002).

Culture independent approaches to microbiology began to gain credence when Norman Pace and his colleagues demonstrated that they could obtain information on microbial diversity by directly analyzing 5S and 16S rRNA gene sequences from the

environment without the need to culture the microorganisms present (Lane *et al.*, 1985; Stahl *et al.*, 1985). This approach was carried out with a number of activated sludge samples where the bacterial community structure was found to consist predominantly of members of the α, ß and γ *Proteobacteria* subclasses (Lu *et al.*, 2006; Snaidr *et al.*, 1997). Other detected phylotypes included *Flavobacterium*, *Comamonadaceae* and *Polyangium*-related bacteria (Shoji *et al.*, 2006) as well as low G+C content Gram positive bacteria (Snaidr *et al.*, 1997). On the other hand, culture-independent analysis of 16S rRNA derived from soils typically results in the identification of diverse taxa such as *Cytophagales* and *Proteobacterium* (Rondon *et al.*, 2000) as well as microorganisms belonging to divisions such as *Verrucomicrobia*, *Firmicutes*, *OP11* and the *Acidobacteria* (Liles *et al.*, 2003). The latter phylum seems to be abundantly represented in the soil but to date only a few of its members have been cultured (Handelsman, 2004). With the emergence of rRNA analysis of uncultured organisms, progress has been made in the study of the diversity of microbes in the environment. However, to learn about the ecological function and the physiology of these organisms, a more integrated approach is necessary.

In the last two decades, an approach called "metagenomics" has been developed to access the genomes of microorganisms that have not been or cannot be cultured. The term, metagenomics, was coined by Handelsman and has been defined as the analysis of the collective genomes in an environmental sample (Handelsman *et al.*, 1998). This approach has been implemented as a means to search the soil metagenome for genes encoding natural products following the realization that by accessing "unknown" soil microbes, one can discover a variety of biotechnologically important natural products (Handelsman *et al.*,

1998). The strategy of "metagenomic mining" begins with DNA isolation from an environmental sample, cloning that DNA into a suitable vector, transforming the clone into a host bacterium to make a metagenomic library and screening the transformants for specific enzymes or metabolites (Handelsman, 2004). The information contained in a metagenomic library can be used to determine community diversity and activity, the presence of specific microorganisms or biosynthetic pathways in a particular environment or to uncover the presence of individual genes (Steele & Streit, 2005).

Metagenomic analysis is not restricted to bacterial populations and has been extended to viruses. Traditionally, the study of new viruses was largely dependent on the success of their amplification in cell culture, overcoming their limited antigenic/serological cross-reactivity and their ability to hybridize to known viral sequences (Delwart, 2007). Currently, viral nucleic acids from biological fluids or environmental samples may be PCR amplified or isolated by plasmid library sequencing. These culture-independent methods for the manipulation of viral DNA constitute the field of viral metagenomics (Delwart, 2007). Characterization of metagenomic clones can be done by either randomly sequencing the clones, screening for phylogenetic markers such as 16S rRNA and *recA* (sequence-driven approach), or by expressing the clones in a bacterial host and testing for specific functions such as enzyme activity or antibiotic production (function-driven approach) (Handelsman, 2004; Payne *et al.*, 2006).

1.2.2 Sequence-driven metagenomics

The application of metagenomics to the study of microbial ecology has facilitated the establishment of a link between the identities of different microorganisms occupying various habitats with the physiological functions encoded by these microorganisms. The goal of the sequence driven approach to metagenomics is to identify the organisms inhabiting the community, infer phylogenetic relationships between them and, ultimately, to determine metabolic functions carried out by these organisms. One large scale metagenomic project examined the human intestinal microbiota (Gill *et al.*, 2006). Around 78 Mbp of intestinal metagenomic DNA and over 2000 16S rDNA sequences were analyzed to determine that the human microbiome contains a variety of genes involved in the metabolism of sugars, amino acids, xenobiotics, methanogenesis and the biosynthetic pathways of vitamins and isoprenoids (Gill *et al.*, 2006).

There are some examples where metagenomic library clones carry genes for the phylogenetic marker, 16S-ribosomal RNA, flanked by genes of physiological and metabolic importance. For example, genes for archaeal RNA helicase and a glutamate semialdehyde aminotransferase were found adjacent to archaeal 16S-rDNA, uncovering information on the genetic organization of an uncultivated archaeon associated with a marine planktonic assemblage (Stein *et al.*, 1996). Another marine metagenomic study led to the discovery of a rhodopsin encoding open reading frame (ORF) of an uncultivated *γ-proteobacterium*, showing this group of prokaryotes can carry out a form of phototrophy previously thought to be unique to archaea (Beja *et al.*, 2000).

Aside from marine environments, the link between phylogeny and gene function has also been explored in soil samples. In one such study, a Bacterial Artifical Chromosome (BAC) clone harbouring the partial genomic sequence of 16S-rDNA affiliated with a subdivision of an uncultured Acidobacterium taxon was found to contain ORFs involved in cell cycling, cell division, folic acid biosynthesis, DNA excision and repair, and several other functions (Liles *et al.*, 2003). In another study, 16S rRNA based phylogenetic screening technology has been combined with fluorescent in situ hybridization (FISH) such that the highly expressed product of 16S rRNA genes is easily detected (Leveau *et al.*, 2004). Although the technique was validated using a large insert library of the soil bacterium *Collimonas fungivorans*, the investigators did not report screening of metagenomic libraries using fluorescence in situ hybridization.

A phylogenetic analysis was also carried out for the microbial community of waste water enriched in phosphorous removal biomass (Yeates & Blackall, 2006). Phosphorous removal from waste water is important as excess Pi incorporated in aquatic systems results in negative effects on water quality by means of eutrophication (Garcia-Martin *et al.*, 2006). In this study, a screen of clones for 16S rRNA genes showed the presence of a variety of uncultured waste water bacteria (Yeates & Blackall, 2006). In some cases where 16S rRNA analysis cannot provide sufficient information, phylogenetic analysis could be carried out by targeting particular sets of genes. For example, a metagenomic library of uncultured methanotrophic bacteria from upland soil was screened by the amplification of genes encoding a methane monooxygenase (Ricke *et al.*, 2005). This work provided information on the evolutionary origin of the uncultured methanotrophs. One of the largest studies of

enhanced biological phosphorous removal (EBPR) from waste water involved the sequence and metabolic reconstruction of an unculturable bacterium, *Candidatus Accumulibacter phosphatis*, a member of the *Rhodocylales* order and a dominant species in the acetate-fed EBPR system (Garcia-Martin *et al.*, 2006). This organism takes up Pi during the aerobic stage of the reaction and stores it in the form of polyphosphate thus, largely depleting Pi from waste water. In the anaerobic phase of the reaction, polyphosphate is used to synthesize ATP to provide energy for carbon storage by means of PHA production (Garcia-Martin *et al.*, 2006).

Aside from sequence based metagenomics geared solely towards phylogenetic reconstruction, there are many examples of projects involving shotgun sequencing of metagenomes that also aim to link phylogenetic data to the metabolic capacity that allows microorganisms to adapt to various environments. One such project is the extensive marine community metagenome project in the Saragaso Sea where novel bacterial phylotypes and gene functions have been discovered (Venter *et al.*, 2004). From the 1.045 billion base-pairs of nonredundant sequence data generated in this project, at least 1800 genomic species were derived including 148 previously unknown bacterial phylotypes (Venter *et al.*, 2004). This project was later extended to become the Sorcerer II Global Ocean Sampling (GOS) expedition in which 6.3 billion bp of marine planktonic microbiota DNA sequence was generated from surface water samples collected from a large area spanning the North Atlantic, Panama Canal, and the South Pacific (Rusch *et al.*, 2007). The dataset showed diversity among 85% of the assembled and 57% of the nonassembled sequence. Using novel comparative genomic and assembly methods, genome structure, evolution as well as

biochemical diversity of genes and gene families were analyzed and metagenomes of abundant but nonclonal organisms were reconstructed (Rusch *et al.*, 2007). Two key findings from this expedition were that within individual ribotypes (elements sharing near identical 16S-rDNA sequence) or across the entire community, microorganisms differentially adapt to conditions such as variable phosphate availability (see section 1.3.3) (Rusch *et al.*, 2007). From this data, 6.12 million proteins were predicted and allocated into approximately 4,000 medium and large-size clusters (Yooseph *et al.*, 2007). This led to the discovery of new protein families, increased the diversity of known protein families and provided information on the evolution of these protein families with specific examples of phosphatases (see section 1.3.4), UV damage-repair enzymes, proteases, glutamine synthetase, RuBisCO, RecA and kinases (Yooseph *et al.*, 2007). Futhermore, the metagenomic approach has proven useful in the reconstruction of full genomes of microorganisms inhabiting extreme environments such as acidic biofilms where partial genomes of bacterial populations were reconstructed followed by a genetic analysis of novel physiological pathways (Tyson *et al.*, 2004). This endeavour provided information about survival strategies employed by microorganisms in an extreme environment.

1.2.3 Function-driven metagenomics

Although whole genome shotgun sequencing efforts and the sequencing clones containing 16SrDNA anchors resulted in the discovery of novel genes, the assignment of ecological roles to these genes is limited by the information contained in the currently available sequence databases. The ability of microorganisms to adapt to a wide range of different environmental conditions is largely attributed to production of catalytic enzymes

and their optimization to be able to carry out biochemical reactions needed for life under the physiochemical conditions of their habitats. Functional probing of the metagenome (function-driven metagenomics) allows for the isolation of the genes/operons that encode a function of interest. By simple plating of the library on agar medium supplemented with an indicator substrate, one can isolate enzymes of biotechnological importance. The underlying principle of the phenotypic screen/selection is that the library confers upon the host a function it cannot perform by complementing a deficiency phenotype due to a mutation in or the absence of the relevant gene.

The ability to isolate compounds by functional screens is determined by the ability of the surrogate host strain to express the coding sequence which is dependent on the presence of a promoter and a ribosome binding site (RBS) upstream of the translation start site. A set of *trans* elements supplied by the surrogate host, such as transcription factors, inducers, chaperonins, cofactors, protein modifying enzymes or a secretion machinery is also required for proper heterologous protein expression. The host strain of choice is often *E. coli* since genetic tools are readily available for cloning environmental DNA and screening metagenomic libraries for novel activities. In typical expression-cloning of the metagenome in *E. coli* three scenarios can occur: independent expression where both the promoter and the RBS are part of the heterologous sequence, a transcriptional fusion where only the RBS is encoded in the insert while the promoter (and possibly a transcriptional terminator) is supplied by the vector, and a translational fusion where both the promoter and the RBS are encoded on the vector. Taking these requirements and the type of vector used for heterologous protein expression into account, it was predicted that only about 40% of the

13

activities encoded by the metagenome can be detected by random cloning in *E. coli* (Gabor *et al.*, 2004a). By expanding the available expression host range, it is possible to increase the proportion of detected activities and to screen for a wider variety of novel functions. Thus, two additional host strains were optimized for expression of metagenomic DNA, *Streptomyces lividans* and *Pseudomonas putida* (Martinez *et al.*, 2004). *Actinomycetes* naturally carry precursors and enzymes, allowing them to express heterologous polyketides and nonribosomal peptides, as they themselves are a rich source of such natural compounds. The pseudomonads have large genomes, granting them a rich metabolic diversity including genes for the degradation of organic pollutants and the production of secondary metabolites such as polyketides and nonribosomal proteins. In addition, many genetic tools have been developed for the pseudomonads, allowing the ability of transformation, conjugation, transposon mutagenesis, and the use of different reporter systems and vectors (Martinez *et al.*, 2004). Heterologous expression of BAC clones carrying genes for antibiotics MG1.1, granaticin and 2,4-diacetylphloroglucinol (DAPG) was successfully demonstrated using the three host strains mentioned above (Martinez *et al.*, 2004).

The screening of compost-derived metagenomic libraries resulted in the identification of proteases using skim milk powder as the protein substrate and screening for clones conferring a clear halo on the milk powder (Gupta *et al.*, 2002b). Similarly, other novel enzymes were identified such as amylases by using soluble starch as the substrate, screening clones that confer a colourless halo on a purple background (Lammle *et al.*, 2007). Esterases and lipases are useful enzymes that hydrolyse short and long chain fatty acids, respectively. Esterases were isolated from metagenomic libraries from a variety of soils using α-napthol

acetate as the indicator substrate and 4-Benzoylamino-2,5-dimethoxybenzenediazonium chloride (Fast Blue RR), a diazonium salt which complexes with the esterase hydrolysis product, α-napthol, thus forming an insoluble brown colour (Kim *et al.*, 2006). The isolation of lipases was accomplished using tributyrin, selecting for clones producing clear halos (Ranjan *et al.*, 2005). Finally, the isolation and identification of complete DNA polymerase I genes and domains was recently accomplished from glacial ice metagenomes (Simon *et al.*, 2009). This was achieved by the complementation of *E. coli* harbouring a temperature sensitive mutation in the 5'-3' exonuclease domain of DNA polymerase I (*polA*) for growth at a low temperature growth (Simon *et al.*, 2009). The utility of functional metagenomics in enzyme discovery has already translated to the development and commercialization of enzymes by the biotechnology industry. For example, the company Verenium, which specializes the development of novel compounds from uncultured microorganisms has a number of commercially available enzymes such as the alpha amylase, Fuelzyme®- LF, a thermostable alpha amylase originating from the metagenome isolated from the deep ocean floor. It is an effective enzyme for starch liquification http://www.verenium.com/specialty-enzymes_products_fuelzyme-lf.asp

Apart from enzymes, the metagenome has proven to be a rich source of novel drugs and compounds with antibacterial capacity. Novel genes encoding proteins that synthesize antibacterial compounds such as indirubin and indigo blue (Lim *et al.*, 2005a), amidases (Gabor *et al.*, 2004b), long-chain N-acyltyrosine antibiotics (Brady *et al.*, 2002), N-acyl amino acid synthase (Brady & Clardy, 2000), and tri-aryl cation antibiotics (Gillespie *et al.*, 2002) were isolated by screening library clones based on functional expression in *E. coli*.

Drug discovery in marine environments has recently received a great deal of attention as marine microorganisms are subjected to a variety of temperature, pressure and nutrient availability in oceanic habitats. From biomining efforts in the last 40 years, about 15,000 structurally diverse natural products were discovered from marine microbes, algae and invertebrates (Li & Qin, 2005). The application of metagenomics to the discovery of biologically active compounds has been applied to bacteria associated with marine invertebrates and sponges (Kennedy *et al.*, 2007). Recombinant expression of metagenomic clones in a suitable surrogate host has been optimized as in the case of the discovery of the cyclic octapeptides patellamide D and ascidiacyclamide from *Prochloron, sp.*, an unculturable endosymbiont of the ascidian *Lisoclinum patella*. The expression of the secondary metabolites was successful using *E. coli* and *Streptomycis lividans* as surrogate hosts (Long *et al.*, 2005).

Besides mining the metagenome for novel compounds such as biocatalysts or antibiotics, functional metagenomics may also be geared towards gaining more knowledge about specific functions carried out by bacteria in various habitats. Functional screens of metagenomic libraries have resulted in the discovery of novel genes encoding physiological functions, thus increasing the understanding of these functions. For example, genes for inducers and inhibitors of quorum sensing (Williamson *et al.*, 2005) and metabolite transport were found by functional screening of libraries constructed from DNA from three different soils (Majernik *et al.*, 2001). Functional screening of soil-derived metagenomic libraries was applied in the study of genes encoding the machinery for bacterial metabolism (Henne *et al.*, 1999; Wang *et al.*, 2006). By screening for utilization of 4-HB as the sole carbon source for

growth, genes for 4-HB dehydrogenase and enoyl-coenzyme A hydratase/isomerase were isolated and characterized (Henne *et al.*, 1999). The metabolism of similar compounds, poly-3-hydroxybutyrate (PHB) was also studied by screening metagenomic libraries derived from complex environments such as soil and activated sludge. Rather than directly screening the libraries for growth on PHB, libraries were screened by phenotypic complementation of *E. coli* mutants and PHB metabolism-deficient mutants of the legume-symbiont, *Sinorhizobium meliloti* (Wang *et al.*, 2006).

1.3 Phosphorous (P) metabolism

1.3.1 Importance of P and its environmental sources

Following carbon and nitrogen, phosphorous (P) is the third most abundant element in the bacterial cell (Wanner, 1996). Elemental P makes up approximately 3% of the dry weight of all living organisms (White & Metcalf, 2007). It is therefore not surprising that P is an essential nutrient for life. Aside from having a key role in the many metabolic pathways in the cell, P is a component of phospholipids, complex carbohydrates such as lipopolysaccharides, RNA, DNA and post-translationally modified proteins (Wanner, 1996; White & Metcalf, 2007).

Phosphorous metabolism is governed by many metabolic pathways including those involved in the acquisition and assimilation of P from the environment. Metabolic pathways of P metabolism are found in many biological processes such as: 1. phosphorous assimilation during exponential growth; 2. scavenging P compounds from the environment while P is limiting; 3. storage of high-energy P compounds while carbon is in excess or during

limitation of biosynthetic capacity; and 4. energy production due to the presence of the high-energy phosphoanhydride bonds (Wanner, 1996). Aside from nutrition and energy conversion, P metabolism is also implicated in cellular signaling processes where regulation of genes and operons are governed by the phosphorylation and dephosphorylation of DNA binding proteins.

Phosphorous is very abundant in the earth's crust and appears as inorganic phosphate minerals and organic phosphate derivatives in rocks and soil (Paytan & McLaughlin, 2007). Phosphorous is deposited into the soil through animal tissue decay and feces, through the weathering of rock and by the incorporation of pesticides, herbicides and P containing fertilizers. Soil borne phosphorous is either assimilated by bacterial and plant cells, or drains into water bodies or is incorporated into sedimentary rock. Organic phosphorous in the form of orthophosphate monoesters, orthophosphate diesters, phosphonates, phosphorous anhydrides and inositol phosphate constitute 30-50% of the total P found in soils (Rodriguez & Fraga, 1999).

In marine cycling, P is primarily transported into the ocean in a dissolved phase via riverine influx but also enters by atmospheric deposition through aerosols, volcanic ash and mineral dust (Paytan & McLaughlin, 2007). The fate of P in oceans is predominantly its deposition in marine sediment after becoming insoluble while a smaller fraction of marine P is taken up through hydrothermal vents that occur on the ocean's floor where interactions occur between seawater and oceanic crusts (Paytan & McLaughlin, 2007). Marine P cycling is largely driven by microorganisms, particularly planktonic cyanobacteria, with specific

emphasis on isolates of *Prochlorococcus* and *Synechococcus*, the major primary producers in the oligotrophic oceans (Moore *et al.*, 2005). These authors elucidated the relationship between the growth phenotype of these strains on various organic P sources, P stress response (marked by alkaline phosphatase activity) and the presence/absence of various P metabolism genes in their respective genomes. In general, P utilization and the P-stress response is variable among different genera of marine cyanobacteria suggesting that certain ecotypes have an advantage under conditions of P limitation (Moore *et al.*, 2005). An additional contribution to our understanding of marine P cycling was the finding of differential abundance of phosphate transport and metabolism genes (most significantly *pstS*, see section 1.3.3) in the metagenome of Caribbean Sea water compared to that of the eastern Pacific sea water (Rusch *et al.*, 2007). The difference was attributed to different conditions of P availability between these two environments, requiring a more enhanced adaptation of one population to limiting P conditions which as a result acquired a larger suite of genes related to P acquisition, P scavenging and regulation of P metabolism.

1.3.2 Types of P compounds

There are three types of P compounds utilized by bacteria, inorganic phosphates, phosphate esters and phosphonates. Inorganic phosphates exist in three chemical forms: inorganic phosphate ion (Pi), pyrophosphate (PPi), and polyphosphate [poly(Pi)]. Of these, Pi is the only form readily taken up by the bacterial cell via transport systems; PPi and poly(Pi) must first be broken down in the periplasm by bacterial alkaline phosphatase (Wanner, 1996). Organic forms of phosphate include ones that are transportable (sn-glycerol-3-phosphate (G3P), glycerophosphoryl diesters, hexose-6-phosphates, phosphoenolpyruvate, 2-

phosphoglycerate, and 3-phosphoglycerate) and non-transportable which include nucleotides, phosphonates (Wanner, 1996). Non transportable organic phosphate compounds enter the periplasm where they are hydrolyzed by periplasmic enzymes resulting in the release of Pi, which is taken up by inner membrane Pi transport systems (Figure 1-1).

Phosphonates are organic compounds that contain a carbon-phophorous (C-P) bond as opposed to the C-O-P ester bond, characteristic of the organophosphates. The first phosphonate compounds to be described were aminomethylphosphonic acid and a series of amino- and amino-substituted alkylphosphonic acids, synthesized by chemists in the middle of the twentieth century (Ternan *et al.*, 1998). The first biogenic phosphonate, 2-aminoethylphosphonate (2-AEP), was isolated from rumen protozoa and was subsequently shown to occur in sea anemones as a phosphonolipid (Ternan *et al.*, 1998). Other naturally occurring phosphonates were discovered more recently and were found to have commercial use and importance. For example, the antibiotics phosphonomycin and bialaphos were isolated from *Streptomyces* cultures (Lee *et al.*, 1995). Exploitation of the useful biological properties of phosphonates led to the production and the use of a wide range of phosphonate-based insecticides, herbicides, fungicides and antibiotics (Quinn, 2007). One widely used, broad spectrum herbicide is glyphosate (N-phosphonomethylglycine), the key ingredient of "Roundup", made and sold by the Monsanto Co. (Balthazor & Hallas, 1986). Glyphosate acts as an effective herbicide by inhibiting the plant enzyme 3-enol-pyruvylshikimate-5-phosphate synthase (EPSP synthase) thereby blocking the synthesis of the aromatic amino acids phenylalanine, tyrosine and tryptophan in plants (Steinrucken & Amrhein, 1980).

20

Excess glyphosate in the soil is degraded rapidly by microorganisms capable of utilizing it as their sole source of P (Liu *et al.*, 1991).

Figure 1-1: Schematic diagram of the assimilation of P by a Gram-negative bacterial cell.

 P compounds are first taken up by diffusion through the outer membrane porins. Porins are water-filled protein channels that take up solutes of various sizes depending on their size cutoff. Organic P compounds are degraded by phosphohydrolase in the periplasm. Depending on Pi availability these compounds are transported across the inner membrane by means of the low affinity-high velocity Pit transport system or the high affinity-low velocity Pst ABC transport system. Phosphonates are transported by an ABC phosphonate transporter and are degraded by a membrane bound C-P lyase protein complex.

outer membrane

periplasm

inner membrane

cytoplasm

Poly-P

PPi

R-O-P

Pi

Poly-P

PPi

R-O-P

Pi

phosphohydrolases

Pi

Pi

high affinity low
velocity Pi
transpoter

Pi

Pi

Pi

R-P

R + Pi

phosphonate
transporter

low affinity high
velocity Pi
transporter

R-P

C-P lyase

R-P

23
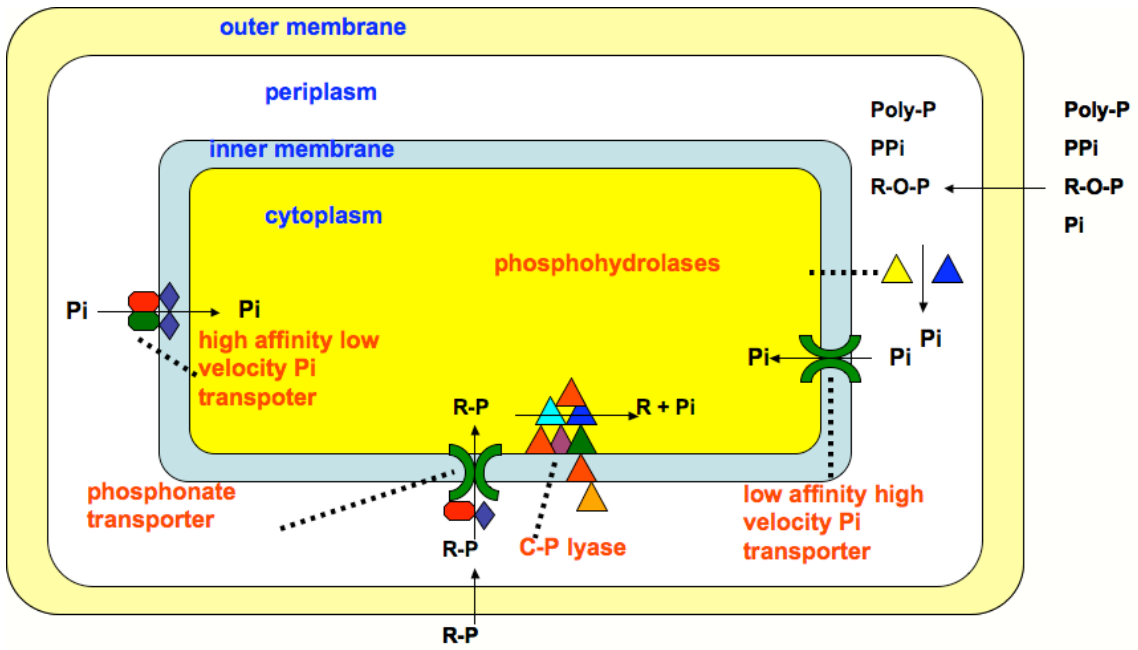
1.3.3 Inorganic phosphate (P_i) uptake

Most cells are highly dependent upon inorganic phosphate for energy. Due to a negative electrochemical potential across the cell membrane and its limited bioavailability owing to poor solubility, anionic Pi cannot enter the cytosol by means of simple diffusion (Werner & Kinne, 2001). Therefore, to ensure that its intracellular pools remain sufficient the cell has evolved active transport mechanisms for P_i: The low-affinity, high-velocity inorganic phosphate transporter (Pit) system and the high-affinity, low-velocity phosphate specific transport system (Pst).

The Pit system is a highly conserved P_i-Na^+ or P_i-H^+ symport system (Bottger & Pedersen, 2005). It is comprised of a single protein component which transports metal phosphates and is active under conditions of excess phosphate (Wanner, 1996). The transport of Pi by this system requires the anion to be complexed with divalent cations and is pH-dependent (Werner & Kinne, 2001). There has been extensive characterization of the *E. coli* Pit transporter, PitA, which is inhibited by arsenate, the energy uncoupler 2-4 dinitrophenol and the sulfhydryl reagent N-ethylmaleimide, and is resistant to sodium cyanide (Willsky & Malamy, 1980). A second Pit transporter, PitB, was shown to be regulated differently, by the environmental Pi concentration (Harris *et al.*, 2001).

The Pst transport system is well characterized in *E. coli*. It is a multi-protein complex spanning the periplasm and the inner membrane. It is a member of the ATP Binding Cassette (ABC) super family of transporters that are present in cells from all domains of life (Wanner, 1996). The complex usually contains four proteins, two of which (PstA and PstC) comprise the trans-membrane transport channel, one ATP binding protein (PstB) and one periplasmic

P$_i$ binding protein (PstS) (Wanner, 1996). The genes for the Pst transport system are found upstream to the gene for the regulatory protein PhoU (*pstSCAB-phoU*) (Steed & Wanner, 1993). The *phoU* gene encoded protein does not effect Pi uptake by the Pst system, however, a mutation in this gene has a deleterious effect on growth (Steed & Wanner, 1993). The mutation inhibits growth only in the presence of a functional Pst system suggesting that in the absence of a functional PhoU, toxic levels of Pi transported by the Pst system accumulate in the cell (Steed & Wanner, 1993). This made it difficult to determine the function of PhoU in *E. coli* until a recent study where the *phoB* promoter was replaced with the *Ptac* promoter, rendering the expression of the Pst system inducible by IPTG, thus alleviating the growth defects caused by the mutation (Rice *et al.*, 2009). The authors presented evidence that PhoU negatively regulates Pi transport by the Pst system at sufficient levels of Pi thus having a role in cellular Pi homeostasis (Rice *et al.*, 2009).

The study of Pst transport systems has also been carried out in Gram positive *Bacillus subtilis* and *Streptococcus pneumoniae* cells. The *B. subtilis* Pst system is composed of the proteins PstSCA,BA,BB where PstBA and PstBB are permease like membrane proteins that assist in the uptake of Pi across the cell membrane (Allenby *et al.*, 2004). The expression of the transport system in *B. subtilis* is regulated at the promoter level by the phosphate concentration (Qi *et al.*, 1997). The Pst system (PstSACB) in *S. pneumoniae* has a wider range of effects on cellular physiology than it does in *E. coli* (Novak *et al.*, 1999). For example when the *S. pneumoniae pstB* gene was disrupted, in addition to a decrease in phosphate uptake, the cell's transformation capability and antibiotic induced autolysis were both decreased (Novak *et al.*, 1999).

1.3.4 Enzymatic degradation of organic phosphate

Only a few organophosphate compounds such as G3P are taken up by a transport system. Most organophosphates such as nucleotides, sugar phosphates, phytic acid, etc. cannot be taken up into the cytoplasm via a membrane transport system. In Gram negative cells, non-transportable organophosphates enter the periplasm with or without the aid of a particular porin and are degraded by secreted periplasmic phosphatase enzymes (Wanner, 1996). These enzymes function as scavengers of organic phosphates by hydrolyzing them, thus releasing Pi and organic by-products (Rossolini *et al.*, 1998). In addition to their roles in bacterial metabolism and nutrition, phosphohydrolases may be involved in microbial virulence as exemplified by the respiratory burst-inhibiting acid phosphatases of *Legionella micadadei* and *Francisella tularensis* (Dowling *et al.*, 1992; Reilly *et al.*, 1996) and the protein-tyrosine phosphatases of *Yersinia spp* (Bliska *et al.*, 1991; Guan & Dixon, 1990). The major bacterial alkaline phosphatase, PhoA, has been extensively studied in *E. coli*. Upon limited Pi availability, *phoA* expression is induced 100-fold in a *phoB*-dependent manner and organic P esters are hydrolysed (Wanner, 1996). It is this type of regulation that has enabled the use of *phoA* as a molecular marker for environmental Pi stress e.g., in marine bacteria (Sebastian & Ammerman, 2009). These organisms are subject to variable Pi concentrations and therefore must adapt accordingly in the event of Pi stress. In a survey of marine bacterial alkaline phosphatases, the prevalence of *phoA* among bacterial isolates and GOS samples was found to be fairly low.

In contrast, a more recently discovered alkaline phosphatase, PhoX, was found to be widely distributed among marine isolates and metagenomic samples (Sebastian &

Ammerman, 2009). PhoX was first identified and genetically characterized in the cholera disease-causing *Vibrio cholerae* as a new monomeric alkaline phosphatase (Majumdar *et al.*, 2005). The enzyme has a different bivalent metal requirement for catalysis (Ca2$^+$) than PhoA which requires the metals Mn^{2+} and Zn^{2+} as co factors (Sebastian & Ammerman, 2009). PhoX was also found to be induced solely by P-starvation and accounted for 90% of the activity in the model marine bacterium, *Silicibacter pomeroyi*. Using metatranscriptomic datasets, it was found that *phoX* expression is more prevalent in oligotrophic marine environments (Sebastian & Ammerman, 2009). PhoX and PhoA share no significant amino acid sequence similarity and have different mechanisms for export into the periplasm. PhoX is transported via the Twin Arginine Transport (Tat) system while processed PhoA is transported to the periplasm via the Sec pathway (Zaheer *et al.*, 2009). Contrary to marine microorganisms, PhoX homologues have been identified in only a few terrestrial microorganisms including *Pseudomonas fluorescens* Pf0-1 (Monds *et al.*, 2006), *Campylobacter jejuni* (Wosten *et al.*, 2006), *Pasturella multocida* X-73 (Wu *et al.*, 2007) and *Sinorhizobium meliloti* (Zaheer *et al.*, 2009); the latter two PhoX enzymes were purified to homogeneity.

The basis for the classification of phosphatases was initially determined by their biochemical and physical properties such as pH optimum, substrate profiles and molecular size, however, with the increasing availability of sequence data, enzyme families can be identified based on conserved motifs (Thaller *et al.*, 1998). Below, two families of phosphohydrolases are discussed in detail: The Nonspecific Acid Phosphatases (NSAPs) and the Nucleotide Pyrophosphatase Phosphodiesterases (NPP).

27

The NSAPs are a group of secreted enzymes with optimal activity in the acidic-neutral pH range which hydrolyze a broad range of organic phosphoesters. Bacterial NSAPs have been divided into three classification groups, A, B and C based on the relatedness of their amino acid sequences and cellular function (Thaller *et al.*, 1994; Thaller *et al.*, 1995b; Thaller *et al.*, 1997a). Members of class A and B are secreted phosphatases of molecular weight 25-27 kDa that are distinguished from one another by having different conserved sequence motifs (Thaller *et al.*, 1994; Thaller *et al.*, 1995a; Thaller *et al.*, 1995b). Unlike members of class A, phosphatases belonging to class B are resistant to depolymerization by SDS, cannot metabolize the chromogenic substrate 5-bromo-3-chloro-indolyl phosphate (BCIP) and are inhibited by EDTA (Uerkvitz & Beck, 1981). Members of class C phosphatases are secreted lipoproteins with a polypeptide component with a molecular weight of approximately 30 kDa (Thaller *et al.*, 1997b). These enzymes share a distant amino acid sequence identity to members of class B and to some plant acid phosphatases (Rossolini *et al.*, 1998).

All members of the class A NSAP family contain the conserved amino acid sequence motifs, K-X(6)-R-P (domain I), P-S-G-H (domain II), and S-R-X(5)-H-X(2)-D (domain III) (Stukey & Carman, 1997). This sequence motif is reflected in the crystal structure of chloroperoxidase which shows the conserved residues form a vanadate-binding enzyme pocket which may act to bind the substrate to enzymes belonging to this family (Stukey & Carman, 1997). The first report of the purification and characterization of bacterial class A NSAPs was PhoN-Se from *Salmonella enterica ser. typhimurium* (Kasahara *et al.*, 1991). This enzyme is a homodimeric protein with 27-kDa subunits and a pH optimum of 5.5

(Weppelman *et al.*, 1977). It hydrolyzes a variety of different substrates including nucleoside mono-, di- and triphosphates, hexose and pentose phosphates; and many others (Kier *et al.*, 1977). Its affinity towards various substrate was found to be comparable to that of PhoN from *E. coli* and its overall reaction velocity is consistent against various hydrolysable substrates (Weppelman *et al.*, 1977). Another example of class A NSAP is from *Zymomonas mobilis* (PhoC-Zm). This protein was discovered to be Pi-irrepressible and it was the first class A NSAP to be sequenced (Pond *et al.*, 1989). Two additional class A NSAP genes with a Pi irrepressible acid phosphatase phenotype were also identified and characterized: *phoC*-Mm from *Morganella morganii* and the *phoN*-Ps from *Providencia stuartii*. The PhoC-Mm enzyme is a homotetrameric protein containing 25-kDa subunits and exhibiting a broad substrate specificity similar to that of the PhoN-Se (Thaller *et al.*, 1994). The NSAP from *P. stuartii* has not been purified and characterized but zymogram analysis and a high amino acid similarity to PhoN-Mm has lead researchers to hypothesize similar biochemical properties as well (Thaller *et al.*, 1995a). Two class A NSAPs from *Shigella flexneri* (PhoN-Sf and Apy-Sf) are also worth noting. PhoN-Sf is active against a number of hexose phosphates and nucleotides and has a pH optimum of 6.6 (Uchiya *et al.*, 1996). Apy-Sf is a 25-kDa monomer that shows a high activity towards nucleoside triphosphates and can also hydrolyze pyrophosphate, and to some extent *pNPP* ; however, it cannot hydrolyze AMP (Bhargava *et al.*, 1995). According to this pattern of substrate preference, the enzyme is considered to be an ATP diphosphohydrolase or apyrase.

The first and the only class A NSAP for which a structure has been determined thus far is the NSAP from *Escherichia blattae* (EB-NSAP) (Ishikawa *et al.*, 2000). By X-ray

structure determination, it was shown that EB-NSAP is a histidine phosphatase, a homohexamer of 150 kDa which shares a conserved sequence motif found in lipid phosphatases and mammalian glucose-6-phosphatases (G6P) (Ishikawa *et al.*, 2000). Its reaction mechanism is consistent with that of the two-step reaction mechanism of histidine phosphatases in that in the first step, nucleophilic attack on the phosphate group is achieved by histidine followed by a leaving group protonation by a different functional group, forming a phosphoenzyme intermediate. In the second step, the phosphoenzyme intermediate is hydrolysed, releasing Pi (Ishikawa *et al.*, 2000). The active site consists of the residues that comprise the three signature domains that characterize NSAPs. The conserved histidine of domain III, His 218, targets the substrate's phosphoryl group to produce a phosphoenzyme intermediate, thus being essential for catalytic activity. The side chain conformation of His 218 is stabilized by forming a hydrogen bond with Asp 193. The conserved histidine of domain II, His 150, functions as the proton donor for the substrate leaving group. Lys 115 and Arg 122 in domain I are involved in keeping the phosphate group proximal to His 218 thus supporting the nucleophilic attack. This stabilizing effect is enhanced by the side chain of Ser 148 and the amide nitrogen atoms of Gly 149 and His 150 which compose domain II. Ser 182 may have an important overall structural role while Arg 183 may have a role in the stabilization the phosphoenzyme intermediate (Ishikawa *et al.*, 2000).

The true identification of the class B NSAP family occurred with the isolation and characterization of the first class B NSAP. This was the minor Pi irrepressible NSAP, NapA-Mm, enzyme isolated from *M. morganii* in addition to the Class A NSAP, PhoC-Mm (Thaller *et al.*, 1995b). The first purified Class B NSAP was isolated from *S. enterica ser.*

typhimurium LT2 and was named AphA-Se (NSAP II) to distinguish it from the PhoN-Se isolated from the same bacterium (Kasahara *et al.*, 1991; Uerkvitz, 1988). A class B NSAP, AphA, was also found to be produced in *E. coli* MG1655. It was purified and determined to be a 100-kDa homotetrameric protein which requires a metal co-factor for activity (Passariello *et al.*, 2006; Thaller *et al.*, 1997a). Finally, a class B NSAP gene from *H. influenzae* was isolated, cloned and expressed, showing acid phosphatase activity and zymogram properties consistent with those of Class B NSAPs (Rossolini *et al.*, 1998).

Class C NSAPs are secreted lipoproteins that have an NSAP activity and are composed of proteins of approximately 30 kDa in size. Members of this NSAP family appear to be related on the sequence level to Class B NSAPs and to some plant acid phosphatases as they have four invariant aspartic acid residues in their conserved domains (Rossolini *et al.*, 1998). The first identified Class C NSAP was the OlpA enzyme of *Chryseobacterium meningosepticum*. This protein is 29 kDa in size and contains a signal peptide typical of bacterial membrane lipoproteins (Passariello *et al.*, 2003).

Many phosphoryl transfer reactions are catalyzed by members of the alkaline phosphatase (AP) superfamily which includes alkaline phosphatase, N-acetylgalactosamine-4-sulfatase, cerebroside sulfatase, phosphopentomutase, 2,3-bisphosphoglycerate-indepenent phosphoglycerate mutase, phosphoglycerol transferase, phosphonate monoesterase, streptomycin-6-phosphate phosphatase and phosphodiesterase/nucleotide pyrophosphatases (NPP) (Galperin *et al.*, 1998). The latter is a subgroup of phosphohydrolases that catalyze phosphate diester hydrolysis (Zalatan *et al.*, 2006a). In eukarya, they are found as both

membrane proteins with an extracellular active site and also as soluble proteins present in body fluids (Bollen *et al.*, 2000). The biochemistry and cellular function of NPPs have been extensively studied in eukaryal systems where it has been shown that they have a role in the cleavage of extracellular phosphate diester signaling molecules (Bollen *et al.*, 2000). The mammalian NPPs were nonetheless found to share a significant similarity to their counterparts from bacteria to plants when examining the catalytic domain, residues involved in metal binding and structural fold. This was backed up by site-directed mutagenesis studies (Gijsbers *et al.*, 2001). Owing to the crystallization of NPP from *Xanthomonas axonpodis*, structural and functional data is available for the elucidation of the role of distinct functional groups in substrate binding interaction and the shift in protein folding for the discrimination between monoester and diester substrates (Zalatan *et al.*, 2006a). The model for the NPP active site involves the coordination of two metal ions by six conserved metal ligands and a threonine mediated nucleophilic attack on the phosphate group. Threonine is covalently bonded to the phosphate group, displacing the leaving group to form a phosphoenzyme intermediate. Subsequently, water is used to hydrolyse the phosphoenzyme intermediate to complete the double displacement mechanism. The two metal ions activate the nucleophile, and stabilize the charge build up on the leaving group (Zalatan *et al.*, 2006a). The active site of NPP consists of residues of Asp 54, Asp 257, His 258, Asp 210, His 363 and His 214 which coordinate two Zn^{2+} ions. Thr 90 acts as the active site nucleophile which covalently bonds with an O from the phosphate group and along with Asn 111 forms a H bond with a non bridging O of the phosphate group. Residues Phe 91, Leu 123, Ser 155, Glu 160, Tyr 174 and Tyr 205 form the R' (second functional group in phosphodiesters) binding pocket. This

serves as a determinant for the specificity of NPP towards phosphodiesters (Zalatan *et al.*, 2006a).

Among the numerous proteins surveyed in the metagenomic GOS expedition, one particular Mg^{2+}- or Mn^{2+}-dependent phosphoserine/phosphothreonine protein phosphatase family was found to be prevalent among many sequenced environmental samples (Yooseph *et al.*, 2007). These sequences have at least seven conserved functional and structural motifs with invariant residues involved in metal and phosphate binding. By means of phylogenetic analysis of the catalytic domain key differences between bacterial and eukaryotic protein phosphatases were found (Yooseph *et al.*, 2007).

1.3.5 Phosphonate uptake and degradation

When introduced into the environment, phosphonates have a short half-life due to their degradation by microorganisms, which are not harmed by them and can use them as an energy source (Singh, 2009). Bacteria have evolved a variety of pathways to use phosphonates as the sole P or C source. Utilization of phosphonates for growth has been reported for a variety of bacterial isolates such as *Arthrobacter spp., Pseudomonas spp., Salmonella spp.*, *E. coli* and the photosynthetic bacterium *Rhodobacter capsulatus* ATCC 23782 (Schowanek & Verstraete, 1990). There have also been reports of phosphonate degradation by soil-borne fungi (Krzysko-Lupicka *et al.*, 1997). Phosphonate utilization has also been demonstrated in various environmental samples including soil, activated sludge and water samples (Schowanek & Verstraete, 1990). Although most of the attention is given to phosphonate degradation, some knowledge of phosphonate transport in *E. coli* has been

33

accumulated. Using TnphoA'-9 transposon mutagenesis, it was determined that phosphonate is most likely taken up by the binding protein-dependent PhnCDE transport system consisting of a permease component (PhnC), a periplasmic binding protein (PhnD) and an integral membrane protein (PhnE) (Metcalf & Wanner, 1993). There are also accounts of phosphonate transport by *Bacillus cereus*. The uptake of the biogenic phosphonate 2-AEP by this organism was demonstrated using radiolabelling experiments and it was found that the transport system is energy dependent and inducible by the substrate (Rosenberg & La Nauze, 1967).

