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\textsubscript{ACX6.13} and PhoN\textsubscript{BCX4.10}. In PhoA\textsubscript{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, \textit{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

I would like to extend special thanks to Dr. Trevor Charles and Dr. Bernard Glick for the dedication, guidance, patience and support throughout this long journey. To Dr. Bernard Duncker and Dr. Owen Ward, thank you for your participation in my advisory committee. Your input and insight has greatly contributed to the success of this work. I would also like to thank Dr. Kirsten Müller and Michael Lynch for their assistance and guidance in phylogenetic analysis.

To my colleagues, Asha Jacob, Louise Belanger and Keith Walsh, thanks very much for sharing ideas, thoughts and suggestions and for your friendship. To Ricardo, Youai, Merav, Zhenyu and Jin, thank you for all your support along the way.

Finally, to my family, my wife Shimrit, my children Idan and Shir, my parents Ora and Amos and my sisters Natalie and Noa - thank you all for the unconditional support, patience and love throughout this adventure.
# Table of contents

List of figures ................................................................. ix
List of tables ........................................................................ xii
List of abbreviations .......................................................... xiii
Claims of contributions to scientific knowledge ......................... xv

Chapter 1: Introduction .......................................................... 1

1.1 Soil and sludge microbial communities ........................................ 1
1.2 Metagenomics ........................................................................... 5
  1.2.1 Overview .............................................................................. 5
  1.2.2 Sequence-driven metagenomics ........................................... 9
  1.2.3 Function-driven metagenomics ............................................. 12
1.3 Phosphorous (P) metabolism .................................................... 17
  1.3.1 Importance of P and its environmental sources ....................... 17
  1.3.2 Types of P compounds ......................................................... 19
  1.3.3 Inorganic phosphate (P<sub>i</sub>) uptake .................................. 23
  1.3.3 Inorganic phosphate (P<sub>i</sub>) uptake .................................. 24
  1.3.4 Enzymatic degradation of organic phosphate ......................... 26
  1.3.5 Phophonate uptake and degradation ...................................... 33
  1.3.6 Phophonate metabolism in *Rhizobia* .................................. 37
1.4 This work ............................................................................... 41

Chapter 2: Materials and methods .............................................. 43

2.1 Bacterial culture and microbiological techniques ....................... 43
  2.1.1 Bacterial strains, plasmids and transposons .......................... 43
  2.1.2 Media, antibiotics and growth conditions ............................ 43
  2.1.3 Environmental samples: ..................................................... 48
2.2 Bacterial genetic techniques ................................................... 49
  2.2.1 Triparental mating / conjugation ........................................ 49
  2.2.2 Screening of metagenomic libraries for phosphatase activity and growth selection on phosphonate .................................................. 49
2.3 Molecular biology techniques ................................................ 50
  2.3.1 Plasmid isolation (alkaline lysis) ......................................... 50
2.3.2 Preparation and transformation of competent \textit{E. coli} cells

2.3.3 DNA library construction

2.3.4 Induction of phosphatase genes in $p\text{ET}30$ series of expression vectors

2.4 DNA manipulation methods

2.4.1 Restriction digestion

2.4.2 Ligation reaction

2.4.3 Dephosphorylation of vector DNA

2.4.4 Agarose gel electrophoresis

2.4.5 DNA amplification by PCR

2.4.6 Cloning of phosphatase genes

2.4.7 DNA sequence determination strategies

2.5 Bioinformatic techniques

2.6 Biochemical techniques

2.6.1 Preparation of periplasmic protein fraction

2.6.2 Phosphatase activity assay

2.6.3 Partial protein purification

2.6.4 Protein determination

2.6.5 Detection of phosphatases by western blotting

Chapter 3: Isolation and characterization of metagenomic phosphatase genes

3.1 Isolation and sequencing of acid phosphatases $\text{phoN}_{\text{ACX6.13}}$ and $\text{phoN}_{\text{BCX4.10}}$ and alkaline phosphatase $\text{phoA}_{\text{ACX6.71}}$

3.2 Characterization of phosphatases

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

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

3.3 Expression and biochemical characterization of phosphatase proteins

Chapter 4: Partial reconstruction of soil and activated sludge metagenomes by phenotypic complementation of phosphorous metabolism-deficient \textit{Sinorhizobium meliloti} mutants

4.1 Isolation, identification and sequencing of library cosmids

4.2 Characterization of phosphate/phosphonate transport genes

4.3 Additional functions in soil and activated sludge metagenomes

4.3.1 Signal transduction and regulatory mechanisms
4.3.2 Nitrogen fixation and assimilation ................................................................. 169
4.3.3 Defense mechanisms: ............................................................................... 174
4.3.4 Housekeeping genes ............................................................................... 177
4.3.5 Proteins of general function ................................................................... 177
4.3.6 Proteins of unknown function ................................................................... 187
Chapter 5: Conclusions .................................................................................. 191
Appendix: Structures of chemicals mentioned in this study ........................... 196
Bibliography .................................................................................................... 198
List of figures

**Figure 1-1:** Schematic diagram of the assimilation of P by a Gram-negative bacterial cell. 22

**Figure 3-1a:** Strategy for subcloning acid and alkaline phosphatases from pACX6.13 and pACX6.71, respectively. .......................................................... 67

**Figure 3-1b:** Strategy for subcloning acid phosphatase from pBCX4.10 ........................................ 69

**Figure 3-2:** Sequence of $\text{phoN}_{\text{ACX6.13}}$ and flanking region in pAR003. .................... 75

**Figure 3-3:** Sequence of $\text{phoN}_{\text{BCX4.10}}$ and its flanking region in pAR005. .................... 78

**Figure 3-4:** Phosphatase activity of periplasmic extracts of *E. coli* DH5α expressing $\text{phoN}_{\text{ACX6.13}}$ (pAR003) and $\text{phoN}_{\text{BCX4.10}}$ (pAR005) at various pHSs. ......................... 81

**Figure 3-5:** Neighbour-joining tree showing non-specific acid phosphatase proteins. ........ 83

**Figure 3-6a:** Multiple amino acid sequence alignment of class A NSAPs from bacterial isolates in comparison with PhoN$_{\text{ACX6.13}}$ and PhoN$_{\text{BCX4.10}}$ .................................................. 86

**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. .......................................................... 88

**Figure 3-6c:** Comparative alignment of consensus sequences of conserved regions in class A NSAPs from cultured and uncultured bacteria. .............................. 90

**Figure 3-7:** Predicted molecular structure of NSAP PhoN$_{\text{ACX6.13}}$. ................................ 98

**Figure 3-8:** Predicted molecular structure of NSAP PhoN$_{\text{BCX4.10}}$. .............................. 100

**Figure 3-9:** Sequence of $\text{phoA}_{\text{ACX6.71}}$ and flanking regions in pAR004. ................. 104
**Figure 3-10:** Phosphatase activity of periplasmic extracts of *E. coli* DH5α expressing *phoA<sub>ACX6.71</sub>* (pAR004) at various pHs. Assays were done in triplicate. ........................ 108