Once taken up by the cell, phosphonates are then cleaved by four possible enzymes or enzyme complexes: phosphonatase, phosphonoacetate hydrolase, phosphonopyruvate hydrolase and CP-lyase (Kononova & Nesmeyanova, 2002). The phosphonatase enzyme was from *B. cereus*, the first reported C-P hydrolyzing enzyme isolated, purified and characterized (La Nauze *et al.*, 1970). The enzyme however was found to be specific to only 2-AEP as it could not degrade any other phosphonates (La Nauze *et al.*, 1970). Another example of a narrow substrate range enzyme was phosphonoacetate hydrolase, isolated from *P. fluorescens 23F* (McMullan & Quinn, 1994). The enzyme was found to be induced by phosphonoacetate and did not require phosphorous starvation for its induction (McMullan & Quinn, 1994). An enzyme for the hydrolysis of L-phosphonoalanine was also isolated from *Burkholderia cepacia* Pal6*,* conferring upon the organism the ability to utilize the substrate as a source of nitrogen, carbon or phosphorous (Ternan & Quinn, 1998). Finally, a large enzyme complex, C-P lyase, was found to be involved in the direct cleavage of the C-P bond to generate inorganic phosphate and the corresponding hydrocarbon. C-P lyase is responsible

34

for the degradation of alkylphosphonates which are commonly used by bacteria as the sole P source (Kononova & Nesmeyanova, 2002). Activity of the enzyme, however, occurs intracellularly and has not yet been detected in cell free extracts. The C-P lyase pathway has been identified in a few Gram positive bacteria such as *Arthrobacter sp. GLP-1* and *B. megaterium* as well as in numerous Gram negative bacteria including *E. coli, Pseudomonas* spp*., Rhizobium sp., Agrobacterium radiobacter*, and *Klebsiella spp* (Kononova & Nesmeyanova, 2002). As phosphonate utilization by bacteria is important both in growth and in global P cycling, it is not surprising that genes for the C-P lyase pathway are distributed among many different and distantly related microorganisms. Phylogenetic analysis of C-P lyase homologues suggests that this degradative pathway has undergone significant lateral gene transfer in the course of its evolution (Huang *et al.*, 2005).

The genetic basis for C-P lyase has been extensively studied in *E. coli* which contains a *phn* operon responsible for phosphonate uptake and utilization. The *phn* operon consists of 14 genes (*phnCDEFGHIJKLMNOP*) (Chen *et al.*, 1990). Gene products PhnG-PhnM are believed to form the membrane-associated C-P lyase required for phosphonate utilization (Metcalf & Wanner, 1993). PhnM has sequence similarity to integral membrane proteins of transporters while PhnL contains nucleotide binding sequences. The two gene products PhnN and PhnP, may be C-P lyase accessory proteins but are not required for phosphonate utilization. Finally, PhnF and PhnO have sequence similarity to regulatory proteins suggesting that they may have a regulatory role and not be necessary for catalysis (Metcalf & Wanner, 1993).

35

C-P lyase has also been studied in *Pseudomonas stuzeri*. This organism was found to have two distinct C-P lyase operons (White & Metcalf, 2004). One operon, the *htxABCDEFGHIJKLMN,* encodes proteins homologous to the constituents of the *E. coli* C-P lyase and the second operon, *phnCDEFGHIJKLMNP,* was identified upon mutation of the *htx* operon. Interestingly, the presence of two C-P lyase pathways does not confer upon *P. stutzeri* the ability to grow on phosphonate sources other than methylphosphonate and aminoethylphosphonate (White & Metcalf, 2004). The occurrence of two C-P lyases also has been demonstrated in *Arthrobacter sp. GLP-1*. This organism uses a different C-P lyase to degrade methane-phosphonate than the one it uses to degrade glyphosate (Kertesz *et al.*, 1991). An additional organism with two degradative pathways for phosphonates is *Enterobacter aerogenes*. Through complementation experiments, a phosphonatase pathway and a C-P lyase pathway were identified, supporting the fact that *Enterobacter aerogenes* has a broader phosphonate substrate utilization range than *E. coli* (Lee *et al.*, 1992). On the other hand, some organisms have a narrower phosphonate utilization range. For example, *Salmonella typhimurium* LT2 can only degrade AEP, consistent with the finding that only the phosphonatase pathway exists in this organism (Jiang *et al.*, 1995). Phosphonate degradation has also been examined in the enteric bacterium *Klebsiella aerogenes* due to its ability to grow on a variety of alkylphosphonates. By heterologously introducing polyphosphate kinase to the organism, it was found that phosphonate utilization is enhanced when Pi is assimilated into polyphosphate (Imazu *et al.*, 1998). Additional enteric bacteria with phosphonate degradation capabilities are *Klebsiella oxytoca, Kluyvera ascorbata* and *Kluyvera cryoscrescens* (Wackett *et al.*, 1987). Phosphonate degradation has also been characterized in

36

the food-borne pathogen, *Campylobacter spp.* Unlike most phosphonate-degrading bacteria, the ability of this organism to degrade phosphonate is not inhibited when Pi levels become elevated (Mendz *et al.*, 2005). In addition, no orthologues to known genes encoding phosphonate-degrading activities were found in the genome of *C. jejuni* suggesting its phosphonate-degrading enzymes are different from known C-P degrading enzymes (Mendz *et al.*, 2005).

The above examples indicate that the degradation of phosphonate is common among microorganisms, but due to the inability to demonstrate cell free activity, our understanding of its biochemical basis is limited. However, the genetic basis of the degradative process is now understood owing to the extensive genetic characterization of *E. coli* phosphonate metabolism. This pioneering work led the way to the genetic characterization of many additional bacterial isolates, as the *phn* operon was acquired through horizontal gene transfer. It is important to note, however, that not all phosphonate-degrading microorganisms have proteins similar to the Phn gene products. Consequently, microbial phosphonate degradation may involve previously uncharacterized pathways, potentially leading to the discovery of biotechnologically important novel proteins or protein complexes.

1.3.6 Phosphonate metabolism in *Rhizobia*

The application of phosphonate-based herbicides in agricultural soil is a continuing process despite their toxicity towards mammals (including humans). Organophosphorous herbicides affect the mammalian nervous system by overstimulation of the nerves, causing convulsion, paralysis and finally death for insects and mammals (Singh & Walker, 2006).

The main toxicological mechanism of organophosphorous herbicides is the irreversible inhibition of acetylcholine esterase, an enzyme responsible for conversion of acetylcholine, the neurotransmitter responsible for muscle contraction into choline and acetyl-coA, thus causing the accumulation of acetylcholine and the resulting permanent muscular contraction (Karp, 2005). It is therefore not surprising that extensive research is continuously done on the biodegradation of phosphonate-based herbicides by soil bacteria. In addition to herbicides, soil bacteria can also degrade phosphonate-based antibiotics produced as secondary metabolites of soil microorganisms. *Rhizobia* are soil bacteria that are known for their ability to symbiotically form nodules and fix nitrogen for leguminous plants. Aside from their symbiotic role, rhizobial phosphonate utilization also has implications in survival in the soil due to antibiotic resistance and agricultural benefits due to herbicide resistance. The ability to degrade phosphonomycin, a phosphonate-based antibiotic produced by members of the *Streptomyces* genus, was demonstrated in *Rhizobium haukauii* PMY1 (McGrath *et al.*, 1998). This activity represents a mechanism for resistance to the antibiotic by *Rhizobium haukauii* PMY1 as well a novel phosphate irrepressible pathway of phosphonate metabolism by *Rhizobium spp* (McGrath *et al.*, 1998). The degradation of phosphonate-based antibiotics and a wide range of additional phosphonates has also been demonstrated in *Agrobacterium radiobacter*. Its broad substrate specificity makes it a good model organism for the study of bacterial C-P lyase (Wackett *et al.*, 1987). The benefits of rhizobial degradation of phosphonate were demonstrated with the application of rhizobial genes in the development of transgenic herbicide tolerant plants. This was shown with transgenic bentgrass (*Agrostis*

stolonifera L) which acquired the glyphosate tolerance gene 5-enol-pyruvylshikimate-3-phosphate synthase (CP4 EPSPS) from *Agrobacterium* CP4 (Watrud *et al.*, 2004).

A number of *Rhizobium* and *Agrobacterium* strains can utilize glyphosate as a sole P source, and glyphosate uptake and degradation was demonstrated in *Sinorhizobium meliloti* revealing that glyphosate is metabolized via the C-P lyase pathway (Liu *et al.*, 1991). Genetic and biochemical characterization of phosphonate and phosphate metabolism in *S. meliloti* later began upon the discovery that the PhoCDET transport system for phosphate/phosphonate is essential for its symbiotic association with alfalfa (*Medicago sativa*) (Bardin *et al.*, 1996).

The role of the transport system in the uptake of phosphonate by *S. meliloti* was indirectly assessed by showing that its Pi transport capacity is inhibited by phosphonates, providing evidence that PhoCDET may transport both Pi and phosphonates (Voegele *et al.*, 1997). Furthermore, a PhoCDET deletion strain grew poorly using methylphosphonate, ethylphosphonate and aminomethylphosphonate as the sole P source, suggesting the possibility of the above phosphonates being taken up by this transport system (Bardin *et al.*, 1996).

However in addition to the *phoCDET* operon, an upstream gene, homologous to the *E. coli phnM*, was deleted. This gene is believed to encode a necessary component of the CP-lyase complex, therefore rather than the absence of *phoCDET*, the absence of *phnM* may account for the above phenotype. This is further supported by the ability of *phoC* and *phoT* single mutants to grow using aminomethylphosphonate and methylphosphonate as sole P

sources (Bardin *et al.*, 1996). The *phoCDET* deletion strain was able to grow using AEP as the sole P source suggesting that its transport is carried out by a system different than PhoCDET and its degradation is carried out by a pathway different than C-P lyase (Bardin *et al.*, 1996).

Taken together with the fact that phosphonate transport by PhoCDET has not been directly assayed, it is not certain that PhoCDET is the sole phosphonate transport system in *S. meliloti*. The symbiotic deficiency of PhoCDET mutants was found to be suppressed by a second mutation mapped to genes encoding a Pit-like transport system for Pi (Bardin *et al.*, 1998). There are no reports of the ability of the *S. meliloti* Pit system to transport phosphonates. In contrast to the PhoCDET system, Pi transport by a third ABC transporter, PstSCAB, was not inhibited by phosphonates suggesting that it is Pi specific (Yuan *et al.*, 2006a). Phosphonate degradation by *S. meliloti* was characterized by cloning the putative C-P lyase pathway *phnGHIJK*, constructing a deletion mutation of this locus, and expressing each of PhnG, PhnH, PhnI, PhnJ and PhnK (Parker *et al.*, 1999). It was found that the mutation in the *phn* locus resulted in the inability of *S. meliloti* to utilize a wide variety of phosphonates commonly degraded by the C-P lyase pathway as the sole P source (Parker *et al.*, 1999). Furthermore, the expression of the PhnG, PhnH and PhnK proteins occurred only when *S. meliloti* was grown with specific phosphonates as the sole P source (Parker *et al.*, 1999). In addition to C-P lyase, *S. meliloti* uses a different enzyme to degrade 2-AEP which was shown by the lack of expression of *phnG,-H* and *–K* genes when grown in the presence of 2-AEP and also by the ability of the *phn* mutant to utilize this compound as the as the sole P source (Parker *et al.*, 1999).

1.4 This work

The vast majority of bacteria (90-99.9%) have not yet been cultured. Consequently, bacterial diversity and metabolic capacity has been greatly underestimated. By applying functional metagenomics a better understanding of their diversity becomes accessible, paving the way to: (i) the discovery of novel genes encoding enzymes of biotechnological importance; and (ii) a better understanding of microbial genetic diversity, population structure and ecological roles microorganisms play in the environment.

One of the key elements metabolized by bacteria is P, a nutrient required for cell structure, metabolic pathways and DNA synthesis. As microbial habitats vary in P bioavailability, so do the means by which microbial populations adapt to such conditions. Through P metabolism bacterial populations in activated sludge and soil have a key role in bioremediation from P-containing pollutants, increasing P availability to plants and are a source of phosphatase enzymes which may be used for various applications. Consequently, the genetic and functional diversity associated with bacterial P metabolism is of great interest.

By employing functional metagenomics, this project addresses P metabolism in microorganisms inhabiting wastewater treatment activated sludge and soil. Metagenomic libraries from the above communities were used for the phenotypic complementation of *E. coli* and *S. meliloti* mutants, resulting in the identification of two acid phosphatases, one alkaline phosphatase, several genes involved in phosphonate metabolism and a large number of other genes, some linked with P metabolism and others with unknown functions. Through phylogenetic analysis and biochemical characterization of some of the proteins, this project

illustrates some of the biodiversity of P metabolism genes in bacteria found in complex communities. Overall, the project demonstrates the effective use of function-driven metagenomics for discovering novel enzymes and deepens our knowledge of an important metabolic process carried out by microorganisms.

Chapter 2: Materials and methods

2.1 Bacterial culture and microbiological techniques

2.1.1 Bacterial strains, plasmids and transposons

The bacterial strains, plasmids and transposons are listed in Table 2-1. Metagenomic library clones isolated by screening for BCIP hydrolysis are denoted as pACX_ and pBCX_ where clones designated with an "A" were identified by screening at pH 7 and clones designated with "B" were identified by screening at pH 5.5. Library clones isolated by phenotypic complementation of *S. meliloti* strain RmF726 are designated pCX4-nF, pCX6-nF and pCX9-nF and library clones isolated by phenotypic complementation of *S. meliloti* strain RmG471 are designated pCX4-nG, pCX6-nG and pCX9-nG. All clones isolated by phenotypic complementation of *S. meliloti* are maintained in *E. coli* DH5α and in either *S. meliloti* mutant RmG471 or RmF726. All strains are preserved as permanenty frozen stocks at -80°C in tryptone yeast extract (TY) broth with 16% DMSO for *S. meliloti* and Luria broth (LB) with 14% DMSO for *E. coli* strains.

2.1.2 Media, antibiotics and growth conditions

S. meliloti was grown on TY (Beringer, 1980) or Mineral Salts (MS) medium at 30°. *E. coli* was grown on LB medium (Sambrook & Russel, 2001) at 37°C. The MS medium was composed of 15 g/l NH_4Cl, 0.24 g/l $CaCl_2 2H_2O$, 0.48 g/l $MgSO_4$, 35 g/l morpholino

Table 2-1: Bacterial strains, plasmids, metagenomic libraries and transposons used in this study

Strain, plasmid, metagenomic library or transposon	Relevant characteristics	Reference or source
Strain		
Sinorhizobium		
meliloti		
Rm1021	SU47 *str-21* (Sm^r)	(Meade *et al.*, 1982)
RmP110	Rm1021 with changed wild type *pstC*	(Yuan *et al.*, 2006a)
RmP636	RmP110 *pit310*::Tn5 *phoC*490ΩSp^r	(Yuan *et al.*, 2006a)
RmF726	ΔΩ5149-5079::Tn5-233 Gm^r-Sp^r, Nm^s	(Charles & Finan, 1991)
RmG471	ΔΩ5033-5025::Tn5 Nm^r Gm^s-Sp^s	(Charles & Finan, 1991)
RmG439	ΔG439 [Δ*phoCDET* HindIII::Nm^r (12 kb)]	(Charles *et al.*, 1991)
RmG490	ndvF-1.7ΩSp^r = *phoC*Ω490	(Charles *et al.*, 1991)
RmG491	ndvF-5.3ΩSp^r = *phoT*Ω490	(Charles *et al.*, 1991)
RmG830	*phoC*Ω490 *sfx1* pit310::Tn5	(Bardin *et al.*, 1998)
RmH838	*phoB*::TnV	(Bardin & Finan, 1998)
RmG762	*phoC*490ΩSp^r *sfx1*	(Bardin *et al.*, 1998)
Escherichia coli		
DH5α	F$^-$ *endA1 hsdR17 (r$_K$$^-m_K$$^+$) supE44 thi-1 recA1gyrA96 relA1Δ(argFo-lacZYA)U169φ80dlacZΔM15,λ$^-$*	(Finan *et al.*, 1986; Hanahan, 1983)
DH5α (pRK600)	Mobilizing strain; Cm^r	(Finan *et al.*, 1986)

Strain, plasmid, metagenomic library or transposon	Relevant characteristics	Reference or source
HB101	*supE44hsdS20(r_B^- m_B^-) recA13 ara-14 proA2 lacY1 galK2 rpsL2 xyl-5 mtl-1 leuB6 thi-1*	(Boyer & Roulland-Dussoix, 1969)
BL21 (DE3)/ pLysS	*F-ompT hsdSB (rB- mb-) gal dcm (DE3)/ pLysS (Cm^R)*	Novagen
MM294A	*pro*-82 *thi-pro*-82 *thi*-1 *hsd*RI7 *sup*E44 *end*AI	(Yarosh *et al.*, 1989)
MT621	MM294A *malF*::Tn*phoA*	(Charles & Finan, 1991)
Metagenomic library		
CX4	pRK7813 clone bank in HB101 constructed with DNA inserts derived from pulp waste activated sludge	(Wang *et al.*, 2006)
CX6	pRK7813 clone bank in HB101 constructed with DNA inserts derived from municipal waste activated sludge	(Wang *et al.*, 2006)
CX9	pRK7813 clone bank in HB101 constructed with DNA inserts derived from soil	(Wang *et al.*, 2006)
Plasmids		
pRK7813	IncP cosmid cloning vector; Tc^R	(Jones & Gutterson, 1987)

Strain, plasmid, metagenomic library or transposon	Relevant characteristics	Reference or source
pUC18	Cloning vector, ColE1, *oriV, bla;* Amp^R	Fermentas
pGEM T-Easy	Cloning vector, Amp^R	Promega
pACX4(6)-xx and pBCX4(6)-xx series	pRK7813 cosmid clones from libraries CX4 or CX6, that utilize BCIP	This study
pAR003	6.5 kb BamHI fragment of pACX6.13 in pUC19	This study
pAR004	3.5 kb BamHI fragment of pACX6.71 in pUC19	This study
pAR005	6.3 kb BamHI-EcoRI fragment of pBCX4.10 in pUC19	This study
pET30 series	Bacterial expression vector, Km^R	Novagen
pAR0032	835 bp PCR product *phoN*$_{ACX6.13}$ from pAR003 in pGEM T-Easy, Km^R	This study
pAR0045	1648 bp PCR product (flanked by NotI) *phoA*$_{ACX6.71}$ from pAR004 in pET30a, Km^R	This study
pBBR1-MCS5	Cloning vector; Gm^r derivative of pBBR1	(Antoine & Locht, 1992)
pCX4-nF(G) series	CX4 cosmid clones that complement RmF726 and/or RmG471 for the utilization of glyphosate as the sole P source	This study

Strain, plasmid, metagenomic library or transposon	Relevant characteristics	Reference or source
pCX6-nF(G) series	CX6 cosmid clones that complement RmF726 and/or RmG471 for the utilization of glyphosate as the sole P source	This study
pCX9-nF(G) series	CX9 cosmid clones that complement RmF726 and/or RmG471 for the utilization of glyphosate as the sole P source	This study
pAR010	3.5 kb HindIII fragment of pCX4-3G in pBBR1-MCS5	This study
pAR011	8.4 kb EcoRI fragment of pCX6-8G in pBBR1-MCS5	This study
pAR012	16 kb BamHI fragment of pCX9-12G in pBBR1-MCS5	This study
pAR013	6 kb BamHI fragment of pCX6-9G in pBBR1-MCS5	This study
Transposon		
EZ::Tn<Kan2>	Km^R	EPICENTRE Technologies

propane sulfonic acid (MOPS), pH 7, 20 mg/ml FeEDTA, 15 mM glucose, $3X10^{-4}$ mg/ml biotin, $1.69X10^{-3}$% thiamine 0.34 g/l glyphosate, trace elements at pH 7.2 as follows: 1 mg/l $FeSO_47H_2O$, 10 µg/l H_3BO_3, 11.2 g/l $MnSO_4H_2O$, 124.6 µg/l $ZnSO_47H_2O$, 78.2 µg/l $CuSO_45H_2O$ and 10ug MoO_3. Agar (Bioshop) was added to the media at a final concentration of 1.5% for *E. coli*. For *S. meliloti*, phosphate free Difco Agar was used at 1.5%.

Antibiotics were obtained from BioShop or Fisher Biochem. All antibiotics were maintained as sterile stock solutions at 4°C. Water-soluble antibiotics were filter sterilized. For *S. meliloti*, antibiotics were routinely used at the following concentrations (µg/ml): Gentamycin sulfate, 20; Neomycin sulfate, 200; Streptomycin sulfate, 200; Tetracycline hydrochloride 10, and spectinomycin, 100. For *E. coli*, antibiotics were routinely used at the following concentrations (µg/ml): Ampicillin, 100; Chloramphenicol, 20; Kanamycin sulfate, 10 (except 50 when expressing pET30 in *E. coli* BL21); Tetracycline hydrochloride, 10. For blue white screening in *E. coli*, using pUC18, pGEM T-Easy or pBBR1-MCS5, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) was used at a concentration of 40 µg/ml. For screening metagenomic libraries for phosphatase activity in *E. coli*, 5-bromo-4-chloro-3-indolyl phosphate

2.1.3 Environmental samples:

Pulp and municipal waste activated sludge samples were previously described (Neufeld *et al.*, 2001). Soil samples were collected from along the bank of Laurel Creek,

University of Waterloo campus, Waterloo, ON, Canada. This soil is characterized by its richness in organic material and its sandy loam.

2.2 Bacterial genetic techniques

2.2.1 Triparental mating / conjugation

Triparental mating/conjugation was done either using liquid cultures of donor, recipient and mobilizer in log phase or mixing colonies of donor, recipient and mobilizer strains on an LB plate. In cases when liquid cultures were used, donor, recipient and mobilizer strains were mixed together, washed 3-4 times with sterile saline solution and resuspended in 50 µl sterile LB. Resuspended mixture was spotted on LB, allowed to dry and incubated at 30°C overnight. The mating spots were resuspended in sterile saline solution and plated on the appropriate selective medium. For conjugation experiments where *E. coli* was the recipient and *S. meliloti* was the donor, *E. coli* transconjugants were selected by incubation at 37°C for 24 hours.

2.2.2 Screening of metagenomic libraries for phosphatase activity and growth selection on phosphonate

Metagenomic libraries were plated on LB supplemented with BCIP and tetracycline, and incubated O/N at 37°C. BCIP hydrolyzing clones were restreaked 3-4 times on LB BCIP to confirm the phenotype. Once confirmed, cosmids were isolated by alkaline lysis method and introduced by transformation to *E. coli* DH5α. *S. meliloti* mutants were screened for the inability to grow on MS medium supplemented with glyphosate as the sole P source. Phenotypic complementation was done by mobilizing the metagenomic libraries into the

growth-deficient mutants by triparental mating. Complemented RmG471 mutants were selected on MS medium with glyphosate, tetracycline and neomycin. Complemented RmF726 mutants were selected on MS medium with glyphosate tetracycline and gentamycin. Complemented mutants were restreaked 3-4 times on MS glyphosate and the appropriate antibiotics to confirm the phenotype. Cosmids from complemented mutants were reintroduced into *E. coli* DH5α by conjugation selecting transconjugants on LB tetracycline while incubated at 37°C to counterselect the *S. meliloti* donor. Cosmids were purified by alkaline lysis method, digested with BamHI to determine the number of unique cosmids, and reintroduced into RmG471 and/or RmF726 by conjugation, selecting on MS medium with glyphosate, tetracycline and either neomycin (if complementing RmG471) or gentamycin (if complementing RmF726). Additional mutants were complemented in a similar manner, selecting transconjugants with the following antibiotics: RmG439, tetracycline and neomycin; RmG490, tetracycline and spectinomycin; RmG491, tetracycline and spectinomycin; RmG830, tetracycline and spectinomycin.

2.3 Molecular biology techniques

2.3.1 Plasmid isolation (alkaline lysis)

An *E. coli* culture (5 ml) was made by inoculation with a single colony and incubated overnight at 37°C. The culture (3 ml) was pelleted by centrifugation at 13,000 rpm in a table top centrifuge (Desaga) for 30 sec. The supernatant was discarded. The pellet was resuspended in 100 μl TEG solution (50 mM glucose, 25 mM Tris-HCl pH 8, 10 mM EDTA pH 8) containing 200 μg/ml RNase (20 μl of 10 mg/ml RNase stock per 1ml of TEG). 200 μl

of alkaline lysis solution (0.2 N NaOH, 1% SDS) was added and the suspension was mixed gently by inverting 6-8 times. Next, 150 μl of neutralization solution (3 M potassium acetate, 11.5% v/v acetic acid) was added and the suspension was mixed by inverting 6-8 times. The suspension was centrifuged at 13000 rpm for 5 minutes and the supernatant was transferred to a new tube. 150 ul of chloroform was added to the supernatant and the mixture was vortexed. The mixture was then centrifuged at 13000 rpm for 3 minutes and the aqueous layer was transferred to a new tube. Two volumes (~800 μl) of cold 95% ethanol were added and the mixture was vortexed. The mixture was then centrifuged at 13000 rpm for 10 minutes and the ethanol containing supernantant was decanted. The pellet was washed with 150 μl 70% ethanol, left to dry and resuspended in 20 μl 10 mM Tris-HCl 1 mM EDTA.

2.3.2 Preparation and transformation of competent *E. coli* cells

Competent *E. coli* cells were prepard using the $CaCl_2$ method based on the protocol described by Cohen *et al.* (Cohen *et al.*, 1972). All centrifugation steps were carried out at 4°C using the 7685C rotor in an IEC 21000R centrifuge. 100 ml culture was grown to mid log phase and cooled on ice. Culture was pelleted by centrifugation at 5000 rpm for 5 min. The pellet was resuspended in cold 50 ml 100 mM $CaCl_2$ and incubated on ice for 30 min. The suspension was then centrifuged at 3500 rpm for 5 min and resuspended gently in cold 10 ml 100 mM $CalCl_2$. Cells were incubated on ice for 24 hours and glycerol was added to 15% final concentration. Cells were stored frozen at -70°C.

Competent cells were transformed as follows: 200 μl of competent cells were thawed on ice. 50 μl of ice-cold 100 mM $CaCl_2$ and 50 μl of competent cells were added to a tube

containing DNA used for transformation, and mixed gently. The mixture was incubated on ice for 30 min. The mixture was then incubated at 42°C for 90 s and immediately transferred to ice for 2 min. 0.5 ml fresh LB broth was added and the transformation mixture was incubated at 37°C for 45 min -1 h. Cells were recovered by centrifuging at 10,000 rpm for 2 min. Supernatant was decanted and the pellet was resuspended in 100 μl LB. 10-100 ul of suspension was plated on the appropriate selective LB plate. Plates were incubated overnight at 37°C.

Electrocompetent cells were prepared using the method based on the protocol described by Hanahan (Hanahan, 1983), as follows: 200 ml culture of *E. coli* was grown at 37°C with agitation to OD 0.6-0.9. The culture was then placed on ice and subsequent steps were carried out while cells were kept at 4°C. The cells were pelleted by centrifugation at 4000 rpm for 25 min. The supernatant was removed and the pellet was resuspended in 40 ml ice-cold deionized H_2O. Cells were centrifuged at 4000 rpm for 25 min. The pellet was resuspended in 25 ml ice-cold deionized H_2O and centrifuged at 4000 rpm for 25 min. The pellet was then resuspended with 4 ml ice-cold 10% glycerol and centrifuged at 4000 rpm for 10 min. The pellet was resuspended in 500 μl x ice-cold 10% glycerol and either used right away for transformation or stored at -70°C.

Electrocompetent cells were transformed as follows: 1-3 μl of DNA used for transformation was mixed with 50 μl of electrocompetent cells and left on ice for 1 min. The mixture was added to a chilled sterile electorporation 0.1 cm cuvette (BioRad) Electroporation was done using the MicroPulser electroporator (BioRad) at 1.8 kV electric

potential, 10 μF capacitance and 200 Ω resistance. After electroporation, cells were recovered by adding 1 ml fresh LB and the mixture was incubated at 37°C for 45 min – 1 hr. The cells were recovered by centrifuging at 10,000 rpm for 2 min. The supernantat was decanted and the pellet was resuspended in 100 μl LB. 10-100 ul of suspension was plated on the appropriate selective LB plate. The plates were incubated overnight at 37°C.

2.3.3 DNA library construction

Cosmid DNA was purified using the "Plasmid Midi Kit" (Qiagen). 8 μg cosmid DNA was partially digested with Bsp1431 for 90 min at 37°C using 0.2 units / μl of the restriction enzyme. The reaction mixture was then inactivated at 80°C for 20 min. Restriction digest products of size ~2kb were extracted from the gel and purified using the "Silica Bead DNA Gel Extraction Kit" (Fermentas) and ligated into dephosphorylated pUC18 vector. Ligation mixture was precipitated in ethanol as follows: Two volumes of 95% ethanol were added to ligation mixture with 2.5 M of ammonium acetate. The mixture was incubated at -20°C for at least 30 min before centrifuging at 130000 rpm for 30 min at 4°C. The supernatant was decanted and 150 μl of 70% ethanol was added to the DNA pellet. The mixture was centrifuged at 13000 rpm for 2 min at 4°C. The DNA pellet was resuspended in 10 μl deionized H_2O. 2 μl was used to transform *E. coli* DH5α by electroporation, selecting using LB with ampicillin and blue white screening using X-gal. 192 recombinant transformants were collected for each cosmid for shotgun sequence assembly.

2.3.4 Induction of phosphatase genes in *pET*30 series of expression vectors

Full length ORFs of phosphatases *phoN*$_{ACX6.13}$ and *phoN*$_{BCX4.10}$ were expressed in *E. coli* BL21 DE3/pLysS by autoinduction (non-IPTG expression). Autoinduction medium (Studier, 2005) contained per litre: 6 g Na$_2$HPO$_4$, 3 g KH$_2$PO$_4$, 20 g tryptone, 5 g yeast extract and 5 g NaCl (pH 7.2). The medium was supplemented with glucose, lactose and glycerol at a final concentration of 0.05%, 0.2% and 0.6%, respectively. Strains were inoculated into this medium (5 ml) and incubated for 16 hours at 37°C. 10 μl of cell induction culture was added to 10 μl 2X loading dye (10% SDS, 28% glycerol, 115 mM tris base, 0.27 mM DTT and bromophenol blue). The mixture was boiled for 10 min, and separated by SDS-PAGE at 10% polyacrylamide concentration. Following electrophoresis, the gel was stained using Coomassie Brilliant Blue and analyzed by Western blotting.

2.4 DNA manipulation methods

2.4.1 Restriction digestion

Restriction enzymes were purchased from Fermentas, New England Biolabs and Roche. Routine digestions were performed in a final volume of 20 μl containing reaction buffer at 1X concentration, DNA (200 – 500 ng) and at least 5 U of the appropriate restriction enzymes. All digestions were carried out at 37°C for 2 – 4 h. Partial digesions using Bsp1431 were performed under different conditions (section 2.3.3). Enzymes were inactivated by incubation at the recommended temperature for the specific enzyme.

2.4.2 Ligation reaction

The vector and insert DNA were combined at a molar ratio of 1:3 at a reaction volume not exceeding 10 µl. The reaction mixture contained reaction buffer at 1 X final concentration (buffer contains ATP) and T4 DNA ligase (Fermentas) (1U). The mixture was incubated at 16°C for 1 h to overnight and was used to transform *E. coli*.

2.4.3 Dephosphorylation of vector DNA

Vector DNA fragments digested with the appropriate restriction enzymes were dephosphorylated using shrimp alkaline phosphatase (Fermentas). The mixture containing the reaction buffer at 1 X concentration, the digested DNA and enzyme in a final volume of 10 µl was incubated at 37°C for 10 min followed by inactivation of the alkaline phosphatase at 65°C for 15 min.

2.4.4 Agarose gel electrophoresis

Gels were prepared with 1 X TAE buffer with an agarose concentration of 0.6-1% DNA was visualized by adding Gel Red stain (Biotium Inc.) at 1 X final concentration before casting the gel. For estimation of size of fragments, the following standard markers were used: Lambda DNA cut with HindIII (Fermentas) and 1 kb ladder (Fermentas).

2.4.5 DNA amplification by PCR

Primers used in this study were obtained from Sigma Genosys and were designed using the program "Amplify 3X" http://engels.genetics.wisc.edu/amplify/. PCR reaction mixtures contained 1 X reaction buffer, 1.5 mM $MgSO_4$, 0.2 mM dNTPs mix, 0.3 µM of each primer, 0.02 U / µl KOD Hot Start DNA polymerase (Novagen) or 0.025 U / µl Taq

DNA polymerase (Fermentas) and 50 ng DNA template. Reaction was carried out in a Mastercycler Personal thermocycler (Eppendorf, model # 5332-45779) using the recommended protocol (thermal activation at 95°C, 2 min; 30 cycles of denaturation at 95°C, for 20 sec; primer annealing at 59°C for 10 sec; extension at 70°C for 15 sec; followed by a 70°C for 10 min). Primer annealing temperatures were varied depending on the melting temperature (Tm) of the primers used. PCR products were visualized by agarose gel electrophoresis where gels were stained with Gel Red.

2.4.6 Cloning of phosphatase genes

To prepare an overexpression construct for acid phosphatase $phoNA_{CX6.13}$, primers pAR003F: 5' ATGTTTTCGCCACGCCAACT 3' and pAR003R2:

5' CTGGCTTTCCGAGCGCTGCG 3', were used to amplify the 835-bp ORF $phoN_{ACX6.13,}$ by PCR, including the start codon but excluding the stop codon, from pAR003. A poly A-tailing reaction was subsequently performed on the PCR product with the reaction mixture containing 7 µl PCR product 1X Taq polymerase buffer (Fermentas), 1 µl 0.2 mM dATP and 1 U Taq polymerase (Fermentas) in 10 µl volume. The mixture was incubated at 70°C for 30 min in a thermocycler (Eppendorf). The A-tailed fragment was then cloned into pGEM T-Easy vector (Promega), to generate construct pAR0032. $phoN_{ACX6.13}$ was cut with EcoRI and ligated into EcoRI cut expression vector pET30b (Novgen) to generate pAR0033 containing $phoN_{ACX6.13}$ in frame with the C-terminal His tag of pET30b. Diagnostic digestion of pAR0033 prepared from *E. coli* DH5α with EcoRI and SphI was performed to check for the presence of the insert and to ascertain its orientation.

To overexpress the alkaline phosphatase encoded by $phoA_{ACX6.71}$, primers designed with a NotI restriction site (pAR004F4: 5' GGGGCGGCCGCATGCGTAATATATTCCTTTC 3' and pAR004R4: 5' GGGGCGGCCGCCCGCACACCGGTTTCTCC 3') were used to PCR-amplify the 1645-bp ORF $phoA_{ACX6.71}$ including the start codon but excluding the stop codon, from pAR004. The PCR product was subsequently digested with NotI and ligated into NotI digested expression vector pET30a to generate construct pAR0045 containing $phoA_{ACX6.71}$ in frame with the C-terminal His tag of pET30a. Diagnostic digestion of pAR0045 prepared from *E. coli* DH5α with NotI and EcoRI was performed to check for the presence of the insert and to ascertain its orientation.

2.4.7 DNA sequence determination strategies

Regions of interest from cosmids were sequenced by *in vitro* transposon mutagenesis. Mutagenesis was performed using the EZ::TN^{TM} <Kan-2> insertion kit (Epicentre Technologies, Madison, WI, USA). Plasmid DNA was prepared by alkaline lysis method, digested with BamHI and HindIII to determine relative location of transposon insertions. DNA sequence determination was performed at the MOBIX Lab, McMaster University, Hamilton, ON, Canada. The primers KAN-2 FP-1 forward primer 5' ACCTACAACAAAGCTCTCATCAACC 3' and KAN-2 RP-1 reverse primer 5' GCAATGTAACATCAGAGATTTTGAG 3', supplied by Epicentre Technologies were used for DNA sequence determination from regions of transposition sites. DNA sequence assembly was done using the SeqMan program from the DNA Star program suite.

Large scale cosmid sequencing was done by generating subclone libraries from the cosmids in pUC18 and sequencing 192 clones from each library using the universal M13 forward and reverse primers. Sequencing service was provided by Agencourt, Bioscience Corporation (Beverley, MA, U.S.A.). DNA sequence was assembled using Vector NTI program (Invitrogen).

2.5 Bioinformatic techniques

Multiple amino acid sequence alignments were constructed using the programs: Bio Edit (Hall, 1999), CLC Free Workbench (Knusden *et al.*, 2005), Clustal W (Larkin *et al.*, 2007), and MUSCLE (Edgar, 2004). Neighbour-joining phylogenetic trees were inferred from multiple sequence alignments using the software PAUP 4 (Swofford, 2002) and were based on the Wheelan and Goldman (WAG) model of protein evolution (Whelan & Goldman, 2001). Prediction of signal peptides was done using the software Signal P 3.0 (Dyrlov Bendtsen *et al.*, 2004) employing the definitions for Gram negative bacteria. Prediction of protein structures was done using Muster (Wu & Zhang, 2008) applying template structures Class A NSAP from *Escherichia blattae* (Ishikawa *et al.*, 2000) and nucleotide pyrophosphatase/phosphodiesterase from *Xanthomonas axonpodis* (Zalatan *et al.*, 2006b). Protein structural modelling was performed using the software "DeepView/ SwissPDBviewer" (Guex & Peitsch, 1997). Promoter prediction was done using web-based tools BPROM (http://linux1.softberry.com/berry) and Visual Footprint (Munch *et al.*, 2005). Gene annotations were performed using "Artemis" (Rutherford *et al.*, 2000). Plasmid and construct drawings were prepared using XPlasMap v. 0.99 (http://www.iayork.com/XPlasMap).

2.6 Biochemical techniques

2.6.1 Preparation of periplasmic protein fraction

E. coli cells were grown overnight to stationary phase in 250 ml flasks in 50 ml batch cultures at 37°C. Cells were collected by centrifugation at 7250 g for 10 minutes, using an IEC 7685C rotor in an IEC 21000R centrifuge, washed twice with 0.01 M Tris-HCl pH 8 and concentrated to an OD_{600nm} of 3 by resuspension in 0.03 M Tris-HCl pH 8 with 20% sucrose. Then, 0.01 M EDTA and 10 µg/ml lysozyme were added. The cell suspension was stirred on ice for 20 minutes. Soluble periplasmic proteins were collected by pelleting the cell debris by centrifugation at 4050 X g, for 15 minutes using the centrifuge and rotor indicated above. The supernatant containing the periplasmic fraction was assayed for phosphatase activity.

2.6.2 Phosphatase activity assay

Enzyme activity was measured by monitoring the rate of the hydrolysis of *p*-nitrophenyl phosphate (PNPP) as previously described (Charles *et al.*, 1991). Buffers used for phosphohydrolase activity assay were as follows: 0.4 M sodium acetate/acetic acid pH 4-4.9, 0.4 M 2-(N-morpholino)-ethanesulfonic acid (MES), pH 5.4-6.7; 1 M 3-(N-morpholino) propanesulfonic acid (MOPS), pH 6.75; 0.4 M tris(hydroxymethyl)- aminoethane (TRIS), pH 6.9-9; 0.4 M 3-[cyclohexylamino]-2-hydroxyl-1-propane-sulfonic acid (CAPSO), pH 9.3-10.1. 5-Bromo-4-chloro-indolyl phosphate (BCIP) (40 µg/ml) was used as a phosphate hydrolysis indicator substrate for metagenomic library screens and PNPP 0.44 mg/ml was used as a substrate for organic phosphate hydrolysis. Briefly, clones were grown overnight at 37°C in 2 ml Luria Bertani (LB) broth with either 100 µg/ml ampicillin or 10 µg/ml tetracycline. Cultures were diluted 1:1 with 2 M tris-HCl, 1 M MOPS or 0.6 M MES buffer,

depending on the pH desired for the assay. The reaction mixture was incubated at 37°C for 10 minutes before the addition of the substrate, pNPP. The reaction was stopped after the desired time period by the addition of 0.33 ml 1 M KOH. For the activity assay in the periplasmic fraction, the 1 ml reaction mixture consisted of 400 µl of the extract, 100 µl PNPP and 500 µl of the appropriate buffer. The amount of *p*-nitrophenol produced in the reaction was quantified by measuring the absorbance at 420 nm and units of specific activity were determined using the following formula:

Specific activity U/OD600 nm = 1000 X ΔOD420 nm / (Δt (min)*OD600 nm), where Δt denotes the reaction incubation time. Assuming a molar extinction coefficient of 16000 $M^{-1}cm^{-1}$ for PNPP (Zhou & Zhang, 1999), U is equal to 0.062 nmol PNPP hydrolysed per minute at a cell OD of 1. For the activity in the periplasmic fraction, units of activity were determined using the following formula:

Activity (U/ml) = 1000 X ΔOD420nm / V (ml) X (Δt (min)), where V denotes the volume of enzyme preparation and Δt denotes the reaction incubation time.

2.6.3 Partial protein purification

Overnight cultures (100 ml) of pAR0045 and pAR0033 in autoinduction medium were harvested by centrifugation at 10,000 X g for 10 min at 4°C. To lyse the cells, for every 1 g of cell pellet, 5 ml of BugBuster Protein Extraction Reagent (Novagen) was added. The cell suspension was shaken for 20 min at room temperature followed by the removal of cell debris by centrifugation at 16,000 X g for 20 min at 4°C. PhoA$_{ACX6.71}$ and PhoN$_{ACX6.13}$ were then partially purified using the Ni-NTA His•Bind purification system (Novagen) following

the recommended protocol for batch purification under native conditions. Fractions were collected and analysed by SDS PAGE.

2.6.4 Protein determination

Protein concentration was assayed using the BioRad protein reagent with bovine serum albumin (BSA) as a protein standard. Reaction mixtures contained: 8 μl sample/standard, 792 μl deionized H_2O and 200 μl of BioRad reagent. The reaction mixture was incubated at room temperature for 5 min and the absorbance of the mixture was determined at 595 nm.

2.6.5 Detection of phosphatases by western blotting

To confirm expression of $PhoN_{ACX6.13}$ and $PhoA_{ACX6.71}$ proteins, detection of the His tag fused to the C-terminal portion of the proteins was performed. Strains were grown in autoinduction medium as described above and cell extracts were loaded onto a SDS polyacrylamide gel. Proteins on the gel were blotted onto PVDF membrane by cold wet transfer at 200 mA for 45 minutes. The extent of transfer was assessed by staining the post-transfer gel with Coomassie Brilliant Blue. Membranes were incubated at 4°C overnight in blocking buffer (15% skim milk in 1X TEN (20 mM Tris-HCl pH 8, 1 mM EDTA and 0.14 M NaCl)). The primary antibody, His Tag Monoclonal Antibody (Novagen), was added at a 1:2000 dilution in blocking buffer and membrane was incubated at room temperature for 2 hours. Membrane was washed in three times 1 X TEN for 10 minutes each wash cycle. Secondary antibody, Alexa Fluor 488 goat-anti-mouse IgG, was added at 1:3000 dilution in blocking buffer and the membrane was subsequently incubated for 2 hours at room

temperature. The membrane was washed three times in 1 X TEN buffer for 10 minutes and twice in deionized water for 10 minutes. Fluorescence detection was done using a Typhoon 9400 scanner.

Chapter 3: Isolation and characterization of metagenomic phosphatase genes

3.1 Isolation and sequencing of acid phosphatases $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$ and alkaline phosphatase $phoA_{ACX6.71}$

To identify novel phosphatase genes from complex communities, metagenomic libraries CX4 and CX6, respectively originating from pulp and municipal waste activated sludge and CX9, a soil-derived metagenomic library were screened *en masse* for phosphohydrolase activity using the chromogenic substrate BCIP. Prior to testing, it was determined that the library host strain, *E. coli* HB101 was not able to hydrolyze BCIP due to the lack of the appropriate phosphohydrolase for its hydrolysis in its genome. To increase the diversity of metagenomic phosphatases, screening was done at two different pHs, 7 and 5.5. A sufficient number of clones was tested from each library to cover at least three times the number of clones in the library. Accordingly, of approximately 15,000 clones screened from each library, 16 positive (blue) colonies were obtained from CX6 at pH 7, and 9 positive colonies at pH 5.5. Four positive colonies were obtained from CX4, two at pH 7 and two at pH 5.5. The phenotype conferred by the cosmids was confirmed by streak purifying the selected clones three times on LB medium supplemented with BCIP. No BCIP hydrolyzing cosmids were obtained from the CX9 library at either of the tested pHs after screening ~70,000 clones.