**Figure 3-11:** Neighbour-joining tree showing type I nucleotide phosphodiesterase/nucleotide pyrophosphatases proteins. .................................................................................. 110

**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. ........................................................................................................ 113

**Figure 3-12b:** Multiple amino acid sequence alignment of alkaline phosphatases and NPPs from metagenomic DNA derived from various environmental samples........ 115

**Figure 3-13:** Predicted molecular structure of phosphatase *PhoA<sub>ACX6.71</sub>*. ........................................ 122

**Figure 3-14a:** Expression of protein *PhoN<sub>ACX6.13</sub>* as detected on a Western blot. ................. 124

**Figure 3-14b:** Expression of protein *PhoA<sub>ACX6.71</sub>* as detected on a Western blot. ................. 126

**Figure 4-1:** Functional distribution of ORFs identified in cosmids pCX4-10F, pCX6-13F, pCX9-45F, pCX4-3G and pCX6-8G................................................................. 132

**Figure 4-2:** Gene map for *gly* locus of cosmids pAR010 (a) and pAR011 (b).......................... 136

**Figure 4-3:** N-terminal and region upstream to ORF *pstS<sub>pAR011</sub>*................................................. 140

**Figure 4-4:** Amino acid sequence alignment of PstS proteins....................................................... 145

**Figure 4-5:** Schematic illustration of bacterial PhoB/PhoR two component regulatory system. .............................................................................................................................. 150

**Figure 4-6:** Map of ORFs identified on cosmid pCX4-10F (pulp waste activated sludge). 158

**Figure 4-7:** Map of ORFs identified on cosmid pCX6-13F (municipal waste activated sludge) .............................................................. 162
Figure 4-8: Map of ORFs identified on cosmid pCX9-45F (soil) ........................................ 165
List of tables

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

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

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

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

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

Table 4-5: Identified genes involved in housekeeping functions

Table 4-6: Identified genes involved in general functions

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

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-AEP:</td>
<td>2-aminoethylphosphonate</td>
</tr>
<tr>
<td>4-HB:</td>
<td>4-hydroxybutyrate</td>
</tr>
<tr>
<td>ABC:</td>
<td>ATP binding cassette</td>
</tr>
<tr>
<td>AMP:</td>
<td>Adenosine monophosphate</td>
</tr>
<tr>
<td>ATP:</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>BAC:</td>
<td>Bacterial artificial chromosome</td>
</tr>
<tr>
<td>BCIP:</td>
<td>5-bromo-4-chloro-3-indolyl phosphate</td>
</tr>
<tr>
<td>CAMERA:</td>
<td>Community cyberinfrastructure for advanced marine microbial ecology research and analysis</td>
</tr>
<tr>
<td>cAMP:</td>
<td>cyclic adenosine monophosphate</td>
</tr>
<tr>
<td>cGMP:</td>
<td>cyclic guanosine monophosphate</td>
</tr>
<tr>
<td>CAPSO:</td>
<td>3-[cyclohexylamino]-2-hydrosyl-1-propane-sulfonic acid</td>
</tr>
<tr>
<td>CP-lyase:</td>
<td>Carbon-phosphorous lyase</td>
</tr>
<tr>
<td>DAPG:</td>
<td>2,4-diacetylphloroglucinol</td>
</tr>
<tr>
<td>DMSO:</td>
<td>Dimethylsulfoxide</td>
</tr>
<tr>
<td>DTT:</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>EBNSAP:</td>
<td><em>Escherichia blattae</em> nonspecific acid phosphatase</td>
</tr>
<tr>
<td>EDTA:</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EPSP:</td>
<td>3-enol-pyruvylshikimate-5-phosphate</td>
</tr>
<tr>
<td>G3P:</td>
<td>Glucose-3-phosphate</td>
</tr>
<tr>
<td>G6P:</td>
<td>Glucose-6-phosphate</td>
</tr>
<tr>
<td>GMP:</td>
<td>Guanosine monophosphate</td>
</tr>
<tr>
<td>GTP:</td>
<td>Guanosine triphosphate</td>
</tr>
</tbody>
</table>
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-indoly1-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 \( \text{phoN}_{\text{ACX6.13}} \) and \( \text{phoN}_{\text{BCX4.10}} \) and one alkaline phosphatase, \( \text{phoA}_{\text{ACX6.71}} \) 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 \( \text{phoN}_{\text{ACX6.13}}, \text{phoN}_{\text{BCX4.10}} \) and \( \text{phoA}_{\text{ACX6.71}} \). PhoB binding sites were predicted upstream to \( \text{phoN}_{\text{ACX6.13}} \) and \( \text{phoA}_{\text{ACX6.71}} \).

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

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 PstSpAR011 and PstSpCX4-10F 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*<sub>pCX4-10F</sub> and *gst2*<sub>pCX4-10F</sub>, a glutathione-dependent formaldehyde dehydrogenase, *adhC*<sub>pCX4-10F</sub> and an S-formylglutathione hydrolase, *fghA*<sub>pCX4-10F</sub>.

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...
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, Phytophthora 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
nitrate (Delgado & Follett, 2002). One of the ways soil microorganisms affect the carbon cycle is by acting as a biological sink to atmospheric CH₄ (Conrad, 1996). There are two known types of Methanotrophic bacteria (MB), types I and II, which are distinguished by their phylogenetic affiliation (\(\gamma\)-Proteobacteria vs. \(\alpha\)-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 \(\alpha\)- and \(\beta\)-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
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 Amplified 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, β 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.,
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 \textit{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 \textit{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 \(\gamma\)-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 Artificial 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 Rhodocyclales 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 Saragasso 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). Furthermore, 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
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-diacytethylphloroglucinol (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 Streptomyces 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-phosphorous (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).
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.
1.3.3 Inorganic phosphate (Pi) 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 Pi: 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 Pi-Na\(^+\) or Pi-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 \textit{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 \textit{et al.}, 2001).

The Pst transport system is well characterized in \textit{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 &
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 Mn²⁺ and Zn²⁺ 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), *Pasteurella 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).
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.
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 irrepresible 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-independent 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
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 

\textit{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 \textit{B. cereus}, the first reported C-P hydrolyzing enzyme isolated, purified and characterized (La Nauze \textit{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 \textit{et al.}, 1970). Another example of a narrow substrate range enzyme was phosphonoacetate hydrolase, isolated from \textit{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 \textit{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
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).
C-P lyase has also been studied in *Pseudomonas stutzeri*. 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 cryos crescens* (Wackett et al., 1987). Phosphonate degradation has also been characterized in
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 convolution, 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 permanently 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°C. \( E. \) coli was grown on LB medium (Sambrook & Russel, 2001) at 37°C. The MS medium was composed of 15 g/l \( \text{NH}_4\text{Cl} \), 0.24 g/l \( \text{CaCl}_2\text{H}_2\text{O} \), 0.48 g/l \( \text{MgSO}_4 \), 35 g/l morpholino
Table 2-1: Bacterial strains, plasmids, metagenomic libraries and transposons used in this study

<table>
<thead>
<tr>
<th>Strain, plasmid, metagenomic library or transposon</th>
<th>Relevant characteristics</th>
<th>Reference or source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rm1021</td>
<td>SU47 str-21 (Sm&lt;sup&gt;+&lt;/sup&gt;)</td>
<td>(Meade et al., 1982)</td>
</tr>
<tr>
<td>RmP110</td>
<td>Rm1021 with changed wild type pstC</td>
<td>(Yuan et al., 2006a)</td>
</tr>
<tr>
<td>RmP636</td>
<td>RmP110 pit310::Tn5 phoC490ΩSp&lt;sup&gt;+&lt;/sup&gt;</td>
<td>(Yuan et al., 2006a)</td>
</tr>
<tr>
<td>RmF726</td>
<td>ΔΩ5149-5079::Tn5-233 Gm&lt;sup&gt;+&lt;/sup&gt;-Sp&lt;sup&gt;+&lt;/sup&gt;, Nm&lt;sup&gt;+&lt;/sup&gt;</td>
<td>(Charles &amp; Finan, 1991)</td>
</tr>
<tr>
<td>RmG471</td>
<td>ΔΩ5033-5025::Tn5 Nm&lt;sup&gt;+&lt;/sup&gt; Gm&lt;sup&gt;-&lt;/sup&gt;-Sp&lt;sup&gt;+&lt;/sup&gt;</td>
<td>(Charles &amp; Finan, 1991)</td>
</tr>
<tr>
<td>RmG439</td>
<td>ΔG439 [ΔphoCDET HindIII::Nm&lt;sup&gt;+&lt;/sup&gt; (12 kb)]</td>
<td>(Charles et al., 1991)</td>
</tr>
<tr>
<td>RmG490</td>
<td>ndvF-1.7ΩSp&lt;sup&gt;+&lt;/sup&gt; = phoCΩ490</td>
<td>(Charles et al., 1991)</td>
</tr>
<tr>
<td>RmG491</td>
<td>ndvF-5.3ΩSp&lt;sup&gt;+&lt;/sup&gt; = phoTΩ490</td>
<td>(Charles et al., 1991)</td>
</tr>
<tr>
<td>RmG830</td>
<td>phoCΩ490 sfxⅠ pit310::Tn5</td>
<td>(Bardin et al., 1998)</td>
</tr>
<tr>
<td>RmH838</td>
<td>phoB::TnV</td>
<td>(Bardin &amp; Finan, 1998)</td>
</tr>
<tr>
<td>RmG762</td>
<td>phoC490ΩSp&lt;sup&gt;+&lt;/sup&gt; sfxⅠ</td>
<td>(Bardin et al., 1998)</td>
</tr>
</tbody>
</table>

**Sinorhizobium meliloti**

- Rm1021
- RmP110
- RmP636
- RmF726
- RmG471
- RmG439
- RmG490
- RmG491
- RmG830
- RmH838
- RmG762

**Escherichia coli**

- DH5<sup>●</sup>
  - F<sup>+</sup> endA1 hsdR17 (rK<sup>+</sup>m<sub>K</sub><sup>+</sup>) supE44 thi-1
  - recA1 gyrA96 relA1Δ(argFo-lacZYA)U169φ80dlacZΔM15,λ<sup>+</sup>
- DH5<sup>●</sup> (pRK600)
  - Mobilizing strain; Cm<sup>+</sup>
- DH5<sup>●</sup> (pRK600)
  - Mobilizing strain; Cm<sup>+</sup>
<table>
<thead>
<tr>
<th>Strain, plasmid, metagenomic library or transposon</th>
<th>Relevant characteristics</th>
<th>Reference or source</th>
</tr>
</thead>
<tbody>
<tr>
<td>HB101</td>
<td>supE44hsdS20(rB&lt;sub&gt;B&lt;/sub&gt;mB&lt;sub&gt;B&lt;/sub&gt;) recA13 ara-14 proA2 lacY1 galK2 rpsL2 xyl-5 mtl-1 leuB6 thi-1</td>
<td>(Boyer &amp; Roulland-Dussoix, 1969)</td>
</tr>
<tr>
<td>BL21 (DE3)/pLysS</td>
<td>F-ompT hsdSB (rB- mb-) gal dcm (DE3)/pLysS (Cm&lt;sup&gt;8&lt;/sup&gt;)</td>
<td>Novagen</td>
</tr>
<tr>
<td>MM294A</td>
<td>pro-82 thi-pro-82 thi-1 hsdR17 supE44 endA1</td>
<td>(Yarosh et al., 1989)</td>
</tr>
<tr>
<td>MT621</td>
<td>MM294A malF::TnphoA</td>
<td>(Charles &amp; Finan, 1991)</td>
</tr>
</tbody>
</table>

**Metagenomic library**

<table>
<thead>
<tr>
<th>Metagenomic library</th>
<th>Relevant characteristics</th>
<th>Reference or source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CX4</td>
<td>pRK7813 clone bank in HB101 constructed with DNA inserts derived from pulp waste activated sludge</td>
<td>(Wang et al., 2006)</td>
</tr>
<tr>
<td>CX6</td>
<td>pRK7813 clone bank in HB101 constructed with DNA inserts derived from municipal waste activated sludge</td>
<td>(Wang et al., 2006)</td>
</tr>
<tr>
<td>CX9</td>
<td>pRK7813 clone bank in HB101 constructed with DNA inserts derived from soil</td>
<td>(Wang et al., 2006)</td>
</tr>
</tbody>
</table>