Cosmids were subsequently introduced by transformation into *E. coli* DH5α, and transformants were selected on LB medium containing tetracycline and screened on BCIP to confirm phosphatase activity. *E. coli* DH5α cannot hydrolyze BCIP due to the lack of the

appropriate phosphohydrolase in its genome. From restriction pattern analysis of the cosmids, three unique restriction pattern groups (A, B and C) were present. Cosmids from positive clones were isolated, assayed for phosphatase activity and subjected to restriction enzyme analysis to determine unique clones (Table 3-1). Cosmids belonging to group "A" originate from the CX4 library and demonstrate optimal phosphatase activity at acidic-neutral pH. A similar activity profile was demonstrated for group "B" cosmids that originated from the CX6 library, suggesting group A and group B cosmids both encode acid-neutral phosphatases. On the other hand, group C cosmids demonstrated optimal phosphatase activity at an alkaline pH, suggesting that they encode an alkaline phosphatase. Consistent with the restriction digest pattern, phosphatase activity conferred by representative clones followed a pattern with respect to their environmental source and optimal pH for activity. Interestingly, the pH used for library screening had no effect on the distribution of cosmids among the different restriction pattern groups which suggests that the phosphatases encoded by these cosmids have activity within a broad pH range. Based on activity profiles acid-neutral phosphatase-encoding cosmids pACX6.13 and pBCX4.10, respectively isolated from CX6 (pH 7) and CX4 (pH 5.5) were chosen along with alkaline phosphatase-encoding pACX6.71, (CX6 at pH 7) for further characterization. To identify metagenomic acid and alkaline phosphatases, the genes were sub-cloned as follows. Library cosmids pACX6.13, pACX6.71 and pBCX4.10 were each digested with BamHI, HindIII and EcoRI (individually) and ligated en-masse to plasmids pUC18 followed by transformation of *E. coli* DH5α, and selected with ampicillin in the presence of BCIP. This resulted in constructs pAR003 containing a 6.5kb BamHI fragment from cosmid pACX6.13, pAR004 containing a 3.5kb

BamHI fragment from pACX6.71 and pAR005 containing a 6.3kb BamHI-EcoRI fragment

from pBCX4.10 (Figure 3-1 a and b).

Table 3-1: Classification and phosphatase activity of BCIP-hydrolyzing metagenomic clones at various pHs

Clone designation	Restriction pattern group	Specific phosphatase activity at various pHs		
		5.5	6.75	8
pACX4.21	A	13	11.6	2.2
pACX4.26	A	10.9	8.5	1.6
pBCX4.10	A	5.89	22.5	1.6
pBCX4.11	A	2.08	6.4	1.5
pACX6.3	B	13.2	11.3	2.4
pACX6.13	B	14.9	8.5	2.1
pACX6.21	B	13.9	14	2.4
pACX6.87	B	12	11.6	4.2
pBCX6.22	B	6.35	7.3	2
pBCX6.26	B	5.79	24.8	2.1
pBCX6.31	B	5.73	24	2.2
pBCX6.35	B	7.14	9.3	3.5
pBCX6.37	B	2.25	10.9	2.2
pBCX6.38	B	5.46	6.7	2.3
pACX6.7	C	18.7	20.6	40
pACX6.29	C	21.9	19.6	66.7
pACX6.57	C	16.4	24.1	59.7
pACX6.62	C	16.4	22.7	79.6
pACX6.71	C	20.7	23.1	46.8
pACX6.80	C	19	23	79.5
pACX6.96	C	16.2	16.9	62.9
pACX6.101	C	19.8	18.4	58.8
pACX6.102	C	18.4	20	28.9
pACX6.107	C	25.2	23.4	100.1
pACX6.109	C	22.3	19.7	39.3
pACX6.111	C	18.7	22.9	28.4
pBCX6.6	C	3.19	13.3	137
pBCX6.9	C	2.98	13.9	76.2

Figure 3-1a: Strategy for subcloning acid and alkaline phosphatases from pACX6.13 and pACX6.71, respectively.

pUC18 (2686 bp)

HindIII
BamHI
EcoRI

bla

rep

BamHI — BamHI

ACX6.71 (3500 bp)

BamHI — BamHI

ACX6.13 (7500 bp)

Digest with BamHI and ligate
into pUC18.

pAR004 (6186 bp)

BamHI

bla

rep

lacZ

BamHI

ACX6.71

pAR003 (10186 bp)

BamHI

bla

lacZ

rep

lacZ

BamHI

ACX6.13
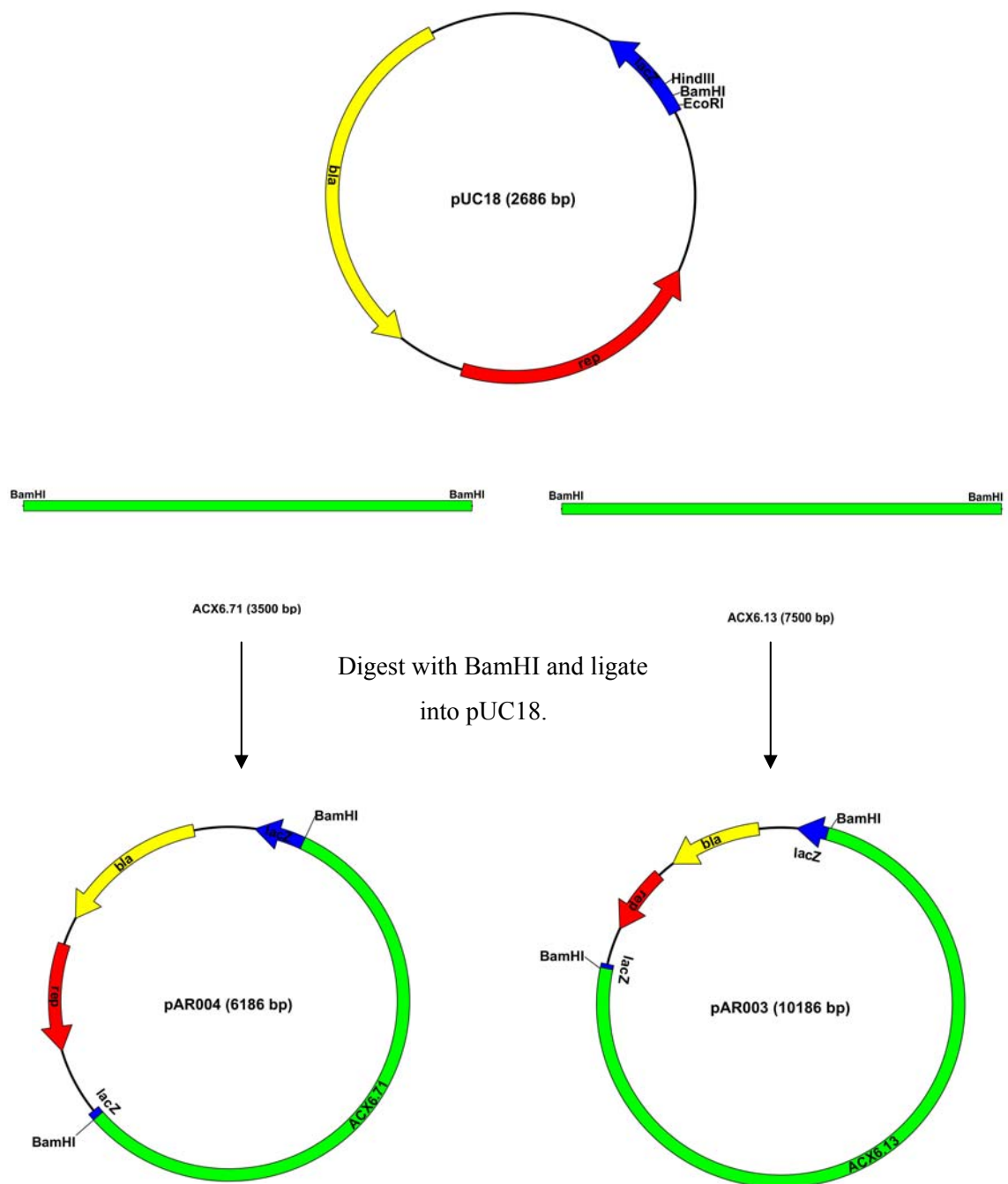
Figure 3-1b: Strategy for subcloning acid phosphatase from pBCX4.10.
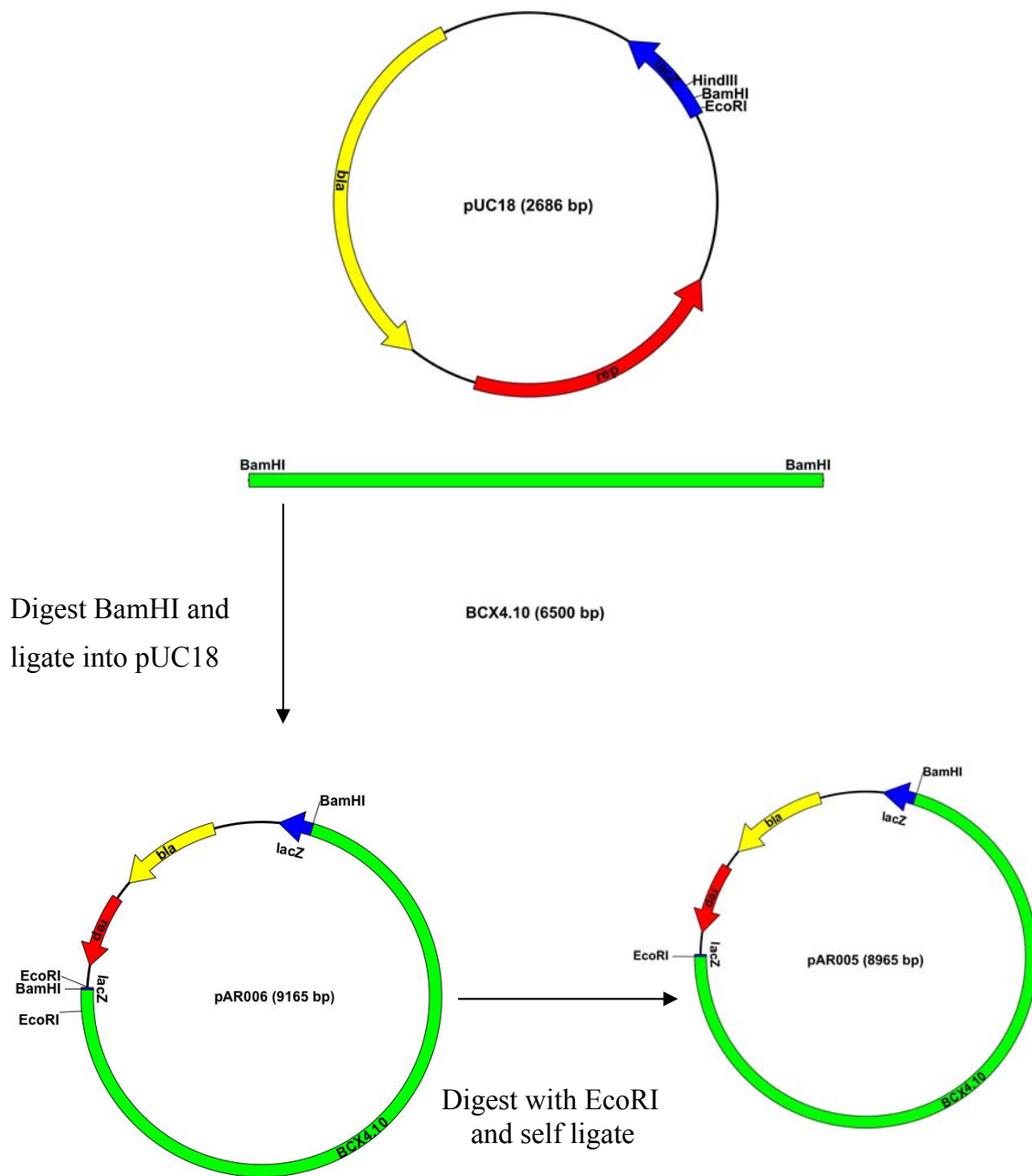
The ORFs responsible for phosphatase activity in pAR003, pAR004 and pAR005 were localized by *in vitro* transposon mutagenesis using EZ-Tn5 in which inserts were selected using kanamycin and screened for phosphatase activity using BCIP. For each construct, approximately 20 insertion clones that failed to utilize BCIP (indicating the presence of a transposon interrupting the phosphatase coding region) were chosen for DNA sequence analysis from the insertion site. About 10 adjacent transposon insertions were chosen for sequence assembly. Using bidirectional primer binding sites of the EZ-Tn5 transposon, sequence contigs of 1474 bp, 2533 bp and 1566 bp were assembled from constructs pAR003, pAR004 and pAR005, respectively. Open reading frames (ORFs) and ribosomal binding sites (RBS) were subsequently predicted for each contig. The DNA sequences were deposited in GenBank under accession numbers DQ303407, DQ303408 and DQ303409 (http://www.ncbi.nlm.nih.gov/Genbank/index.html)

3.2 Characterization of phosphatases

3.2.1 *phoN*$_{ACX6.13}$ and *phoN*$_{BCX4.10}$ are new members of the NSAP family

DNA sequence analysis revealed the ORFs in pAR003 (from pACX6.13, municipal waste) and pAR005 (from pBCX4.10, pulp waste) responsible for the phosphohydrolase activity are predicted to encode proteins belonging to the acid phosphatase family (smart00014) based on the Conserved Domain Database (CDD) from NCBI (Marchler-Bauer *et al.*, 2005). These ORFs were thus designated *phoN*$_{ACX6.13}$ and *phoN*$_{BCX4.10}$, respectively. The two ORFs were predicted to have Ribosome Binding Sites (RBSs) upstream of their translation start sites and contained signature sequence domains, KXXXXXXRP (Domain I),

PSGH (Domain II) and SRXXXXXHXXXD (Domain III) (Figures 3-2 and 3-3), typically found in a family of glucose-6-phosphatases, NSAPs and lipid phosphatases (Stukey & Carman, 1997). Upstream of each NSAP, lie genes transcribed in the opposite direction. The N-terminal portion of a putative thioesterase superfamily ORF was located 104 bp upstream of $phoN_{ACX6.13}$ (Figure 3-2) and the N-terminal portion of a putative $uvrD$ ORF was located 144 bp upstream of $phoN_{BCX4.10}$ (Figure 3-3). To determine the occurrence of promoters, the sequences upstream of genes $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$ were analysed for regulatory protein binding sites using BPROM (http://linux1.softberry.com) and Visual Footprint (Munch *et al.*, 2005). Putative *E. coli* class σ_{70} -10 and -35 recognition sites were predicted upstream of a transcription start site located 71 bp upstream of the translation start site of $phoN_{BCX4.10}$ (Figure 3-3). In the case of $phoN_{ACX6.13}$, a putative *E. coli* class σ_{70} -10 recognition sequence was identified 31 bp upstream of the translation start site, but no -35 recognition sequence was identified. However, recognition sequences for *E. coli* CytR and GlnG regulators were identified 104 bp and 51 bp, respectively upstream of the $phoN_{ACX6.13}$ translation start site (Figure 3-2). GlnG is a response regulator which affects the transcription of certain nitrogen metabolism genes (Magasanik, 1989). CytR is a repressor of CRP-directed transcription in the CytR regulon, affecting expression of genes such as *udpP* (uridine phosphorylase) and *tsx* (nucleoside uptake protein), (Jorgensen *et al.*, 1998). A putative PhoB binding site ("pho" box) was identified upstream of $phoN_{ACX6.13}$.

An NCBI BlastX search revealed acid phosphatases $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$, respectively share 59% and 62% amino acid identity with the acid phosphatase of *β-proteobacterium Ralstonia eutropha* H16 (accession no YP_840758.1) and a 67% amino acid

identity with each other. The two NSAPs also share homology with acid phosphatases from the environmental genome database generated from the Global Ocean Sample (GOS) expedition. The gene $phoN_{ACX6.13}$ shares 48% amino acid identity with acid phosphatase JCVI_PEP_1112727259730 from ocean metagenome sample GS117b near St. Anne Island and 41% identity with acid phosphatase JCVI_PEP_1113373111550 from sludge metagenome Nine Springs Waste Water Treatment Plant, U.S.A. On the other hand, acid phosphatase *phoNBCX4.10* has an amino acid identity of 49% with Class A NSAP JCVI_PEP_1105081317249 from the ocean metagenome from the Saragasso Sea and 44% identity with acid phosphatase JCVI_PEP_1113373111550 from the metagenome of sludge from Nine Spring Waste Water Treatment Plant, U.S.A

To assess whether $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10}$ are indeed secreted non-specific acid phosphatases, the leader peptide sequence was first predicted using the online tool, SignalP3.0 Server (Bendtsen *et al.*, 2004). Analysis of the first 70 N-terminal amino acids of $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10}$ indicated the probability of these sequences containing a signal peptide was 0.996 and 1.0, respectively, with the signal peptide cleavage sites in $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10}$ being most likely Ala21-Cys22 and Ala29-Pro30, respectively.

In order to determine the pH range for activity of $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10}$, periplasmic extracts of *E. coli* DH5α expressing pAR003 and pAR005 were assayed for activity at a wide pH range (Figure 3-4). The two proteins showed marked differences in phosphatase activity with respect to pH range. The activity of $PhoN_{ACX6.13}$ peaked at pH 4.4 while $PhoN_{BCX4.10}$ was optimally functional at a broader range of pH 5.4-7.9. The above results suggest that both $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$ encode secreted acid-neutral

phosphatases. To assess the phylogenetic relationship between PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ and other bacterial NSAPs a neighbour-joining phylogenetic tree (Figure 3-5), was constructed using the conserved catalytic domains of NSAPs from a number of demonstrated and predicted bacterial acid phosphatases. NSAPs were divided into three major clusters plus a fourth cluster that included only the acid phosphatase from *Zymomonas mobilis*. Cluster I contains primarily NSAPs of enteric bacteria belonging to the γ-*Proteobacteria* class. Most of the members of this cluster have been experimentally determined. Conversely, about half of the orthologues of Cluster II are non-enteric γ-*Proteobacteria* while the remaining orthologues are represented by α-, β-, δ- *Proteobacteria* and *Cyanobacterial* classes. The NSAPs derived from uncultured bacteria from activated sludge, PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ along with the putative acid phosphatase from *P. fluorescens* H16, form Cluster III. It should be noted that only with the exception of the acid phosphatase from *Zymomonas mobilis*, all of the experimentally determined NSAPs were isolated from enteric γ-*Proteobacteria* suggesting the occurrence of two novel clusters of undescribed NSAPs including the two NSAPs originating from uncultured bacteria.

Figure 3-2: Sequence of *phoN*_{ACX6.13} and flanking region in pAR003.

Shaded sequence corresponds to predicted RBS. ⇑ denotes the predicted site of signal peptide cleavage. Amino acid sequences of *phoN*_{ACX6.13} and thioesterase superfamily protein ORFs are indicated. The latter is transcribed on the antisense strand (not shown). Underlined amino acid sequences correspond to Class A NSAP signature domains I, II and III. Boxed promoter recognition sites are denoted as follows:"pho box" (purple) *E. coli* class σ_{70} -10 (red), CytR (yellow) and GlnG (blue).

TGACTTTAACCACCATCGTGCCGCGCTTACGCTCGGAATACACCTCGACATCGACGGTGATGGAGGTGCGGCCGACGCGCTTGATCG
 V K V V M T G R K R E S Y V E V D V T I S T R G V R K I

TCGCGTAGAAGCTCACCACGTCGCCGATCTGCACGGGTTGCTTAAACAAAAACTGATTCACTGCCACCGTCGCGATGCGCCCACGCG
 T A Y F S V V D G I Q V P Q K F L F Q N V A V T A I R G R

CGCGGCGCACCGCAGGCACGCTACCCGCCAAGTCCACTTGCGACATGATCCAACCGCCGAAGATGTCGCCGTTGGCGTTGGCGTCGG
 A R R V A P V S G A L D V Q S M I W G G F I D G N A N A D

<mark style="background:purple">CTGGCATCGGAATGACCC</mark>GCAACGCAAGGTATTTGTCGGTGGGGAGTTGCAGCATGGCATGGTCTCGAAACA<mark style="background:yellow">TGGAAACTGGAT</mark>GAT
 A P M P I V R L A L Y K D T P L Q L M A H D R F M

                    ← Thioesterase superfamily protein ORF

ACGGAAATACGGAGCATTGACACAACGGGCGCGCCGGG<mark style="background:blue">TGCTGG</mark>CGACTCGT<mark style="background:red">TATCCT</mark>TGTGCTCAATCGACCGAGCT<mark style="background:gray">AGGAGGA</mark>CA

CCATGTTTTCGCCACGCCAACTCTCATTGAGTATCGCCGCAACTGTCTCGCTCACGCTGTGCGCTTGCAGCACGGCCCCGAAGCCGC
  M F S P R Q L S L S I A A T V S L T L C A S T A P K P P

 phoN$_{ACX6.13}$ ORF →   ⇑

CCACGACGGTCGCGGAGGTCGGCGAATTTCGTGCGGGGTCGGGTTACGTCAATGGTTATCTTGATCGCAAATCGCTGCCGAATAGTT
  T T V A E V G E F R A G S G Y V N G Y L D R K S L P N S L

TGGCGTTGTTGCCGCCTCCGCCCGCAGAGGGCTCTGCGCGCGCTGCCTCCGATCTGGAAACGCATCGCACCACGAGAAGTCTGCGCG
  A L L P P P P A E G S A R A A S D L E T H R T T R S L R D

ACACGCCGCGCTGGACGCTGGCGACCCGCGACGCCAACCTGAAATTCCCAGCCGCCGCCGATACGTTTTCGTGTGCGCTCGACGCTC
  T P R W T L A T R D A N L K F P A A A D T F S C A L D A P

CGATCAGCGCGGAATCCACTCCGCACCTGAACATGCTGCTGCGCCGCACCTTGCTCGACGCGGGGCTCTCGACGTATGCCGCGAAAG
  I S A E S T P H L N M L L R R T L L D A G L S T Y A A <u>K D</u>

Domain I
ATCGCTATAACCGTCCAAGACCATTCGTCGCGCTAAACGAAACCATGTGTACTCCCGCCGAGGACGCAACGCTGCGCAAAGACGGCT
 <u>R Y N R P</u> R P F V A L N E T M C T P A E D A T L R K D G S

        Domain II
CGTATCCGTCGGGCCACGCCGCGCTGGGGTGGGCGTGGGCTCTGGTGCTCGCTGAGATTGCGCCGGAGCGCGCCGATGTGTTGATTG
 Y <u>P S G H</u> A A L G W A W A L V L A E I A P E R A D V L I A

                    Domain III
CACGTGGGTTGGCGTTTGGTCAGAGCCGCGTGATTTGCGGAGTGCACTGGCAAAGCGACGTTGAAGCGGGACGGACGATGGGCGCGG
 R G L A F G Q <u>S R V I C G V H W Q S D</u> V E A G R T M G A A

CAGCGGTGGCGCGGCTGCATGCGAACGAGGTGTTTCGCGCGCAACTCGCGGCGGCGAAAGAAGAGGTCCAGAGCGCAAGAGCGGCGG
 A V A R L H A N E V F R A Q L A A A K E E V Q S A R A A G

```
GCAAGAAGCCCGCGCTCGATTGCAAAACCGAGGCCGCAGCGCTCGGAAGCCAGTAGGCAAATGGTGGCCGCGGTGCGGCCACCGTGC
  K  K  P  A  L  D  C  K  T  E  A  A  A  L  G  S  Q  *


AACGGCGAATTAGCCGAGTTGCGCCAGCATGTTTTCTGCGAACGAGCAGTTCGCGGCGCCGACGCCTTCAACGTTGACTTGCTTCAC


TACGCCGTTGTCGACGAGAAGCGAGAAGCGCTGGCAGCGAACGCCCATGCCCTTGGCAACTAGATCAAGCTCAAGGCCGAGC
```

Figure 3-3: Sequence of *phoN*_{BCX4.10} and its flanking region in pAR005.

Shaded sequence corresponds to the predicted RBS. ⇑ denotes the predicted site of signal peptide cleavage. Amino acid sequences of *phoN*_{BCX4.10} and *uvrD* ORFs are indicated. The latter is transcribed on the antisense strand (not shown). Underlined amino acid sequences correspond to Class A NSAP signature domains I, II and III. *E. coli* class σ_{70} -10 and -35 recognition sequences are boxed in red and the *phoN*_{BCX4.10} transcription start site is denoted by ↪

TTGAGGCCCAGCAACTGGCCGTCCTCGCGCAGCAGGCGCACGCCCAGCGCGTGGAAGGTGCAGATCATCACTTTCTTGGCGTCGCGC
　L　G　L　L　Q　G　D　E　R　L　L　R　V　G　L　A　H　F　T　C　I　M　V　K　K　A　D　R

CCCACCAGCTGGCTGGCGCGCTCGCGCATCTCGGCGGCCGCCTTGTTGGTGAAGGTGATGGCCGCAATGCGCTCAGCCGCCAAACCG
　G　V　L　Q　S　A　R　E　R　M　E　A　A　A　K　N　T　F　T　I　A　A　I　R　E　A　A　L　G

GCCTTTATCAGCCGCCCGATCTTGTGCGTGATCACCCGCGTCTTGCCCGAACCGGCGCCCGCCAGCACCAGACAGGGGCCGCCCAGG
　A　K　I　L　R　G　I　K　H　T　I　V　R　T　K　G　S　G　A　G　A　L　V　L　C　P　G　G　L

TGGTTCACCGCTTCGAGCTGGGCAAGGTTCAGGCCGGGACTGGCGGAGGACGACATGGGGCGGGAGGGCG`TTGCAC`ACCAAGATGGA
　H　N　V　A　E　L　Q　A　L　N　L　G　P　S　A　S　S　S　M　　　　　　　　　　　　-35

<div align="center">← uvrD ORF</div>

　　　　　　　　　　　　　　　　　　　　↱
GCG`GCGAATGAT`ACCGGTGGGGTTTGTCGCACCGGCTACAGCCGCCGTGCAGCCAAGGCGTCTAAGCTGAGCTGGCCCAAGTTCCCT
　　-10

CGCCCCCACCGACCC`AGGAGA`CCTGCCATGACCTCGTTTGCCTTCCGTCTGTTGCCGCTGCTCGCCGCCGCCGCCCTGCTCACCGCC
　　　　　　　　　　　　　　　　M　T　S　F　A　F　R　L　L　P　L　L　A　A　A　A　L　L　T　A

<div align="center">phoN_{BCX4.10} ORF →</div>

TGCGCCACCCGGACCGGCGCGGTCGCGCCGCCACCGACCGATCCGGCCCAGGTCGGCGAGATCCGTGCTGGCTCGGGCATCCTCACG
　C　A　T　R　T　G　A　V　A　P　P　P　T　D　P　A　Q　V　G　E　I　R　A　G　S　G　I　L　T
　　　　　　　　　　　　　⇑

GGCTACCTGGCACGCAACCAGCTGCCCGACAGCCTGGCGCTGTTGGGCGCCCCCCCTGCGCTGGGCTCGCCGCAGGCCGCAACGGAC
　G　Y　L　A　R　N　Q　L　P　D　S　L　A　L　L　G　A　P　P　A　L　G　S　P　Q　A　A　T　D

GAGGCGGTGTTCCGCCAGACCCGTTCGCTCGAAGGCGGGCCGCGCTGGAAACTCGCGGCGAGGGACGCCAACCTGCAGTTCCCTGAA
　E　A　V　F　R　Q　T　R　S　L　E　G　G　P　R　W　K　L　A　A　R　D　A　N　L　Q　F　P　E

GCGGCCACCGCGTTCTCCTGCGCCATGGGCATGGACATGGACGCCAACCGCACCCCCCACCTGAACATGCTGCTGCGCCGCACCTTG
　A　A　T　A　F　S　C　A　M　G　M　D　M　D　A　N　R　T　P　H　L　N　M　L　L　R　R　T　L

<div align="center">Domain I</div>
ACCGACGCAGGCCTGTCCACCTACCGGGCGAAGGACCACTACAAGCGCACGCGGCCGTTTGTCGAACTCAAGACCGGCACCTGCACC
　T　D　A　G　L　S　T　Y　R　A　<u>K　D　H　Y　K　R　T　R　P</u>　F　V　E　L　K　T　G　T　C　T

<div align="center">Domain II</div>
CCGGACGAGGAAGCCCGGCTGGCCAAGGACGGCTCCTATCCGTCGGGGCATTCGGCGCTGGGTTGGGCCTGGGCGCTGGTGCTGACC
　P　D　E　E　A　R　L　A　K　D　G　S　Y　<u>P　S　G　H</u>　S　A　L　G　W　A　W　A　L　V　L　T

<div align="center">Domain III</div>
GAACTGGCGCCCGAGCGCGCCGACACCCTGCTCGCGCGTGGGCACGCCTTCGGCCAGAGCCGCGTGATCTGCGGGGTGCACTGGCAA
　E　L　A　P　E　R　A　D　T　L　L　A　R　G　H　A　F　G　Q　<u>S　R　V　I　C　G　V　H　W　Q</u>

<div align="center">79</div>

AGTGATGTGGACCAGGGACGCATCATGGGGGCGGCCGCGGTGGCGCGCGGCTGCATGCCGATGCCGCATTCCAGGCCCAGCTGCAGGCG
<u>S D</u> V D Q G R I M G A A A V A R L H A D A A F Q A Q L Q A

GCGCGCGCCGAAATCGCCGCGCAGCGGGCTGCCGGCGCCAGCCCGCACGCCGACTGTCAGGCCGAGGCGACGGCGCTCATGCCCCGG
A R A E I A A Q R A A G A S P H A D C Q A E A T A L M P R

TAGCGCGCCGGCGGCCGGGCCACGGGCCCGGCACTTGAGACAATCGCCCCGTGCTCAACATCCTGCTCGTCACCTTTCCCTTCTTTG
*

CCCTCGTGCTCGCCGGCTTTGTGGCCGCGCACCGGCGCATGCTGCCGCTCGAAGCGATTCCTGGGCTCAACGGATTCGTGCTGTTCT

TCGCGCTGCCCTGCATGCTCTACCGCTTTGGCTCCACCACGCCGATCGCGCAACTGCTCGACGCCTCGGTGGCCGGCGTCTACCTGC

Figure 3-4: Phosphatase activity of periplasmic extracts of *E. coli* DH5α expressing *phoN*$_{\text{ACX6.13}}$ (pAR003) and *phoN*$_{\text{BCX4.10}}$ (pAR005) at various pHs.

Assays were done in triplicate. The apparent broad pH optimum for *phoN*$_{\text{BCX4.10}}$ may be due to substrate depletion resulting in activity being outside the linear range.
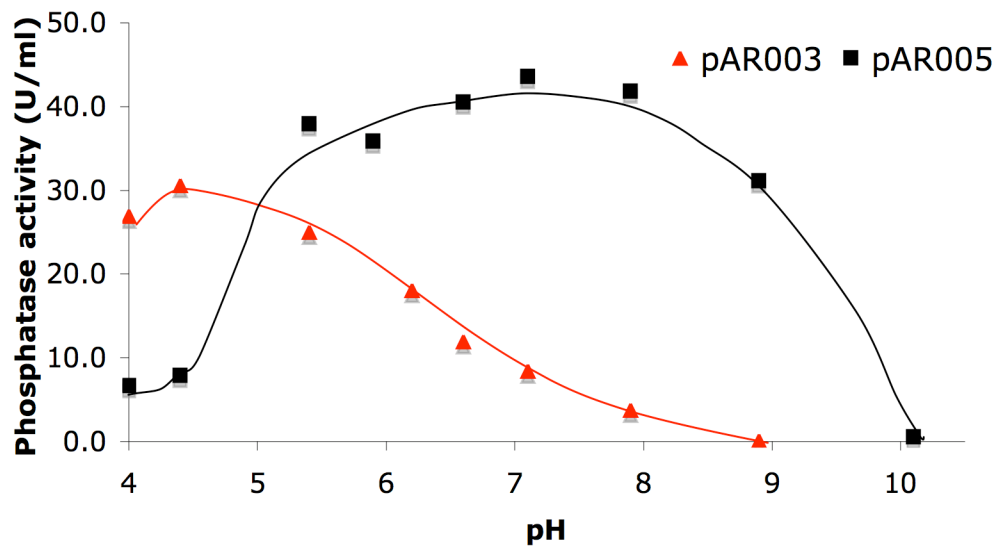
Figure 3-5: Neighbour-joining tree showing non-specific acid phosphatase proteins.

Pfl-2, putative acid phosphatase 2, *Pseudomonas fluorescens* Pf-5; *Sty*-acphos, acid phosphatase, *Salmonella typhimurium*; *Sty*-PhoN, nonspecific acid phosphatase precursor, *Salmonella enterica* subsp. enterica serovar Paratyphi A str. ATCC 9150; *Pin*, acid phosphatase (PiACP) *Prevotella intermedia*; *Ebl,* acid phosphatase, *Escherichia blattae*; *Rpl,* acid phosphatase, *Raoultella planticola*; *Eae*, acid phosphatase, *Enterobacter aerogenes*; *Pst,* PhoN acid phosphatase, *Providencia stuartii*; *Mmo*, acid phosphatase, *Morganella morganii*; *Zmo*, acid phosphatase *Zymomonas mobilis*; *Mca*, acid phosphatase, *Methylococcus capsulatus* str. Bath; *Pma*-MIT 9313, PA-phosphatase-like phosphoesterase *Prochlorococcus marinus* str. MIT 9313; *Dps*-2, acid phosphatase 2, *Desulfotalea psychrophila* LSv54; *Dps*-1, acid phosphatase 1 *Desulfotalea psychrophila*; LSv54; *Pfl*-1, putative acid phosphatase 1 *Pseudomonas fluorescens* Pf-5; *Reu*, acid phosphatase, *Ralstonia eutropha* JMP134; *Xca,* acid phosphatase *Xanthomonas campestris* pv. vesicatoria str. 85-10; *Xax*, phosphatase precursor, *Xanthomonas axonopodis* pv. citri str. 306; *Xor*, phosphatase precursor, *Xanthomonas oryzae* pv. oryzae KACC10331; Boxed: *phoN*$_{ACX6.13}$- class A non-specific acid phosphatase isolated from the metagenome of municipal waste activated sludge; *phoN*$_{BCX4.10}$-class A nonspecific acid phosphatase isolated from the metagenome of pulp waste activated sludge. ★ denotes isolates for which acid phosphatase activity was determined experimentally.

Cluster I

Cluster III

Cluster IV

Cluster II
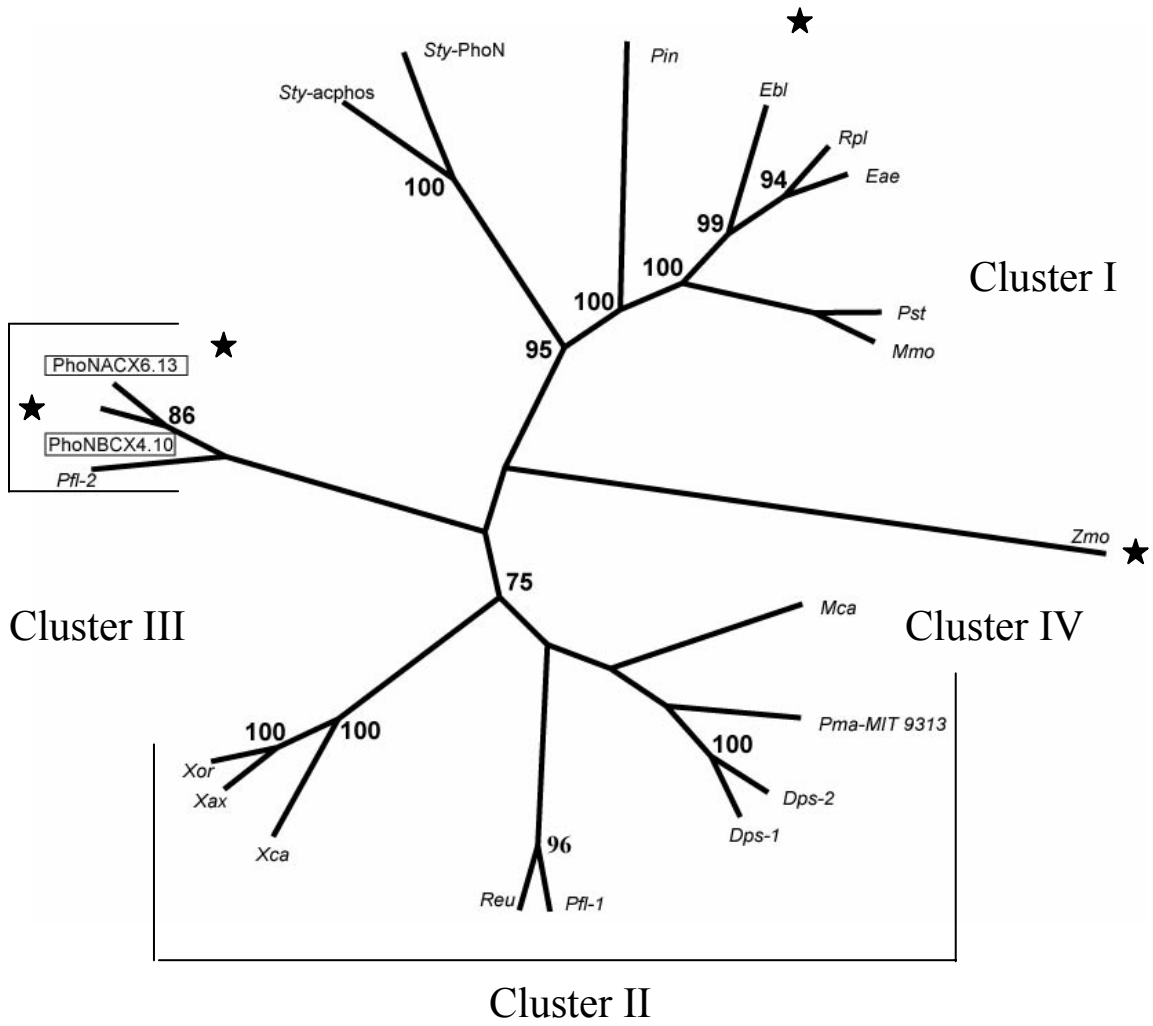
To examine the diversity of bacterial NSAPs with respect to primary structure, the pattern of conservation in the 79 residues comprising the conserved region spanning the three signature motif domains was compared using the protein sequences of demonstrated and predicted NSAPs from bacterial isolates (Figure 3-6a) and those of predicted NSAPs from the CAMERA database for sequences derived from environmental samples (Seshadri *et al.*, 2007) (Figure 3-6b). Upon examination of the consensus sequences in NSAPs from isolates and environmental samples (Figure 3-6c) only 48% of the residues, a majority of which comprise the invariant signature motifs, were identical or similar in both sets of consensus sequences. Of the remaining residues, a majority (29%) were conserved only in the predicted NSAPs from environmental samples compared to 5% conserved only in NSAPs from bacterial isolates while 2.5% of the residues were not similar but were conserved in both types of NSAPs. The differential residue conservation in NSAPs derived from environmental samples compared to bacterial isolates suggests functional diversity which may be linked to differences in enzyme kinetics, substrate specificity or even regulation mechanisms among NSAPs.

Figure 3-6a: Multiple amino acid sequence alignment of class A NSAPs from bacterial isolates in comparison with PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$.

GenBank accession numbers are indicated to the left of the strain name. Conserved residues are highlighted and include those (in bold letters) that make up the signature sequence motifs. Class A NSAP signature sequence comprising Domain I, Domain II and Domain III are marked with ●, ■ and ▲, respectively. Consensus sequence residues are denoted as follows: χ, polar amino acids DEKRNQ; Ψ, non polar amino acids ILVMFY; ν, acidic amino acids DE; σ, small amino acids GAST; ., no particular amino acid.

```
gi|11610581|  Raoultella planticola            KEKYMRIRPFAFYGVSTCNTTEQDKISKNGSYPSGHTSIGWATALVLAEINPQRQNEILKRGYELGESRVICGYHWQSD
gi|11610579|  Enterobacter aerogenes           KEKYMRIRPFAFYGVSTCNTTEQDKISKNGSYPSGHTSIGWATALVLAEINPQRQNEILKRGYELGESRVICGYHWQSD
gi|6009517|   Escherichia blattae              KDHYMRIRPFAFYGVSTCNTTEQDKISKNGSYPSGHTSIGWATALVLAEINPQRQNEILKRGYELGQSRVICGYHWQSD
gi|44464|     Morganella morganii              KEHYMRIRPFAFYGTETCNTKDQKKISTNGSYPSGHTSIGWATALVLAEVNPANQDAIIERGYQLGQSRVICGYHWQSD
gi|45862|     Providencia stuartii             KEKYMRIRPFAFYGVATCNTKDQDKISKNGSYPSGHTAIGWASSALVLSEINPENQDKILKRGYELGQSRVICGYHWQSD
gi|3599684|   Prevotella intermedia            KDHYMRVRPFAFYNEMTCNPEQQELSTNGSYPSGHTAIGWATALVLSEINIDRQNEILERGYQMGQSRVICGYHWQSD
gi|56416115|  Salmonella enterica ATCC 9150    KKYYMRTRPFVLFNHSTCRPEDENTIRKDGSYPSGHTAYSTLLALVISQARPERAQELARRGWEFGQSRVICGAHWQSD
gi|16762998|  Salmonella enterica CT18         KKYYMRTRPFVLFNHSTCRPEDENTIRKDGSYPSGHTAYSTLLALVISQARPERAQELARRGWEFGQSRVICGAHWQSD
gi|78050038|  Xanthomonas campestris           KARYQRPRPFMRNAQPMCTPADDAAIRKNGSYPSGHTAIGWTWGLIISEIAPAHRDALLARGRAFGDSRLVCNVHWQSD
gi|21245081|  Xanthomonas axonpodis            KARYQRPRPFMRNAQPMCTPDDAAIRKNGSYPSGHTAIGWTWGLIISEIAPAHRDALLARGRAFGDSRLVCNVHWQSD
gi|58584256|  Xanthomonas oryzae               KARYQRPRPFMRNAQPMCTPDDAAIRKNGSYPSGHTAIGWTWGLIISEIAPAHRDALLARGRAFGDSRLVCNVHWQSD
gi|53803760|  Methylococcus capsulatus         RNYYRRPRPFMSNQEPICTPAEQARLEENGSYPSAHGAIGMTWALILAGLSPERAAEILARGRAFGLSRVVCNTHWYSD
gi|51246217|  Desulfotalea psychrophila1       KNHYRRTRPFVLNGEPTCTPTKEAHMKKSGSFPSGHTAIGWAWALIIVEVAPEQTDAILARGWAFGQSRIICNVHWQSD
gi|51246207|  Desulfotalea psychrophila2       KNYRRTRPFVVNGEPVCTPQQEEQIKKSGSYPSGHTSIGWAWTLIVEVSPEQTDAILARGWAFGQSRIVCNAHWQSD
gi|124022056| P marinus MIT 9303               KNNYNRRPFLVNKKMICSPQDTAELTKDPSYPSGHTAIGWAWALIISEMVPERRDQILARGLAFGESRHLCNVHWYSD
gi|33863836|  P marinus MIT 9313               KNNYNRRPFLVNKKMICSPQDTAELTKDPSYPSGHTAIGWAWALIISEMVPERRDQIIARGLAFGESRHLCNVHWYSD
gi|73538365|  Ralstonia eutropha JMP134        KDRYNRTRPFVATSTSSCTPADEAKIAKDGSYPSGHSAAGWAWALVITEIAPDRADAIIMRGRAFGQSRIACNVHWQSD
gi|70728379|  P fluorescens Pf-5 2             KKRYQRTRPFVYNQQPSCTPEDEAKLAADGSYPSGHSAIGWAWALLLSELSPEHGNALWARGRAYSQSRMVCNVHWYSD
gi|83583480|  PhoN_ACX6.13                     KDRYNRPRPFVALNETMCTPAEDATIRKDGSYPSGHAALGWAVIAEIAPERADVIIARGLAFGQSRVICGVHWQSD
gi|83583484|  PhoN_BCX4.10                     KDHYKRTRPFVELKTGTCTPDEEARIAKDGSYPSGHSALGWAWALVITEIAPERADTILARGHAFGQSRVICGVHWQSD
Consensus                                      KYχ.Y.R.RPF.......C...χχ..L.KχGSYPSGHχOAψψGWO.oLψL.Eψ.PχΧ.χ.ψψ.RG..ψGχSRψψχC.ψHWQSD
```

Figure 3-6b: Multiple amino acid sequence alignment of class A NSAPs from various environmental samples and metagenomic DNA derived from pulp and municipal waste activated sludge.