**Plasmids**

<table>
<thead>
<tr>
<th>Plasmids</th>
<th>Relevant characteristics</th>
<th>Reference or source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pRK7813</td>
<td>IncP cosmid cloning vector; Tc&lt;sup&gt;R&lt;/sup&gt;</td>
<td>(Jones &amp; Gutterson, 1987)</td>
</tr>
<tr>
<td>Strain, plasmid, metagenomic library or transposon</td>
<td>Relevant characteristics</td>
<td>Reference or source</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>--------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>pUC18</td>
<td>Cloning vector, CoIE1, oriV, bla; Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Fermentas</td>
</tr>
<tr>
<td>pGEM T-Easy</td>
<td>Cloning vector, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Promega</td>
</tr>
<tr>
<td>pACX4(6)-xx and pBCX4(6)-xx series</td>
<td>pRK7813 cosmid clones from libraries CX4 or CX6, that utilize BCIP</td>
<td>This study</td>
</tr>
<tr>
<td>pAR003</td>
<td>6.5 kb BamHI fragment of pACX6.13 in pUC19</td>
<td>This study</td>
</tr>
<tr>
<td>pAR004</td>
<td>3.5 kb BamHI fragment of pACX6.71 in pUC19</td>
<td>This study</td>
</tr>
<tr>
<td>pAR005</td>
<td>6.3 kb BamHI-EcoRI fragment of pBCX4.10 in pUC19</td>
<td>This study</td>
</tr>
<tr>
<td>pET30 series</td>
<td>Bacterial expression vector, Km&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Novagen</td>
</tr>
<tr>
<td>pAR0032</td>
<td>835 bp PCR product &lt;i&gt;phoN&lt;sub&gt;ACX6.13&lt;/sub&gt;&lt;/i&gt; from pAR003 in pGEM T-Easy, Km&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This study</td>
</tr>
<tr>
<td>pAR0045</td>
<td>1648 bp PCR product (flanked by NotI) &lt;i&gt;phoA&lt;sub&gt;ACX6.71&lt;/sub&gt;&lt;/i&gt; from pAR004 in pET30a, Km&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This study</td>
</tr>
<tr>
<td>pBBR1-MCS5</td>
<td>Cloning vector; Gm&lt;sup&gt;R&lt;/sup&gt; derivative of pBBR1</td>
<td>(Antoine &amp; Locht, 1992)</td>
</tr>
<tr>
<td>pCX4-nF(G) series</td>
<td>CX4 cosmid clones that complement RmF726 and/or RmG471 for the utilization of glyphosate as the sole P source</td>
<td>This study</td>
</tr>
<tr>
<td>Strain, plasmid, metagenomic library or transposon</td>
<td>Relevant characteristics</td>
<td>Reference or source</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>pCX6-nF(G) series</td>
<td>CX6 cosmid clones that complement RmF726 and/or RmG471 for the utilization of glyphosate as the sole P source</td>
<td>This study</td>
</tr>
<tr>
<td>pCX9-nF(G) series</td>
<td>CX9 cosmid clones that complement RmF726 and/or RmG471 for the utilization of glyphosate as the sole P source</td>
<td>This study</td>
</tr>
<tr>
<td>pAR010</td>
<td>3.5 kb HindIII fragment of pCX4-3G in pBBR1-MCS5</td>
<td>This study</td>
</tr>
<tr>
<td>pAR011</td>
<td>8.4 kb EcoRI fragment of pCX6-8G in pBBR1-MCS5</td>
<td>This study</td>
</tr>
<tr>
<td>pAR012</td>
<td>16 kb BamHI fragment of pCX9-12G in pBBR1-MCS5</td>
<td>This study</td>
</tr>
<tr>
<td>pAR013</td>
<td>6 kb BamHI fragment of pCX6-9G in pBBR1-MCS5</td>
<td>This study</td>
</tr>
</tbody>
</table>

Transposon

| EZ::Tn<Kan2> | Km<sup>R</sup> | EPICENTRE Technologies |

47
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 FeSO₄7H₂O, 10 μg/l H₃BO₃, 11.2 g/l MnSO₄H₂O, 124.6 μg/l ZnSO₄7H₂O, 78.2 μg/l CuSO₄5H₂O and 10μg MoO₃. 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 \textit{E. coli} DH5α by conjugation selecting transconjugants on LB tetracycline while incubated at 37°C to counterselect the \textit{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 \textit{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 \( \mu \)l TEG solution (50 mM glucose, 25 mM Tris-HCl pH 8, 10 mM EDTA pH 8) containing 200 \( \mu \)g/ml RNase (20 \( \mu \)l of 10 mg/ml RNase stock per 1ml of TEG). 200 \( \mu \)l
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 μl 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 supernatant 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 prepared using the CaCl₂ 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 CaCl₂ 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 CaCl₂. 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 CaCl₂ 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 H2O. Cells were centrifuged at 4000 rpm for 25 min. The pellet was resuspended in 25 ml ice-cold deionized H2O 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 electroporation 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 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. 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 13000 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 H2O. 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 pET30 series of expression vectors

Full length ORFs of phosphatases \( \text{phoN}_{\text{ACX6.13}} \) and \( \text{phoN}_{\text{BCX4.10}} \) were expressed in \( E. \text{coli} \) BL21 DE3/pLysS by autoinduction (non-IPTG expression). Autoinduction medium (Studier, 2005) contained per litre: 6 g \( \text{Na}_2\text{HPO}_4 \), 3 g \( \text{KH}_2\text{PO}_4 \), 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 \( \mu \)l of cell induction culture was added to 10 \( \mu \)l 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 \( \mu \)l 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 digestions 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 MgSO₄, 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 \textit{phoN}_{ACX6.13}, primers pAR003F: 5’ ATGTTTTCGCCACGCAACT 3’ and pAR003R2: 5’ CTGGCTTTCCGAGCGCCTCG 3’, were used to amplify the 835-bp ORF \textit{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. \textit{phoN}_{ACX6.13} was cut with EcoRI and ligated into EcoRI cut expression vector pET30b (Novgen) to generate pAR0033 containing \textit{phoN}_{ACX6.13} in frame with the C-terminal His tag of pET30b. Diagnostic digestion of pAR0033 prepared from \textit{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’ GGGGCAGCCGACATGCAGTAATATATTTCTTTT 3’ and pAR004R4: 5’ GGGGCAGCCGGCCGCACACCGGTTCCTCC 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 \( E. coli \) DH5\( \alpha \) 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’ ACCTACAAACAAAGCTCTCATCAACC 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: BioEdit (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](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](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 \( \Delta \)OD420 nm / (\( \Delta \)t (min)*OD600 nm), where \( \Delta \)t 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 \( \Delta \)OD420nm / V (ml) X (\( \Delta \)t (min)), where V denotes the volume of enzyme preparation and \( \Delta \)t 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\textsubscript{ACX6.71} and PhoN\textsubscript{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 H₂O 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

<table>
<thead>
<tr>
<th>Clone designation</th>
<th>Restriction pattern group</th>
<th>Specific phosphatase activity at various pHs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5.5</td>
</tr>
<tr>
<td>pACX4.21</td>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>pACX4.26</td>
<td>A</td>
<td>10.9</td>
</tr>
<tr>
<td>pBCX4.10</td>
<td>A</td>
<td>5.89</td>
</tr>
<tr>
<td>pACX4.11</td>
<td>A</td>
<td>2.08</td>
</tr>
<tr>
<td>pACX6.3</td>
<td>B</td>
<td>13.2</td>
</tr>
<tr>
<td>pACX6.13</td>
<td>B</td>
<td>14.9</td>
</tr>
<tr>
<td>pACX6.21</td>
<td>B</td>
<td>13.9</td>
</tr>
<tr>
<td>pACX6.87</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>pBCX6.22</td>
<td>B</td>
<td>6.35</td>
</tr>
<tr>
<td>pBCX6.26</td>
<td>B</td>
<td>5.79</td>
</tr>
<tr>
<td>pBCX6.31</td>
<td>B</td>
<td>5.73</td>
</tr>
<tr>
<td>pBCX6.35</td>
<td>B</td>
<td>7.14</td>
</tr>
<tr>
<td>pBCX6.37</td>
<td>B</td>
<td>2.25</td>
</tr>
<tr>
<td>pBCX6.38</td>
<td>B</td>
<td>5.46</td>
</tr>
<tr>
<td>pACX6.7</td>
<td>C</td>
<td>18.7</td>
</tr>
<tr>
<td>pACX6.29</td>
<td>C</td>
<td>21.9</td>
</tr>
<tr>
<td>pACX6.57</td>
<td>C</td>
<td>16.4</td>
</tr>
<tr>
<td>pACX6.62</td>
<td>C</td>
<td>16.4</td>
</tr>
<tr>
<td>pACX6.71</td>
<td>C</td>
<td>20.7</td>
</tr>
<tr>
<td>pACX6.80</td>
<td>C</td>
<td>19</td>
</tr>
<tr>
<td>pACX6.96</td>
<td>C</td>
<td>16.2</td>
</tr>
<tr>
<td>pACX6.101</td>
<td>C</td>
<td>19.8</td>
</tr>
<tr>
<td>pACX6.102</td>
<td>C</td>
<td>18.4</td>
</tr>
<tr>
<td>pACX6.107</td>
<td>C</td>
<td>25.2</td>
</tr>
<tr>
<td>pACX6.109</td>
<td>C</td>
<td>22.3</td>
</tr>
<tr>
<td>pACX6.111</td>
<td>C</td>
<td>18.7</td>
</tr>
<tr>
<td>pBCX6.6</td>
<td>C</td>
<td>3.19</td>
</tr>
<tr>
<td>pBCX6.9</td>
<td>C</td>
<td>2.98</td>
</tr>
</tbody>
</table>
**Figure 3-1a:** Strategy for subcloning acid and alkaline phosphatases from pACX6.13 and pACX6.71, respectively.
Digest with BamHI and ligate into pUC18.
Figure 3-1b: Strategy for subcloning acid phosphatase from pBCX4.10.
Digest BamHI and ligate into pUC18

Digest with EcoRI and self ligate
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 \textit{phoN}_{\text{ACX6.13}} and \textit{phoN}_{\text{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 \textit{et al.}, 2005). These ORFs were thus designated \textit{phoN}_{\text{ACX6.13}} and \textit{phoN}_{\text{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, KXXXXXXXRP (Domain I),
PSGH (Domain II) and SRXXXXXXHXXXD (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 \(\text{phoN}_{\text{ACX6.13}}\) (Figure 3-2) and the N-terminal portion of a putative \(\text{uvrD}\) ORF was located 144 bp upstream of \(\text{phoN}_{\text{BCX4.10}}\) (Figure 3-3). To determine the occurrence of promoters, the sequences upstream of genes \(\text{phoN}_{\text{ACX6.13}}\) and \(\text{phoN}_{\text{BCX4.10}}\) were analysed for regulatory protein binding sites using BPROM (http://linux1.softberry.com) and Visual Footprint (Munch et al., 2005). Putative \(\text{E. coli}\) class \(\sigma_70\) -10 and -35 recognition sites were predicted upstream of a transcription start site located 71 bp upstream of the translation start site of \(\text{phoN}_{\text{BCX4.10}}\) (Figure 3-3). In the case of \(\text{phoN}_{\text{ACX6.13}}\), a putative \(\text{E. coli}\) class \(\sigma_70\) -10 recognition sequence was identified 31 bp upstream of the translation start site, but no -35 recognition sequence was identified. However, recognition sequences for \(\text{E. coli}\) CytR and GlnG regulators were identified 104 bp and 51 bp, respectively upstream of the \(\text{phoN}_{\text{ACX6.13}}\) 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 \(\text{udpP}\) (uridine phosphorylase) and \(\text{tsx}\) (nucleoside uptake protein), (Jorgensen et al., 1998). A putative PhoB binding site (“pho” box) was identified upstream of \(\text{phoN}_{\text{ACX6.13}}\).

An NCBI BlastX search revealed acid phosphatases \(\text{phoN}_{\text{ACX6.13}}\) and \(\text{phoN}_{\text{BCX4.10}}\), respectively share 59% and 62% amino acid identity with the acid phosphatase of \(\beta\)-\text{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 \( \text{phoN}_{\text{ACX6.13}} \) 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 \( \text{phoNBCX4.10} \) 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 \( \text{PhoN}_{\text{ACX6.13}} \) and \( \text{PhoN}_{\text{BCX4.10}} \) 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 \( \text{PhoN}_{\text{ACX6.13}} \) and \( \text{PhoN}_{\text{BCX4.10}} \) indicated the probability of these sequences containing a signal peptide was 0.996 and 1.0, respectively, with the signal peptide cleavage sites in \( \text{PhoN}_{\text{ACX6.13}} \) and \( \text{PhoN}_{\text{BCX4.10}} \) being most likely Ala21-Cys22 and Ala29-Pro30, respectively.

In order to determine the pH range for activity of \( \text{PhoN}_{\text{ACX6.13}} \) and \( \text{PhoN}_{\text{BCX4.10}} \), periplasmic extracts of \( E. \text{coli} \) DH5\( \alpha \) 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 \( \text{PhoN}_{\text{ACX6.13}} \) peaked at pH 4.4 while \( \text{PhoN}_{\text{BCX4.10}} \) was optimally functional at a broader range of pH 5.4-7.9. The above results suggest that both \( \text{phoN}_{\text{ACX6.13}} \) and \( \text{phoN}_{\text{BCX4.10}} \) encode secreted acid-neutral
phosphatases. To assess the phylogenetic relationship between \( \text{PhoN}_{\text{ACX6.13}} \) and \( \text{PhoN}_{\text{BCX4.10}} \) 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 \( \text{Zymomonas mobilis} \). Cluster I contains primarily NSAPs of enteric bacteria belonging to the \( \gamma \)-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 \( \gamma \)-Proteobacteria while the remaining orthologues are represented by \( \alpha \)-, \( \beta \)-, \( \delta \)-Proteobacteria and Cyanobacterial classes. The NSAPs derived from uncultured bacteria from activated sludge, \( \text{PhoN}_{\text{ACX6.13}} \) and \( \text{PhoN}_{\text{BCX4.10}} \) along with the putative acid phosphatase from \( \text{P. fluorescens} \) H16, form Cluster III. It should be noted that only with the exception of the acid phosphatase from \( \text{Zymomonas mobilis} \), all of the experimentally determined NSAPs were isolated from enteric \( \gamma \)-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. ꚦ 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 $\sigma_70$ -10 (red), CytR (yellow) and GlnG (blue).
GCAAGAAGCCGGCTCGATTTGGCAAAACGAGCCGGCGCAGCTCGGAAGCCAGTAGGCAAATGGTGCCCGCGGTGGCCACCGTGC
KKPALKDCKTEAAAALGSQ*

AACGGCAATTACCGAATTCGGGCAAGCATGTTTCTGGAACGCAGCATCCCGGGCAGGCGGCACSCCTTCAACGGACTGGCTTCAC

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

Shaded sequence corresponds to the predicted RBS. \( \uparrow \) denotes the predicted site of signal peptide cleavage. Amino acid sequences of \( \text{phoN}_{\text{BCX4.10}} \) and \( \text{uvrD} \) 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 \( \sigma_70 \) -10 and -35 recognition sequences are boxed in red and the \( \text{phoN}_{\text{BCX4.10}} \) transcription start site is denoted by \( \rightarrow \).
Figure 3-4: Phosphatase activity of periplasmic extracts of *E. coli* DH5α expressing $\text{pho}N_{\text{ACX6.13}}$ (pAR003) and $\text{pho}N_{\text{BCX4.10}}$ (pAR005) at various pHs.