GenBank accession numbers are indicated to the left of the strain name. Conserved residues are highlighted and include those (in bold letters) that make up the signature sequence motifs. Class A NSAP signature sequence comprising Domain I, Domain II and Domain III are marked with ●, ■ and ▲, respectively. Consensus sequence residues are denoted as follows: χ, polar amino acids DEKRNQ; Ψ, non polar amino acids ILVMFY; ν, acidic amino acids DE; σ, small amino acids GAST; . , no particular amino acid.

```
JCVI_PEP_1112708827952  92  KNNYQRPRPFMVNGKPICTPDEETLLRGDGSYPSGHTAAGWAWALVLSELAPERRDAILARGIEYGKSRYICNVHWLSD    8
JCVI_PEP_1105081317249  70  KNNYQRPRPFMVNGKPICTPDEETLLRGDGSYPSGHTAAGWAWALVLSELAPERRDAILARGIEYGKSRYICNVHWLSD   54
JCVI_PEP_1105133616393 170  KNNYQRPRPFMVNGKPICTPDEETLLRGDGSYPSGHTAAGWAWALVLSELAPERRDAILARGIEYGKSRYICNVHWLSD    8
JCVI_PEP_1105126720529  70  KINYQRPRPFMINGKPICTPDEETLLRGDGSYPSGHTAAGWAWALVLSELAPERRDAILARGIEYGKSRYICNVHWLSD   30
JCVI_PEP_1112708436832 113  KNHYQRPRPFMVNGEAICTPEETLLRSDGSYPSGHTAAGWAWALVLSEIAPDRRDLLARGMEYGKSRYICNVHWLSD     44
JCVI_PEP_1112727259730 156  KKAHQRPRPFLINGEAICTPDDRELLETDGSYPSGHSAIGWGWALLSQLAPDKAEALLARGRAYARSRMVCNVHWMTD    54
JCVI_PEP_1105084320587 153  KKAHQRPRPFLINGEAICTPDDRELLETDGSYPSGHSAIGWGWALLSQLAPERAEALLARGRDYTRSRMVCNVHWMSD    55
JCVI_PEP_1105112589987 128  KKAYQRSRPFLVNNQAICSPEERTDLETDGSYPSGHSAVGWGWALILSQLAPERAEAIISRGRDYARSRMVCNVHWMSD  55
JCVI_PEP_1105143789097 137  KRAYQRQRPFHINGGEVCTPDDQEMLETDGSYPSGHSAIGWGWTLTLSQLAPERAEALITRGRAYVQSRMVCNVHWLSD 102
JCVI_PEP_1112716360714 156  KNVYQRPRPFLINGEEICTPGELELLETDGSYPSGHSAVGWGWALVLSQLAPERAEQILARGRAYAQSRIVCNVHWKSD   17
JCVI_PEP_1105140234991  56  KLRYQRARPFVENGAPTCTPEDEADLRGNGSYPSGHTAIGWLWALVLTQAAPERTDALLARGRNFGESRLVCNVHWQSD  88
JCVI_PEP_1113954913676 184  KLRYQRARPFVENGAPTCTPEDEADLRGNGSYPSGHTAIGWLWALVITQAAPERTDALLARGRNFGESRLVCNVHWQSD  51
JCVI_PEP_1113373111550 143  KLRYQRARPFVENGAPTCTPEDEADLRGNGSYPSGHTAIGWLWALVLTQAAPERTDALLARGRNFGESRLVCHVHWQSD  60
JCVI_PEP_1113953986902 148  KDFFSRARPYIANPGVHPSVPLNPK-DGYDSYPSGHSTFGYMTAILLAQMVPEKRDAVFARGREYGENRVVDGVHYPSD  89
JCVI_PEP_1105133809433  68  KDRYNRPRPFVALNETMCTPAEDATLRKDGSYPSGHAALGWAWALVLAEIAPERADVLIARGLAFGQSRVICGVHWQSD  38
phoNACX6.13            143  KDHYKRTRPFVELKTGTCTPDEEARLAKDGSYPSGHSALGWAWALVITELAPERADTLLARGHAFGQSRVICGVHWQSD  55
phoNBCX4.10            146                                                                                          55
Consensus                   Kψ.YQR.RPFψ.NG..ψCTPvvχ..LχoDGSYPSGHσ A.GW.WALVLOχψAPvR.vAGLARG.χψoχSRψψCNVHW.SD
                            ●●●●●●●●●●●●●         ■■■■■■                                        ▲▲▲▲▲▲▲▲▲▲▲▲
```

89

Figure 3-6c: Comparative alignment of consensus sequences of conserved regions in class A NSAPs from cultured and uncultured bacteria.

Residues are denoted as follows: χ, polar amino acids DEKRNQ; Ψ, non polar amino acids ILVMFY; ν, acidic amino acids DE; σ, small amino acids GAST; ., no particular amino acid.

Kχ.**Y**.R.**RPF**........C...χχ...L.KχχGSY**PSGH**σA$^\psi$GWσ.σLμL.Eψ.Pχχ.χ.ψΨ.RG..$^\psi$Gχ**SR**$^{\psi\psi}$C.$^\psi$**HW**QS   cultured

Kχ.YQR.**RPF**Ψ.NG..$^\psi$CTPννχ..Lχσ DGSY**PSGH**σA.GW.WALVLσχ^ψAPνR.νΑσLARG.$\chi^{\psi\sigma}\chi$**SR**$^{\psi\psi}$CNVHW.**SD**   uncultured

To study the molecular structure of these proteins, models of PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ were predicted (Figures 3-7 and 3-8). Both proteins have similar secondary structure composed of 12 α-helices and 2 β-sheets oriented such that they enclose the residues involved in essential catalytic roles and maintenance of stability. The conserved residues comprising signature domains I, II, III (indicated in blue, green and red, respectively), were shown in similar proteins to be part of the active site (Ishikawa *et al.*, 2000). The conserved histidine of domain III, His 218 (His 221 in PhoN$_{BCX4.10}$), is predicted to target the substrate's phosphoryl group to produce a phosphoenzyme intermediate, thus being essential for catalytic activity. The side chain conformation of His 218 is stabilized by forming a Hydrogen bond with Asp 222 (Asp 225 in PhoN$_{BCX4.10}$). The conserved histidine of domain II, His 179 (His 182 in PhoN$_{BCX4.10}$), may function as the proton donor for the leaving group. Lys 144 (Lys 147 in PhoN$_{BCX4.10}$) and Arg 151 (Arg 154 in PhoN$_{BCX4.10}$) in domain I may be involved in maintaining the orientation of the phosphate group in relation to His 179 thus supporting the nucleophilic attack. This stabilizing effect is enhanced by the side chain of Ser 177 (Ser 180 PhoN$_{BCX4.10}$) and the amide nitrogen atoms of Gly 178 (Gly 181 in PhoN$_{BCX4.10}$) and His 179 (His 182 PhoN$_{BCX4.10}$) which compose domain II. Two additional domain III residues are also conserved. Ser 211 (Ser 214 in PhoN$_{BCX4.10}$) may have an important overall structural role while Arg 212 (Arg 215 in PhoN$_{BCX4.10}$) may have a role in the stabilization the phosphoenzyme intermediate.

Due to a wide range of pH optima of members of the class A NSAP family, PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ are considered acid/neutral phosphatases. The overall functional range of pH exhibited by PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ is consistent with the optimal pHs of

known secreted NSAPs from various bacterial isolates. The two NSAPs differ in that PhoN$_{ACX6.13}$ exhibits a sharp acidic pH optimum while the acidic-neutral PhoN$_{BCX4.10}$ has a broad range of functional pH. The broad pH range for activity demonstrated by PhoN$_{BCX4.10}$ is in contrast to the distinct pH optima observed in the six experimentally determined NSAPs described in Figure 3-11. The apparent broad pH optimum for *phoN*$_{BCX4.10}$ may be due to substrate depletion resulting in activity being outside the linear range. The pH optimum exhibited by PhoN$_{ACX6.13}$ is somewhat more acidic than that of the experimentally determined NSAPs and is comparible to the pH optimum of the NSAP from *Prevotella intermedia* (Chen *et al.*, 1999).

Signal peptide sequences predicted for PhoN$_{ACX6.13}$ and PhoN$_{BCX4.10}$ both include alanine as the last residue which is consistent with the signal peptides of the acid phosphatases of *E. blattae* (Ishikawa *et al.*, 2000), *Prevotella intermedia* (Chen *et al.*, 1999) and *Salmonella typhimurium* (Makde *et al.*, 2003). The number of residues in the predicted signal peptide of PhoN$_{ACX6.13}$ is similar to that of the signal peptides of the above mentioned three acid phosphatases, while the 29-residue length of the predicted signal peptide of PhoN$_{BCX4.10}$ is longer than that of the signal peptides of the above three acid phosphatases but still within the range typical of leader peptides (von Heijne, 1985). Apart from having optimal phosphatase activity at the acidic-neutral pH range, affiliation with the class A NSAP family is indicated by the presence of the conserved PAP2 acid phosphatase domain and signature sequence domains, KXXXXXXRP (domain 1), PSGH (domain 2) and SRXXXXXHXXXD (domain 3). These domains were previously identified as motif

consensus sequences in a group of phosphatases consisting of lipid phosphatases, glucose-6-phosphatase and a number of NSAPs (Stukey & Carman, 1997).

From the analysis of regions upstream to $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$ putative recognition sites for DNA binding proteins such as RNA polymerases or proteins that bind to RNA polymerase subunits, were identified, revealing possible mechanisms for transcriptional regulation of the two NSAPs. Due to the lack of knowledge of the microorganism of origin, it is difficult to predict which of the possible promoter sequences drive the expression of $phoN_{ACX6.13}$ and $phoN_{BCX4.10}$ necessitating the consideration of additional potential promoters. Since phosphatase expression is generally regulated by external Pi concentration, it would be expected that the PhoB-PhoR two component regulatory system is primarily involved in transcription-level regulation of the phosphatase gene. The presence of a putative "pho box" upstream to $phoN_{ACX6.13}$ suggests the possibility of PhoB dependent regulation. Since the association of PhoB with σ_{70} factor of RNA polymerase has been previously demonstrated in *E. coli* (Kumar *et al.*, 1994), it is not suprising to find σ_{70} binding sites in the vicinity of the pho box in $phoN_{ACX6.13}$. The absence of a "pho box" upstream to $phoN_{BCX4.10}$ does not necessarily mean that PhoB dependent regulation does not occur. It was previously demonstrated that no PhoB binding site was predicted in the upstream region of *E. coli* and *Sinorhizobium meliloti* polyphophosphate kinase (*ppk*) but promoter fusion assays still showed the gene was still induced by low Pi concentration as were most members of the *pho* regulon (Kornberg, 1999; Yuan *et al.*, 2006a; Yuan *et al.*, 2006b). The presence of additional promoters suggests the possible involvement of other modes of regulation. Further

experiments such as transcriptional fusion assays are needed to confirm the involvement of PhoB in phosphatase regulation.

Based on the previously determined structure of the NSAP from *E. blattae*, the molecular structure of the core polypeptide of NSAPs $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10}$ was predicted. Although a pairwise BLAST alignment showed that a 50% amino acid identity is shared among each of metagenomically derived NSAPs with *E. blattae* NSAP, the identity occurred throughout the area spanning the signature motifs which were demonstrated to be important for catalysis in *E. blattae* (Ishikawa *et al.*, 2000). From the predicted molecular structure of $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10,}$ the enzymes take on a globular shape arranged such that the catalytic sites are embedded in the inner core but are accessible to the substrate. This type of arrangement is typical of many enzymes. The predicted structures show an arrangement of the signature sequence residues within a close distance to each other and in positions consistent with the reaction mechanism of the NSAP from *E. blattae*. Therefore, a reasonable prediction of the formation of a phosphoenzyme intermediate during the nucleophilic attack of the substrate's phosphoryl group, followed by the protonation of the leaving group, can be made and is carried out by the concerted action of the signature motif residues.

Although the new class A NSAPs identified here carry the motif sequences of phosphatases, they exhibit a low amino acid identity to class A NSAPs whereas the highest identity shared is with the acid phosphatase protein of *R. eutropha* H16. The distance-based phylogenetic analysis indicates this phenomenon by showing $PhoN_{ACX6.13}$ and $PhoN_{BCX4.10}$ on a separate clade from almost all the other known class A NSAPs. From a previous survey

of sequences of acid phosphatases from bacterial isolates, the group of class A NSAPs was defined. We show here that by looking at similar phosphatases from uncultured bacteria, we find a new protein sequence diversity illustrated by the differential residue conservation between the proteins from cultured and uncultured bacteria. By expanding the collection of sequence data through a metagenomic analysis, we may be able to learn more about how bacterial phosphatases have evolved to adapt to various environmental conditions with respect to P availability.

Functional annotation is currently available for a large number of NSAPs belonging to the enteric bacterial class of γ-*Proteobacteria* such as *Salmonella typhimurium* (Makde *et al.*, 2003), *Prevotella intermedia* (Chen *et al.*, 1999), *E. blattae* (Ishikawa *et al.*, 2000) and *Morganella morganii* (Thaller *et al.*, 1994), in addition to the acid phosphatase of the α-*Proteobacterium Zymomonas mobilis* ZM4 (Pond *et al.*, 1989). The two remaining clusters contain bacterial isolates for which the class A NSAPs have yet to be experimentally determined. The implications of this are that there are a large number of NSAPs that remain to be extensively characterized, with potential for learning more about the functional characteristics and biochemical diversity of this group of enzymes.

3.2.2 PhoA$_{ACX6.71}$ is a novel member of the nucleotide pyrophosphatse (NPP) phosphodiesterase enzyme family

The 2533-bp sequence from pAR004, derived from municipal waste activated sludge metagenomic cosmid pACX6.71, contains an ORF that predicts a protein of 549 amino acids in length that is homologous to proteins of the phosphodiesterase–nucleotide pyrophosphatase (NPP) family (pfam016633.11) based on the NCBI CDD (Marchler-Bauer

et al., 2005). The gene was designated *phoA*$_{ACX6.71}$. As was shown with the acid phosphatases, *phoA*$_{ACX6.71}$ also contains a RBS upstream of the translation start site. In addition, like many other members of the NPP family, PhoA$_{ACX6.71}$ contains six conserved aspartic acid and histidine residues that form two metal binding domains and a conserved threonine residue that makes up the catalytic centre (Figure 3-9). To determine possible promoter sequences upstream of *phoA*$_{ACX6.71}$, 476 bp upsteam of the translation start site were analysed for regulatory protein recognition sequences. Seven possible promoter sequences were determined including *E. coli* class σ_{70} -10 and -35 binding sites and a "pho box". A potential transcription start site was located 37 bp upstream of the ORF translation start site (Figure 3-9). Other possible promoters include recognition sites for a NAGC-like transcriptional regulator MIc, known for regulating genes involved in carbon metabolism (Kim *et al.*, 1999). A putative recognition site was also found for an inversion stimulation factor, Fis. This protein is known to regulate a variety of genes and operons including the activation of amino acid transport and the repression of nitrite reductase and NADH dehydrogenase (Kim *et al.*, 1999; Travers *et al.*, 2001). A possible recognition sequence for *E. coli* FadR was also identified. This protein is known to regulate many genes and operons involved in long chain fatty acid transport, activation and beta-oxidation (Campbell & Cronan, 2001). Additional promoters that were identified include binding sites for the *E. coli* MalT regulator, which affects the transcription of maltose regulon genes (Schlegel *et al.*, 2002), and *E. coli* GlnG, a response regulator which affects the transcription of certain nitrogen metabolism genes (Magasanik, 1989).

Figure 3-7: Predicted molecular structure of NSAP PhoN$_{ACX6.13}$.

The structure was predicted using the computer program MUSTER, threading the known structure of the class A NSAP of *Escherichia blattae* (Ishikawa *et. al*, 2000). The conserved residues in each of domain I (blue), domain II (green) and domain III (red) of the NSAP signature motifs are highlighted. The structure was predicted based on the template structure of the NSAP from *E. blattae* (Ishikawa *et al.*, 2000).
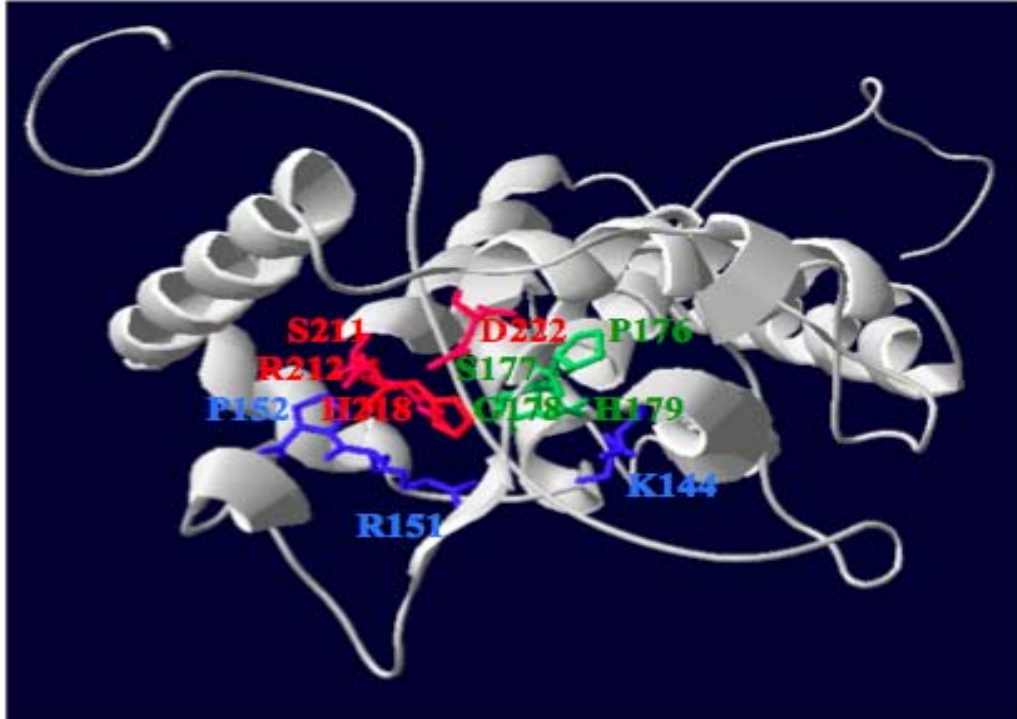
Figure 3-8: Predicted molecular structure of NSAP PhoN$_{BCX4.10}$.

The conserved residues in each of domain I (blue), domain II (green) and domain III (red) of the NSAP signature motifs are highlighted. The structure was predicted based on the template structure of the NSAP from *E. blattae* (Ishikawa *et al.*, 2000).

Since PhoA$_{ACX6.71}$ is predicted to be a secreted alkaline phosphatase, it was of interest to examine its N terminal sequence for the occurrence of a signal peptide and to know the functional pH range of the enzyme. An analysis of the first 70 N-terminal amino acids of PhoA$_{ACX6.71}$ using SignalP3.0 Server (Bendtsen *et al.*, 2004) indicated that the protein contains a signal peptide with a cleavage site at Ala20-Glu21. The residues that make up the signal peptide consist of positively charged amino acids, followed by hydrophobic non-polar amino acids, ending with a small side chain amino acid, alanine, at the cleavage site. The occurrence of this primary structure suggests that the N-terminal part of PhoA$_{ACX6.71}$ is indeed a signal peptide. Consistent with the possibility that PhoA$_{ACX6.71}$ is a secreted alkaline phosphatase, an optimal pH of 9.5 was determined for enzyme activity of periplasmic extracts of DH5α (pAR004) (Figure 3-10).

By amino acid sequence comparison, PhoA$_{ACX6.71}$ showed the highest identity with the predicted protein from the type I phosphodiesterase/nucleotide pyrophosphatase superfamily of a marine γ-proteobacterium HTCC2148 (accession no YP_002652938.1) at 50% and no significant similarity to any orthologues of PhoX, a recently discovered class of alkaline phosphatases (Sebastian & Ammerman, 2009). PhoA$_{ACX6.71}$ shares an amino acid identity of 57% with a putative protein of the Type I phosphidesterase/nucleotide pyrophosphatase superfamily, identified from the metagenome of a bacterial symbiont of a gutless worm originating from the Meditarranean Sea (JCVI_PEP_1113958955152). Aside from the very low amino acid identity to known bacterial alkaline phosphatases, the novelty of PhoA$_{ACX6.71}$ became apparent when it was placed on a neighbour-joining phylogenetic tree with alkaline phosphatases of known bacterial isolates (Figure 3-11). Most representative

alkaline phosphatases and NPPs form three major clusters while proteins from two acidobacterial isolates of *Solibacter usitatus Ellin 6076* formed Cluster V as well as the proteins from the *Bacteriodetes* isolates, *Bacteroides thetaiotaomicron* VPI-5482 and *Bacteroides fragiles* NCTC 9343 (Cluster IV). Alkaline phosphatases and NPPs from members of the *Flavobacteria* class including the experimentally determined Ca^{2+} ATPase of *Myroides odoratus* and Pi irrepressible alkaline phosphatase of *Chryseobacterium meningosepticum* predominantly make up Cluster I. NPP from the uncultured marine bacterium Ant29B7 is also affiliated with Cluster I. Cluster II contains exclusively isolates belonging to the α-*Proteobacterial* class. Isolates belonging to the γ-*Proteobacterial* class and activated sludge-derived PhoA$_{ACX6.71}$ make up Cluster III. This analysis demonstrates that PhoA$_{ACX6.71}$ is clustered together with alkaline phosphatases from isolates of the γ-*Proteobacterial* class: *C. psychrerythraea* 34H, and strains from the *Vibrio* and *Legionella*.

Figure 3-9: Sequence of *phoA*_{ACX6.71} and flanking regions in pAR004.

Amino acid sequence of the *phoA*_{ACX6.71} ORF is indicated. Recognition sequences of *E. coli* class regulatory proteins are boxed as follows: σ70 -35 and -10 (red), MIc (turquoise), Fis (blue), FadR (yellow), MalT (pink) and GlnG (green). "Pho box" is underlined. Putative transcription start site is denoted by ↪. Bases shaded in grey correspond to a predicted RBS. ⇑ denotes the putative signal peptide cleavage site. Residues shaded in gold are predicted to be involved in Zn^{2+} binding. The reaction centre threonine (Thr 73) is highlighted in red.

GAGCAGTTCGCGGTAGTCATTGTAGGAGGCGTCCGCGCCGAGTCCGGCGAGATCCGCGCCCTGTGCCCGGCGGGCCGCGTCCACATC

GCACACGGCCACAATCTGCACTTCGGGCTTTGCCTTGAAGGCGGCCAGGTTGCTCATGCCGCGGTTGCCGGTGCCGATGCACCCTAC

GGTGATGCGCTCGCTGGGGGCCACGGTGCCATCGAGGCCGAGGGCGCTGGCGGGGATGATCATGGGGGCCGCTATCGCGGCGGTGGT

GGCGAGGAACTGGCGGCGGGTGGGGTTGGGCGACTTGGGCATGGAAGAAAACTCCTGAGCAGGGCAATAAGCAGAAAGGGTTAGCAT

TGCTGGCCGGGTGCCCGGGGTCGGACCGCGCAATCCCGAAGGGACAGCATAACATAATCTGAACAATTTCAAAGTATGCTATGTTGT
                                                      -35   -10

CAAAAAATACTGTTTCTCCATCCCGCTCAGGAGGTCATTTCATGCGTAATATATTCCTTTCCGCACTGGTCGCGTTTGCTTTCGCGG
                                      M R N I F L S A L V A F A F A A

CCCTGGCGTCGGCGGAGACACCCGTCTGGTTCTGTTCATCGCGGTAGATCAATTGCGCGGAGACATGCCCATACGTTATCAGGAGC
 L A S A E T P R L V L F I A V D Q L R G D M P I R Y Q E R
                                              ⇑

GCTTTGGCGAAGGCGGCTTTCGTTATCTCATGGAGCAGGGCTTTCACTATCGCAACGCCCACTACAGGCACTCCAATACCTTCACGG
 F G E G G F R Y L M E Q G F H Y R N A H Y R H S N T F T A

CGGTGGGTCATGCCACGCTGGCCACCGGGGGGCAATACGCCGCAGCACGGCATTGTGGGGAATGAGTGGTTCGACCGGGCGAGCGGCT
 V G H A T L A T G G N T P Q H G I V G N E W F D R A S G S

CCGGCATGAACGCGGTGGAGGATCAGGCGCATCCGCTGCTCGGTGTGGAAGGGGCTGCGGGCGAGGGGCGCTCGCCGCGAAACCTGA
 G M N A V E D Q A H P L L G V E G A A G E G R S P R N L T

CCTCGAGCACCTTCGGTGACGAGCTGGTGCTGGCCAGCGGCGGCGCTTCGCGCGTGTTTGGCGTGTCCATCAAGGACAGGGGCGCGG
 S S T F G D E L V L A S G G A S R V F G V S I K D R G A V

TGATCCTCGGCGGGCACCTGGGGAAGAGCTTCTGGTATGAGCCGGAGTCGGGGCGCTTTGTGACGAGTTCCTTCTACTACGCGGAGT
 I L G G H L G K S F W Y E P E S G R F V T S S F Y Y A E Y

ACCCGGCGTGGGTGAGCGCCTGGAACGGGGCGAAGCCCGCAGACGCGTGGCTGGGTGCGTCGTGGTCGCTCCTGAATCCGGTGGATA
 P A W V S A W N G A K P A D A W L G A S W S L L N P V D T

CTTACATCTACGGGGCGCAGGATGATCGGCCCGAGGAGCGGCCCTACAAGGCCATGGGCAATACCTTTCCCCATCCGCTGCCCGCCG
 Y I Y G A Q D D R P E E R P Y K A M G N T F P H P L P A E

```
AAGCGGGGAAGGACTACTATGGCAATCTGCGCCTTCACACCGATGGGCGACGTGCTCACGCTCCAGTTTGCCCGGGAACTGATGCGCG
 A   G   K   D   Y   Y   G   N   L   R   F   T   P   M   G   D   V   L   T   L   Q   F   A   R   E   L   M   R   A


CGGAGGAAGTGGGGCAGCGGGACGCCACGGACATCCTGACTATCAGCCTTTCGGCAACGGACTACATCGGCCACGCCTTCGGGCCGA
 E   E   V   G   Q   R   D   A   T   D   I   L   T   I   S   L   S   A   T   D   Y   I   G   H   A   F   G   P   N


ACAGCTTGGAGTCGGAAGACAACCAGCTCCAGCTCGACCGGAGCCTCGCCGAGTTCTTTGCGGAGGTGGACCGGCTGGTGGGGCTCG
 S   L   E   S   E   D   N   Q   L   Q   L   D   R   S   L   A   E   F   F   A   E   V   D   R   L   V   G   L   D


ATCGCACGCTGATCGTGCTCAGTTCGGACCACGGCATCGACGAGATCCCGGAGTATACGCAGCACCTGGGATGTGACGCCGGGCGTC
 R   T   L   I   V   L   S   S   D   H   G   I   D   E   I   P   E   Y   T   Q   H   L   G   C   D   A   G   R   H


ACGTGCCCGAGGAATTCATTGCGGCGGCGAATGGGGCGCTCAAGGCGCGCTTCGGTATCGCGGACGACCTGGTGTTGACCTTTCAGA
 V   P   E   E   F   I   A   A   A   N   G   A   L   K   A   R   F   G   I   A   D   D   L   V   L   T   F   Q   N


ATCCGAGTCTGTATCTGGACGAGGCCCGGGTTCAGGCGCTGGGCCTGGCGCTGCCCGAGGTCGAGCGGGCCCTGGCCGACGCCATGG
 P   S   L   Y   L   D   E   A   R   V   Q   A   L   G   L   A   L   P   E   V   E   R   A   L   A   D   A   M   V


TGGCCCTGCCCGGCTTCGATATGGCGGTGACGCGCAGCGACCTTCTGGAAGGCCGGGTTCCGAATACGAAGGTCATGGACATGGTGA
 A   L   P   G   F   D   M   A   V   T   R   S   D   L   L   E   G   R   V   P   N   T   K   V   M   D   M   V   T


CGCGCGCCTTCCACCCGAAGCGATCGGGCAATGTACTGGTGGTGCAGTCGCCCTCGTGGTACCTCTACCCCGAAGCCCAGAAATATG
 R   A   F   H   P   K   R   S   G   N   V   L   V   V   Q   S   P   S   W   Y   L   Y   P   E   A   Q   K   Y   A


CCGCCATGCACGGATCGCCCTACAGCTACGACACCTATGTGCCCATCTTCTTTGCGGGGCCCGGCGTTGTGTCGGGTGCAAGCCAGC
 A   M   H   G   S   P   Y   S   Y   D   T   Y   V   P   I   F   F   A   G   P   G   V   V   S   G   A   S   Q   R


GTCCTGTGGCGCCCGAGGATGTcGCCTCAAcGATAACGGCCTACCTGGGCATCAAGCCGCCGTCGGGTAACATGGGCAATCCCCTGC
 P   V   A   P   E   D   V   A   S   T   I   T   A   Y   L   G   I   K   P   P   S   G   N   M   G   N   P   L   L


TGGAGGTGGTGGGTGCGGGAGAAACCGGTGTGCGGTGAATTTGTGCGCACCGTTGGGCGTGGTAGAATATGAGGAGCCGGGCACGTC
 E   V   V   G   A   G   E   T   G   V   R   *


GCCCGATTCTGGCGAGGACCTGTCATGGATACCACAATCGAACGTGAAGCGCAGACCACCCTGGCTCAACTGCTGGAGCGCGTGGCG


CAGGGCGAACGGATCACCATTACGGAGAACGGCGTGCCCGTGGCGGTGATGAACTCAGCAGAACCCGCGAGCAGCACCAGGAACTCC
```

GATCGTCAACGAGAAATTCGTCAGGCGATTGAAGACATCAAGGAACTCCGAAAAGGTCTCACGCTTGGTCCGGATCTAACGATCAGA

GACTTGATTGAGGAAGGCAGGCGTTAGTTGGACGGAATTGTTCTGGATAATTCCATTATCGTCGCATGGTGTTTTGAAGACGAGATA

AGCGAGTACG

Figure 3-10: Phosphatase activity of periplasmic extracts of *E. coli* DH5α expressing

phoA$_{\text{ACX6.71}}$ (pAR004) at various pHs. Assays were done in triplicate.
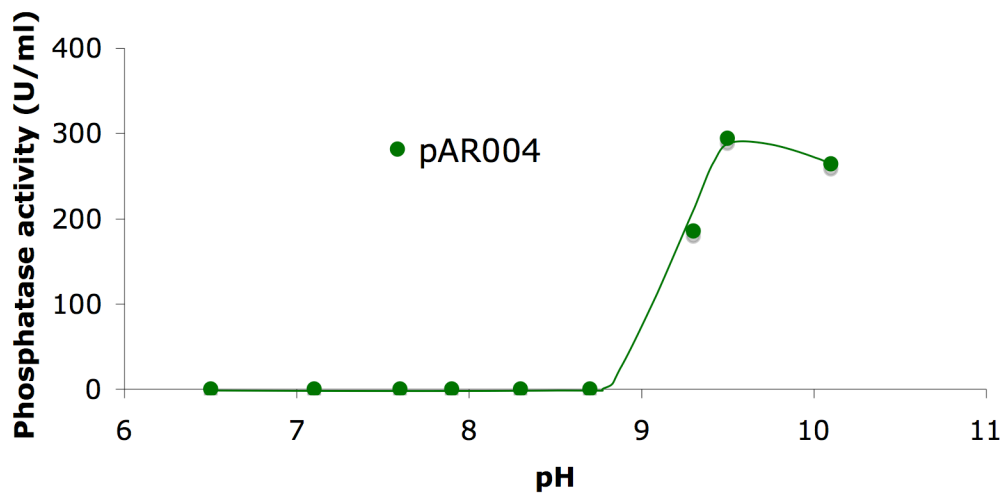
Figure 3-11: Neighbour-joining tree showing type I nucleotide phosphodiesterase/ nucleotide pyrophosphatases proteins.

CP34H, *Colwellia psychrerythraea* 34H; V Sp-MED222, *Vibrio sp.* MED222; VV-YJ106, *Vibrio vulnificus* YJ016; VV-CMCP6, *Vibrio vulnificus* CMCP6; BF-NCTC, *Bacteroides fragilis* NCTC 9343; BT-VPI-5482, *Bacteroides thetaiotaomicron* VPI-5482; RBHP-HTCC2501*, Robiginitalea biformata* HTCC2501 hypothetical protein; UC-29B7, uncultured marine bacterium Ant29B7; TSp-MED152, *Tenacibaculum sp.* MED152; PI-23P, *Polaribacter irgensii* 23-P; CA-HT, *Croceibacter atlanticus* HTCC2559; FSp-MED217, *Flavobacterium sp.* MED217; RB-HTCC2501, *Robiginitalea biformata* HTCC2501; FB-HTCC21270, *Flavobacteriales bacterium* HTCC2170; SE-PCC7942, *Synechococcus elongatus* PCC7942; CM, *Chryseobacterium meningosepticum*; Ca-Mo, Ca^{2+} ATPase *Myroides odoratus*; SUE6076a, *Solibacter usitatus* Ellin 6076a; SUE6076b, *Solibacter usitatus* Ellin 6076b; CC-CB15, *Caulobacter crescentus* CB15; ZmSuZM4, *Zymomonas mobils subsp* ZM4; NA-DSM12444, *Novosphingobium aromaticivorans* DSM 12444; SA-RB2256, *Sphingopyxis alaskensis* RB2256; RSp-MED193, *Roseobacter sp.* MED193; LP-Lens, *Legionella pneumophila* str. Lens; LP-Paris, *Legionella pneumophila* str. Paris; Boxed: $PhoA_{ACX6.71}$ -phosphodiesterase isolated from the metagenome of municipal waste activated sludge; ★ denotes isolates for which alkaline phosphatases or phosphodiesterase activity were determined experimentally.

Cluster I

RB-HTCC2501 FB-HTCC2170 CM ★

Fsp-MED217

CA-HT

PI-23P

Tsp-MED152

UC-29B7

Ca-Mo ★

SE-PCC7942 ★

RBHP-HTCC2501

98

100

98

88

SUE-6076a

Cluster V

SUE-6076b

BT-VPI-5482 100

BF-NCTC-9343 Cluster IV

94

98

78

CC-CB15

100

PhoA_ACX6.71

★

100

100

VV-CMCP6

VV-YJ106

Vsp-MED

222 CP34H

100

100

LP-Lens LP-Paris LPHP-Paris

RSp-MED193

ZMSu-ZM4

★

NA-DSM

1244

SA-RB2256
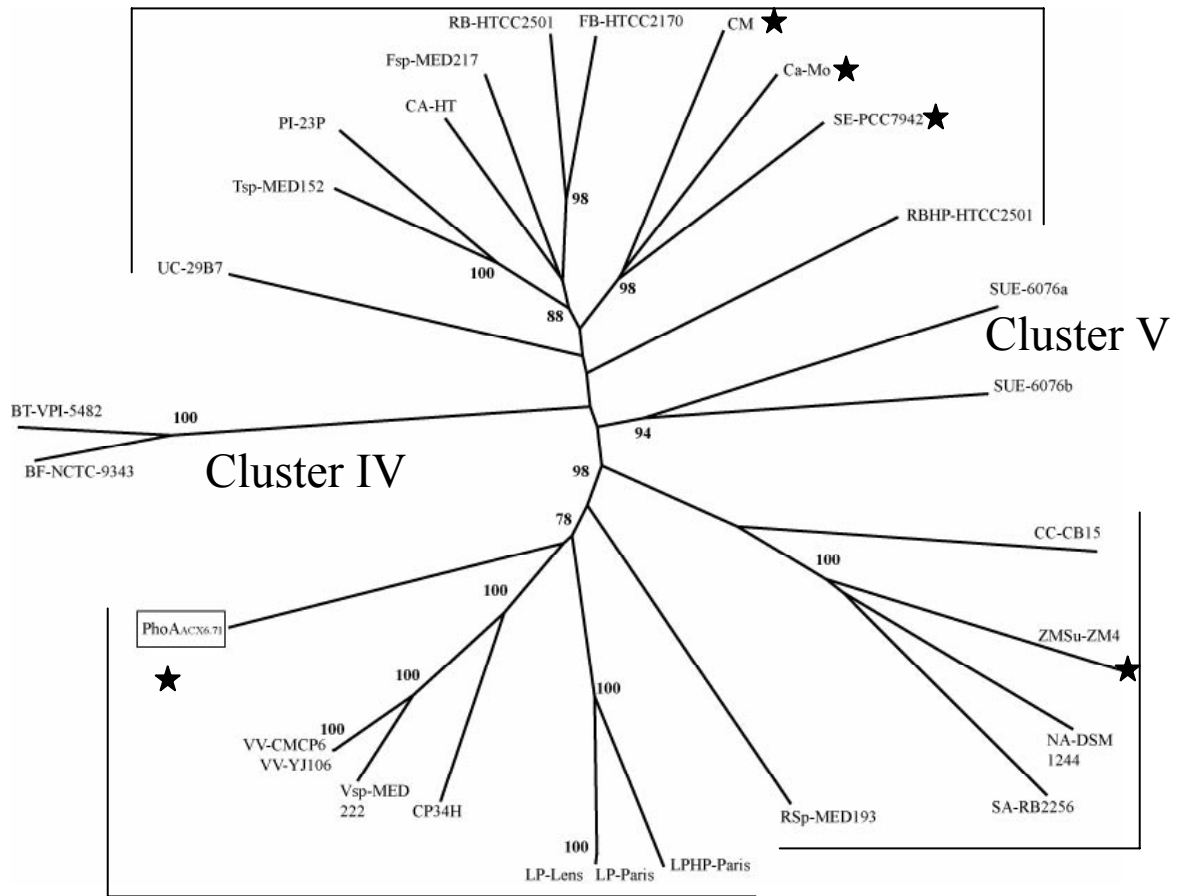
Cluster II

Cluster III

To illustrate the diversity of NPPs and alkaline phosphatases, multiple amino acid sequence alignments of conserved regions of proteins from cultured and uncultured bacteria were constructed (Figures 3-12a and b). These regions included key aspartic acid and histidine residues with potential roles in metal binding, and a reaction centre threonine. Upon comparison of the two alignments there are many differences in the number and the type of conserved residues with a larger number of conserved residues occurring in the proteins from uncultured bacteria. One key difference between the aligned regions of proteins from cultured and uncultured bacteria is the absence of the region in the C-terminus that contains the conserved histidine in the proteins from uncultured bacteria.

Figure 3-12a: Multiple amino acid sequence alignment of alkaline phosphatases and NPPs from bacterial isolates and metagenomic DNA derived from municipal waste activated sludge.