Assays were done in triplicate. The apparent broad pH optimum for $\text{pho}N_{\text{BCX4.10}}$ 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.

Cluster I

Cluster II

Cluster III

Cluster IV
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\textsubscript{ACX6.13} and PhoN\textsubscript{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 $\bullet$, $\blacksquare$ and $\blacklozenge$, respectively. Consensus sequence residues are denoted as follows: $\chi$, polar amino acids DEKRNQ; $\Psi$, non polar amino acids ILVMFY; $\nu$, acidic amino acids DE; $\sigma$, small amino acids GAST; $.$, no particular amino acid.
**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.
**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.RF....C...χχ..L.KχGSYPGψGHαΑψΓψδ.αΛψL.Eψ.Pχχ.χ.ψΨ.RG..ψGχSRΨΨC.ΨHψQS cultured
Kχ.YQR.RPFψ.NG..ψCTΡννχ..LχΟDGYPGψGHαΑ.GW.WALVLψΛψΑPνR.νΑαΛΑRG.χΨΟχχSRΨΨCNΨΗΨΗ.ΗΨD uncultured
To study the molecular structure of these proteins, models of PhoN\textsubscript{ACX6.13} and PhoN\textsubscript{BCX4.10} were predicted (Figures 3-7 and 3-8). Both proteins have similar secondary structure composed of 12 $\alpha$-helices and 2 $\beta$-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 \textit{et al.}, 2000). The conserved histidine of domain III, His 218 (His 221 in PhoN\textsubscript{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\textsubscript{BCX4.10}). The conserved histidine of domain II, His 179 (His 182 in PhoN\textsubscript{BCX4.10}), may function as the proton donor for the leaving group. Lys 144 (Lys 147 in PhoN\textsubscript{BCX4.10}) and Arg 151 (Arg 154 in PhoN\textsubscript{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\textsubscript{BCX4.10}) and the amide nitrogen atoms of Gly 178 (Gly 181 in PhoN\textsubscript{BCX4.10}) and His 179 (His 182 PhoN\textsubscript{BCX4.10}) which compose domain II. Two additional domain III residues are also conserved. Ser 211 (Ser 214 in PhoN\textsubscript{BCX4.10}) may have an important overall structural role while Arg 212 (Arg 215 in PhoN\textsubscript{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\textsubscript{ACX6.13} and PhoN\textsubscript{BCX4.10} are considered acid/neutral phosphatases. The overall functional range of pH exhibited by PhoN\textsubscript{ACX6.13} and PhoN\textsubscript{BCX4.10} is consistent with the optimal pHs of
known secreted NSAPs from various bacterial isolates. The two NSAPs differ in that PhoN\textsubscript{ACX6.13} exhibits a sharp acidic pH optimum while the acidic-neutral PhoN\textsubscript{BCX4.10} has a broad range of functional pH. The broad pH range for activity demonstrated by PhoN\textsubscript{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 \textit{phoN\textsubscript{BCX4.10}} may be due to substrate depletion resulting in activity being outside the linear range. The pH optimum exhibited by PhoN\textsubscript{ACX6.13} is somewhat more acidic than that of the experimentally determined NSAPs and is comparable to the pH optimum of the NSAP from \textit{Prevotella intermedia} (Chen \textit{et al.}, 1999).