GenBank accession numbers are indicated to the left of the strain name abbreviation. Conserved residues are highlighted and include those (in bold letters) that are involved in metal binding (underlined with •) and the reaction centre threonine (underlined with ★). Strain names are denoted as follows*: C. me, Chryseobacterium meningosepticum*; *P. sp*, *Polaribacter sp.* MED152; *M. od, Myroides odoratus*; *S. us*1, *Solibacter usitatus* Ellin 6076a; *S. us2, Solibacter usitatus* Ellin 6076b; *L. pn* Lens, *Legionella pneumophila* str. Lens; *L. pn* Phil; *Legionella pneuomophila* str. Philadelphia; *C. sp, Caulobacter sp.* K31; *C. cr* CB15, *Caulobacter crescentus* CB15; *Z. mo* Zm4, *Zymomonas mobilis* Zm4; *N. ar* 12444, *Novosphingobium aromaticivorans* DSM 12444; *S. al*, *Sphingopyxis alaskensis* RB2256; *C.ps* 34H, *Colwellia psychrerythraea* 34H; PhoA, PhoA$_{ACX6.71}$. Consensus sequence residues are denoted as follows: χ, polar amino acids DEKRNQ; Ψ, non polar amino acids ILVMFY; ν, acidic amino acids DE; σ, small amino acids GAST; ., no particular amino acid.

```
gi|88886019|C. me            108  KSVYCTSDETVQPVGT--TSNSVGQHSPRNIWSTTVTDQLGLATNFTSKVVGVSLKDRASILPAGHNPTGAFWFD (96)  QMADAAIDGNQMGVDD--ITDFLTV
gi|226377317|P. sp           106  KTLYCVDDANYETVGN---SGKAGKKSPHRLFTSTLADQLKLHQVSQGKSIGIAIKDRSAILPAGHFANAAYWFD (97)  DFAKAAILGEKLGKSK--HTDFLAI
gi|927036|M. od              128  QNMYCTQDDSVQAVGG---EGKVGQQSPKNLLVSTITDQLKLATNFQSKVIGIAIKDRGGLPAGHFANAAAWLD  (95)  DFAKRAIENEQLGNNPLQVTDFLAV
gi|116619950|S. us1          114  KTWTSVSDDTERLLGG----QDIAGNSPRRLLVSTVGDEMKRGLPDHPKVIGISLKDRASILPSGHMADAAYWYD (72)  QFVEAVRREKLGQRE--ATDLLTV
gi|116623345 S. us2          100  KSTSAVSDDSVKVVGA----GGGASASPHLLLAPTVGDELKAATGGRALVIGISQKDRSAILPAGAKANAAFWFD (64)  GIAERALEAEQLGKHD--DPDLLSV
gi|54294741|L. pn Lens       111  KLKYCMEDLQAKILPTVHTHIEISGRSPRNLKPSTLSDEIILAN--KGKVFGVSLKDRAAITLAGH-AGKAFWFD (89)  DFAKQLLVNEKLGKTT-GQTDYLAI
gi|52842079|L. pn Phil       103  KLKYCMEDLQAKILPTVHTHIEISGRSPRNLKPSTLSDEIILAN--KGKAFGVSLKDRAAITLAGH-AGKAFWFD (89)  DFAKQLLVNEKLGKTT-GQTDYLAI
gi|167645342|C. sp K31       115  KKTYCLDDPSVTLANND--PSGGGRLASPKLLMAETYGDWLKAVSP-KSRVYAVSGKDRGAINMAGHKADGFWLE (91)  EAAQALRDAFKLGDGE--ATDVLTI
gi|16126700|C. cr CB15       120  KTTYCLADPSVTLADD-----PKGRAVSPANMVATTYGDWLKAVSP-NSRVFGVSGKDRGAITMSGHAADGQFWYQ(103) ELATALREAYGLGDGP--QVDVLAI
gi|56551834|Z. mo ZM4        139  KKKYCSEDPSLSADPQ------NYQPSVHYLKVPTLGDRMKKANP-HSRVISVAGKDRAAIMMGGHMTDQIWFWS  (82)  DIAIGLIDELKLGHGN--APDLLTV
gi|87200803|N. ar12444       128  KQVYCAEDTGKRAAGS----KDYVASPIHLLVPTLGEWMKRANP-AARNIAVSGKDRGALMMGGHDTDQVYWK    (80)  DLAVKLVDEQKLGRGA-VPDLLAV
gi|103485542|S. al           126  KRVYCAEDETVPDTSS----GSGKYAPSVKHLLVPTLGDLMKARDP-QAQVVSVAGKDRAAIMMGGHRADETLWLA(81)  ATAAALRQARRMQOGS--GTDLLII
gi|71280023|C. ps 34H        127  SADAGVNQKTEIDPTQ--RTAKSSGRSPNNIQVSTFSDELKIASNGKAKVFGVSVKDRGAISMAGHMGKAFWFSK (93)  DFAKAVVNNEQLSQDE--VTDYLSV
gi|83655662|PhoA_ACX6.71     102  SGMNAVEDOAHPLLGV--EGAAGEGRSPRNLTSSTFGDELVLASGGASRVFGVSIKDRGAVILGGH-LGKSFWPE (94)  QFARELMRAEEVGQRD--ATDILTI
Consensus                         K..ψYC.χD.σ...........SP..Lψ..Tψσp.ψ.......σχVψσψσ.KDRσσI..σGH..χσ.ψWψχ            χψAχ..ψ..χχψGχ.....DψLσψ
```
                                                                    ★   △

```
gi|88886019|C. me            NLASTDYVGHNFGPNSIEVEDTYLRLDRDLADFFNNLDKKVGKGNYLVFLSADHGA (108)  SGDIQIISKDGYLSAYSKKGTT----HSVW 57
gi|226377317|P. sp           SFSSTDYIGHQYGPASVEIEDTYLKLDRDLSDFFKFLDNKVGKDNYTLFLTADHAA (109)  SGDVMLI-PYPATLSRSKTGTS----HGSG 57
gi|927036|M. od              SLSSTDYIGHQFAINSIEIEDTYLALDQNIGKGNYTLFLSADHGA            (110)  SGAIQLI-LEPQWFSGSKDGKGTT--HGSW 57
gi|116619950|S. us1          SFSPNDPIGHTYGPDDPRAHDISLQTDRALAKLFQLDRALAKLFDQMQHVLVLTADHGV (108) SGDLEII-MDPYWARVATGTT----HGTP 64
gi|116623345 S. us2          SLSATDYVGHDYGPDSPEEKEALMETDRDLDHLLQAVDRQVGLANTLVVITSDHGV (103)  AADLVFI-PEPNWIFFDGNYT----HSTT 57
gi|54294741|L. pn Lens       SFSAVDAIGHQFGPNSLEAEDNLIALDDTLSQLFKTIDKEVGLNNTLIITADHGV (109)  AGDVYIV--QPPYQSHGAKSEDRVA--HGSP 65
gi|52842079|L. pn Phil       SFSAVDAIGHQFGPNSLEAEDNLIALDDTLSQLFKTIDKEVGLNNTLIITADHGV (109)  AGDVYIV--QPPYQSHGAKSEDRVA--HGSP 65
gi|167645342|C. sp K31       SLSATDFIGHRYGARGPEMCDQVLRLDDRLGVFLGSLDKVKG--QVLVVITADHGG (120)  VGDVLVA-FQPYQTLAA-AGTTYVASHGSP 59
gi|16126700|C. cr CB15       SLSATDAVGHHRYGTRGPEMCDQIARLDERLGAFLKGLEKVKG--GVLVALSADHGG (120) VGDIVVA-FRPDRVPAT-AGPTYVSTHGSP 59
gi|56551834|Z. mo ZM4        SLSATDAVGHAYGTEGAEMCSQMAGLDDNIARIIAALDSN-GV-PYVLVLTADHGG (117)  SGDLIVI--LKPRVTPIAKP-VSYVATHGSA 65
gi|87200803|N. ar12444       SLSATDYVGHATGEGAEMCIQLAQLDLALGDFLASLDKR-GI-DYAVVLTADHGG (116)   SGDFVVM--LKRSVVAIPTPRAGYVATHGSP 66
gi|103485542|S. al           GLSATDVVGHGTGTEGSEMCLQLAGLDRELGDFLARLDAT-GM-NYAVVLTADHGG (113)  SGDFIVA--LKPRVTPIAEPGVGYVATHGSV 64
gi|71280023|C. ps 34H        SFSSTDYVSHIFGPSLEAEDNMLRLDRTLANLLTFIDKKVGLENTLIVLSADHGA (112)  SGDVYVV--LEAHRFVADMEGLTVAATHGSP 58
gi|83655662|PhoA_ACX6.71     SLSATDYIGHAFGPNSLESEDNQLQLDRSLAEFFAEVDRLVGLDRTLIVLSSDRGI(111)  SGNLVVV--QSPSWYLYP-EAQKYAAMHGSP 63
Consensus                    SψσSσTDψψGH.ψG.χσ.E..χ.ψψσψLDχ.Lσχψψ...ψDχ..G..χ.           σGDψψσψψ..χP................HGσ.
```
    △   △   ★   △△   △

114

Figure 3-12b: Multiple amino acid sequence alignment of alkaline phosphatases and NPPs from metagenomic DNA derived from various environmental samples.

Numbers at the beginning and the end of the sequences denote the number of residues before the beginning and after the end of the alignment portion, respectively. Numbers in brackets denote the number of residues omitted from the alignment. Conserved residues are highlighted and include (in bold letters) those that are involved in metal binding (underlined with •) and the reaction centre threonine (underlined with ★). Consensus sequence residues are denoted as follows: χ, polar amino acids DEKRNQ; Ψ, non polar amino acids ILVMFY; ν, acidic amino acids DE; σ, small amino acids GAST; ., no particular amino acid.

```
JCVI_PEP_11139958955152   71  FVYNTEDKRHHIIGREAKAHEGVSPRNLLASTISDELVHTGGRSRAFSVSVKDRGAILP-GGHAGKAFWFS-KSTGEFVTSTYYY
JCVI_PEP_1113957163738    37  FVYNTEDKRHHIIGREAKAHEGVSPRNLLASTISDELVHTGGRSRAFSVSVKDRGAILP-GGHAGKAFWFS-KSTGEFVTSTYYY
JCVI_PEP_11139600577716   66  FVYNTEDKRHHIIGREAKAHEGVSPRNLLASTISDELVHTGGRSRAFSVSVKDRGAILP-GGHAGKAFWFS-KSTGEFVTSTYYY
JCVI_PEP_1113523889380    86  DVYCCDDPSVQLLGLPAA--KGVSPHKLLVETVGDVMKLASGNRSKVFGVAIKDRVAMLMAGHNADTAFWLEDNL---WTTSTYYR
JCVI_PEP_1105127036099    35  NKYCVDDKRFNSVGTSGEE-GEKSPKSMLVTTIGDEMKLSNNNHSKVISIALKDRSAIIPGGHKANGAYWID--NQGKWITSSYYM
JCVI_PEP_1105142559785    12  IFIGVGDDQFKTIGSDNNS-GKKSPRKMLTSTFGDELRLSSMFRSKVIGISLKDRAAILTAGHTANYALWLD-SESGNFITSSYYT
JCVI_PEP_1105103328953    77  EIYCVDDFLQKSVGTYSDE-GQKSPKNLQTTTISDENRISSNFKAKTFSVALKDRSAVLSGGHTSEMSYWFNAKNEGKFITSSYYM
JCVI_PEP_1105089570043    61  EIYCVDDFLQKSVGTLSDE-GQKSPKNLQTTTISDENRISSNFKAKTFSVALKDRSAVLSGGHTSEMSYWFNAKNEGKFITSSYYM
JCVI_PEP_1105134367270    34  SVYCVDDENYNTIGSSTNS-GKKSPSKLLTTTLSDENRIATNFKGTIGVAIKDRGSILSSGHTANAAYWFD-SREGNFISSTYYM
JCVI_PEP_1112715314422    16  MVYCVEDGNVKIVGGVGNK-EPMSPKNLISTTVTDELELATNFRSKVIGVAIKDRGSILPAGHAGDAAYWE-AGSGNFSSTWYMV
JCVI_PEP_1105091331755    89  SMYCVEDKSVKTVGSDSHD-GEMSPFNLVSNTVTDELRLATNFRSKVIGISIKDRGAILPAGHSANAAYWYD-ESKGVFISSTYYM
JCVI_PEP_1105103198155    79  MIYCAEDSLVSHVGGTPEN-GKISPRNLLTSTITDELRLSTGKRAKVVGISIKDRGASLPAGHTGD--AYWD-DVNGDWMTSTYYK
JCVI_PEP_1105131869587    52  EVYCAEDMDVEPLGTIESS-SKRSPKNIRALTFSDGIKLYTNNKGKTYGVSIKDRGAIFPAGHMADGAFWLD--ASMHFVSSSYYM
JCVI_PEP_1105075816231    51  EMYCAEDTTVSPIGTIESS-SRRSPKNLRALTFSDGIKLHTNHRGKNYGVSVKDRGAIFPAGHFADGAFWLD--NKLHFVSSSYYG
JCVI_PEP_1112733327466   116  EMYCAEDTTVTPVGTVEKS-SRRSPKNLRALTFSDGIKLHTNKQGKSFGVSVKDRGAIFPAGHFADGAYWLD--GGMHFVSSSYYS
JCVI_PEP_1105139982877   103  GMNAVEDQAHPLLGVEGAAGEGRSPRNLTSSITFGDELVLASGGASRVFGVSIKDRGAVIL-GGHLGKSFWYE-PESGRFVTSSFYY
PhoA_ACX6.71
Consensus             .φφC..φD.....φG...χ......SPχχLφφφoTφφoDEφχφ.σ...χφK.φφφφKDRGAφφ.oGH.σχ.oφWφχ....GχFφoSoYYφ
                                         Δ                          ★
```

```
JCVI_PEP_11139958955152  (89)  DFAKALVEHEKLGQNDNGAPDYLQISFSSSTDYIGHLFGPSSLETEDNILRLDRTLAELFQFVD-EKVGLDKTLILLSADHG   21
JCVI_PEP_1113957163738   (89)  DFAKALVEHEKLGQNDNGAPDYLQISFSSSTDYIGHLFGPSSLETEDNILRLDRTLAELFQFVD-EKVGLDKTLILLSADHG   29
JCVI_PEP_11139600577716  (89)  DFAKALVEHEKLGQNDNGAPDYLQISFSSSTDYIGHLFGPSSLETEDNILRLDRTLAELFQFVD-EKVGLDKTLILLSADHG   37
JCVI_PEP_1113523889380   (89)  EAARAIVEGEKLGLDE--VTDLLCINLSSNDFVGHAFGPHSLEVEDMVYRTDRQLGEFLRWLD-QTIGVGNWSFALTADHG    61
JCVI_PEP_1105127036099   (89)  DFAQKAIIKERLGQDK--YTDLLAISFSPIDYIGHFYGPHSIEMEDAIIQLDDDLSSFFTFID-KYIGKE-VLYFLTADHG    58
JCVI_PEP_1105142559785   (89)  DLAKALIKEEELGNDD--YTDLLSISLSSTDYIGHKFGVDSKEIHDTYIRLJDKDLDLFIIYLD-EIIGIENLLFFLTSDHG   20
JCVI_PEP_1105103328953   (89)  DFALNAIDYENLGKDE--ITDFLHIGFSSADHLGHRFGSNSVEVLDMYIRLDKNIEDILNYLD-KNVGEKNVIVYLTSDHG    12
JCVI_PEP_1105089570043   (89)  DFALNAIDYENLGKDE--ITDFLHIGFSSADHLGHRFGSNSVEVLDMYIRLDKNIEDILNYLD-KNVGEKNVIVYLTSDHG    56
JCVI_PEP_1105134367270   (89)  DFSLQAVESEDLGKDY--TTDFLLIDYSSTDYVGHQFGANSKEVQDTYIRLDKEIERLLDYFD-KHIGKDQYTLFLSADHG    37
JCVI_PEP_1112715314422   (89)  DFSLQAAESENLGKDH--TTDFLLIDYSSTDYVGHQFGANSKEVQDTYIKLDKDLANFDFLD-KHIGKDQYTLFLSADHG   104
JCVI_PEP_1105091331755   (89)  EFSKTAVQAEOMGVDS--IPDFIAISLSSPDYIGHAFGPNSVEVEDNYIKLDKDLANFFDFLD-KKVGKGNVLFFLTADHA    7
JCVI_PEP_1105103198155   (89)  EMAKAAAINEKMGKSQ--ETDFLAVSISSPDYAGHTFGPNAIELEDTYIKLDKDIEQFLNFLD-TQVGKGNVLLFLTADHG   47
JCVI_PEP_1105131869587   (89)  DFALAAIEGEKLGQRE--VTDFIAISFSSPDYIGHRFGPSSVEVEDNVIMLDRELAGFFTYLD-QKLGKGEVIVFISADHA   70
JCVI_PEP_1105075816231   (89)  EMALLLEKEELGEDE--FTDFLGVSFSSTDYAGHTWGVRSREVHDMYLKDRQLRQLIKALD-KKVGRDQYIIYLTADHG    26
JCVI_PEP_1112733327466   (89)  DMALLLLEKEELGKDK--YTDFLGVSFSSTDYAGHITFGVRSREVHDMYLKDRQLQLIKKLD-KEVGRDAVIVYLTADHG    12
JCVI_PEP_1105139982877   (89)  DMAELILDEEELGQDE--FTDFLGVSFSSPDVAGHTWGGSIARSTRYVFKTRCTVGSIYSYPKIKKVGDGSISSISQPDHG  203
PhoA_ACX6.71
Consensus             vφφσ...φφχ.EχφφGχχχ...TDφL.φφφSS.DφφφGH.φGP.SqEφ.SφφEφ.D.φφφχLDχχφχXφφφχζφφD.χ..φG.χXφφφφφLφφDHG
                                 Δ                        Δ                                                    ΔΔ
```

To elucidate the overall structure and the arrangement of key conserved residues in the protein, a model of molecular structure was predicted for PhoA$_{ACX6.71}$. The model suggests that the protein is globular in structure, composed of 13 α-helices and 19 β-sheets with a cleft formed in its inner core. Conserved residues previously demonstrated as involved in metal binding and catalysis in *Xanthomonas axonpodis* NPP (Zalatan *et al.*, 2006a) are shown to be located in the inner core of PhoA$_{ACX6.71}$. The conserved residues include six amino acids which may coordinate two Zn^{2+} ions (shown in blue, Figure 3-13), Thr 73 (shown in green, Figure 3-13) possibly having a key catalytic role, and Asn 94 (shown in red, Figure 3-13) believed to interact with one of the δ-oxygens of the phosphate group.

From the metagenome of municipal waste activated sludge, a new alkaline phosphatase, *phoA*$_{ACX6.71}$, was cloned and expressed. Alkaline phosphatases have an affiliation with the NPP family due the conservation of residues involved in metal binding and catalysis (Bollen *et al.*, 2000). With the exception of the experimentally determined PafA of *Chrysiobacterium meningosepticum* (Berlutti *et al.*, 2001), Ca^{2+} ATPase of *Myroides odoratus* (Peiffer *et al.*, 1996), PhoV of *Synechococcus sp.* PCC7942 (Wagner *et al.*, 1995) and alkaline phosphatase (PhoD) of *Zymomonas mobilis* (Gomez & Ingram, 1995), all of which have been demonstrated to encode the enzyme, all of the known phosphodiesterases used in the phylogenetic analysis are putative.

From the characterization of PhoA$_{ACX6.71}$, we suggest that the protein contains an N-terminal signal peptide with a primary structure consistent with that of bacterial signal peptides. Its optimal phosphatase activity at an alkaline pH is consistent with the optimal pH of 9 reported for the Pi-irrepressible alkaline phosphatase, PafA from *Chryseobacterium*

meningosepticum (Berlutti *et al.*, 2001), and is narrow in range compared to the pH range of 7-10 for the alkaline phosphatase PhoV of *Synechococcus sp*. PCC7942 (Wagner *et al.*, 1995). Phosphatase activity data, and the occurrence of an NPP conserved domain in PhoA$_{ACX6.71}$, provide evidence for the identification of a gene belonging to the newly described NPP family. Phylogenetic analysis shows that *phoA*$_{ACX6.71}$ appears in the same clade as phosphodiesterases that are found in members of the γ-*Proteobacteria* but, nonetheless, its amino acid sequence is divergent from these known phosphodiesterases as it only shares a 42% identity with its closest ortholog, the predicted alkaline phosphatase of *Colwellia psychrerythraea* 34-H.

Based on the previously determined molecular structure of the NPP of *X. axonpodis*, conserved aspartic acid and histidne residues that may be involved in the coordination of two bivalent zinc metal ions were identified. An asparagine residue that may bridge with the substrate's O atom via a H-bond and a conserved reaction centre threonine that may be part of the catalytic functioning of the enzyme were also identified. These residues have been experimentally shown to be involved in the binding of these metals and in forming the reaction centre in NPP from *X. axonpodis* (Zalatan *et al.*, 2006b), the Pi-irrepressible periplasmic alkaline phosphatase PafA from *C. meningosepticum* (Berlutti *et al.*, 2001) and the membrane-bound alkaline phosphatise, PhoD of *Zymomonas mobilis* ZM4 (Gomez & Ingram, 1995). Although key metal binding residues are conserved, overall, PhoA$_{ACX6.71}$ has a remarkably low amino acid sequence identity with known NPPs. Furthermore, the missing histidine residue from NPPs from uncultured bacteria, typically conserved in known bacterial

and mammalian NPPs is another example of the elusive diversity which cannot be accounted for when strictly employing culture dependent studies.

Here, novel phosphohydrolase genes from uncultured bacteria originating from pulp and from municipal waste activated sludge have been isolated. We provide here the first example of a functional analysis of phosphatases derived from uncultured bacteria by demonstrating their functionality by overexpression and partial purification as well as by a preliminary characterization at the protein sequence level.

3.3 Expression and biochemical characterization of phosphatase proteins

Of all putative NSAPs, alkaline phosphatases and NPPs included in the phylogenetic analysis, only a fraction of them have been demonstrated experimentally to be biochemically functional. In many cases, there are difficulties in demonstrating activity in newly identified enzymes due to the inability of the host to express these enzymes due to incompatible protein synthesis machinery, poor promoter recognition by host transcriptional machinery, or the toxicity of the protein product to the host cell. We show here the first successful attempt at demonstrating the partial purification and activity of phosphatases from the metagenome. To assess the functionality and to obtain protein in sufficient quantities, ORFs $phoN_{ACX6.13}$ and $phoA_{ACX6.71}$ were cloned into pET30b and pET30a, respectively and the proteins were overexpressed. ORF sequences of $phoN_{ACX6.13}$ and $phoA_{ACX6.71}$ lacking the C terminal stop codon, were amplified from pAR003 and pAR004, respectively and cloned as EcoRI and NotI fragments into dephosphorylated and digested pET30b and pET30a. The proteins generated contained a C-terminal His-tag. Induction of the proteins in an *E. coli* BL21 background was carried out in autoinduction medium. Protein expression was confirmed by

Western analysis, with the detection of a ~48.1 kDa protein for PhoN$_{ACX6.13}$ (Figure 3-14q). The size of the detected protein is much larger than the predicted 29.5 kDa and the size corresponding to the protein with its signal peptide (37 kDa). This suggests the possibility of the lack of cleavage of the signal peptide. The expression of PhoA$_{ACX6.71}$ resulted in the detection of two peptides of molecular weights of ~91 kDa and ~81 kDa (Figure 3-14b). The detection of two peptides may be attributed to degradation and removal of ~10 kDa by proteases. The apparent molecular weight was higher than the predicted ~60 kDa, again, possibly due to the lack of cleavage of the signal peptide.

The discrepancy between the expected molecular size of the enzyme and its estimated size of the detected protein may be attributed to inefficient binding of the protein to SDS, causing the protein to have a lower net negative charge than it would otherwise have, resulting in retardation of its migration on SDS polyacrylamide gel. Inefficient SDS binding may have occurred due to the significantly high composition (41% for PhoN$_{ACX6.13}$ and 44% for PhoA$_{ACX6.71}$) of polar residues. Another possible reason could be the high Lysine + Proline composition (10% for PhoN$_{ACX6.13}$ and 8% for PhoA$_{ACX6.71}$) which could increase the rigidity of the protein affecting its binding to the detergent (Kirkland *et al.*, 1998). The lack of cleavage of the signal peptide and the subsequent fusion of the protein with the N-terminal His tag may also account for the larger size of the detected protein.

To check the catalytic ability of the overexpressed proteins, phosphatase activity was assayed against the substrate pNPP. A cell extract of *E. coli* BL21 expressing pAR0033 (PhoN$_{ACX6.13}$ in pET30b) exhibited 641 units of activity (U), 320-fold higher than the activity

yielded by the cell extract of the negative control *E. coli* BL21 expressing pET30b. Similarly, cell extracts of *E. coli* BL21 expressing pAR0044 (PhoA$_{ACX6.71}$ in pET30a) exhibited 76.7U of activity, 240-fold higher than the activity yielded by the negative control cell extract of *E. coli* BL21 expressing pET30a. PhoA$_{ACX6.71}$ and PhoN$_{ACX6.13}$ were partially purified. 5.1-fold purification was achieved for the class A NSAP PhoN$_{ACX6.13}$ while a 6.2-fold purification was achieved for alkaline phosphatase PhoA$_{ACX6.71}$.

Figure 3-13: Predicted molecular structure of phosphatase PhoA$_{ACX6.71}$.

Conserved residues involved in Zn^{2+} coordination are shown in blue. Reaction centre Threonine is shown in green. Asn 94, which may be involved in H-bonding with a phosphoryl oxygen in the phosphate group. The structure was predicted based on the template structure of the NPP from *X. axonpodis* (Zalatan et al., 2006b).
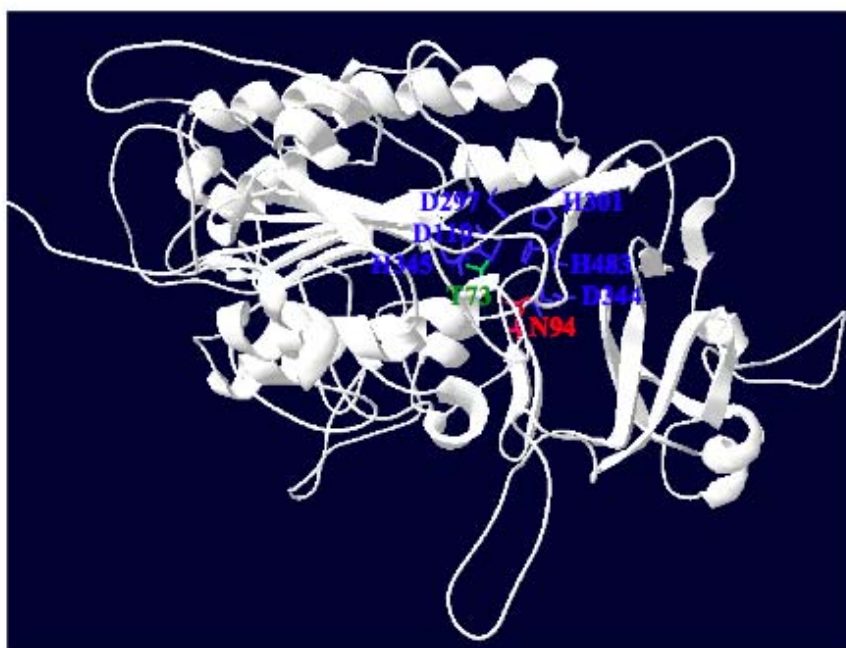
Figure 3-14a: Expression of protein PhoN$_{ACX6.13}$ as detected on a Western blot.

Lane 1, *E. coli* BL21 pAR0033 (pET30b with PhoNACX6.13) protein extract; Lane 2, *E. coli* BL21 pET30b protein extract.

Protein marker size
(kDa)

170
130
100

70
55

40

35

25
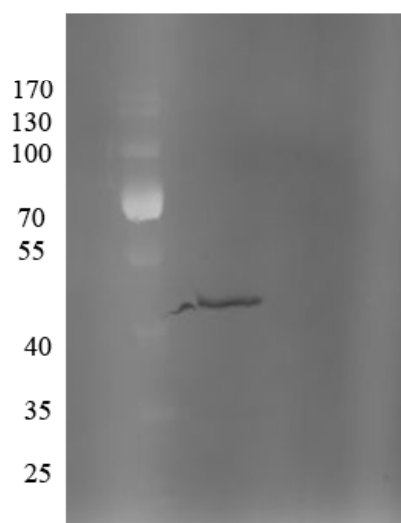
1   2

Figure 3-14b: Expression of protein PhoA$_{ACX6.71}$ as detected on a Western blot.

Lane 1, *E. coli* BL21 pAR0045 (pET30a with PhoA$_{ACX6.71}$) protein extract; Lane 2, *E. coli* BL21 pET30a protein extract.

Protein marker size (kDa)

170
130
100
70

55

40

35

25

1   2
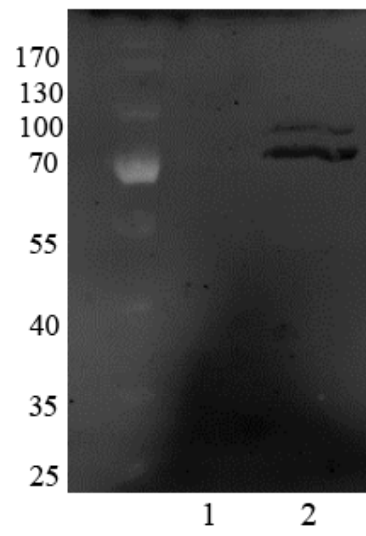
Chapter 4: Partial reconstruction of soil and activated sluge metagenomes by phenotypic complementation of phosphorous metabolism-deficient *Sinorhizobium meliloti* mutants

4.1 Isolation, identification and sequencing of library cosmids

The alfalfa symbiont, *Sinorhizobium meliloti* was used as a host for screening metagenomic libraries for phosphonate metabolism genes. The genome of *S. meliloti* is encoded on a chromosomal element and two megaplasmids pSymA and pSymB. To identify the gene loci required for *S. meliloti* growth on phosphonate, various existing mutants were screened for their inability to utilize glyphosate as the sole P source (Gly⁻). Six *S. meliloti* mutant strains were found to have a Gly⁻ phenotype (Table 2-1), due to a deficiency in genes required for either transport or degradation of phosphonate. Two of these mutants, RmF726 and RmG471, have large deletions in the pSymB megaplasmid. RmF726 lacks a 220 kb region that includes the *phnGHIJKL* operon for C-P lyase. RmG471 lacks a 84 kb region including the *phnM* and *phoCDET* operons (Charles & Finan, 1991). Other mutants, with disruptions in up to four genes were also screened. RmG439 (*phoCDET*), RmG490 (*phoC*), RmG491 (*phoT*), and the double mutant RmG830 (*phoCpit*) were Gly⁻. Wild type *S. meliloti* strains Rm1021, RmP110 and mutants RmH838 (*phoB*::TnV), RmP636 (RmP110 *pit310*::tn5 *phoC*490) and RmG762 (*phoC*490 *pit310*::Tn5) had a Gly⁺ phenotype.

To isolate cosmids encoding phosphonate metabolism genes, metagenomic libraries CX4 (pulp waste activated sludge), CX6 (municipal waste activated sludge) and CX9 (organic rich sandy loam surface soil) were screened by phenotypic selection of clones conferring phosphonate growth upon RmF726 and RmG471. By transferring each library to

each mutant, via triparental mating, and selecting for the ability of the transconjugants to utilize glyphosate as the sole P source, a total of eight unique cosmids were isolated (Table 4-1). Four cosmids, pCX4-10F, pCX6-13F, pCX9-45F and pCX6-8G, representing each of the metagenomes used in this study, complemented both RmF726 and RmG471, while the remaining cosmids, pCX4-3G, pCX6-9G, pCX9-2G and pCX9-12G, complemented only RmG471. The ability of some cosmids to complement both mutants reveals some information about the genomic arrangement of the source organism. It is likely that transport and degradation genes are located within 40 kb in the metagenome, an arrangement quite different than in *S. meliloti* where the two loci are separated by at least 500 kb.

The function-driven approach to metagenomics was used to further investigate this feature and to explore the regions flanking P metabolism genes. This allowed the examination of functional diversity in soil and wastewater treatment plant activated sludge-inhabiting organisms. To study features of activated sludge and soil metagenomes and to identify the gene loci required for growth on glyphosate, a major portion of cosmids pCX4-10F, pCX6-13F and pCX9-45F was sequenced. The above cosmids were chosen due to their ability to complement both Gly$^-$ mutants, in loci required for transport and degradation. Small insert libraries were constructed from each cosmid and sequence contigs were generated.

Instead of sequencing the entire cosmids, the loci required for growth on glyphosate were identified and sequenced in cosmids pCX4-3G and pCX6-8G by constructing subclone libraries and selecting subclones pAR010 (pCX4-3G) and pAR011 (pCX6-8G), for their ability to complement RmG471. *In vitro* transposon mutagenesis was performed on pAR010

129

and pAR011 and the sequence of the loci required for glyphosate growth was assembled from insertion clones with abolished ability to complement RmG471.

A total of 92,896 bp of sequence, consisting of 92 protein coding sequences (CDS) 83 of which are complete open reading frames (ORFs) were assembled from 5 cosmids as follows: 23,833 bp from cosmid pCX4-10F, 34,486 bp from pCX6-13F and 27,695 bp from pCX9-45F, 1,746 bp from pCX4-3G (subcloned in pAR010) and 5,129 bp from pCX6-8G (subcloned in pAR011). It should be noted that the each cosmid contains about 40,000 bp of metagenomic DNA. Of the total number of ORFs identified 70% are predicted to be involved in specific cellular functions while only a general function was predicted for 16% of the ORFs. The function associated with the remaining 14% of the ORFs is unknown. A large portion of the ORFs (25%) encode functions involving phosphorous transport and metabolism (Figure 4-1).

Table 4-1: Summary of metagenomic clones complementing *S. meliloti* **mutants RmF726 and RmG471**

Library	Source	No of unique clones from RmF726 complementation	No of unique clones from RmG471 complementation
CX4	Pulp waste activated sludge	1	1
CX6	Municipal waste activated sludge	1	2
CX9	Soil	1	2

Figure 4-1: Functional distribution of ORFs identified in cosmids pCX4-10F, pCX6-13F, pCX9-45F, pCX4-3G and pCX6-8G.
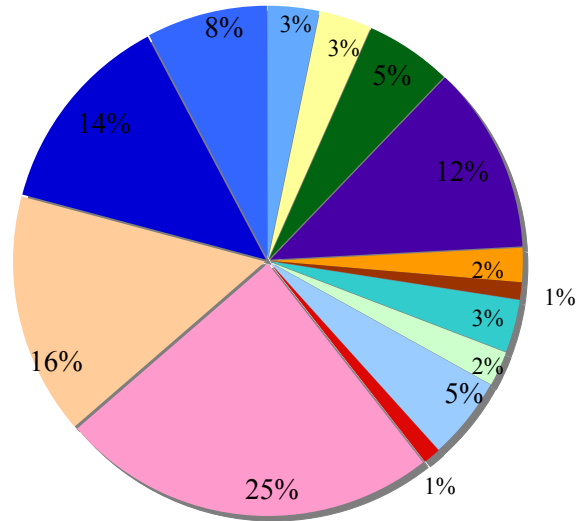
Translation, ribosomal structure and biogenesis

Replication, recombination and repair

Defense mechanisms

Signal transduction and regulatory mechanisms

Cell wall/membrane/envelope biogenesis

Posttranslational modification, protein turnover, chaperones

Carbohydrate transport and metabolism

Amino acid transport and metabolism

Nucleotide transport and metabolism

Lipid transport and metabolism

Phosphorous transport and metabolism

General function prediction only

Unknown Function

Nitrogen metabolism

4.2 Characterization of phosphate/phosphonate transport genes

The assembled metagenomes harbour many genes predicted to be involved in a wide variety of cellular functions, 25% of which are phosphate transport and metabolism (Figure 4-1). The Gly⁻ phenotype shown by all mutants, except for RmF726, is apparently associated with the knock out or disruption of the *phoCDET* operon. ORFs identified from pAR010 and pAR011 are putative operons for phosphate transport. 1747 bp of sequence were assembled for pAR010. GenBank searches revealed two ORFs, *orfA*$_{pAR010}$ and *pitA*$_{pAR010}$ separated by 11 bp (Figure 4-2a). The gene *orfA*$_{pAR010}$ encodes a 215-amino acid protein, homologous to a number of *Rhizobiacae* family pit accessory proteins such as *Rhizobium sp*. NGR234, accession no YP_002827978, *Sinorhizobium meliloti* SMc02862, accession no NP_384293, and *Sinorhizobium medicae*, accession no YP_001329050. Due to incomplete sequencing of pAR010, the second ORF, *pitA*$_{pAR010}$, was only sequenced up to 1014 bp. The ORF is homologous to inorganic phosphate transporter (pit) proteins of a number of organisms such as *Rhizobium etli* CFN 42, accession no YP_468484, *Phenylobacterium zucineum* HLK1, accession no YP_002132152, and *Rhizobium leguminosarum bv*. Trifolii WSM1325, accession no ZP_02293977. The G+C contents of *orfA*$_{pAR010}$ and *pitA*$_{pAR010}$ genes are 60.5% and 66%, respectively. A strong bias for G or C at the third position of each codon was shown for both genes at 87.4% and 81.6% for *orfA*$_{pAR010}$ and *pitA*$_{pAR010}$, respectively. The strong G+C bias in the third position of each codon and the overall G+C in the two genes is consistent with the G+C content and codon bias displayed across the entire *S. meliloti* genome (Muto & Osawa, 1987). This suggests that *orfA*$_{pAR010}$ and *pitA*$_{pAR010}$ were expressed in *S. meliloti*. Potential ribosome binding sites (RBS) were found upstream to the translation
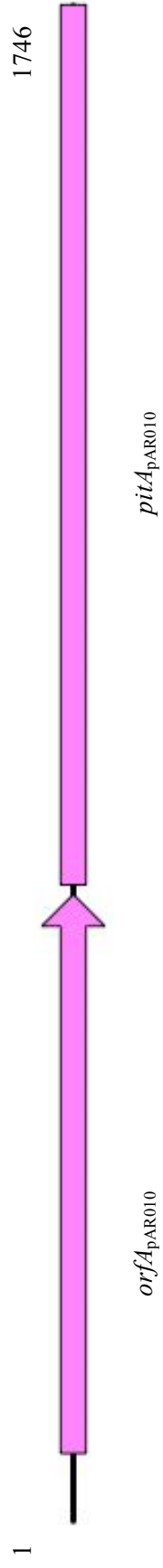
start sites as follows: GGGA-N$_4$-ATG for $orfA_{pAR010}$ and GAGGA-N$_{53}$-ATG for $pitA_{pAR011}$. The region upstream to $orfA_{pAR010}$ did not have a "pho box" suggesting this transport system may be regulated in a PhoB-independent manner.

The nucleotide sequence of the gly locus of pAR011 revealed 1 partial and 5 complete ORFs (Figure 4-2b). The 339 bp partial reading frame, Orf13$_{pAR011}$, is comprised of the 113 amino acid C terminal portion of a putative phosphoglucosamine mutase, an enzyme involved in the interconversion of glucosamine-6-phosphate and glucosamine-1-phosphate in the biosynthetic pathway of UDP-N-acetylglucosamine, an essential precursor to the components of the cell membrane. A BLAST X search (Altschul *et al.*, 1997), revealed homology between the C-terminal part of Orf13$_{pAR011}$ to that of members of the *Burkholderiales* family including *Leptothrix cholodnii* SP-6, accession no YP_001791845.1, the iron reducing *Rhodoferax ferrireducens* T118, accession no YP_523269.1 and *Polaromonas naphthalenivorans* CJ2, accession no YP_982832.1.
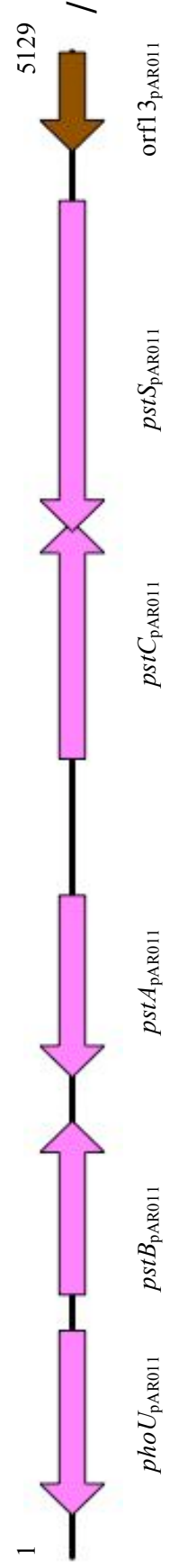
Figure 4-2: Gene map for *gly* locus of cosmids pAR010 (a) and pAR011 (b).

P metabolism genes are illustrated in pink. Orf13$_{pAR011}$ is a putative phosphoglucosamine mutase, belonging to functional category "cell wall/membrane/envelope biogenesis" (brown)

a)

1746

$pitA_{pAR010}$

$orfA_{pAR010}$

1

b)

5129

orf13$_{pAR011}$

$pstS_{pAR011}$

$pstC_{pAR011}$

$pstA_{pAR011}$

$pstB_{pAR011}$

$phoU_{pAR011}$

1

137

The 5 ORFs downstream from the glucosamine mutase gene encode a *pstSCAB-phoU* operon where (*pstSCAB*)$_{pAR011}$ encodes a phosphate transport system and *phoU*$_{pAR011}$ encodes an uptake regulator. A BLAST X search of the 5 ORFs comprising the *pstSCAB-phoU* genes for phosphate transport revealed homology of 3 of the genes to the *pst* gene cluster of the methyl tert-butyl ether-degrading *Beta-Proteobacterium Methylibium petroleiphilum* PM1 (Kane *et al.*, 2007) as follows: PstS$_{pAR011}$ (77% a.a. identity), PstA$_{pAR011}$ (78% a.a. identity) PstB$_{pAR011}$ (90% a.a. identity). PstC$_{pAR011}$, on the other hand showed an amino acid identity of 90% to the PstC protein of the *β-Proteobacterium Variovorax paradoxus* S110 and PhoU$_{pAR011}$ showed an amino acid identity of 85% to the PhoU of the *β-Proteobacterium Leptothrix cholodnii* SP6. Potential RBSs were found upstream of the translation start site of each gene as follows: GAAGG(N7)ATG upstream of *pstS*$_{pAR011}$, GGAAGG(N6)ATG upstream of *pstC*$_{pAR011}$, GAAG(N38)ATG upstream of *pstA*$_{pAR011}$, AAGGA(N6)ATG upstream of *pstB*$_{pAR011}$ and AAGGA(N7)ATG upstream of *phoU*$_{pAR011}$. A putative "pho box" was found 76 bp upstream to the translation start site of *pstS*$_{pAR011}$ (Figure 4-3), suggesting the transport system may be regulated in a PhoB-dependent manner.

Unexpectedly, no PhoCDET transport systems were identified in either pAR010 or pAR011. Alternatively, Pit or Pst transport systems of metagenomic origin complement the *gly* growth phenotype of *Sinorhizobium meliloti*. To verify that the *orfA-pit* and *pstSCAB-phoU* systems complement the growth deficiency of *S. meliloti* on glyphosate in the absence of a functional PhoCDET or Pit transport systems, the ability of pAR010 (*orfA-pit*) and pAR011 (*pstSCAB*) to complement a number of *phoCDET* and *pit* mutants was assessed. In addition to complementing RmG471, the *orfA*$_{pAR010}$-*pitA*$_{pAR010}$ operon encoded on pAR010

was found to complement a *pit-phoC* double mutant for growth on glyphosate. It also complemented *phoC* (RmG490) and *phoT* (RmG491) single mutants, and *phoCDET* mutant, RmG439. The ability of *orfA*$_{pAR010}$-*pitA*$_{pAR010}$ to complement the *phoCDET* mutant, RmG439, single mutants RmG490 (*phoC*), and RmG491 (*phoT*), and the *phoCpit* double mutant RmG830, suggests that the operon is rescuing the ability of *phoCDET*-deficient *S. meliloti* to grow on glyphosate as the sole P source, thus serving a dual role as a phosphonate and phosphate transporter. The presence of more than one P transport system in bacteria as in the case of *E. coli* (Willsky & Malamy, 1980), *Acinetobacter johnsonii* (van Veen *et al.*, 1993), *Bacillus cereus* (Rosenberg & La Nauze, 1967) and *S. meliloti* (Bardin *et al.*, 1998) is not unusual since bacteria adjust to variable P conditions and availability by employing transport systems that vary in efficiency, specificity and affinity.

Figure 4-3: N-terminal and region upstream to ORF *pstS*_{pAR011}.

Predicted "pho box" is highlighted in purple. Underlined sequence denotes the first nucleotides of the ORF.

AGGTTGTCATTGGGTTGTCACATGGCCTTTCTAAAGTGACGCACGTCACGGGTTTCCCC

GTTGTCTGCGTCTTGACGAAAGGTTCAACCATGCTGTTTTCCATC

                                                 M   L   F   S   I

Complementation tests demonstrate that OrfA$_{pAR010}$-PitA$_{pAR010}$ may be functioning as a replacement transport system in the absence of a functional PhoCDET transport system. A survey of the region upstream to *orfA*$_{pAR010}$ did not reveal evidence of regulatory mechanism equivalent to those of the *S. meliloti* OrfA-Pit system. The failure to detect a "pho box" upstream to *orfA*$_{pAR010}$ does not agree with the finding that a pho box is present upstream to the *S. meliloti orfA-pit* operon and that PhoB acts as a negative regulator of the OrfA-Pit transport system in *S. meliloti* (Yuan *et al.*, 2006b). The failure to detect a PhoB binding site upstream to *orfA*$_{pAR010}$ suggests that *orfA*$_{pAR010}$-*pitA*$_{pAR010}$ may be regulated by the external phosphate concentration in a PhoB-independent manner as was shown with the polyphosphate kinase (*ppk*) gene of *S. meliloti* (Yuan *et al.*, 2006b). The failure to detect a "pho box" could also be due to an insufficient amount of DNA sequence determination in that region as "pho boxes" may occur between -500 and +100 nucleotides relative to the translation start site (Yuan *et al.*, 2006b).

A second P transport system and a downstream uptake regulator *pstSCAB-phoU* were also able to complement *gly⁻* growth phenotypes of various *S. meliloti* mutants. Contrary to *orfA*$_{pAR010}$-*pitA*$_{pAR010}$, the pAR011-encoded *pstSCAB-phoU,* isolated from a municipal waste activated sludge metagenome, was not able to complement the *phoC-pit* double mutant, RmG830. However, growth of each of the *phoC* single mutant, RmG490, the *phoT* single mutant, RmG491 and *phoCDET* mutant, RmG439 on glyphosate was complemented by pAR011. These results show the new Pst transport system is used to take up phosphonate in the absence of the PhoCDET system but cannot substitute for the deficiency of the Pit system.