Signal peptide sequences predicted for PhoN\textsubscript{ACX6.13} and PhoN\textsubscript{BCX4.10} both include alanine as the last residue which is consistent with the signal peptides of the acid phosphatases of \textit{E. blattae} (Ishikawa \textit{et al.}, 2000), \textit{Prevotella intermedia} (Chen \textit{et al.}, 1999) and \textit{Salmonella typhimurium} (Makde \textit{et al.}, 2003). The number of residues in the predicted signal peptide of PhoN\textsubscript{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\textsubscript{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, KXXXXXXXRP (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 \( \text{phoN}_{ACX6.13} \) and \( \text{phoN}_{BCX4.10} \) 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 \( \text{phoN}_{ACX6.13} \) and \( \text{phoN}_{BCX4.10} \) 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 \( \text{phoN}_{ACX6.13} \) suggests the possibility of PhoB dependent regulation. Since the association of PhoB with \( \sigma_{70} \) factor of RNA polymerase has been previously demonstrated in \( E. \ coli \) (Kumar et al., 1994), it is not surprising to find \( \sigma_{70} \) binding sites in the vicinity of the pho box in \( \text{phoN}_{ACX6.13} \). The absence of a “pho box” upstream to \( \text{phoN}_{BCX4.10} \) 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 \( Sinorhizobium \ meliloti \) polyphosphatase 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<sub>ACX6.13</sub> and PhoN<sub>BCX4.10</sub> 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<sub>ACX6.13</sub> and PhoN<sub>BCX4.10</sub>, 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<sub>ACX6.13</sub> and PhoN<sub>BCX4.10</sub> 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 PhoAACX6.71 is a novel member of the nucleotide pyrophosphatase (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
The gene was designated \( \text{pho}A_{\text{ACX6.71}} \). As was shown with the acid phosphatases, \( \text{pho}A_{\text{ACX6.71}} \) also contains a RBS upstream of the translation start site. In addition, like many other members of the NPP family, \( \text{PhoA}_{\text{ACX6.71}} \) 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 \( \text{pho}A_{\text{ACX6.71}} \), 476 bp upstream of the translation start site were analysed for regulatory protein recognition sequences. Seven possible promoter sequences were determined including \( E. \text{coli} \) class \( \sigma_{70} \) -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 Mim, 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 \( E. \text{coli} \) 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 \( E. \text{coli} \) MalT regulator, which affects the transcription of maltose regulon genes (Schlegel et al., 2002), and \( E. \text{coli} \) 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. blatta*e (Ishikawa *et al.*, 2000).
Since PhoA<sub>ACX6.71</sub> 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<sub>ACX6.71</sub> 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<sub>ACX6.71</sub> is indeed a signal peptide. Consistent with the possibility that PhoA<sub>ACX6.71</sub> 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<sub>ACX6.71</sub> 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<sub>ACX6.71</sub> shares an amino acid identity of 57% with a putative protein of the Type I phosphodiesterase/nucleotide pyrophosphatase superfamily, identified from the metagenome of a bacterial symbiont of a gutless worm originating from the Mediterranean Sea (JCVI_PEP_1113958955152). Aside from the very low amino acid identity to known bacterial alkaline phosphatases, the novelty of PhoA<sub>ACX6.71</sub> 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 *Bacteroidetes* 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 irrepresible 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 $\alpha$-*Proteobacterial* class. Isolates belonging to the $\gamma$-*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 $\gamma$-*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: $\sigma$70 -35 and -10 (red), Mlc (turquoise), Fis (blue), FadR (yellow), MalT (pink) and GlnG (green). “Pho box” is underlined. Putative transcription start site is denoted by $\rightarrow$. Bases shaded in grey correspond to a predicted RBS. $\Uparrow$ denotes the putative signal peptide cleavage site. Residues shaded in gold are predicted to be involved in $\text{Zn}^{2+}$ binding. The reaction centre threonine (Thr 73) is highlighted in red.
AAGCGGGGAAGGACTACTATGGCAATCTGCGCTTCACACCGATGGGCGACGTGCTCACGCTCCAGTTTGCCCGGGAACTGATGCGCGG
AGGAAAGTTGGGCAACGCGGACCCAGAGCATCCTGAGATATCGGCAAGCAGACTACATGGCAGCCCGCTCCGAGGCTCTTGGGAGG
AGGATGGAGGACGCCACGGACATCCTGACTATCAGCCTTTCGGCAACGGACTACATCGGCCACGCCTTCGGGCCGA
EVEVGQRDADTDILTISLSATYIGAFGPN
ACAGCTTGGAGTCCGAGAACACAAACAGCAGTCCGAGCTACGGAGCATCCTGCCGAACTGAGTCTGGGAGGGACTGGAGA
SLSEDNQLQLDRSLAEFPFAEVDRLVGLD
ATCGCAAGCTTGCTGATCTGCTGACGGACAGCTGGAGATCTGCAAGGATGGTACGC
RTLIVLDDLSSHDIEIPEYTQHLCGCDAGRH
ACGTGGCCAGGAATTCTATTGTCGGCGCGAAATGGGCGCTGGTATCGGACAGCTGGACCGGCTCGTCTGTA
VEEFIAAAANGLKARFGBIDDDLVLTFQN
ATCCGAGATCTGATCTGAGCGAGCGAGCCTGCAGTCCGAGGATGGAGA
PSLYLDEARVQALGLALPEVERALADAMV
TGCCCTGCGCCGCCCTGATAATGGCGTGACGGCGACGGACCTGCTGAGAGGGTTCG
ALPGFDMAVSTRSLLEGRVPNTKVMDMVT
CGCGCAGCTTCACCACCGAGCGATCGGGGCAATGTTGCTGCTTGGGCAATCGGAGCGGCA
RAFHPKRSGSNLVQLVQVQSPSWLYPEAQA
CGCGCATGCAGGATGGCGTGCTTGGCAGCTGTTGCGGTGGTGGTAACATGGGATACCATGGGCAATCCCTTG
AMGSPSYDTRYVPIFFAGPGGVSGASQR
GTCTCTGTGGCGCAGAGGATGTCGCTCTACAACGCGACCCTATGTTGCGCGCAGCACCTTGGTGGTGCTGACG
PVAPEDVESTITAYLGIKPSPSGMNMPPLL
TGAGGTTGGTGGGTGGGTGGGAGAACCCTGTGGTGGTGAATTTGTCGACCGTGGTGGTGGGATAGAAP
EVVAGETGV*
GCCCGATTCTGGGCAAGACCTGCACTGATGACCCCTACCGGACAGCGGAGACGCGGAGCCGACGTC
CAGGCGAACGGAATCCATCCATACCCAGGAGAACACGCGGCGCCCGTGGCGCTGGAATCTCAGCAAGAACCCGCAGACAGCACCGA
GATCGTCAACGAGAAAATCAGTCAGGATTGAAGACATCAAAGGAACCTCCGAAAAAGTCTCAGCTTGCTCCGGATCTAAGATCAGA

GACTTGATTGAAAGGCAAGACCAGCTTAATTGGGACCCATTGCTTCTGCGATAATTCATATCGTCAATGFGTTTTGAAAGACGAGATA

AGCAGAGTACG
**Figure 3-10:** Phosphatase activity of periplasmic extracts of *E. coli* DH5α expressing *phoA*<sub>ACX6.71</sub> (pAR004) at various pHs. Assays were done in triplicate.
**Figure 3-11:** Neighbour-joining tree showing type I nucleotide phosphodiesterase/nucleotide pyrophosphatases proteins.

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. us1*, *Solibacter usitatus* Ellin 6076a; *S. us2*, *Solibacter usitatus* Ellin 6076b; *L. pn* Lens, *Legionella pneumophila* str. Lens; *L. pn* Phil; *Legionella pneumophila* 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: \(\chi\), polar amino acids DEKRNQ; \(\Psi\), non polar amino acids ILVMFY; \(\nu\), acidic amino acids DE; \(\sigma\), small amino acids GAST; ., no particular amino acid.
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.
To elucidate the overall structure and the arrangement of key conserved residues in the protein, a model of molecular structure was predicted for PhoA\textsubscript{ACX6.71}. The model suggests that the protein is globular in structure, composed of 13 $\alpha$-helices and 19 $\beta$-sheets with a cleft formed in its inner core. Conserved residues previously demonstrated as involved in metal binding and catalysis in \textit{Xanthomonas axonopodis} NPP (Zalatan \textit{et al.}, 2006a) are shown to be located in the inner core of PhoA\textsubscript{ACX6.71}. The conserved residues include six amino acids which may coordinate two Zn\textsuperscript{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 $\delta$-oxygens of the phosphate group.

From the metagenome of municipal waste activated sludge, a new alkaline phosphatase, \textit{phoA}\textsubscript{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 \textit{et al.}, 2000). With the exception of the experimentally determined PafA of \textit{Chrysiobacterium meningosepticum} (Berlutti \textit{et al.}, 2001), Ca\textsuperscript{2+} ATPase of \textit{Myroides odoratus} (Peiffer \textit{et al.}, 1996), PhoV of \textit{Synechococcus sp.} PCC7942 (Wagner \textit{et al.}, 1995) and alkaline phosphatase (PhoD) of \textit{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\textsubscript{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 \textit{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<sub>ACX6.71</sub>, provide evidence for the identification of a gene belonging to the newly described NPP family. Phylogenetic analysis shows that phoA<sub>ACX6.71</sub> 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 histidine 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<sub>ACX6.71</sub> 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 \( \text{phoN}_{\text{ACX6.13}} \) and \( \text{phoA}_{\text{ACX6.71}} \) were cloned into pET30b and pET30a, respectively and the proteins were overexpressed. ORF sequences of \( \text{phoN}_{\text{ACX6.13}} \) and \( \text{phoA}_{\text{ACX6.71}} \) 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 \textit{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<sub>ACX6.71</sub> 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<sub>ACX6.71</sub> and PhoN<sub>ACX6.13</sub> were partially purified. 5.1-fold purification was achieved for the class A NSAP PhoN<sub>ACX6.13</sub> while a 6.2-fold purification was achieved for alkaline phosphatase PhoA<sub>ACX6.71</sub>. 
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\textsubscript{ACX6.13} as detected on a Western blot.