The identification of a "pho box" upstream of $pstS_{pAR011}$ is not unusual. It suggests the regulation of Pst expression is dependent on PhoB, however, additional experiments such as electrophoretic mobility assays are required to confirm this dependence. PhoB dependent regulation of *pstSCABphoU* encoded on pAR011 by external phosphate concentration is possible as the availability of glyphosate as the sole P source is equivalent to Pi free medium, which is a condition for PhoB-dependent activation of the Pst system. Such activation was demonstrated in *S. meliloti* where two "pho boxes" were present upstream to *pstS* and PhoB was shown to be required for phosphate dependent regulation of the Pst transport system (Yuan *et al.*, 2006a).

The finding that the Gly⁻ phenotype of RmG471 was complemented by the *pstSCAB-phoU* genes encoded on pAR011 is rather unusual because like the Pit transport system, there are no reports of phosphonate transport via the Pst system, as it is designed to specifically take up Pi. In fact, it was demonstrated in *S. meliloti* that the PstSCAB transport system is not competitively inhibited by phosphonate (Yuan *et al.*, 2006a). Furthermore, RmG471 contains a *pstSCAB* in the chromosome but this conventional Pi uptake system cannot substitute for PhoCDET in the uptake of phosphonate while $(pstSCAB\text{-}phoU)_{pAR011}$ appears to have a broader substrate specificity.

Aside from $(pstSCAB\text{-}phoU)_{pAR011}$ additional Pst transport systems were found in soil and sludge metagenomes, appearing with the regulatory protein PhoU. From the metagenomes of soil and activated sludge, two *pstS* ORFs, four each of *pstC, pstA* and *pstB*, and two *phoU* ORFs were identified. Since the uptake of glyphosate by the Pst system is

proposed, the sequences of a number of identified metagenomic PstS proteins were examined and compared with known bacterial PstS. To analyse the substrate binding of PstS, amino acid sequences of pulp waste activated sludge derived, $PstS_{pAR011}$ and $PstS_{pCX4-10F}$ were aligned with PstS proteins of the plague causing γ-*Proteobacterium, Yersinia pestis* (YpPstS), (Tanabe *et al.*, 2007), *E. coli* (EcPBP), (Yao *et al.*, 1996) and *Mycobacterium tuberculosis*, (MtPstS) (Vyas *et al.*, 2003), for which the structures have been determined (Figure 4-4). $PstS_{pCX4-10F}$ and $PstS_{pAR011}$ have a mutual amino acid identity of 60% and a similarity of 76%. $PstS_{pAR011}$was found to have a 59% amino acid identity and 74% similarity to *Y. pestis* PstS while $PstS_{pCX4-10F}$ was found to have a 55% identity and 72% similarity to *Y. pestis* PstS. A lower amino acid identity was observed between the two activated sludge derived PstS proteins and MtPstS of *Mycobacterium tuberculosis* (35% identity for $PstS_{pAR011}$ and $PstS_{pCX4-10F}$ with MtPstS). Aside from the overall sequence identity, $PstS_{pAR011}$ and $PstS_{pCX4-10F}$ differ from MtPstS in that Ser 35 from which the OH-group is a H-bond donor, thus involved in substrate binding in MtPstS is replaced with Ala in $PstS_{pAR011}$ and $PstS_{pCX4-10F}$. The occurrence of Ala in place of a well-conserved Ser residue eliminates one structure stabilizing H-bond, resulting in a change in the folding of the protein that allows the binding of a variety of substrates. Furthermore, Asp 168, one of the four residues involved in pairs of H-bonds with the substrate in MtPstS is replaced with Thr in $PstS_{pAR011}$ and $PstS_{pCX4-10F}$ (Figure 4-4). This change may also allow less rigidity in the folding such that other substrates may bind to the protein.

Figure 4-4: Amino acid sequence alignment of PstS proteins.

Serine and aspartic acid residues demonstrated to be important for catalysis in *Mycobacterium tuberculosis* PstS1 are highlighted red and blue, respectively. Taxon names are denoted as follows: YpPstS: *Yersinia pestis* PstS; Ecpbp: *E. coli* periplasmic binding protein; MtPstS: *Mycobacterium tuberculosis* PstS1. * denotes invariant conserved residues

```
pAR011pstS          22 LAQDVTGAGATFPAPLYAKWADAYNKS-TGVKMNYQSVGSGAGLKQIRSKTVDFGASDMPLKDED
pCX4-10FpstS        23 QAADITGAGATFPYPAYAKWAEAYKAQ-GGAALNYQSIGSGGGIKQIRAKTVDFGASDMPLKPEE
YpPstS  gi|218931093|  25 AAASLTGAGATFPAPVYAKWADSYQKE-TGNKINYQGIGSSGGVKQIIANTVDFGASDAPLTDEK
EcPBP   gi|89110279|   25 AEASLTGAGATFPAPVYAKWADTYQKE-TGNKVNYQGIGSSGGVKQIIANTVDFGASDAPLSDEK
MtPstS  gi|41353623|   26 SPVTLAETGSTLLYPLFNLWGPAFHERYPNVTITAQGTGSGAGIAQAAAGTVNIGASDAYLSEGD
                           *  *    *       *          *  **  *  *    **  **** *


pAR011pstS             LAK-DGLVQFPTVIGGVVPVVNIAGIQPGQIKLTGQVLGDIYLGKITKWSDPALAALNPGLKLPDADI
pCX4-10FpstS           LKA-DGMIQFPAIIGGVVPVLNVEGLQPGQLKLTGQLLADIYLGKVKKWDDAPIAALNPGVKLPSTQI
YpPstS  gi|218931093|  LAT-EGLFQFPTVIGGVVLAVNIPGIKSGELTLDGKTLGDIYLGTVKKWNDPAIVKLNPGVKLPDQNI
EcPBP   gi|89110279|   LAQ-EGLFQFPTVIGGVVLAVNIPGLKSGELVLDGKTLGDIYLGKIKKWDDEAIAKLNPGLKLPSQNI
MtPstS  gi|41353623|   MAAHKGLMNIALAISAQQVNYNLPGV-SEHLKLNGKVLAAMYQGTIKTWDDPQIAALNPGVNLPGTAV
                          .*        *          *  *      *  *  * *      *  *         ****  **


pAR011pstS             SPVRRADGSGITFIFTNYLSKVNPD-WKAKVGEGTAVNWPT---GAGGKGNEGVAAFVQRLPNSIGY
pCX4-10FpstS           AVVRRSDGSGISFLFTTYLSQVSAE-WKEKVGAGTAVSWPA---GVGGKGNEGVASYVQSIKGSIGY
YpPstS  gi|218931093|  AVVRRADGSGISFVFTSYLAKVNAE-WKEKVGAGSTVNWPT---GLGGKGNDGIAAFVQRLPGSIGY
EcPBP   gi|89110279|   AVVRRADGSGISFVFTSYLAKVNEE-WKNNVGTGSTVKWPI---GLGGKGNDGIAAFVQRLPGAIGY
MtPstS  gi|41353623|   VPLHRSDGSGITFLFTQYLSKQDPEGWGKSPGFGTTVDFPAVPGALGENGNGGMVTGCAETPGCVAY
                          * **** * ** **         *     * * * *      * ** *            *


pAR011pstS             LEYAY---VKQNKLSYTLMKNRDGNFVPPSDTAFKAAAAGADW--KSSFYQITTDQPGKD
pCX4-10FpstS           VEYAY---AKRNKMTYALLKNKDGNFVAPDDETFAAAAASADWAKAPAFGEVLTDEPGKN
YpPstS  gi|218931093|  VEYAY---AKQNNLAYTKLISADGKPVSPTEHSFSSAAKGVDW--SKSFAQDLTNQKGDD
EcPBP   gi|89110279|   VEYAY---AKQNNLAYTKLISADGKPVSPTEENFANAAKGADW--SKTFAQDLTNQKGED
MtPstS  gi|41353623|   IGISFLDQASQRGLGEAQLGNSSGNFLLPDAQSIQAAAAGFASKTPANQAISMIDGPAPD
                           *       *          **


pAR011pstS             AWPITNPTYIMMHKVQEKPAQATAAMKFFDWAYGNGDKMA--DDLDYVPL 318
pCX4-10FpstS           SWPITGASFILMHANPEKPERTLEVLKFFSWAFKNGGKLA--TELDYVPM 321
YpPstS  gi|218931093|  VWPITSTTFILVHKEQKNAANGTEVLKFFDWGYTHGAKQA--NELDYATL 321
EcPBP   gi|89110279|   AWPITSTTFILIHKDQKKPEQGTEVLKFFDWAYKTGAKQA--NDLDYASL 321
MtPstS  gi|41353623|   GYPIINYEYAIVNNRQKDAATAQTLQAFLHWAITDGNKASFLDQVHFQPL 357
                          **                     *   *    * *
```

PstA and PstC are transmembrane proteins, usually hydrophobic, and are essentially the permease components of the Pst phosphate transport system. From the metagenome of municipal waste activated sludge, two of each of *pstA* (*pstA*$_{pCX6-13F}$ and *pstA*$_{pAR011}$) and *pstC* (*pstC*$_{pCX6-13F}$ and *pstC*$_{pAR011}$) were identified while one of each *pstA* and *pstC* were identified from the metagenomes of pulp waste activated sludge (*pstA*$_{pCX4-10F}$ and *pstC*$_{pCX4-10F}$) and soil (*pstA*$_{pCX9-45F}$ and *pstC*$_{pCX9-45F}$). As previously illustrated in *E. coli*, a functional PstC permease requires charged residues Arg 237 and Glu 241 in one of its transmembrane helices for phosphate transport (Cox *et al.*, 1989). All PstC proteins identified in this study contain the equivalent of Arg 237 and Glu 241 with the exception of PstC$_{pCX9-45F}$ which has valine and leucine residues in place of the respective arginine and glutamate residues required for phosphate transport while PstC$_{pAR011}$ has a glycine instead of the required glutamine. Similarly it was shown in *E. coli* that in order to be functional, one of the transmembrane helices of the PstA protein requires Arg 220 (Cox *et al.*, 1989). Each of the metagenomic PstA have the required arginine residue, equivalent to Arg 220 of the *E. coli* PstA.

PstB is a hydrophilic cytoplasmic protein which assists in phosphate transport by being associated with the inner membrane proteins PstC and PstA and by binding and hydrolyzing ATP to drive the energy dependent transport process. Four *pstB* ORFs, *pstB*$_{pCX4-10F}$, *pstB*$_{pCX6-13F}$, *pstB*$_{pCX9-45F}$ and *pstB*$_{pAR011}$ were identified from the pulp and municipal waste activated sludge and soil metagenomes. Each of the four *pstB* ORFs is homologous to the *E. coli pstB* and all but *pstB*$_{pAR011}$ contain the conserved glycine and lysine residues, equivalent to Gly 48 and Lys 49, required for phosphate transport (Cox *et al.*, 1989). In

pstB$_{pAR011}$ the equivalent of Gly 48 and Lys 49 are replaced with alanine and serine, respectively.

PhoU is a negative regulator of the Pst transport system as it was demonstrated in *E. coli phoU* mutants that when intracellular levels of Pi are sufficient, the elevated levels of expression of the PstSCAB2 transport system occur and toxic levels of phosphate accumulate in the cell (Rice *et al.*, 2009). From the metagenome of pulp waste activated sludge and soil, three new homologues of the *phoU* gene, *phoU*$_{pCX4-10F}$, *phoU*$_{pAR011}$ and *phoU*$_{pCX9-45F}$ were identified (Figures 4-6, 4-2b and 4-8). PhoU$_{pAR011}$ shares 85% amino acid identity with the putative PhoU from the *β-proteobacterium Leptothrix cholodnii* SP-6. However PhoU$_{pAR011}$ shares only 52% amino acid identity over 100 N-terminal residues with experimentally determined PhoU protein from *E. coli* (EcPhoU) and shares an even smaller amino acid identity with *Aquifex aeolicus* (PhoUAQUE) (36% over 93 N-terminal residues) and *Thermotoga maritima* MSB8 (Tm1743) (16% over 93 N-terminal residues), for which protein molecular structure data is available (Liu *et al.*, 2005; Oganesyan *et al.*, 2005). The nearest homologue to PhoU$_{pCX4-10F}$ is the putative PhoU from the *β-proteobacterium Lutiella nitroferrum* 2002 at 42% amino acid identity while compared to EcPhoU an even lower amino acid identity is observed at 32% across the entire length of the protein. Compared to the structurally determined PhoUAQUE and Tm1743, PhoU$_{pCX4-10F}$ showed an amino acid identity of 25% and 24%, respectively. Similarly, PhoU$_{pCX9-45F}$ also shares a 41% amino acid identity with its nearest homologue, the *β-proteobacterium Dechlormonas aromatica* RCB while 20% amino acid identity was observed against Tm1743 and no significant amino acid identity with EcPhoU and PhoUAQUE. These results suggest the regulatory mechanism of
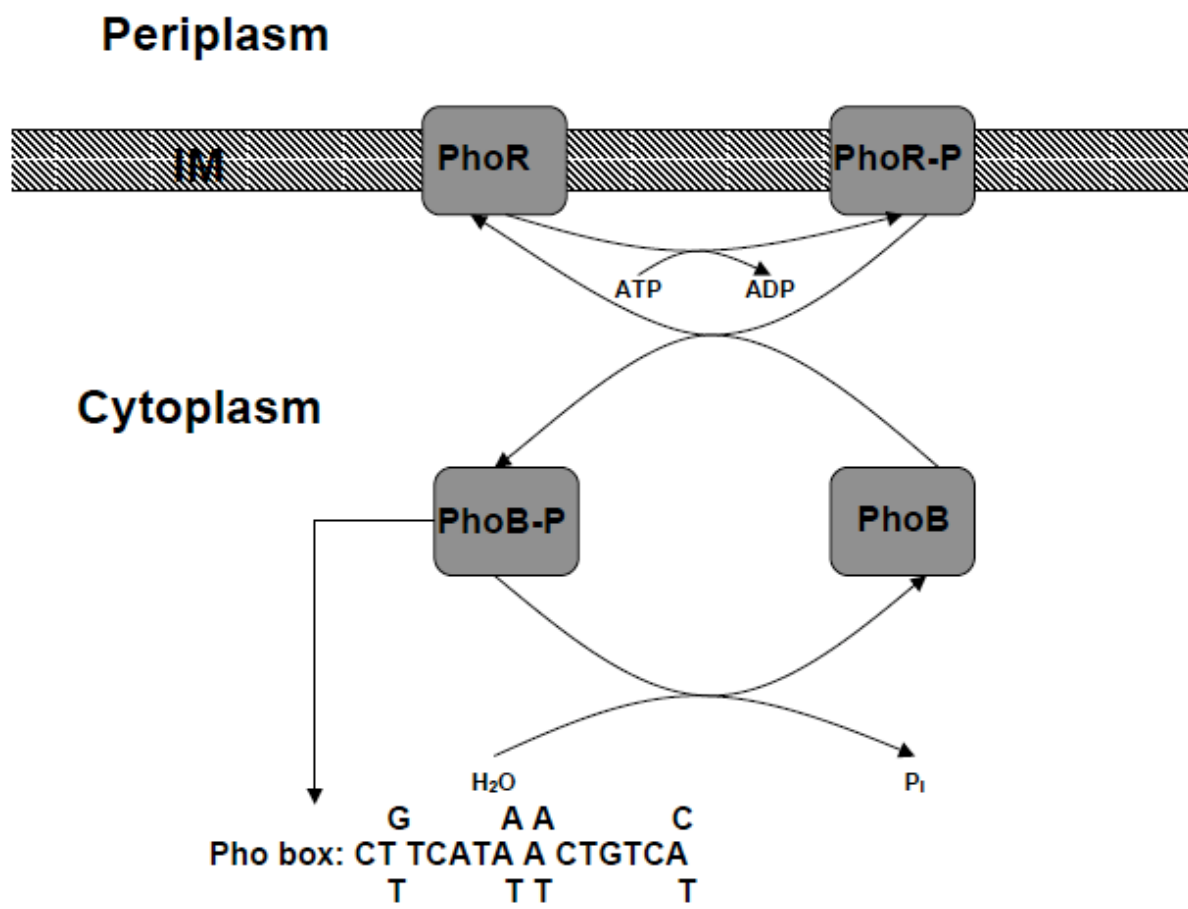
148

phosphate transport by PhoU$_{pAR011}$, PhoU$_{pCX4-10F}$ and PhoU$_{pCX9-45F}$ differs from the known regulatory mechanism demonstrated in EcPhoU, PhoUAQUE and Tm1743.

A *phoB/phoR* two-component regulatory system controls a large number of genes that are part of the phosphate (*pho*) regulon. Genes of *pho* regulon are controlled by the concerted action of the PhoR and PhoB proteins. When the external Pi concentration is limiting, PhoR is autophosphorylated and subsequently transfers its phosphoryl group to the response regulator PhoB which binds to conserved DNA sequences ("pho-boxes") in the promoter region of genes under the control of the *pho* regulon (Figure 4-5). A putative two component system *phoR*$_{pCX4-10F}$-*phoB*$_{pCX4-10F}$ was identified from pulp waste activated sludge metagenome (Figure 4-6). The genes *phoB*$_{pCX4-10F}$ and *phoR*$_{pCX4-10F}$ are found as an operon in which the stop codon of *phoB*$_{pCX4-10F}$ overlaps with the start codon of *phoR*$_{pCX4-10F}$. The response regulator, *phoB*$_{pCX4-10F}$ has an amino acid identity of 71% with the putative *phoB* of the *β-proteobacterium Laribacter hongkongensis* HLHK9 (accession no YP_002794170.1) while *phoR*$_{pCX4-10F}$ shares a lower amino acid identity (44%) with the putative *phoR* of *β-proteobacterium Dechloromonas aromatica* RCB (accession no YP_286767.1).

The inability of RmF726 to grow on glyphosate is likely due to the deletion of the *phnGHIJKL* operon for C-P lyase. Previous work demonstrates the requirement of the *phn* operon by *S. meliloti* to utilize phosphonates as the sole P source (Parker *et al.*, 1999). Surprisingly, no C-P lyase or similar genes for phosphonate degradation were identified in any sequenced portions of the cloned metagenome. There are two possible explanations for this finding. The first is the possibility that gene loci other than *phnGHIJKL* are required to restore the Gly$^+$ phenotype of RmF726.

Figure 4-5: Schematic illustration of bacterial PhoB/PhoR two component regulatory system.

When the external Pi concentration is limiting, PhoR is autophosphorylated and then phosphorylates PhoB which in turn binds to PhoB-binding DNA sequences or pho boxes, activating and in some cases repressing the transcription of pho regulon genes. When the external concentration is sufficient, PhoR remains phosphorylated and PhoB is dephosphorylated, rendering it inactive.

Periplasm

IM   PhoR   PhoR-P

ATP   ADP

Cytoplasm

PhoB-P   PhoB

H₂O   Pᵢ

Pho box:
```
        G         A A         C
    CT TCATA A CTGTCA
        T         T T         T
```

The locus responsible for phosphonate degradation may reside in one or a group of genes for which the function could not be predicted by homology searches. Second, since not all of the metagenome encoded in pCX4-10F, pCX6-13F and pCX9-45F was sequenced, the complementing region was not identified.

Bacteria inhabiting waste treatment plants are constantly exposed to a variety of organic P sources, requiring the ability to efficiently take up the readily available forms of P to survive. To increase uptake efficiency, activated sludge microorganisms evolved P transporters that take up more than one type of P. To confirm this hypothesis, mutations can be made in the various pst systems found in the metagenome, such that only one of the transport systems is functional. The ability to take up multiple forms of phosphates can be easily assessed by performing transport assays. Aside from *S. meliloti* PhoCDET, there aren't any known dual-role P transporters from bacterial isolates. This illustrates the utility of the metagenomic approach in identifying novel functions.

4.3 Additional functions in soil and activated sludge metagenomes

Phenotypic selection employed here resulted in the recovery of a large number of genes putatively involved in P transport and some regulatory proteins related to P metabolism. Aside from P metabolism, the metagenomes of activated sludge and soil carries an extensive complement of genes with regulatory and housekeeping roles (DNA replication, transcription, translation), cellular defense mechanisms, nutrient metabolism, vitamin biosynthesis and membrane structural elements.

It is not surprising that most of the genes identified have low amino acid identity to genes from known bacterial isolates. This illustrates the high genetic diversity that characterizes activated sludge and soil habitats. It also shows the databases available to date that are based on cultured microbes, are fairly limited in the number of available gene sequences and do not reflect the genetic diversity in such complex microbial communities.

4.3.1 Signal transduction and regulatory mechanisms

Approximately 12% of the ORFs may be involved in various regulatory and signal transduction mechanisms, including two component regulatory systems, c-AMP and c–diGMP-mediated signal transduction systems, a member of the deoR family of transcriptional regulators and a nitrogenase enzyme complex repressor (Table 4-2). The putative two-component regulatory systems include a PAS/PAC sensor hybrid histidine kinase, identified from pulp waste activated sludge, by homology to a putative PAS/PAC sensor hybrid histidine kinase *Methylobacterium nodulans ORS 2060* (accession no YP_002499709.1) at 37% a.a identity (Figure 4-6). This histidine kinase component employs a PAS sensor domain, a protein that uses haeme, flavin and 4-hydroxycinnamyl chromophore as cofactors, and a PAC protein at the C-terminus of the PAS protein is believed to assist in the folding of the PAS domain. Interestingly, no response regulator component was found adjacent to the histidine kinase. Alternatively, an oxidoreductase, transcribed in the opposite direction is found upstream to the histidine kinase and a hypothetical protein is found downstream.

A heavy metal sensor histidine kinase and response regulator, composing the two component regulatory system was also identified. This system controls the efflux of copper,

silver, cadmium and/or zinc thus providing the cell with resistance to these metals. The heavy metal sensory histidine kinase is autophosphorylated when the internal heavy metal concentration is above a threshold level and in turn phosphorylates the cognate response regulator which activates the transcription of the heavy metal efflux system. The heavy metal response regulator cztR_silR_copR$_{pCX9\text{-}45F}$ and the sensor histidine kinase cztS_silS_copS$_{pCX9\text{-}45F}$ were found as two ORFs separated by 50bp (Figure 4-8).

Table 4-2: Identified genes involved in signal transduction and regulation

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX4-10F (6-2)	orf4$_{pCX4-10F}$	PAS/PAC sensor hybrid histidine kinase	*M. nodulans ORS 2060*	37	302	143
			A. cryptum JF5	36	350	139
			Azoaracus sp. BH72	34	288	137
pCX9-45F (17)	cztR_silR_copR$_{pCX9-45F}$	Two component transcriptional regulator, winged helix family	*G. bemidjiensis*	63	224	279
			G. lovleyi	63	220	269
			G. metaloreducins	50	226	221
pCX9-45F (17)	cztS_silS_copS$_{pCX9-45F}$	Heavy metal sensor signal transduction histidine kinase	*G. bemidjiensis*	31	205	90.5
			G. lovleyi	26	215	78.6
			S. thermophilus	32	206	70.1
pCX6-13F (1)	*adc*$_{pCX6-13F}$	Adenylate cyclase	*D. aromatica RCB*	45	487	408
			A. aromaticum EbN	37	506	255
			Azoarcus sp. BH72	36	497	252

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX6-13F (1)	$gdc_{pCX6-13F}$	Diguanylate cyclase/phosphodiesterase	*D. aromatica* RCB	51	592	571
			J. sp. Marseille	49	576	568
			H. arsenicoxydans	48	577	555
pCX6-13F (1)	$draT_{pCX6-13F}$	Hypothetical protein- NAD+- dinitrogen- reductase ADP-D- ribosyltransferase activity	*D. aromatica* RCB	65	328	420
			A. vinosum DSM 180	57	268	313
			Magnetococcus sp. MC-1	51	266	281
pCX6-13F (1)	$deoR_{pCX6-13F}$	Regulatory functions: DNA interactions	*M. xanthus* DK 16-22	56	211	235
			A. dehalogenans 2-CP1	56	216	229
			Anaeromyxobacter sp. K	56	216	228

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX6-13F (4)	*Orf12* $_{pCX6-13F}$	Cyclic nucleotide-binding	*P. pacifica* SIR-1	69	185	260
			H. ocraceum DSM 14365	53	215	229
			B. marina DSM 3645	33	202	148
pCX4-10F (13-1)	*phoR*$_{pCX4-10F}$	Phosphate regulon histidine kinase	*D. aromatica* RCB	44	409	293
			Methylovorus sp. SIP 3-4	45	408	287
			Azoarcus sp. BH72	43	409	278
pCX4-10F (13-1)	*phoB*$_{pCX4-10F}$	Phosphate regulon response regulator	*L. hongkongensis* HLH-K9	71	170	216
			A. aromaticum EbN1	69	170	213
			Thauera sp. MZ1T	68	170	208

Figure 4-6: Map of ORFs identified on cosmid pCX4-10F (pulp waste activated sludge)

Direction of transcription is indicated by arrows. Numbers indicate nucleotide position on the contig. Incomplete ORFs are interrupted by /. Colours indicate functional categories as follows: Pale orange: replication, recombination and repair; Pale blue, nucleotide transport and metabolism; Green, defense mechanisms; Turquoise, carbohydrate transport and metabolism; Pink, phosphorous transport and metabolism; Purple, signal transduction and regulatory mechanisms; Dark orange; proteins with predicted general functions; Dark blue, proteins of unknown function. Undesignated ORFs are described as follows: $Orf1_{pCX4-10F}$ protein of unknown function; $Orf2_{pCX4-10F}$, probable transmembrane permease; $Orf3_{pCX4-10F}$ protein of unknown function; $Orf4_{pCX4-10F}$, putative histidine kinase sensor protein; $Orf5_{pCX4-10F}$, hypothetical protein; $Orf6_{pCX4-10F}$, putative outer membrane protein.
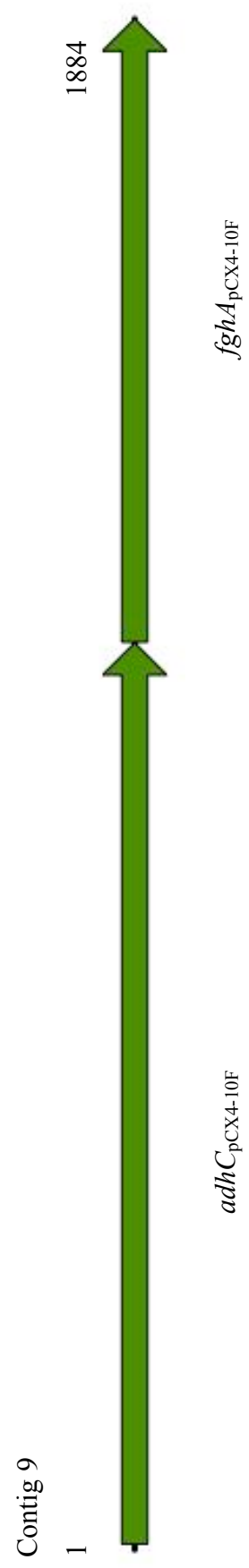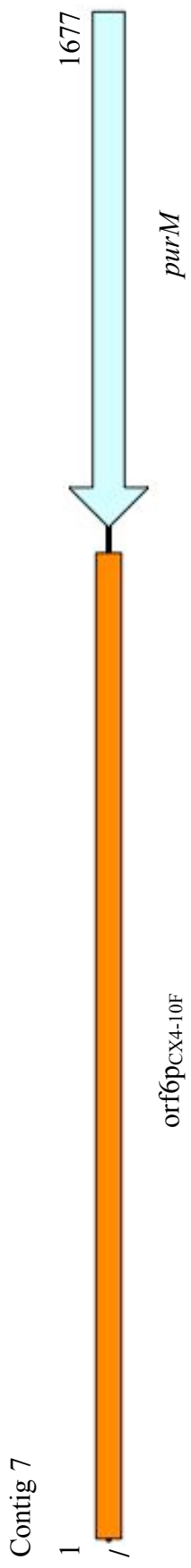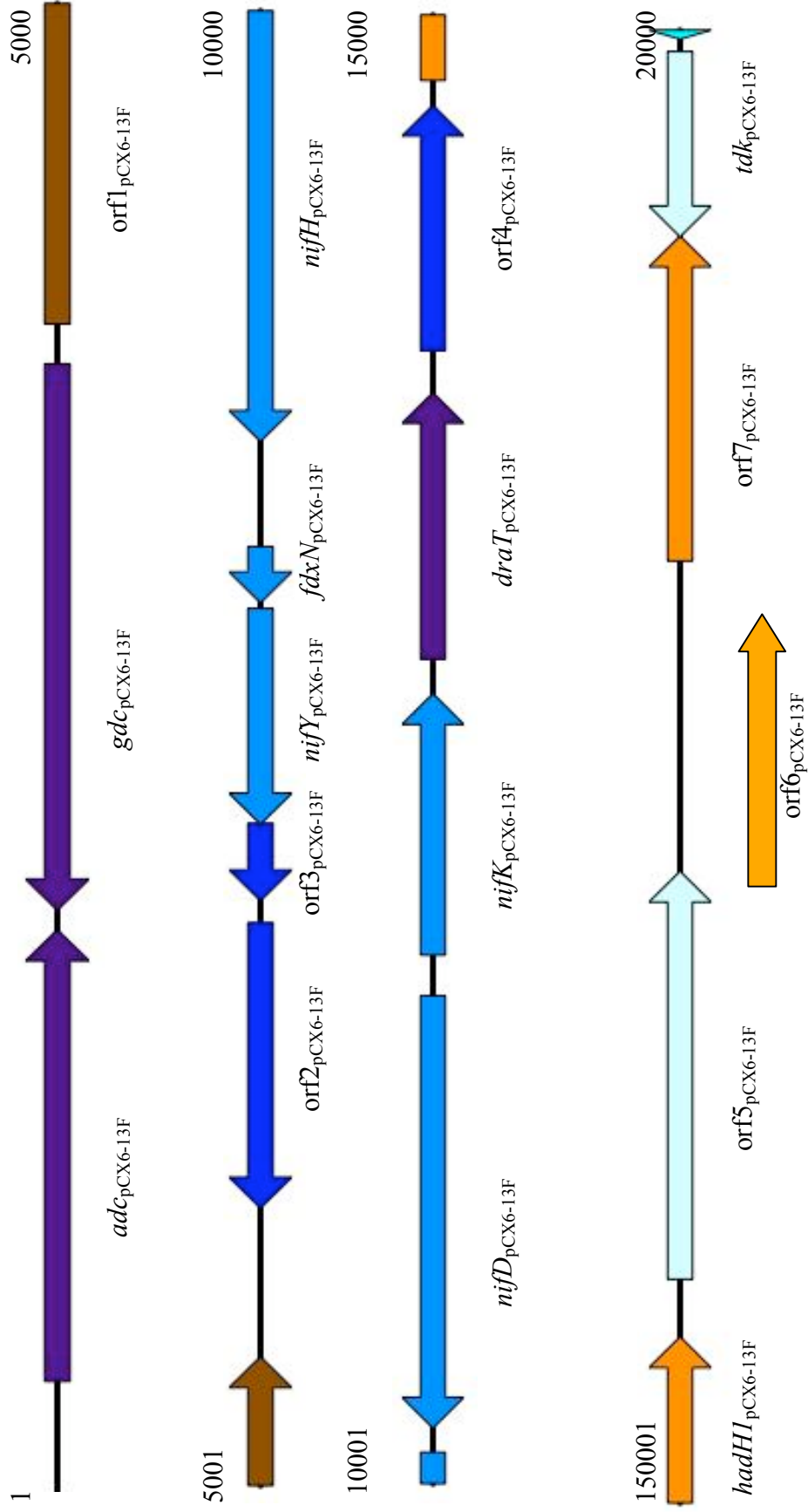
Contig 10

Contig 13-1

159

Contig 6-2

$gst2_{pCX4-10F}$ $orf3_{pCX4-10F}$ $orf4_{pCX4-10F}$ $orf5_{pCX4-10}$

1   5000

Contig 14

$pstS_{pCX4-10F}$ $pstC_{pCX4-10F}$ $pstA_{pCX4-10F}$ $pstB_{pCX4-10F}$ $phoU_{pCX4-10F}$

1   4676

160

Contig 7

1

1677

orf6p_{pCX4-10F}

purM

Contig 9

1

1884

*adhC*_{pCX4-10F}

*fghA*_{pCX4-10F}

161
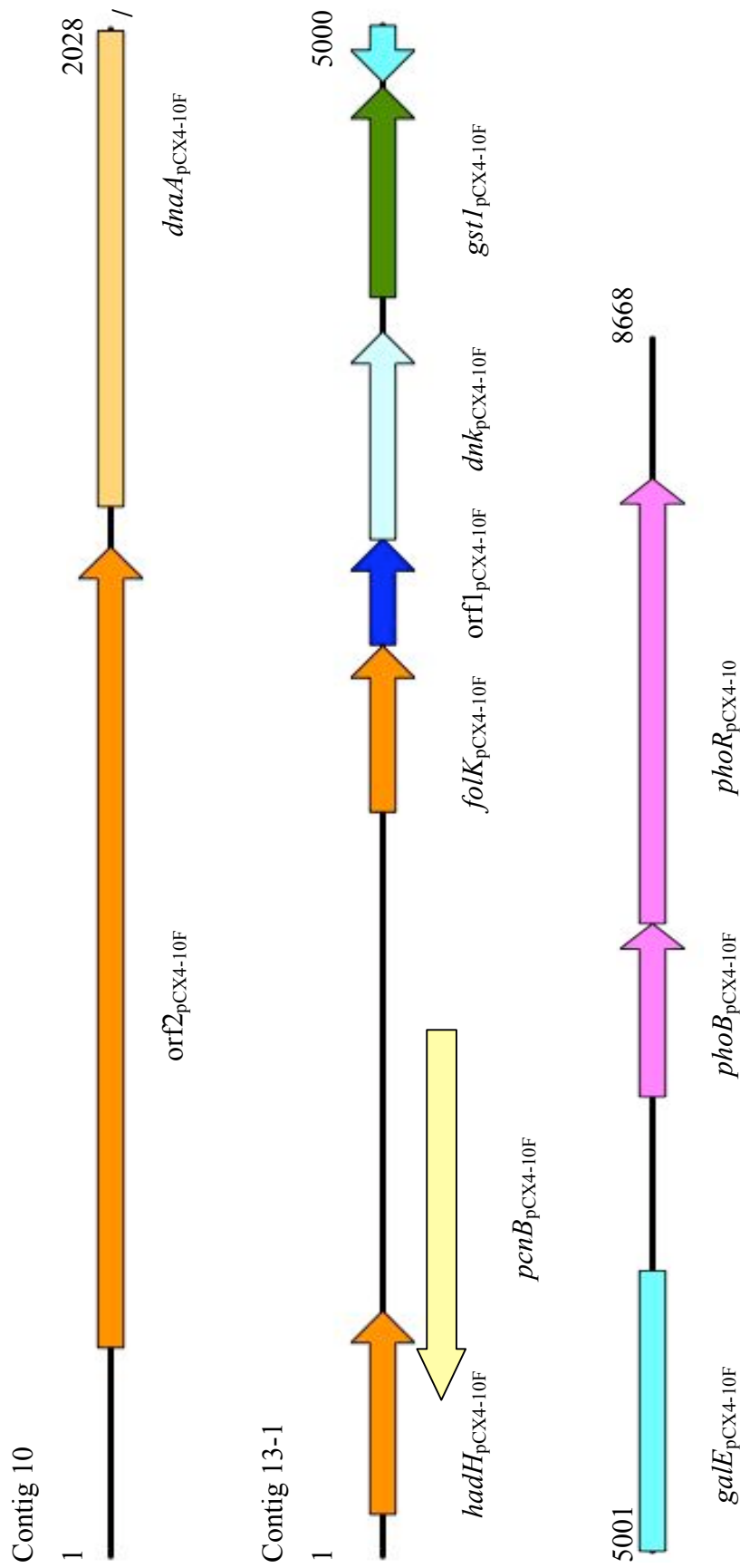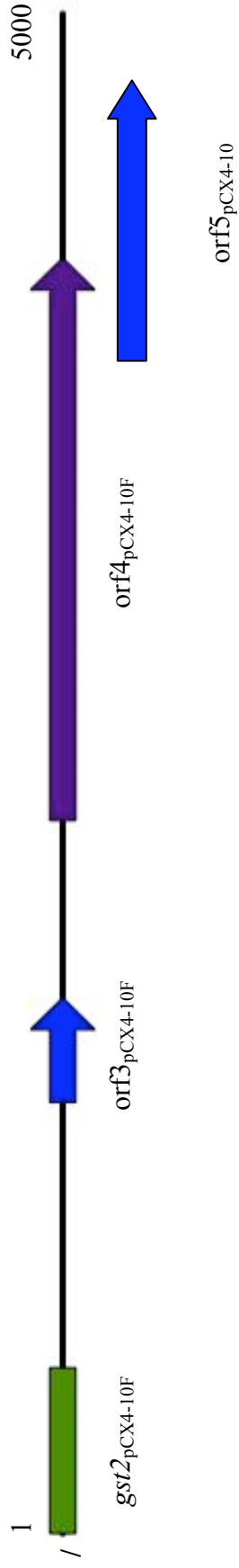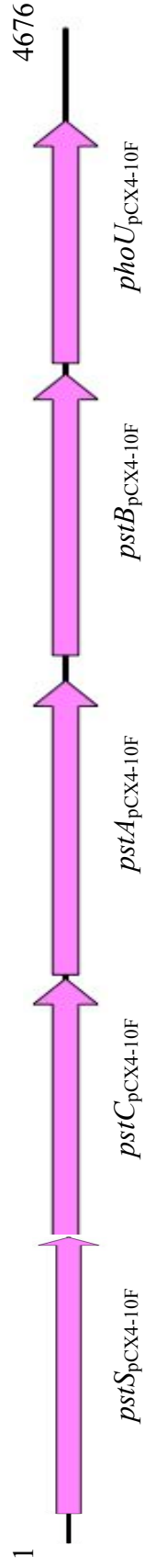
Figure 4-7: Map of ORFs identified on cosmid pCX6-13F (municipal waste activated sludge)

Direction of transcription is indicated by arrows. Numbers indicate nucleotide position on contig. Incomplete ORFs are interrupted by /. Colours indicate functional categories as follows: Pale orange: replication, recombination and repair; Pale blue, nucleotide transport and metabolism; Green, defense mechanisms; Aqua blue, nitrogen transport and metabolism; Pink, phosphorous transport and metabolism; Purple, signal transduction and regulatory mechanisms; Red, fatty acid transport and metabolism; Dark orange, proteins with predicted general functions; Dark blue, proteins of unknown function. Undesignated ORFs are described as follows: $Orf1_{pCX6-13F}$, putative phosphoglucosamine mutase; $Orf2_{pCX6-13F}$, hypothetical protein; $Orf3_{pCX6-13F}$, no significant hits; $Orf4_{pCX6-13F}$, hypothetical protein $Orf5_{pCX6-13F}$, putative metal dependent phosphohydrolase; $Orf6_{pCX6-13F}$, putative myo-inositol 1-monophosphatase; $Orf7_{pCX6-13F}$, putative DNA binding protein; $Orf8_{pCX6-13F}$, capsular exopolysaccharide family protein; $Orf9_{pCX6-13F}$, putative class I chitinase; $Orf11_{pCX6-13F}$, hypothetical protein $Orf12_{pCX6-13F}$, putative cyclic nucleotide binding protein; $Orf13_{pCX6-13F}$, hypothetical protein; $Orf14_{pCX6-13F}$, acyl-CoA dehydrogenase family protein.
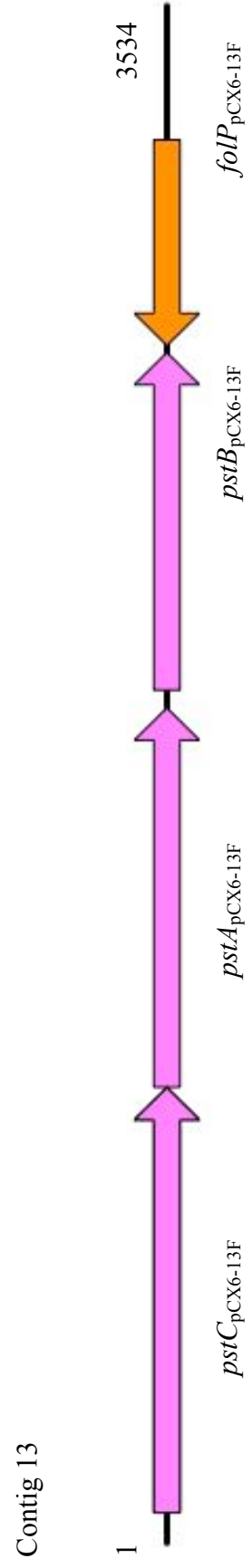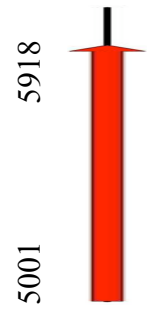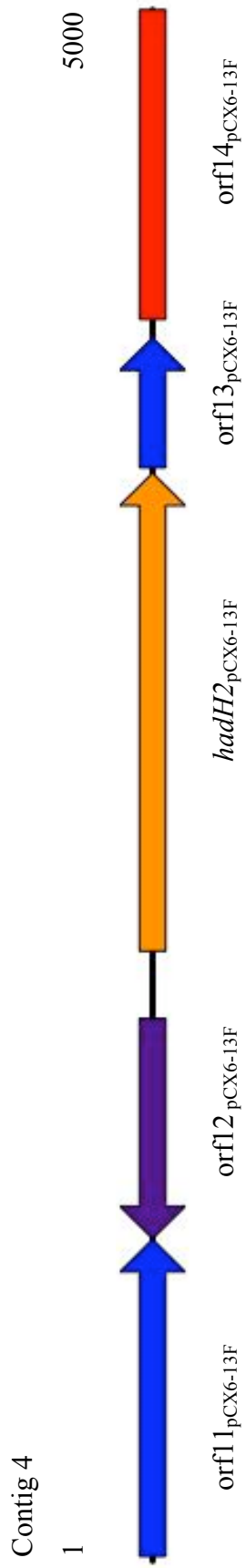
Contig 1

orf1_pCX6-13F
gdc_pCX6-13F
adc_pCX6-13F
nifH_pCX6-13F
fdxN_pCX6-13F
nifY_pCX6-13F
orf3_pCX6-13F
orf2_pCX6-13F
orf4_pCX6-13F
draT_pCX6-13F
nifK_pCX6-13F
nifD_pCX6-13F
tdk_pCX6-13F
orf7_pCX6-13F
orf6_pCX6-13F
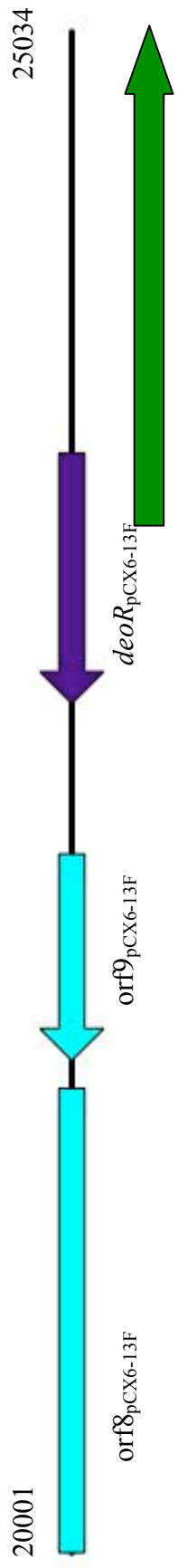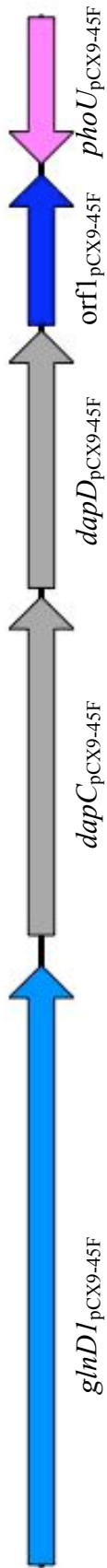orf5_pCX6-13F
hadH1_pCX6-13F

163

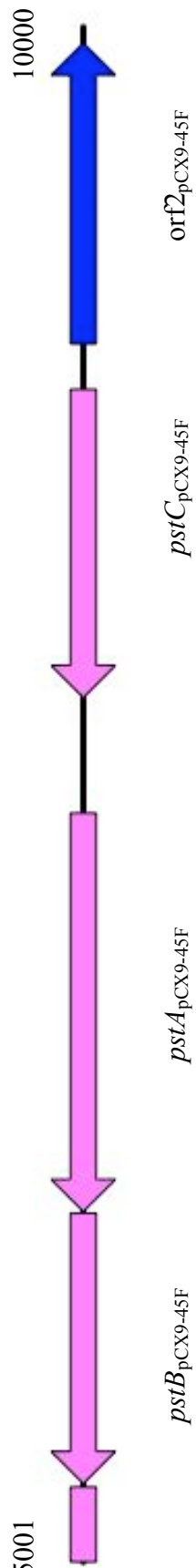**Figure 4-8:** Map of ORFs identified on cosmid pCX9-45F (soil)

Direction of transcription is indicated by arrows. Numbers indicate nucleotide position on contig. Incomplete ORFs are interrupted by /. Colours indicate functional categories as follows; Grey, amino acid transport and metabolism; Pale orange, replication recombination and repair; Sea green, translation ribosomal structure and biogenesis; Aqua blue, nitrogen transport and metabolism; Brown, post-translational modification, protein turnover and chaperones; Pale blue, nucleotide transport and metabolism; Pink, phosphorous transport and metabolism; Purple, signal transduction and regulatory mechanisms; Dark orange, proteins with predicted general functions; Dark blue, proteins of unknown function. Undesignated ORFs are described as follows:Orf1$_{pCX9-45F}$, protein of unknown function; Orf2$_{pCX9-45F}$, protein of unknown function; Orf3$_{pCX9-45F}$, unknown protein; Orf4$_{pCX9-45F}$, unknown protein; Orf5$_{pCX9-45F}$, ATP/GTP binding protein; Orf6$_{pCX9-45F}$, TonB-dependent receptor plug domain; Orf7$_{pCX9-45F}$, putative serine-threonine protein kinase; Orf8$_{pCX9-45F}$, protein of unknown function.
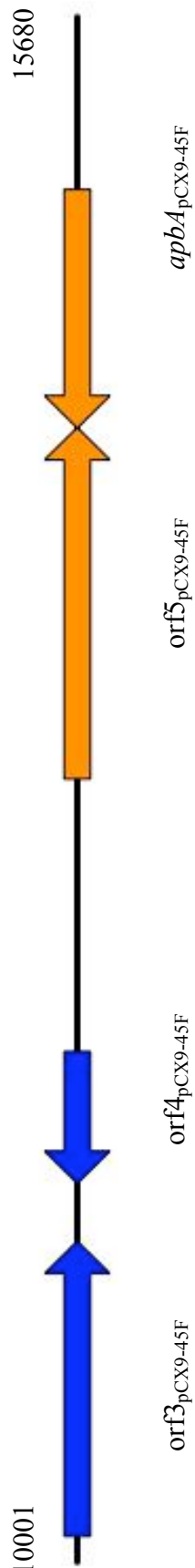
Contig 10-3-19

$glnD1_{pCX9-45F}$

$dapC_{pCX9-45F}$

$dapD_{pCX9-45F}$

orf1$_{pCX9-45F}$

$phoU_{pCX9-45F}$

$pstB_{pCX9-45F}$

$pstA_{pCX9-45F}$

$pstC_{pCX9-45F}$

orf2$_{pCX9-45F}$

orf3$_{pCX9-45F}$

orf4$_{pCX9-45F}$

orf5$_{pCX9-45F}$

$apbA_{pCX9-45F}$

5001

5001

10001

10000

15680

Contig 12

5000

$tsf_{pCX9-45F}$

$pyrH_{pCX9-45F}$

orf6$_{pCX9-45F}$

5001

7443

$rpsB_{pCX9-45F}$

$map_{pCX9-45F}$

$glnD2_{pCX9-45F}$

Contig 17

1934

$cztS\_silS\_copS_{pCX9-45F}$

$cztR\_silR\_copR_{pCX9-45F}$

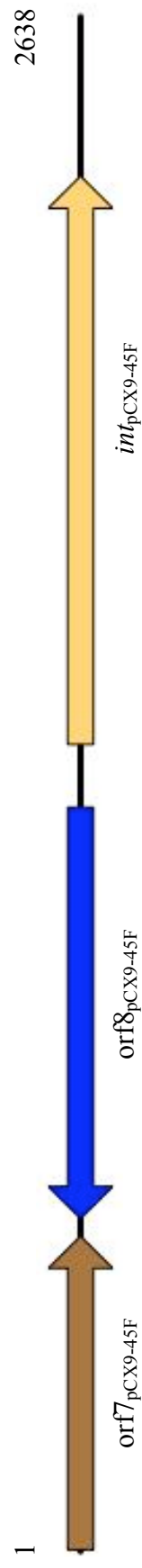$pabB_{pCX9-45F}$

Contig 18

2638

$int_{pCX9-45F}$

orf8$_{pCX9-45F}$

orf7$_{pCX9-45F}$

167

Two additional genes involved in signal transduction were found in the metagenome of municipal waste activated sludge. A putative adenylate cyclase, $adc_{pCX6-13F}$ and a putative diguanylate cyclase/phosphodiesterase $gdc_{pCX6-13F}$ were found adjacent to each other, transcribed in opposite directions with their stop codons separated by 64 bp (Figure 4-6). Adenylate cyclase catalyses the conversion of ATP to c-AMP, which is part of a cascade leading to the transcriptional regulation of many genes. Diguanylate cyclases/phosphodiesterases catalyse the synthesis and/or degradation of second messenger cyclic-di-GMP, an intracellular signaling molecule regulating various functions such as exopolysaccharide synthesis, biofilm formation, motility and cell differentiation (Chan *et al.*, 2004). Bacterial diguanylate cyclases/phosphodiesterases typically contain a GGDEF domain which participates in diguanylate cyclase activity and an EAL domain which is responsible for phosphodiesterase activity (Rao *et al.*, 2008). $adc_{pCX6-13F}$ and $gdc_{pCX6-13F}$ are homologues of the respective putative adenylate cyclase and diguanylate cyclase/phosphodiesterase from the *β-proteobacterium Dechloromonas aromatica* RCB with adenylate cyclase homologues sharing 45% amino acid identity and the diguanylate cyclase/phosphodiesterase homologues share 51% amino acid identity. $gdc_{pCX6-13F}$ is predicted to have an N-terminal EAL domain of 234 a.a. and a C-terminal GGDEF domain of 103 a.a (Marchler-Bauer *et al.*, 2009). Interestingly, residues comprising the conserved G-G-D-E-F motif, contain Asp509 in place of Glu and in the EAL domain, the putative protein has Leu 179 and Glu 181 in place of the cognate Phe and Thr to make up the conserved D-D-F-G-T-G motif, found in most active EAL domains (Schmidt *et al.*, 2005). This sequence variation suggests a new mechanism for

cyclic di GMP production or degradation which may impact on its regulatory activity. The EAL domain contains the E-X-L signature motif where X is a hydrophobic glycine residue.

Nitrogenase activity is regulated post translationally by reversible ADP-ribosylation of dinitrogenase reductase by means of ADP-ribosyl transferase-dinitrogenase reductase-activating glycohydrolase (DRAT-DRAG) regulatory system (Zhang *et al.*, 2001). Under ammonium-sufficient conditions, DRAT catalyses the transfer of ADP-ribose from NAD to Arg-101 residue of dinitrogen reductase, resulting in the inactivation of that enzyme. The reverse process of dinitrogenase reductase activation, characterized by the removal of the ADP-ribose from dinitrogenase reductase is carried out by the DRAG protein (Zhang *et al.*, 2001). A putative DRAT, *draT*$_{pCX6-13F}$ was identified upstream to *nifH*$_{pCX6-13F}$, whose protein product it may be a target for DRAT regulation (Figure 4-7). Both NifH$_{pCX6-13F}$ and the putative NifH from *D. aromatica RCB* contain the motif Gly-Arg-Gly-Val-Iso-Thr previously shown to be the site of ADP-ribose binding in dinitrogen reductase from *Rhodobacter capsulatus* (Jouanneau *et al.*, 1989). DraT$_{pCX6-13F}$ shares 65% amino acid identity with the putative DRAT of *Dechloromonas aromatica* RCB.

4.3.2 Nitrogen fixation and assimilation

Since soil microorganisms occupy habitats where nitrogen is not readily available, genes for atmospheric nitrogen fixation and assimilation often contained by these organisms ensure their survival in such complex communities. On the other hand, activated sludge habitats have sufficient fixed nitrogen readily available for assimilation. The abundance of nitrogen fixation and assimilation related genes found in the activated sludge metagenomes may be explained by the occurence of some of these organisms in oligotrophic lakes where

nitrogen availability is low. From the metagenomes of municipal waste activated sludge and soil, a nitrogenase enzyme complex and nitrogen metabolism genes were identified (Table 4-3). The nitrogenase complex is encoded on the operon *nifHDKfdxNnifY* preceded by a putative upstream dinitrogenase-reductase ADP-D- ribosyltransferase (DRAT) negative regulator, *draT*$_{pCX6-13F}$ (Figure 4-7). Nitrogenase catalyses the energy demanding reduction of atmospheric nitrogen to ammonia (i.e., nitrogen fixation) (Burris, 1991). The nitrogenase enzyme is composed of two components: a dinitrogenase composed of a $\alpha_2\beta_2$ tetramer (component I encoded by *nifK*$_{pCX6-13F}$ and *nifD*$_{pCX6-13F}$) and a dinitrogenase reductase homodimer (component II encoded by *nifH*$_{pCX6-13F}$). Nitrogenase activity is supported by an iron-molybdenum cofactor (*nifY*$_{pCX6-13F}$) that binds to the active site of the dinitrogenase enzyme and an indole pyruvate ferredoxin oxidoreductase (*fdxN*$_{pCX6-13F}$) which mediates electron transfer in metabolic reactions. The *nif* genes identified here are homologues of the putative *nif* genes of the *β-proteobacterium Dechloromonas aromatica* RCB with amino acid identities ranging from 77-93%.

In addition to nitrogen fixation, two genes involved in the regulation of nitrogen assimilation were also identified. Partial ORFs corresponding to two genes, *glnD1*$_{pCX9-45F}$ and *glnD2*$_{pCX9-45F}$, were identified from the metagenome of soil (Figure 4-8). The product of the *glnD* gene is a dual uridylyltransferase/uridylyl-removing enzyme (Utase-UR) whose possible role is as sensor of intracellular glutamine concentration and response by uridylylation or deuridylylation of the PII, a protein which in turn regulates a number of genes (*ntrBC, nifA, glnA*) involved in nitrogen fixation and assimilation (Zhang *et al.*, 2001). GlnD1$_{pCX9-45F}$ and GlnD2$_{pCX9-45F}$ are homologues of halophilic *γ-Proteobacteria*

Halorhodospira halophila SL1 and *Chromohalobacter salexigens* DSM 3043. The abundance of nitrogen metabolism genes is consistent with the expected nitrogen fixation capabilities of bacteria inhabiting soil and sludge communities.

Table 4-3: Identified genes involved in nitrogen fixation and assimilation

Cosmid clone (contig)	Gene identified	Similar protein	Organism	Identity (%)	Region of identical amino acids	Score
pCX6-13F (1)	$nifH_{pCX6-13F}$	Nitrogenase iron protein subunit	D. aromatica RCB	93	287	515
			Azoarcus sp. BH72	87	287	493
			T. turnerae T7901	83	283	473
pCX6-13F (1)	$nifD_{pCX6-13F}$	Nitrogenase molybdenum-iron protein alpha chain: nitrogenase component I, alpha chain	D. aromatica RCB	90	288	197
			Azoarcus sp. BH72	85	288	181
			D. dadantii Ech 703	78	288	159
pCX6-13F (1)	$nifK_{pCX6-13}$	Nitrogenase molybdenum-iron protein beta chain	D. aromatica RCB	85	397	415
			Azoarcus sp. BH72	76	397	355
			Azoarcus sp.	77	397	355
pCX6-13F (1)	$fdxN_{pCX6-13F}$	4Fe-4S ferredoxin, iron-sulfur binding	D. aromatica RCB	94	17	40.8
			R. capsulatus	70	17	35.4
			R. phaeroides KD131	70	17	33.9

Cosmid clone (contig)	Gene identified	Similar protein	Organism	Identity (%)	Region of identical amino acids	Score
pCX6-13F (1)	$nifY_{pCX6-13F}$	Dinitrogenase iron-molybdenum cofactor biosynthesis	D. *aromatica* RCB	77	243	320
			Azoarcus sp. BH72	52	211	193
			A. vinelandii DJ	49	207	176
pCX9-45F (10-3-19)	$glnD1_{pCX9-45F}$	UTP-GlnB uridylyltransferase, GlnD	*H. halophila* SL1	55	620	662
			A. ehrlichii MLHE-1	54	637	654
			N. mobilis Nb-231	53	634	637
pCX9-45F (12)	$glnD2_{pCX9-45F}$	UTP-GlnB (protein PII) uridylyltransferase, GlnD	*C. salexigens* DSM 3043	45	177	152
			A. ehrlichii MLHE-1	46	166	152
			Beggiatoa sp. PS	48	171	151

4.3.3 Defense mechanisms:

Many sludge and many soil communities are characterized by their exposure to various organic compounds, deposited by means of contamination from industrial origin. To protect their cells from such toxic compounds, soil and sludge microorganisms evolved mechanisms for the degradation and detoxification of a variety of organic compounds. This is reflected in the relative abundance of genes involved in glutathione-mediated degradation of organic compounds (Table 4-4). Two genes containing glutathione-S-transferase domains, $gst1_{pCX4-10F}$ and $gst2_{pCX4-10F}$, a glutathione-dependent formaldehyde dehydrogenase, $adhC_{pCX4-10F}$ and an S-formylglutatione hydrolase, $fghA_{pCX4-10F}$ were identified from the metagenome of pulp waste activated sludge (Figure 4-6). The latter two genes, found adjacent to each other, are predicted to play a key role in formaldehyde detoxification. Glutathione dependent formaldehyde dehydrogenase oxidizes S-hydroxymethyl-glutathione, a product of reaction of formaldehyde with glutathione, to S-formylglutathione, using NAD^+ as a cofactor. S-formylglutathione is then hydrolysed to formic acid and glutathione and formate by S-formylglutathione hydrolase, completing the degradation of formaldehyde. Aside from detoxification, bacteria have evolved additional defense mechanisms to ensure survival. From the metagenome of municipal waste activated sludge, the C-terminal domain of a DNA modification methyltransferase-related ORF, dmm_{pCX6-F} was identified. Across 557 amino acids of the C- terminal domain of $Dmm_{pCX6-13F}$, a 33% identity to the putative DNA modification methyltransferase protein of the radiation resistant *Deinococcus radiodurans* R1, suggesting that $Dmm_{pCX6-13F}$ is divergent from known DNA modification methyltransferases.

174

Table 4-4: Identified genes involved in cellular defense mechanisms

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX4-10F (13-1)	$gst1_{pCX4-10F}$	Glutathione S-transferase domain protein	*Acidovorax sp.* JS 42	68	228	325
			Diaphoro-bacter sp. TPSY	68	228	322
			A. avenae subsp. citrulli AAC00-1	66	229	316
pCX9-45F (6-2)	$gst2_{pCX4-10F}$	Glutathione S-transferase domain protein	*V. para-doxus* S110	61	178	215
			Janintho-bacterium sp. Marseille	56	189	202
			gst12 S. meliloti 1021	62	180	199
pCX4-10F (9)	$adhC_{pCX4-10F}$	Glutathione dependent formaldehyde dehydrogenase	*N. spumigena* CCY9419	82	369	585
			Lyngbia sp. PCC8106	81	369	582
			P. lavamentivo rans DS-1	83	369	582
pCX4-10F (9)	$fghA_{pCX4-10F}$	S-formyl-glutathione hydrolase	*P. zucineum* HLK1	61	256	280
			M. maris MCS10	63	256	278
			γ proteo-bacterium HTCC5015	64	244	276

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX6-13 (1)	$dmm_{pCX6-13F}$	DNA modification methyltransferase -related protein	*Rosieflexus sp*. RS-1	35	493	242
			D. radiodurans R1	33	489	236
			Lyngbia sp. PCC8106	31	506	230

4.3.4 Housekeeping genes

Approximately 6% of the genes identified encode functions related to the management of the cell's genetic information such as translation, ribosomal structure and biogenesis, as well as DNA replication and repair (Table 4-5). A putative inactivator of the chromosomal replication initiator ($hda_{pCX4-10F}$, Figure 4-6) and an integrase ($int_{pCX9-45F}$, Figure 4-8) were identified from pulp waste activated sludge and soil metagenome. The nearest homologue to $hda_{pCX4-10F}$ is the putative replication initiator inhibitor ebA7107 of the *β-Proteobacterium Aromatoleum aromaticum* EbN at 49% amino acid identity. Similarly, $int_{pCX9-45F}$ shares a 50% amino acid identity with the integrase Tgr7_0692 from *γ-Proteobacterium thioalkalivibrio sp*. HL-EbGR7. In addition, a poly(A) polymerase ($pcnB_{pCX4-10F}$, Figure 4-6) that catalyses the addition of poly(A) tails to mature mRNA to prepare it for degradation (Yehudai-Resheff & Schuster, 2000) was identified from the metagenome of pulp waste activated sludge. Three additional genes were identified from the soil metagenome. They included a ribosomal protein S2 ($rpsB_{pCX9-45F}$) and a translational elongation factor ($tsf_{pCX9-45F}$), which may constitute an operon (Figure 4-8). Upstream of $rpsB_{pCX9-45F}$, a methionine aminopeptidase, $map_{pCX9-45F}$ was identified; this gene is transcribed in the opposite direction to the potential *tsf-rpsB* operon (Figure 4-8).

4.3.5 Proteins of general function

About 16% of the genes identified in this metagenomic study encode functions for which a specific cellular role is not known (Table 4-6). These include 3 putative members of the "haloacid dehalogenase-like" (HAD) hydrolases superfamily (Figures 4-6 and 4-8). HAD hydrolases are enzymes that cleave carbon-halogen bond in halogenated aliphatic

hydrocarbons, thus aid in cellular detoxification of such compounds (Koonin & Tatusov, 1994). Most HAD hydrolases, however, carry out phosphoryl transfer reactions via an active site aspartate nucleophile. Due to a wide substrate specificity, HAD hydrolases have diverse biochemical roles from signal transduction, to DNA repair to secondary metabolism (Tremblay *et al.*, 2006). Each HAD like hydrolase identified here represents a different subfamily that is defined according to different structural folds and location of catalytic motifs. The gene *hadH*$_{pCX4-10F}$ is homologous to many HAD hydrolases of subfamily IB while *hadH1*$_{pCX6-13F}$ is homologous with many subfamily IIA HAD like hydrolase and *hadH2*$_{pCX6-13F}$ is homologous with many subfamily IG HAD like hydrolases. Subfamilies I and II differ in the location of a variable cap domain, which interacts with the substrate leaving group, with respect to the three catalytic motifs (Tremblay *et al.*, 2006). Members of superfamily IB include phosphoserine phosphatase-like proteins but no evidence of such activity is presented in its members. Subfamily IIA HAD hydrolases are characterized by the occurrence of a cap domain between the second and third core catalytic motifs. Among the experimentally determined HAD hydrolases of this family are *E. coli* K12 NagD, a protein involved in N-acetylglucosamine metabolism, for which structural data is available (Tremblay *et al.*, 2006). HadH1$_{pCX6-13F}$ has conserved nucleophilic core domain motif I residues Asp 9 and Asp 11, equivalent to NagD. It also has a conserved threonine equivalent to Thr 42 from motif II. The two aspartic acid residues are predicted to be involved in the acid/base catalysis of phosphoryl transfer while Thr 42 may form H-bonds with the substrate phosphoryl group.

Table 4-5: Identified genes involved in housekeeping functions

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX4-10F (10)	$hda_{pCX4-10F}$	Chromosomal replication initiator	*A. romaticum* ebA1	49	67	57.8
			Thauera sp. MZ1T	45	74	55.1
			Janintho-bacterium sp. Marseille	49	73	53.1
pCX9-45F (18)	$int_{pCX9-45F}$	Integrase, catalytic region	*Thioalkylvibrio sp.* HL-EbGR7	50	259	251
			R. leguminosarum bv trifolii WSM1325	46	254	219
pCX4-10F (13-1)	$pcnB_{pCX4-10F}$	Poly(A)-polymerse	*B. petrii* DSM 12804	63	354	432
			B. duloso AUO158	56	447	429
			B. avium 197N	63	354	426
pCX9-45F (12)	$rpsB_{pCX9-45F}$	Ribosomal protein S2	*Thioalkali-vibrio sp.* K90mix	63	259	329
			Thioalkali-vibrio sp. HL-EbGR7	62	244	328
			A. aromaticum EbA1	64	264	324

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX9-45F (12)	$tsf_{pCX9-45F}$	Translation elongation factors Ts	*M. thiooxidans* DMS010	55	293	290
			F. philomiragia	57	285	283
			N. oceanii ATCC 19707	55	292	280
pCX9-45F (12)	$map_{pCX9-45F}$	Methionine aminopeptidase	*Thioalkali-vibrio sp.* HL-EbGR7	71	235	368
			S. maltophilia R551-3	69	242	363
			S. maltophilia K279A	68	242	363

Table 4-6: Identified genes involved in general functions

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX4-10F (13-1)	$hadH_{pCX4-10F}$	Haloacid dehalogenase like hydrolase, subfamily IB	*B. phymatum* STM815	62	222	295
			B. glumae BGR1	63	224	290
			R. metallo-durans CH34	62	222	287
pCX6-13F (1)	$hadH1_{pCX6-13F}$	Haloacid dehalogenase like hydrolase, subfamily IIA	*M. mobilis* JW8	49	178	183
			M. flagellatus KT	46	171	174
			Methylo-vorus sp. SIP 3-4	47	177	179
pCX6-13F (4)	$hadH2_{pCX6-13F}$	Haloacid dehalogenase like hydrolase, subfamily IG, 5' nucleotidase	*P. pacifica* SIR-1	58	515	416
			A. dehalo-genans 2CP-C	45	493	408
			Anaero-myxobacter SP. K	44	492	407
pCX4-10F (13-1)	$folK_{pCX4-10F}$	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	*M. petro-liphilum* PM-1	58	150	149
			V. paradoxus S110	58	129	142
			B. petrii DSM 12804	57	153	142

.

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX6-13F (13)	$folP_{pCX6-13F}$	Dihydropteroate synthase	*D. aromatica* RCB	63	134	175
			Thauera sp. MZ1T	62	134	174
			C. japonicus Ueda107	64	136	172
pCX6-13F (1)	orf5$_{pCX9-45F}$	Metal dependent phospho-hydrolase	*D. acidovorans* SPH-1	31	358	93.6
			C. testosteroni KF1	36	280	89.4
			A. delafieldii 2AN	32	365	89.0
pCX9-45F (10-13-1)	$apbA_{pCX9-45F}$	2-dehydropantoate 2-reductase	*S. sviceus* ATCC 29083	48	399	187
			B. cavernae DSM 12333	43	278	187
			C. acidiphila DSM 44928	42	282	178
pCX4-10F (10)	orf2$_{pCX4-10F}$	ABC transporter permease	*R. picketii* 12D	53	266	205
			R. solana-cearum GMI1000	52	266	199
			R. solana-cearum UW552	52	266	198
pCX4-10F (7)	orf6$_{pCX4-10F}$	Surface antigen (D15)	*M. mobilis* JLW8	31	296	142
			R. eutropha JMP134	28	306	140
			L. hong-kongensis HLHK9	31	302	134

182

Cosmid clone (contig)	Gene identified	Similar protein	Organism	% identity	Region of identical amino acids	Score
pCX6-13F (1)	orf6$_{pCX6-13F}$	Putative myo-inositol 1-mono-phosphatase	*Janibacter sp.* HTCC 2649	52	231	232
			C. flavigena DSM 20109	50	256	216
			B. cavernae DSM 1233	53	204	211
pCX6-13F (1)	orf7$_{pCX6-13F}$	Putative DNA-binding protein	*Janibacter sp.* HTCC 2649	45	351	216
			K. radio-tolerans SRS 30216	44	322	203
			B. faecium DSM 4810	38	337	159
pCX9-45F (17)	*pabB*$_{pCX9-45F}$	Para-aminobenzoate synthase subunit I	*D. baculatum* DSM 4028	72	42	70.9
			A. maxima CS328	65	42	67
			A. dehalo-genans 2CP-C	58	42	57

Interestingly, HadH1$_{pCX6-13F}$ lacks a positive charge supplied by the equivalent of NagD Lys 176 to orient the Asp 9 nucleophile via salt bridging and provide electrostatic shielding for the transfer of the phosphoryl group. Alternatively, Lys 176 of motif III in HadH1$_{pCX6-13F}$ is replaced with asparganine, bearing a polar amino group. HadH1$_{pCX6-13F}$ also lacks the equivalent residues of NagD Asp 201 and Asp 206, shown to be involved in Mg^{2+} cofactor binding to the core domain. This suggests that HadH1$_{pCX6-13F}$ activity may not involve metal cofactor binding. HadH2$_{pCX6-13F}$ is homologous to members of subfamily IG of HAD hydrolases and has significant amino acid identity to eukaryotic 5'-nucleotidases specific for purines (IMP and GMP). The nearest homologue to HadH2$_{pCX6-13F}$ is the predicted HAD superfamily (subfamily IG) hydrolase, 5'-nucleotidase of the *δ-Proteobacterium Plesiocystis pacifica* SIR-1 (accession no ZP_01905669) at 58% amino acid identity. To date, HAD superfamily hydrolases of subfamily IG have only been characterized in eukaryotic systems as there are no reports of experimental determination of bacterial HAD superfamily hydrolases belonging to subfamily IG.

Most microorganisms are not able to take up exogenous folic acid, an essential co-factor required for the synthesis of various amino acids, purines, and other important nutrients. Folic acid is therefore synthesized de novo by microorganisms. Two genes, *folK*$_{pCX4-10F}$ and *folP*$_{pCX6-13F}$, encoding enzymes involved in folic acid biosynthesis were identified from pulp and municipal waste activated sludge metagenomes (Figure 4-6 and 4-7). *folK*$_{pCX4-10F}$ encodes a putative 2-amino-4-hydroxy-6-methyldihydropteridine diphosphokinase (HPPK), which catalyses the attachment of pyrophosphate to 2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine to produce 2-amino-7,8-dihydro-4-

hydroxy-6-(diphospho-oxymethyl) pteridine, an intermediate metabolite in the branch leading to 7,8-dihydrofolate (Talarico *et al.*, 1991). HPPK has received a great deal of attention because it has proven to be an effective target for the development antimicrobial agents in the way of inhibition of folic acid biosynthesis (Yan *et al.*, 2001). The nearest homologue to FolK$_{pCX4-10F}$ is the putative HPPK of the *β-Proteobacterium Methylibium petrophilum* PM1 (accession no YP_001022207) at 58%. The overall amino acid sequence identity of FolK$_{pCX4-10F}$ compared to the well-characterized *E. coli* HPPK (accession no 21238998) is only 40%. However, of the nine residues involved in substrate binding, FolK$_{pCX4-10F}$ contains the equivalent of six residues (Gly 8, Thr 42, Pro 43, Asn 55, Asp 95 and Phe 123). Conserved residues Gly 14 and Asn 59 (equivalent to *E. coli* HPPK Gly 8 and Asn 55, respectively) may be crucial in substrate binding as glycine is the only residue that can induce the correct confirmation of the asparganine residue for the formation of H bond with the HP substrate (Yan *et al.*, 2001). Of the nine residues involved in substrate binding in *E. coli* HPPK, Gly 8, Leu 45, Tyr 53, Trp 89 and Phe 123 are the source of hydrophobic interactions (Yan *et al.*, 2001). Equivalent residues in FolK$_{pCX4-10F}$ include only Gly 14 and Phe 128 (corresponding to Gly 8 and Phe 123 in *E. coli* HPPK). Leu 45, Tyr 53 and Trp 89 are replaced in FolK$_{pCX4-10F}$ as follows: Val 51, Phe 59 and Asn 94.

The second gene, *folP*$_{pCX6-13F}$, encodes a putative dihydropteroate synthase (DHPS), an enzyme which catalyzes the condensation of p-aminobenzoic acid (pABA) into 7,8-dihydropteroate, feeding into the pathway leading to 7,8-dihydrofolate. Due to its important role in folate synthesis and the fact that humans lack the enzyme, DHPS is widely investigated as a potential target for drug development against infection by pathogenic

185

bacteria such as *Mycobacterium tuberculosis* and *Staphylococcus aureus* (Baca *et al.*, 2000). FolP$_{pCX6-13F}$ protein's nearest homologue is the putative dihydropteroate synthase of the chancroid disease-causing *γ-Proteobacterium Haemophilus ducreyi* (accession number NP_873180) (Morse, 1989) where the 155 C-terminal amino acids have only 56% identity to the C-terminal portion of FolP$_{pCX6-13F}$. An even lower amino acid sequence identity was observed between FolP$_{pCX6-13F}$ and the structurally determined DHPS enzymes of *E. coli* (accession no 157829902), *Mycobacterial tuberculosis* (accession no 11514674) and *Staphylococcus aureus* (accession no 3212427) at 48%, 38% and 33%, respectively. Another interesting characteristic of the FolP$_{pCX6-13F}$ protein is its small size (157 a.a.) compared to an almost double size (270-280 amino acids) *E. coli*, *M. tubercu*losis and *S. aureus* DHPS homologues. This suggests a different structure to the protein and perhaps a different mode of substrate or analogue binding, possibly making this gene a good candidate for the development of new antibacterial drug.

Aside from folic acid biosynthesis genes, a gene involved in thiamine biosynthesis, *apbA*$_{pCX9-45F}$ was identified from the metagenome of soil (Figure 4-8). The gene encodes a putative 2-dehydropantoate 2-reductase, an enzyme which catalyses the NADPH-dependent conversion of ketopantoate to pantoate. This step is part of the alternative pyrimidine biosynthesis (pathway) leading to the production of thiamine (Frodyma & Downs, 1998). ApbA$_{pCX9-45F}$ shares at most 48% amino acid identity over 280 C-terminal amino acids with several members of the *Actinobacteria* class including *Streptomyces sviceus* ATCC 29083 (accession no YP_002207713), *Beutenbergia cavernae* DSM 12333 (accession no YP_002884032.1), *Catenulispora acidiphila* DSM 44928 (accession no ZP_04375198),

Kineococcus radiotolerans SRS30216 (accession no YP_001364031.1) and *Mycobacterium smegmatis str*. MC2 (accession no YP_887238.1). The fact that the nearest homologues of ApbA$_{pCX9-45F}$ are members of the *actinobacteria* class is consistent with the gene originating from a soil environment. The gene did not bear any significant similarity to the experimentally determined *apbA* from *Salmonella typhimurium* (Frodyma & Downs, 1998).

4.3.6 Proteins of unknown function

Approximately 14% of the ORFs identified have unknown cellular functions (Table 4-7). The amino acid identity to hypothetical proteins in bacterial isolates ranges from 30-90% compared to 31-78% identity with proteins identified from metagenomic samples. The hypothetical protein encoded by Orf1$_{pCX4-10F}$ shared the highest amino acid identity (90%) with the hypothetical protein HEAR2639 of the metalloresistant *β-Proteobacterium Herminiimonas arsenicoxydans* (Figure 4-6). This organism was isolated from arsenic contaminated industrial waste activated sludge environment (Muller *et al.*, 2006). Since Orf1$_{pCX4-10F}$ originated from an uncultured bacterium from an environment with similar characteristics, it is not surprising that the genomes of this organism and *H. arsenicoxydans* share common sets of genes, expecially those geared towards adapting to the complex physiochemical conditions imposed by the activated sludge environment. It is worth noting that within the CAMERA environmental database, Orf2$_{pCX6-13F}$ and Orf4$_{pCX6-13F}$ both derived from the metagenome of municipal waste activated sludge, each share the highest amino acid identity with ORFs derived from the sewage treatment plant activated sludge metagenome (Figure 4-7). Similarly, Orf1$_{pCX9-45F}$ and Orf8$_{pCX9-45F}$, both derived from the metagenome of soil, each share the highest amino acid identity with ORFs derived from farm soil (Figure 4-

8). This consistency shows the occurence of unknown genes with presumably similar functions, in similar environments. Some of the identified ORFs showed no significant hits in one or either database. Three of the ORFs, $Orf3_{pCX4-10F}$, $Orf3_{pCX9-45F}$ and $Orf4_{pCX9-45F}$ (Figures 4-6 and 4-8) have no homologues present in the NCBI database for bacterial isolates while $Orf3_{pCX6-13F}$, $Orf5_{pCX6-13F}$, $Orf3_{pCX9-45F}$ and $Orf4_{pCX9-45F}$ (Figures 4-7 and 4-8) have no homologues in the CAMERA database for environmental samples.

Table 4-7: List of genes with no identified function

CAMERA accession numbers are given for environmental samples' peptides. Sample IDs are as follows: GS123-Open ocean water sampled from the Indian Ocean in international waters between Madagascar and South Africa. GS000a-Open ocean water sampled from Saragasso Sea (Station 13), Bermuda. GS025- Fringing reef water sampled from the Eastern Tropical Pacific, Dirty Rock, Cocos Island, Costa Rica. GS020- Fresh water sampled from Panama Canal, Lake Gatun, Panama. SLUDGE_AUS- Sewage sludge collected from sequencing batch reactor at Thornside Sewage Treatment Plant, Brisbane, Australia. FARM SOIL- Clay loam surface soil collected from a farm in Waseca County, Minnesota. GS032-Water sample collected from a mangrove at Isabella Island, off the coast of Ecuador.

Contig	Orf designation	Best isolate homologue	Identity (%)	Best homologues from environmental samples	Identity (%)
pCX4-10F 13-1	Orf1$_{pCX4-10F}$	*Herminiimonas aresenicoxydans* HEAR2639	90	JCVI_PEP_1112736227834 GS123	78
pCX4-10F 6-2	Orf5$_{pCX4-10F}$	*Methylobacillus fagellatus* KT Mfla_1127	30	JCVI_PEP_1105110335817 GS000a	31
pCX4-10F 6-2	Orf3$_{pCX4-10F}$	-	-	JCVI_PEP_1105160931511 GS025	32
pCX6-13F-1	Orf2$_{pCX6-13F}$	*Dechloromonas aromatica* RCB Daro_1407	74	JCVI_PEP_1113974750160 SLUDGE_ AUS	45
pCX6-13F-1	Orf3$_{pCX6-13F}$	*Dechloromonas aromatica* RCB Daro_1409	79	-	-
pCX6-13F-1	Orf4$_{pCX6-13F}$	*Dechloromonas aromatica* RCB Daro_1418	83	JCVI_PEP_1113368395762 SLUDGE_ AUS	60
pCX6-13F-4	Orf5$_{pCX6-13F}$	*Plesiocystis pacifica* S-R1 PPSIR1_23339	40	-	-
pCX6-13F-4	Orf6$_{pCX6-13F}$	*Candidatus Nitrospira defluvii* UPF0047	82	JCVI_PEP_1105154612789 GS032	64
pCX9-	Orf1$_{pCX9-45F}$	*Desulfococcus*	49	JCVI_PEP_1113522379096	66

Contig	Orf designation	Best isolate homologue	Identity (%)	Best homologues from environmental samples	Identity (%)
45F		*deovorans* Hxd3 Dole_1205		FARM SOIL	
10-3-19 CX9-45F	Orf2$_{pCX9-45F}$	*Beggiatoa sp.* PSZP_01999465	31	JCVI_PEP_1105138462911 GS020	41
10-3-19 CX9-45F	Orf3$_{pCX9-45F}$	-	-	-	-
10-3-19 CX9-45F	Orf4$_{pCX9-45F}$	-	-	-	-
10-3-19 CX9-45F-18	Orf8$_{pCX9-45F}$	*Solibacter usitatus* Ellin6076 Acid_4938	36	JCVI_PEP_1113521680492 FARM SOIL	46

Chapter 5: Conclusions

A culture-independent approach was implemented to study phylogenetic diversity and gene function in bacteria inhabiting complex microbial communities. Bacterial P metabolism has been extensively characterized in cultivated microbes, on the genetic and biochemical level, and a great deal of knowledge has been accumulated with respect to P acquisition, transport, degradation and assimilation. However, since the proportion of cultured microbes on earth is relatively small, much more can be learned examining the mechanisms of P metabolism by microbes that have not yet been cultured. The investigations summarized here have uncovered functional and phylogenetic diversity that would not have likely been accounted for using culture dependent approaches.

The function-driven approach to metagenomics was successfully applied in the investigation of bacterial P metabolism, uncovering new variations of the functional machinery used by bacteria to transport and degrade the essential nutrient. By phenotypic complementation of *E. coli* and *S. meliloti*, about 10 new phosphorous metabolism genes were identified from metagenomic libraries derived from pulp and municipal waste activated sludge and soil communities. Over 80 additional genes were identified by sequencing 3 cosmids that complemented *S. meliloti* for growth on glyphosate. Since the selection methods used were directed towards the degradation and transport of P compounds to support growth, it is not surprising that about a quarter of genes identified are related to P metabolism.

Three phosphohydrolases were cloned, sequenced and characterized in detail. These included two NSAPs and one alkaline phosphatase belonging to the NPP family. The two NSAPs appear to be phylogenetically related and show a similar amino acid conservation pattern. However, they have different pH-rate profiles and their expression seems to be driven by different promoters. It is quite surprising that no PhoB binding sites were found upstream of NSAP $phoN_{BCX4.10}$ as pho regulon genes are generally regulated in a PhoB-dependent manner. The novelty of the two NSAPs is demonstrated by their appearance as a single clade in the phyologenetic tree. These differences show how using a culture-independent strategy uncovers proteins with new characteristics. The functionality of the NSAPs was supported by a molecular structure analysis as the secondary structure and key residues involved in catalysis were identified. The position and orientation of the residues making up the signature domains, in the elucidated structures, was consistent with their role in substrate binding and catalysis.

The new alkaline phosphatase, $phoA_{ACX6.71}$ identified from the metagenome of municipal waste activated sludge belongs to the NPP superfamily, a broad group that includes phosphohydrolases from bacterial and eukaryal origin. This is the first report of a functional enzyme from this family, isolated using a metagenomic approach. From the protein sequence, it is apparent that this enzyme is only distantly related to any known bacterial phosphatase. From a combination of sequence analysis and structural modeling, a reaction centre residue, a residue that stabilizes substrate-enzyme binding, and residues involved in metal coordination were identified and their orientation within the globular protein was shown to be consistent with the protein's functionality. The partial purification

and expression of phosphatases, identified by a metagenomic approach, is reported for the first time. Further biochemical characterization of the enzymes may reveal a broad substrate range and interesting kinetic properties such as a higher kcat/Km value than the ones reported for known enzymes in the same class. This demonstrates the power of activity screening as genes encoding the function of interest are more easily detected than by shotgun sequencing. Since the phosphatases described here have been identified by activity screening, it is not suprising to have identified new genes that demonstrate functional capabilities.

With the exception of $phoN_{ACX6.13}$, $phoN_{BCX4.10}$ and $phoA_{ACX6.71}$, all of the new genes in this study were identified by growth selection of bacterial mutants. 92 bacterial genes from soil and sludge were identified. The most unusual finding of this portion of the work is that a large number of genes for Pit and Pst transport systems for phosphate, were identified even though a phosphonate compound was used in the growth selection. The sequence analysis of PstS proteins from the metagenome compared to known PstS proteins, shows the lack of conservation of key residues in the metagenomically derived PstS, suggesting that these residues have evolved to be able to bind phosphonate in addition to or instead of phosphate. The results raise the possibility that the Pit and Pst transporters, identified from the metagenome, are able to take up phosphonate. However, functional characterization of these proteins is required to confirm this hypothesis. In addition to transport proteins, a number of *phoU* regulators of P transport and a two-component regulatory system *phoR-phoB* were identified. Surprisingly, no typical phosphonate degradation genes were identified, suggesting a new mechanism for the process, perhaps by means of a new hydrolytic phosphatase with properties allowing it to overcome the recalcitrance of the carbon-

phosphorous bond. This mechanism could be easily identified by performing transposon mutagenesis on each of pCX4-10F, pCX6-13F and pCX9-45F, selecting for a number of insertions that abolish the glyphosate growth complementation of RmF726. The locus encoding the degradation machinery could then be identified by sequencing the insertion clones.

Aside from P metabolism, genes encoding additional cellular functions were identified and annotated. A relatively large portion of these genes (12%) is involved in regulatory functions and signal transduction mechanisms. This is somewhat expected as nutritional and biophysical conditions are changing in complex communities such as soil and sludge, requiring the fine tuning of regulatory mechanisms to control the expression of the appropriate genes for the appropriate conditions. Other genes found in relatively large abundance are predicted to be involved in cellular defense mechanisms, primarily detoxification. This is expected as the exposure to toxic compounds may occur in contaminated and sludge communities, requiring efficient mechanisms of cellular detoxification. In soils, compounds occurring in plant root exudates may be toxic to bacteria, requiring the ability of the cell to defend against their accumulation. It is not surprising that 30% of the genes identified, either do not encode any known functions or encode general functions such as transport permeases, haloacid dehalogenase hydrolases, or cell surface antigens with no known targets. Since the majority of the microbes in the biosphere have not been identified by culturing, a large portion of their genetic information remains unknown. With sequence-driven metagenomics, a great deal of this unknown DNA is being discovered. However, the rate of accumulation of new DNA sequences is much higher compared to the

rate at which the functions encoded in these sequences is determined. Therefore the use of the more directed function-driven metagenomics helps elucidate the roles of these newly discovered genes in the environment.

In the short time the field of metagenomics has emerged, great advances have been demonstrated in the field of microbiology, allowing for the discovery of new bacterial lineages, functions and metabolic capacity and the bioprospecting for enzymes of biotechnological importance. In the era where our society is searching for alternative biological resources for fuel, environmentally friendly household reagents or new ways for organopollutant bioremediation, microorganisms are the ultimate source for these desired functions. With the advancement of highthroughput screening, the identification of novel functions and activities in uncultured microbes has become more rapid and convenient. Furthermore, with the advancement of DNA sequencing technologies, obtaining large scale sequence data is gradually becoming more within reach and less expensive. Consequently, the metagenome of microbial communities can be easily reconstructed and a large number of new phylotypes can be easily identified. Thus, with the help of metagenomics, medical advancements can be made by studying the human microbiome, agricultural advancements can be made by examing microbial communities associated with plants and the treatment of water systems can be improved by exploring the microbial communities in activated sludge.
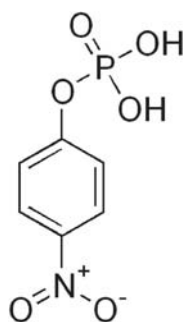
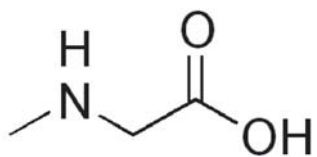# Appendix: Structures of chemicals mentioned in this study
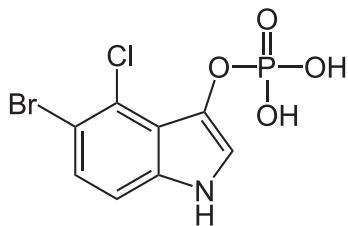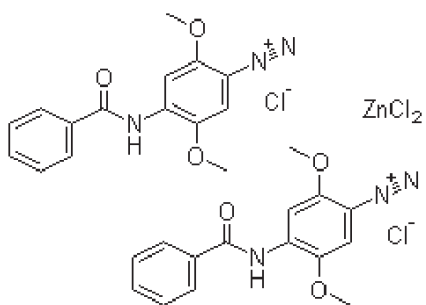
5-bromo-4-chloro-indolyl phosphate (BCIP)
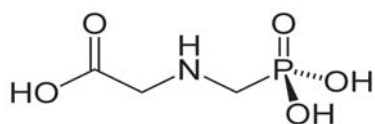
N-phosphonomethyl glycine (Glyphosate)

p-nitrophenyl phosphate (pNPP)

Sarcosine

4-Benzoylamino-2,5-dimethoxybenzenediazonium

chloride hemi(zinc chloride) salt (Fast Blue- RR): Taken

from www.chemblink.com

Bibliography

Allenby, N. E. E., O'Connor, N., Pragai, Z. & other authors (2004). Post-transcriptional regulation of the *Bacillus subtilis pst* operon encoding a phosphate-specific ABC transporter. *Microbiology* **150**, 2619-2628.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**, 3389-3402.

Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**, 143-169.

Antoine, R. & Locht, C. (1992). Isolation and molecular characterization of a novel broad-host-range plasmid from *Bordetella bronchiseptica* with sequence similarities to plasmids from gram-positive organisms. *Mol Microbiol* **6**, 1785-1799.

Baca, A. M., Sirawaraporn, R., Turley, S., Sirawaraporn, W. & Hol, W. G. J. (2000). Crystal structure of *Mycobacterium tuberculosis* 6-hydroxymethyl-7,8-dihydropteroate synthase in complex with pterin monophosphate: new insight into the enzymatic mechanism and sulfa-drug action. *Journal of Molecular Biology* **302**, 1193-1212.

Balthazor, T. M. & Hallas, L. E. (1986). Glyphosate-degrading microorganisms from industrial activated sludge. *Appl Environ Microbiol* **51**, 432-434.

Bardin, S., Dan, S., Osteras, M. & Finan, T. (1996). A phosphate transport system is required for symbiotic nitrogen fixation by *Rhizobium meliloti*. *J Bacteriol* **178**, 4540-4547.

Bardin, S. D. & Finan, T. M. (1998). Regulation of phosphate assimilation in *Rhizobium (Sinorhizobium) meliloti*. *Genetics* **148**, 1689-1700.

Bardin, S. D., Voegele, R. T. & Finan, T. M. (1998). Phosphate assimilation in *Rhizobium (Sinorhizobium) meliloti*: identification of a pit-like gene. *J Bacteriol* **180**, 4219-4226.

Beja, O., Aravind, L., Koonin, E. V. & other authors (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902-1906.

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

Beringer, J. E. (1980). The development of *Rhizobium* genetics. *J Gen Microbiol* **116**, 1-7.

Berlutti, F., Passariello, C., Selan, L., Thaller, M. C. & Rossolini, G. M. (2001). The *Chryseobacterium meningosepticum* PafA enzyme: prototype of a new enzyme family of prokaryotic phosphate-irrepressible alkaline phosphatases? *Microbiology* **147**, 2831-2839.

Bhargava, T., Datta, S., Ramachandran, V., Ramakrishnan, R., Roy, K. R., Sankaran, K. & K., S. Y. V. B. (1995). Virulent *Shigella* codes for a soluble apyrase: Idenitification, characterization and cloning of the gene. *Curr Sci* **68**, 293-300.

Bliska, J. B., Guan, K. L., Dixon, J. E. & Falkow, S. (1991). Tyrosine phosphate hydrolysis of host proteins by an essential *Yersinia* virulence determinant. *Proc Natl Acad Sci U S A* **88**, 1187-1191.

Bollen, M., Gijsbers, R., Ceulemans, H., Stalmans, W. & Stefan, C. (2000). Nucleotide pyrophosphatases/phosphodiesterases on the move. *Crit Rev Biochem Mol Biol* **35**, 393-432.

Bond, P., Hugenholtz, P., Keller, J. & Blackall, L. (1995). Bacterial community structures of phosphate-removing and non-phosphate- removing activated sludges from sequencing batch reactors. *Appl Environ Microbiol* **61**, 1910-1916.

Bottger, P. & Pedersen, L. (2005). Evolutionary and experimental analyses of inorganic phosphate transporter PiT family reveals two related signature sequences harboring highly conserved aspartic acids critical for sodium-dependent phosphate transport function of human PiT2 *FEBS Journal* **272**, 3060-3074.

Boyer, H. W. & Roulland-Dussoix, D. (1969). A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. *J Mol Biol* **41**, 459-472.

Brady, S. F. & Clardy, J. (2000). Long-chain N-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. *J Am Chem Soc* **122**, 12903-12904.

Brady, S. F., Chao, C. J. & Clardy, J. (2002). New natural product families from an environmental DNA (eDNA) gene cluster. *J Am Chem Soc* **124**, 9968-9969.

Burris, R. H. (1991). Nitrogenases. *J Biol Chem* **266**, 9339-9342.

Butler, J. L., Williams, M. A., Bottomley, P. J. & Myrold, D. D. (2003). Microbial community dynamics associated with rhizosphere carbon flow. *Appl Environ Microbiol* **69**, 6793-6800.

Campbell, J. W. & Cronan, J. E., Jr. (2001). *Escherichia coli* FadR positively regulates transcription of the fabB fatty acid biosynthetic gene. *J Bacteriol* **183**, 5982-5990.

Chan, C., Paul, R., Samoray, D., Amiot, N. C., Giese, B., Jenal, U. & Schirmer, T. (2004). Structural basis of activity and allosteric control of diguanylate cyclase. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 17084-17089.

Charles, T. C. & Finan, T. M. (1991). Analysis of a 1600-kilobase *Rhizobium meliloti* Megaplasmid using defined deletions generated *in vivo*. *Genetics* **127**, 5-20.

Charles, T. C., Newcomb, W. & Finan, T. M. (1991). *ndvF*, a novel locus located on megaplasmid pRmeSU47b (pEXO) of *Rhizobium meliloti*, is required for normal nodule development. *J Bacteriol* **173**, 3981-3992.

Chen, C., Ye, Q., Zhu, Z., Wanner, B. & Walsh, C. (1990). Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the *phn* (*psiD*) genes involved in alkylphosphonate uptake and C-P lyase activity in *Escherichia coli* B. *J Biol Chem* **265**, 4461-4471.

Chen, G., Zhu, H. & Zhang, Y. (2003). Soil microbial activities and carbon and nitrogen fixation. In *Research in Microbiology*, pp. 393-398: Elsevier Science.

Chen, X., Ansai, T., Awano, S., Iida, T., Barik, S. & Takehara, T. (1999). Isolation, cloning, and expression of an acid phosphatase containing phosphotyrosyl phosphatase activity from *Prevotella intermedia*. *J Bacteriol* **181**, 7107-7114.

Cohen, S. N., Chang, A. C. & Hsu, L. (1972). Nonchromosomal antibiotic resistance in bacteria: genetic transformation of *Escherichia coli* by R-factor DNA. *Proc Natl Acad Sci U S A* **69**, 2110-2114.

Conrad, R. (1996). Soil microorganisms as controllers of atmospheric trace gases (H_2, CO, CH_4, OCS, N_2O, and NO). *Microbiol Rev* **60**, 609-640.

Cox, G. B., Webb, D. & Rosenberg, H. (1989). Specific amino acid residues in both the PstB and PstC proteins are required for phosphate transport by the *Escherichia coli* Pst system. *J Bacteriol* **171**, 1531-1534.

Delgado, J. A. & Follett, R. F. (2002). Carbon and nutrient cycles. *J Soil Water Conserv* **57**, 455-464.

Delwart, E. L. (2007). *Viral metagenomics*: John Wiley & Sons, Ltd.

Dowling, J. N., Saha, A. K. & Glew, R. H. (1992). Virulence factors of the family *Legionellaceae*. *Microbiol Rev* **56**, 32-60.

Dyrlov Bendtsen, J., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* **340**, 783-795.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.

Finan, T. M., Kunkel, B., De Vos, G. F. & Signer, E. R. (1986). Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. *J Bacteriol* **167**, 66-72.

Frodyma, M. E. & Downs, D. (1998). ApbA, the ketopantoate reductase enzyme of *Salmonella typhimurium* is required for the synthesis of thiamine via the alternative pyrimidine biosynthetic pathway. *J Biol Chem* **273**, 5572-5576.

Gabor, E. M., Alkema, W. B. L. & Janssen, D. B. (2004a). Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**, 879-886.

Gabor, E. M., de Vries, E. J. & Janssen, D. B. (2004b). Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environ Microbiol* **6**, 948-958.

Galperin, M. Y., Bairoch, A. & Koonin, E. V. (1998). A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases. *Protein Sci* **7**, 1829-1835.

Garcia-Martin, H., Ivanova, N., Kunin, V. & other authors (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**, 1263-1269.

Gijsbers, R., Ceulemans, H., Stalmans, W. & Bollen, M. (2001). Structural and catalytic similarities between nucleotide pyrophosphatases/phosphodiesterases and alkaline phosphatases. *J Biol Chem* **276**, 1361-1368.

Gill, S. R., Pop, M., DeBoy, R. T. & other authors (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355-1359.

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., Clardy, J., Goodman, R. M. & Handelsman, J. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**, 4301-4306.

Gomez, P. F. & Ingram, L. O. (1995). Cloning, sequencing and characterization of the alkaline phosphatase gene (*phoD*) from *Zymomonas mobilis*. *FEMS Microbiology Letters* **125**, 237-245.

Guan, K. L. & Dixon, J. E. (1990). Protein tyrosine phosphatase activity of an essential virulence determinant in *Yersinia*. *Science* **249**, 553-556.

Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.

Gupta, A., Meyer, J. M. & Goel, R. (2002a). Development of heavy metal-resistant mutants of phosphate solubilizing *Pseudomonas sp*. NBRI 4014 and their characterization. *Curr Microbiol* **45**, 323-327.

Gupta, R., Beg, Q. K. & Lorenz, P. (2002b). Bacterial alkaline proteases: molecular approaches and industrial applications. *Appl Microbiol Biotechnol* **59**, 15-32.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and anlysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**, 95-98.

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* **166**, 557-580.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**, R245-249.

201

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**, 669-685.

Harris, R. M., Webb, D. C., Howitt, S. M. & Cox, G. B. (2001). Characterization of PitA and PitB from *Escherichia coli*. *J Bacteriol* **183**, 5008-5014.

Henne, A., Daniel, R., Schmitz, R. A. & Gottschalk, G. (1999). Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol* **65**, 3901-3907.

Huang, J., Su, Z. & Xu, Y. (2005). The evolution of microbial phosphonate degradative pathways. *Journal of Molecular Evolution* **61**, 682-690.

Imazu, K., Tanaka, S., Kuroda, A., Anbe, Y., Kato, J. & Ohtake, H. (1998). Enhanced utilization of phosphonate and phosphite by *Klebsiella aerogenes*. *Appl Environ Microbiol* **64**, 3754-3758.

Ishikawa, K., Mihara, Y., Gondoh, K., Suzuki, E.-i. & Asano, Y. (2000). X-ray structures of a novel acid phosphatase from *Escherichia blattae* and its complex with the transition-state analog molybdate. *EMBO J* **19**, 2412-2423.

Jeon, C. O., Lee, D. S. & Park, J. M. (2003). Microbial communities in activated sludge performing enhanced biological phosphorus removal in a sequencing batch reactor. *Water Research* **37**, 2195-2205.

Jiang, W., Metcalf, W., Lee, K. & Wanner, B. (1995). Molecular cloning, mapping, and regulation of Pho regulon genes for phosphonate breakdown by the phosphonatase pathway of *Salmonella typhimurium* LT2. *J Bacteriol* **177**, 6411-6421.

Jones, J. D. & Gutterson, N. (1987). An efficient mobilizable cosmid vector, pRK7813, and its use in a rapid method for marker exchange in *Pseudomonas fluorescens* strain HV37a. *Gene* **61**, 299-306.

Jorgensen, C. I., Kallipolitis, B. H. & Valentin-Hansen, P. (1998). DNA-binding characteristics of the *Escherichia coli* CytR regulator: a relaxed spacing requirement between operator half-sites is provided by a flexible, unstructured interdomain linker. *Mol Microbiol* **27**, 41-50.

Jouanneau, Y., Roby, C., Meyer, C. M. & Vignais, P. M. (1989). ADP-ribosylation of dinitrogenase reductase in *Rhodobacter capsulatus*. *Biochemistry* **28**, 6524-6530.

Kane, S. R., Chakicherla, A. Y., Chain, P. S. G. & other authors (2007). Whole-genome analysis of the methyl tert-butyl ether-degrading *beta-Proteobacterium Methylibium petroleiphilum* PM1. *J Bacteriol* **189**, 1931-1945.

Karp, G. (2005). *Cell and Molecular Biology - Concepts and Experiments*, 4th edn. Hoboken, NJ, U.S.A.: John Wiley and Sons.

Kasahara, M., Nakata, A. & Shinagawa, H. (1991). Molecular analysis of the *Salmonella typhimurium phoN* gene, which encodes nonspecific acid phosphatase. *J Bacteriol* **173**, 6760-6765.

Kennedy, J., Marchesi, J. & Dobson, A. (2007). Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Applied Microbiology and Biotechnology* **75**, 11-20.

Kertesz, M., Elgorriaga, A. & Amrhein, N. (1991). Evidence for two distinct phosphonate-degrading enzymes (C-P lyases) in Arthrobacter sp. GLP-1. *Biodegradation* **2**, 53-59.

Kier, L. D., Weppelman, R. & Ames, B. N. (1977). Resolution and purification of three periplasmic phosphatases of *Salmonella typhimurium*. *J Bacteriol* **130**, 399-410.

Kim, S. Y., Nam, T. W., Shin, D., Koo, B. M., Seok, Y. J. & Ryu, S. (1999). Purification of Mlc and analysis of its effects on the pts expression in *Escherichia coli. J Biol Chem* **274**, 25398-25402.

Kim, Y.-J., Choi, G.-S., Kim, S.-B., Yoon, G.-S., Kim, Y.-S. & Ryu, Y.-W. (2006). Screening and characterization of a novel esterase from a metagenomic library. *Protein Expression and Purification* **45**, 315-323.

Kirk, J. L., Beaudette, L. A., Hart, M., Moutoglis, P., Klironomos, J. N., Lee, H. & Trevors, J. T. (2004). Methods of studying soil microbial diversity. *J Microbiol Methods* **58**, 169-188.

Kirkland, T. N., Finley, F., Orsborn, K. I. & Galgiani, J. N. (1998). Evaluation of the proline-rich antigen of *Coccidioides immitis* as a vaccine candidate in mice. *Infect Immun* **66**, 3519-3522.

Knusden, B., Knusden, T., Flensborg, M. & other authors (2005). CLC free workbench, pp. A free workbench for basic bioinformatics. Aarhus, Denmark.

Kononova, S. V. & Nesmeyanova, M. A. (2002). Phosphonates and their degradation by microorganisms. *Biochemistry (Mosc)* **67**, 184-195.

Koonin, E. V. & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J Mol Biol* **244**, 125-132.

Kornberg, A. (1999). Inorganic polyphosphate: a molecule of many functions. *Prog Mol Subcell Biol* **23**, 1-18.

Krzysko-Lupicka, T., Strof, W., Kubs, K., Skorupa, M., Wieczorek, P., Lejczak, B. & Kafarski, P. (1997). The ability of soil-borne fungi to degrade organophosphonate carbon-to-phosphorus bonds. *Appl Microbiol Biotechnol* **48**, 549-552.

Kumar, A., Grimes, B., Fujita, N., Makino, K., Malloch, R. A., Hayward, R. S. & Ishihama, A. (1994). Role of the sigma 70 subunit of *Escherichia coli* RNA polymerase in transcription activation. *J Mol Biol* **235**, 405-413.

La Nauze, J. M., Rosenberg, H. & Shaw, D. C. (1970). The enzymic cleavage of the carbon-phosphorus bond: purification and properties of phosphonatase. *Biochim Biophys Acta* **212**, 332-350.

Lammle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H. & Rupp, S. (2007). Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J Biotechnol* **127**, 575-592.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L. & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* **82**, 6955-6959.

Larkin, M. A., Blackshields, G., Brown, N. P. & other authors (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Lee, K. S., Metcalf, W. W. & Wanner, B. L. (1992). Evidence for two phosphonate degradative pathways in *Enterobacter aerogenes*. *J Bacteriol* **174**, 2501-2510.

Lee, S. H., Hidaka, T., Nakashita, H. & Seto, H. (1995). The carboxyphosphonoenolpyruvate synthase-encoding gene from the bialaphos-producing organism *Streptomyces hygroscopicus*. *Gene* **153**, 143-144.

Leveau, J. H. J., Gerards, S., de Boer, W. & van Veen, J. A. (2004). Phylogeny-function analysis of (meta)genomic libraries: screening for expression of ribosomal RNA genes by large-insert library fluorescent *in situ* hybridization (LIL-FISH). *Environ Microbiol* **6**, 990-998.

Li, X. & Qin, L. (2005). Metagenomics-based drug discovery and marine microbial diversity. *Trends Biotechnol* **23**, 539-543.

Liles, M. R., Manske, B. F., Bintrim, S. B., Handelsman, J. & Goodman, R. M. (2003). A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* **69**, 2684-2691.

Lim, H. K., Chung, E. J., Kim, J.-C., Choi, G. J., Jang, K. S., Chung, Y. R., Cho, K. Y. & Lee, S.-W. (2005a). Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Appl Environ Microbiol* **71**, 7768-7777.

Lim, Y. W., Lee, S. A., Kim, S. B., Yong, H. Y., Yeon, S. H., Park, Y. K., Jeong, D. W. & Park, J. S. (2005b). Diversity of denitrifying bacteria isolated from Daejeon Sewage Treatment Plant. *J Microbiol* **43**, 383-390.

Liu, C. M., McLean, P. A., Sookdeo, C. C. & Cannon, F. C. (1991). Degradation of the herbicide glyphosate by members of the family *Rhizobiaceae*. *Appl Environ Microbiol* **57**, 1799-1804.

Liu, J., Lou, Y., Yokota, H., Adams, P. D., Kim, R. & Kim, S.-H. (2005). Crystal structure of a PhoU protein homologue: A new class of metalloprotein containing multinuclear iron clusters. *J Biol Chem* **280**, 15960-15966.

Long, P. F., Dunlap, W. C., Battershill, C. N. & Jaspars, M. (2005). Shotgun cloning and heterologous expression of the patellamide gene cluster as a strategy to achieving sustained metabolite production: WILEY-VCH Verlag.

Lu, S., Park, M., Ro, H. S., Lee, D. S., Park, W. & Jeon, C. O. (2006). Analysis of microbial communities using culture-dependent and culture-independent approaches in an anaerobic/aerobic SBR reactor. *J Microbiol* **44**, 155-161.

Magasanik, B. (1989). Regulation of transcription of the *glnALG* operon of *Escherichia coli* by protein phosphorylation. *Biochimie* **71**, 1005-1012.

Majernik, A., Gottschalk, G. & Daniel, R. (2001). Screening of environmental DNA libraries for the presence of genes conferring Na(+)(Li(+))/H(+) antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. *J Bacteriol* **183**, 6645-6653.

Majumdar, A., Ghatak, A. & Ghosh, R. K. (2005). Identification of the gene for the monomeric alkaline phosphatase of *Vibrio cholerae* serogroup O1 strain. *Gene* **344**, 251-258.

Makde, R. D., Kumar, V., Rao, A. S., Yadava, V. S. & Mahajan, S. K. (2003). Purification, crystallization and preliminary X-ray diffraction studies of recombinant class A non-specific acid phosphatase of *Salmonella typhimurium*. *Acta Crystallogr D Biol Crystallogr* **59**, 515-518.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F. & other authors (2005). CDD: a Conserved Domain Database for protein classification. *Nucl Acids Res* **33**, D192-196.

Marchler-Bauer, A., Anderson, J. B., Chitsaz, F. & other authors (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucl Acids Res* **37**, D205-210.

Martinez, A., Kolvek, S. J., Yip, C. L. T., Hopke, J., Brown, K. A., MacNeil, I. A. & Osburne, M. S. (2004). Genetically modified bacterial strains and novel Bacterial Artificial Chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. *Appl Environ Microbiol* **70**, 2452-2463.

McGrath, J. W., Hammerschmidt, F. & Quinn, J. P. (1998). Biodegradation of phosphonomycin by *Rhizobium huakuii* PMY1. *Appl Environ Microbiol* **64**, 356-358.

McMullan, G. & Quinn, J. P. (1994). *In vitro* characterization of a phosphate starvation-independent carbon-phosphorus bond cleavage activity in *Pseudomonas fluorescens* 23F. *J Bacteriol* **176**, 320-324.

Meade, H. M., Long, S. R., Ruvkun, G. B., Brown, S. E. & Ausubel, F. M. (1982). Physical and genetic characterization of symbiotic and auxotrophic mutants of *Rhizobium meliloti* induced by transposon Tn5 mutagenesis. *J Bacteriol* **149**, 114-122.

Mendz, G. L., Mégraud, F. & Korolik, V. (2005). Phosphonate catabolism by *Campylobacter spp. Archives of Microbiology* **183**, 113-120.

Metcalf, W. W. & Wanner, B. L. (1993). Mutational analysis of an *Escherichia coli* fourteen-gene operon for phosphonate degradation, using TnphoA' elements. *J Bacteriol* **175**, 3430-3442.

Monds, R. D., Newell, P. D., Schwartzman, J. A. & O'Toole, G. A. (2006). Conservation of the Pho regulon in *Pseudomonas fluorescens* Pf0-1. *Appl Environ Microbiol* **72**, 1910-1924.

Moore, L. R., Ostrowski, M., Scanlan, D. J., Feren, K. & Sweetsir, T. (2005). Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. *Aquatic Microbial Ecology* **39**, 257-269.

Morris, C. E., Bardin, M., Berge, O. & other authors (2002). Microbial biodiversity: approaches to experimental design and hypothesis testing in primary scientific literature from 1975 to 1999. *Microbiol Mol Biol Rev* **66**, 592-616.

Morse, S. A. (1989). Chancroid and *Haemophilus ducreyi*. *Clin Microbiol Rev* **2**, 137-157.

Muller, D., Simeonova, D. D., Riegel, P., Mangenot, S., Koechler, S., Lievremont, D., Bertin, P. N. & Lett, M. C. (2006). *Herminiimonas arsenicoxydans sp. nov*, a metalloresistant bacterium. *Int J Syst Evol Microbiol* **56**, 1765-1769.

Munch, R., Hiller, K., Grote, A., Scheer, M., Klein, J., Schobert, M. & Jahn, D. (2005). Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics* **21**, 4187-4189.

Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* **84**, 166-169.

Neufeld, J. D., Driscoll, B. T., Knowles, R. & Archibald, F. S. (2001). Quantifying functional gene populations: comparing gene abundance and corresponding enzymatic activity using denitrification and nitrogen fixation in pulp and paper mill effluent treatment systems. *Can J Microbiol* **47**, 925-934.

Novak, R., Cauwels, A., Charpentier, E. & Tuomanen, E. (1999). Identification of a *Streptococcus pneumoniae* gene locus encoding proteins of an ABC phosphate transporter and a two-component regulatory system. *J Bacteriol* **181**, 1126-1133.

Oganesyan, V., Oganesyan, N., Adams, P. D., Jancarik, J., Yokota, H. A., Kim, R. & Kim, S. H. (2005). Crystal structure of the "PhoU-like" phosphate uptake regulator from *Aquifex aeolicus*. *J Bacteriol* **187**, 4238-4244.

Parker, G. F., Higgins, T. P., Hawkes, T. & Robson, R. L. (1999). *Rhizobium (Sinorhizobium) meliloti phn* genes: Characterization and identification of their protein products. *J Bacteriol* **181**, 389-395.

Passariello, C., Schippa, S., Iori, P., Berlutti, F., Thaller, M. C. & Rossolini, G. M. (2003). The molecular class C acid phosphatase of *Chryseobacterium meningosepticum* (OlpA) is a broad-spectrum nucleotidase with preferential activity on 5'-nucleotides. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1648**, 203-209.

Passariello, C., Forleo, C., Micheli, V., Schippa, S., Leone, R., Mangani, S., Thaller, M. C. & Rossolini, G. M. (2006). Biochemical characterization of the class B acid phosphatase (AphA) of *Escherichia coli* MG1655. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1764**, 13-19.

Paterson, E. (2003). Importance of rhizodeposition in the coupling of plant and microbial productivity. *European Journal of Soil Science* **54**, 741-750.

Payne, G. W., Ramette, A., Rose, H. L., Weightman, A. J., Jones, T. H., Tiedje, J. M. & Mahenthiralingam, E. (2006). Application of a *recA* gene-based identification approach to the maize rhizosphere reveals novel diversity in *Burkholderia species*. *FEMS Microbiology Letters* **259**, 126-132.

Paytan, A. & McLaughlin, K. (2007). The oceanic phosphorus cycle. *Chemical Reviews* **107**, 563-576.

Peiffer, W. E., Desrosiers, M. G. & Menick, D. R. (1996). Cloning and expression of the unique Ca^{2+}-ATPase from *Flavobacterium odoratum*. *J Biol Chem* **271**, 5095-5100.

Pond, J. L., Eddy, C. K., Mackenzie, K. F., Conway, T., Borecky, D. J. & Ingram, L. O. (1989). Cloning, sequencing, and characterization of the principal acid phosphatase, the $phoC^{+}$ product, from *Zymomonas mobilis*. *J Bacteriol* **171**, 767-774.

Qi, Y., Kobayashi, Y. & Hulett, F. M. (1997). The *pst* operon of *Bacillus subtilis* has a phosphate-regulated promoter and is involved in phosphate transport but not in regulation of the *pho* regulon. *Journal of Bacteriology* **179**, 2534-2539.

Quinn, J. P. A. N. K. N. A. C. J. W. M. (2007). New ways to break an old bond: the bacterial carbon-phosphorus hydrolases and their role in biogeochemical phosphorus cycling. *Environmental Microbiology* **9**, 2392-2400.

Ranjan, R., Grover, A., Kapardar, R. K. & Sharma, R. (2005). Isolation of novel lipolytic genes from uncultured bacteria of pond water. *Biochemical and Biophysical Research Communications* **335**, 57-65.

Rao, F., Yang, Y., Qi, Y. & Liang, Z.-X. (2008). Catalytic mechanism of cyclic di-GMP-specific phosphodiesterase: a study of the EAL domain-containing RocR from *Pseudomonas aeruginosa*. *J Bacteriol* **190**, 3622-3631.

Reilly, T. J., Baron, G. S., Nano, F. E. & Kuhlenschmidt, M. S. (1996). Characterization and sequencing of a respiratory burst-inhibiting acid phosphatase from *Francisella tularensis*. *J Biol Chem* **271**, 10973-10983.

Rice, C. D., Pollard, J. E., Lewis, Z. T. & McCleary, W. R. (2009). Employment of a promoter-swapping technique shows that PhoU modulates the activity of the PstSCAB2 ABC Transporter in *Escherichia coli*. *Appl Environ Microbiol* **75**, 573-582.

Ricke, P., Kube, M., Nakagawa, S., Erkel, C., Reinhardt, R. & Liesack, W. (2005). First genome data from uncultured upland soil cluster alpha methanotrophs provide further evidence for a close phylogenetic relationship to *Methylocapsa acidiphila* B2 and for high-affinity methanotrophy involving particulate methane monooxygenase. *Appl Environ Microbiol* **71**, 7472-7482.

Rodriguez, H. & Fraga, R. (1999). Phosphate solubilizing bacteria and their role in plant growth promotion. *Biotechnol Adv* **17**, 319-339.

Rondon, M. R., August, P. R., Bettermann, A. D. & other authors (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**, 2541-2547.

Rosenberg, H. & La Nauze, J. M. (1967). The metabolism of phosphonates by microorganisms. The transport of aminoethylphosphonic acid in *Bacillus cereus*. *Biochim Biophys Acta* **141**, 79-90.

Rossolini, G. M., Schippa, S., Riccio, M. L., Berlutti, F., Macaskie, L. E. & Thaller, M. C. (1998). Bacterial nonspecific acid phosphohydrolases: physiology, evolution and use as tools in microbial biotechnology. *Cell Mol Life Sci* **54**, 833-850.

Rusch, D. B., Halpern, A. L., Sutton, G. & other authors (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**, e77.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945.

Sambrook, J. & Russel, D. W. (2001). *Molecular cloning- A laboratory manual*, 3rd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Schlegel, A., Bohm, A., Lee, S. J., Peist, R., Decker, K. & Boos, W. (2002). Network regulation of the *Escherichia coli* maltose system. *J Mol Microbiol Biotechnol* **4**, 301-307.

Schmidt, A. J., Ryjenkov, D. A. & Gomelsky, M. (2005). The ubiquitous protein domain EAL is a cyclic diguanylate-specific phosphodiesterase: enzymatically active and inactive EAL domains. *J Bacteriol* **187**, 4774-4781.

Schowanek, D. & Verstraete, W. (1990). Phosphonate utilization by bacterial cultures and enrichments from environmental samples. *Appl Environ Microbiol* **56**, 895-903.

Sebastian, M. & Ammerman, J. W. (2009). The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. *Isme J*.

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. (2007). CAMERA: A community resource for metagenomics. *PLoS Biology* **5**, e75.

Shoji, T., Nittami, T., Onuki, M., Satoh, H. & Mino, T. (2006). Microbial community of biological phosphorus removal process fed with municipal wastewater under different electron acceptor conditions. *Water Sci Technol* **54**, 81-89.

Simon, C., Herath, J., Rockstroh, S. & Daniel, R. (2009). Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl Environ Microbiol* **75**, 2964-2968.

Singh, B. K., Millard, P., Whiteley, A. S. & Murrell, J. C. (2004). Unravelling rhizosphere-microbial interactions: opportunities and limitations. *Trends Microbiol* **12**, 386-393.

Singh, B. K. & Walker, A. (2006). Microbial degradation of organophosphorus compounds *FEMS Microbiology Reviews* **30**, 428-471.

Singh, B. K. (2009). Organophosphorus-degrading bacteria: ecology and industrial applications. *Nat Rev Micro* **7**, 156-164.

Smith, N. R., Yu, Z. & Mohn, W. W. (2003). Stability of the bacterial community in a pulp mill effluent treatment system during normal operation and a system shutdown. *Water Res* **37**, 4873-4884.

Snaidr, J., Amann, R., Huber, I., Ludwig, W. & Schleifer, K. (1997). Phylogenetic analysis and *in situ* identification of bacteria in activated sludge. *Appl Environ Microbiol* **63**, 2884-2896.

Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. (1985). Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl Environ Microbiol* **49**, 1379-1384.

Steed, P. M. & Wanner, B. L. (1993). Use of the *rep* technique for allele replacement to construct mutants with deletions of the *pstSCAB-phoU* operon: evidence of a new role for the PhoU protein in the phosphate regulon. *J Bacteriol* **175**, 6797-6809.

Steele, H. L. & Streit, W. R. (2005). Metagenomics: advances in ecology and biotechnology. *FEMS Microbiology Letters* **247**, 105-111.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**, 591-599.

Steinrucken, H. C. & Amrhein, N. (1980). The herbicide glyphosate is a potent inhibitor of 5-enolpyruvylshikimic acid-3-phosphate synthase. *Biochemical and Biophysical Research Communications* **94**, 1207-1212.

Studier, F. W. (2005). Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207-234.

Stukey, J. & Carman, G. M. (1997). Identification of a novel phosphatase sequence motif. *Protein Sci* **6**, 469-472.

Swofford, D. L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Sunderland, Massachusetts: Sinauer Associates.

Talarico, T. L., Dev, I. K., Dallas, W. S., Ferone, R. & Ray, P. H. (1991). Purification and partial characterization of 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase and 7,8-dihydropteroate synthase from *Escherichia coli* MC4100. *J Bacteriol* **173**, 7029-7032.

Tanabe, M., Mirza, O., Bertrand, T., Atkins, H. S., Titball, R. W., Iwata, S., Brown, K. A. & Byrne, B. (2007). Structures of OppA and PstS from *Yersinia pestis* indicate variability of interactions with transmembrane domains. *Acta Crystallographica Section D* **63**, 1185-1193.

Ternan, N. G., Mc Grath, J. W., Mc Mullan, G. & Quinn, J. P. (1998). Review: Organophosphonates: occurrence, synthesis and biodegradation by microorganisms. *World Journal of Microbiology & Biotechnology* **14**, 635-647.

Ternan, N. G. & Quinn, J. P. (1998). In vitro cleavage of the carbon-phosphorus bond of phosphonopyruvate by cell extracts of an environmental *Burkholderia cepacia* isolate. *Biochem Biophys Res Commun* **248**, 378-381.

Thaller, M. C., Berlutti, F., Schippa, S., Lombardi, G. & Rossolini, G. M. (1994). Characterization and sequence of PhoC, the principal phosphate-irrepressible acid phosphatase of *Morganella morganii*. *Microbiology* **140 (Pt 6)**, 1341-1350.

Thaller, M. C., Berlutti, F., Schippa, S., Iori, P., Passariello, C. & Rossolini, G. M. (1995a). Heterogeneous patterns of acid phosphatases containing low-molecular-mass polypeptides in members of the family *Enterobacteriaceae*. *International Journal of Systematic Bacteriology* **45**, 225-261.

Thaller, M. C., Lombardi, G., Berlutti, F., Schippa, S. & Rossolini, G. M. (1995b). Cloning and characterization of the NapA acid phosphatase/phosphotransferase of *Morganella morganii*: identification of a new family of bacterial acid-phosphatase-encoding genes. *Microbiology* **141 (Pt 1)**, 147-154.

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

Thaller, M. C., Schippa, S., Iori, P., Berlutti, F. & Rossolini, G. M. (1997b). Cloning of a *Chryseobacterium meningosepticum* acid phosphatse-encoding gene: Identification of a family of outer membrane bacterial phosphatases. In *97th General Meeting of American Society for Microbiology*. Miami Beach, .

Thaller, M. C., Schippa, S. & Rossolini, G. M. (1998). Conserved sequence motifs among bacterial, eukaryotic, and archaeal phosphatases that define a new phosphohydrolase superfamily. *Protein Sci* **7**, 1647-1652.

Torsvik, V., Ovreas, L. & Thingstad, T. F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science* **296**, 1064-1066.

Travers, A., Schneider, R. & Muskhelishvili, G. (2001). DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. *Biochimie* **83**, 213-217.

Tremblay, L. W., Dunaway-Mariano, D. & Allen, K. N. (2006). Structure and activity analyses of *Escherichia coli* K-12 NagD provide insight into the evolution of biochemical function in the haloalkanoic acid dehalogenase superfamily. *Biochemistry* **45**, 1183-1193.

Tyson, G. W., Chapman, J., Hugenholtz, P. & other authors (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43.

Uchiya, K., Tohsuji, M., Nikai, T., Sugihara, H. & Sasakawa, C. (1996). Identification and characterization of *phoN*-Sf, a gene on the large plasmid of *Shigella flexneri* 2a encoding a nonspecific phosphatase. *J Bacteriol* **178**, 4548-4554.

Uerkvitz, W. & Beck, C. F. (1981). Periplasmic phosphatases in *Salmonella typhimurium* LT2. A biochemical, physiological, and partial genetic analysis of three nucleoside monophosphate dephosphorylating enzymes. *J Biol Chem* **256**, 382-389.

Uerkvitz, W. (1988). Periplasmic nonspecific acid phosphatase II from *Salmonella typhimurium* LT2. Crystallization, detergent reactivation, and phosphotransferase activity. *J Biol Chem* **263**, 15823-15830.

van Veen, H., Abee, T., Kortstee, G., Konings, W. & Zehnder, A. (1993). Mechanism and energetics of the secondary phosphate transport system of *Acinetobacter johnsonii* 210A. *J Biol Chem* **268**, 19377-19383.

Venter, J. C., Remington, K., Heidelberg, J. F. & other authors (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Voegele, R. T., Bardin, S. & Finan, T. M. (1997). Characterization of the *Rhizobium (Sinorhizobium) meliloti* high- and low-affinity phosphate uptake systems. *J Bacteriol* **179**, 7226-7232.

von Heijne, G. (1985). Signal sequences. The limits of variation. *J Mol Biol* **184**, 99-105.

Vyas, N. K., Vyas, M. N. & Quiocho, F. A. (2003). Crystal structure of *M. tuberculosis* ABC phosphate transport receptor: Specificity and charge compensation dominated by ion-dipole interactions. **11**, 765-774.

Wackett, L. P., Shames, S. L., Venditti, C. P. & Walsh, C. T. (1987). Bacterial carbon-phosphorus lyase: products, rates, and regulation of phosphonic and phosphinic acid metabolism. *J Bacteriol* **169**, 710-717.

Wagner, K. U., Masepohl, B. & Pistorius, E. K. (1995). The cyanobacterium *Synechococcus sp.* strain PCC 7942 contains a second alkaline phosphatase encoded by *phoV*. *Microbiology* **141 (Pt 12)**, 3049-3058.

Wang, C., Meek, D. J., Panchal, P., Boruvka, N., Archibald, F. S., Driscoll, B. T. & Charles, T. C. (2006). Isolation of Poly-3-Hydroxybutyrate metabolism genes from complex microbial communities by phenotypic complementation of bacterial mutants. *Appl Environ Microbiol* **72**, 384-391.

Wanner, B. L. (1996). Phosphorus assimilation and control of the phosphate regulon. In *Escherichia coli and Salmonella, Cellular and Molecular Biology*, pp. 1357-1381. Edited by F. Neidhardt. Washington, D.C.: ASM Press.

Watrud, L. S., Lee, E. H., Fairbrother, A., Burdick, C., Reichman, J. R., Bollman, M., Storm, M., King, G. & Van de Water, P. K. (2004). Evidence for landscape-level, pollen-mediated gene flow from genetically modified creeping bentgrass with CP4 EPSPS as a marker. *Proc Natl Acad Sci U S A* **101**, 14533-14538.

Weller, D. M., Raaijmakers, J. M., Gardener, B. B. M. & Thomashow, L. S. (2002). Microbial population responsible for soil suppressiveness to plant pathogens. *Annual Review of Phytopathology* **40**, 309-348.

Weppelman, R., Kier, L. D. & Ames, B. N. (1977). Properties of two phosphatases and a cyclic phosphodiesterase of *Salmonella typhimurium*. *J Bacteriol* **130**, 411-419.

Werner, A. & Kinne, R. K. H. (2001). Evolution of the Na-Pi cotransport systems. *Am J Physiol Regul Integr Comp Physiol* **280**, R301-312.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-699.

White, A. K. & Metcalf, W. W. (2004). Two C--P lyase operons in *Pseudomonas stutzeri* and their roles in the oxidation of phosphonates, phosphite, and hypophosphite. *J Bacteriol* **186**, 4730-4739.

White, A. K. & Metcalf, W. W. (2007). Microbial metabolism of reduced phosphorus compounds. *Annual Review of Microbiology* **61**, 379-400.

Williamson, L. L., Borlee, B. R., Schloss, P. D., Guan, C., Allen, H. K. & Handelsman, J. (2005). Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl Environ Microbiol* **71**, 6335-6344.

Willsky, G. R. & Malamy, M. H. (1980). Characterization of two genetically separable inorganic phosphate transport systems in *Escherichia coli*. *J Bacteriol* **144**, 356-365.

Wosten, M. M., Parker, C. T., van Mourik, A., Guilhabert, M. R., van Dijk, L. & van Putten, J. P. (2006). The *Campylobacter jejuni* PhosS/PhosR operon represents a non-classical phosphate-sensitive two-component system. *Mol Microbiol* **62**, 278-291.

Wu, J. R., Shien, J. H., Shieh, H. K., Hu, C. C., Gong, S. R., Chen, L. Y. & Chang, P. C. (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.

Wu, S. & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547-556.

Xu , J. (2006). Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Molecular Ecology* **15**, 1713-1731.

Yan, H., Blaszczyk, J., Xiao, B., Shi, G. & Ji, X. (2001). Structure and dynamics of 6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase. *Journal of Molecular Graphics and Modelling* **19**, 70-77.

Yao, N., Ledvina, P. S., Choudhary, A. & Quiocho, F. A. (1996). Modulation of a salt link does not affect binding of phosphate to its specific active transport receptor. *Biochemistry* **35**, 2079-2085.

Yarosh, O. K., Charles, T. C. & Finan, T. M. (1989). Analysis of C4-dicarboxylate transport genes in *Rhizobium meliloti*. *Mol Microbiol* **3**, 813-823.

Yeates, C. & Blackall, L. L. (2006). Construction and analysis of a metagenomic library from an enhanced biological phosphorus removal biomass. *Water Sci Technol* **54**, 277-284.

Yehudai-Resheff, S. & Schuster, G. (2000). Characterization of the *E. coli* poly(A) polymerase: nucleotide specificity, RNA-binding affinities and RNA structure dependence. *Nucl Acids Res* **28**, 1139-1144.

Yooseph, S., Sutton, G., Rusch, D. B. & other authors (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**, e16.

Yuan, Z.-C., Zaheer, R. & Finan, T. M. (2006a). Regulation and properties of PstSCAB, a high-affinity, high-velocity phosphate transport system of *Sinorhizobium meliloti*. *J Bacteriol* **188**, 1089-1102.

Yuan, Z. C., Zaheer, R., Morton, R. & Finan, T. M. (2006b). Genome prediction of PhoB regulated promoters in *Sinorhizobium meliloti* and twelve *proteobacteria*. *Nucleic Acids Res* **34**, 2686-2697.

Zaheer, R., Morton, R., Proudfoot, M., Yakunin, A. & Finan, T. M. (2009). Genetic and biochemical properties of an alkaline phosphatase PhoX family protein found in many bacteria. *Environmental Microbiology* **9999**.

Zalatan, J. G., Fenn, T. D., Brunger, A. T. & Herschlag, D. (2006). Structural and functional comparisons of nucleotide pyrophosphatase/phosphodiesterase and alkaline phosphatase: Implications for mechanism and evolution. *Biochemistry* **45**, 9788-9803.

Zhang, Y., Pohlmann, E. L., Halbleib, C. M., Ludden, P. W. & Roberts, G. P. (2001). Effect of PII and its homolog GlnK on reversible ADP-ribosylation of dinitrogenase reductase by heterologous expression of the *Rhodospirillum rubrum* dinitrogenase reductase ADP-ribosyl transferase-dinitrogenase reductase-activating glycohydrolase regulatory system in *Klebsiella pneumoniae*. *J Bacteriol* **183**, 1610-1620.

Zhou, B. & Zhang, Z. Y. (1999). Mechanism of mitogen-activated protein kinase phosphatase-3 activation by ERK2. *J Biol Chem* **274**, 35526-35534.