Lane 1, \textit{E. coli} BL21 pAR0033 (pET30b with PhoNACX6.13) protein extract; Lane 2, \textit{E. coli} BL21 pET30b protein extract.
Protein marker size
(kDa)
Figure 3-14b: Expression of protein \( \text{PhoA}_{\text{ACX6.71}} \) as detected on a Western blot.

Lane 1, \textit{E. coli} BL21 pAR0045 (pET30a with \( \text{PhoA}_{\text{ACX6.71}} \)) protein extract; Lane 2, \textit{E. coli} BL21 pET30a protein extract.
Chapter 4: Partial reconstruction of soil and activated sludge 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 pit\(_{310}\)::tn5 phoC490) and RmG762 (phoC490 pit\(_{310}\)::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
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 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

<table>
<thead>
<tr>
<th>Library</th>
<th>Source</th>
<th>No of unique clones from RmF726 complementation</th>
<th>No of unique clones from RmG471 complementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CX4</td>
<td>Pulp waste activated sludge</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CX6</td>
<td>Municipal waste activated sludge</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CX9</td>
<td>Soil</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 4-1: Functional distribution of ORFs identified in cosmids pCX4-10F, pCX6-13F, pCX9-45F, pCX4-3G and pCX6-8G.
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\textsubscript{pAR010} and pitA\textsubscript{pAR010} separated by 11 bp (Figure 4-2a). The gene orfA\textsubscript{pAR010} encodes a 215-amino acid protein, homologous to a number of Rhizobiaceae 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\textsubscript{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, Phenyllobacterium zucineum HLK1, accession no YP_002132152, and Rhizobium leguminosarum bv. Trifolii WSM1325, accession no ZP_02293977. The G+C contents of orfA\textsubscript{pAR010} and pitA\textsubscript{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\textsubscript{pAR010} and pitA\textsubscript{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\textsubscript{pAR010} and pitA\textsubscript{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$_{\text{pAR010}}$ and GAGGA-N$_{53}$-ATG for pitA$_{\text{pAR011}}$. The region upstream to orfA$_{\text{pAR010}}$ 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$_{\text{pAR011}}$, 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$_{\text{pAR011}}$ 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<sub>pAR011</sub> is a putative phosphoglucomutase, belonging to functional category “cell wall/membrane/envelope biogenesis” (brown)
<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>orfA</td>
<td>1746</td>
</tr>
<tr>
<td>pitA</td>
<td>5129</td>
</tr>
<tr>
<td>phoU</td>
<td>1</td>
</tr>
<tr>
<td>pstB</td>
<td>5129</td>
</tr>
<tr>
<td>pstA</td>
<td>5129</td>
</tr>
<tr>
<td>pstC</td>
<td>5129</td>
</tr>
<tr>
<td>pstS</td>
<td>5129</td>
</tr>
<tr>
<td>orf13</td>
<td>5129</td>
</tr>
<tr>
<td>pheU</td>
<td>5129</td>
</tr>
<tr>
<td>Br</td>
<td>1</td>
</tr>
</tbody>
</table>
The 5 ORFs downstream from the glucosamine mutase gene encode a *pstSCAB-phoU* operon where *(pstSCAB)*\textsubscript{pAR011} encodes a phosphate transport system and *phoU*\textsubscript{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\textsubscript{pAR011} (77% a.a. identity), PstA\textsubscript{pAR011} (78% a.a. identity) PstB\textsubscript{pAR011} (90% a.a. identity). PstC\textsubscript{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\textsubscript{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*\textsubscript{pAR011}, GGAAGG(N6)ATG upstream of *pstC*\textsubscript{pAR011}, GAAG(N38)ATG upstream of *pstA*\textsubscript{pAR011}, AAGGA(N6)ATG upstream of *pstB*\textsubscript{pAR011} and AAGGA(N7)ATG upstream of *phoU*\textsubscript{pAR011}. A putative “pho box” was found 76 bp upstream to the translation start site of *pstS*\textsubscript{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*\textsubscript{pAR010-*pitA*\textsubscript{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*\textsubscript{pAR010}-*pitA*\textsubscript{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.
Complementation tests demonstrate that OrfA\textsubscript{pAR010}-PitA\textsubscript{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 \textit{orfA}\textsubscript{pAR010} did not reveal evidence of regulatory mechanism equivalent to those of the \textit{S. meliloti} OrfA-Pit system. The failure to detect a “pho box” upstream to \textit{orfA}\textsubscript{pAR010} does not agree with the finding that a pho box is present upstream to the \textit{S. meliloti orfA-pit} operon and that PhoB acts as a negative regulator of the OrfA-Pit transport system in \textit{S. meliloti} (Yuan \textit{et al.}, 2006b). The failure to detect a PhoB binding site upstream to \textit{orfA}\textsubscript{pAR010} suggests that \textit{orfA}\textsubscript{pAR010}-PitA\textsubscript{pAR010} may be regulated by the external phosphate concentration in a PhoB-independent manner as was shown with the polyphosphate kinase (\textit{ppk}) gene of \textit{S. meliloti} (Yuan \textit{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 \textit{et al.}, 2006b).

A second P transport system and a downstream uptake regulator \textit{pstSCAB-phoU} were also able to complement \textit{gly} growth phenotypes of various \textit{S. meliloti} mutants. Contrary to \textit{orfA}\textsubscript{pAR010}-PitA\textsubscript{pAR010}, the pAR011-encoded \textit{pstSCAB-phoU}, isolated from a municipal waste activated sludge metagenome, was not able to complement the \textit{phoC-pit} double mutant, RmG830. However, growth of each of the \textit{phoC} single mutant, RmG490, the \textit{phoT} single mutant, RmG491 and \textit{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 \( \text{pstS}_{\text{pAR011}} \) 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 \( \text{pstSCABphoU} \) 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 \( \text{S. meliloti} \) where two “pho boxes” were present upstream to \( \text{pstS} \) 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 \( \text{pstSCAB-phoU} \) 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 \( \text{S. meliloti} \) that the PstSCAB transport system is not competitively inhibited by phosphonate (Yuan et al., 2006a). Furthermore, RmG471 contains a \( \text{pstSCAB} \) in the chromosome but this conventional Pi uptake system cannot substitute for PhoCDET in the uptake of phosphonate while \( (\text{pstSCAB-phoU})_{\text{pAR011}} \) appears to have a broader substrate specificity.

Aside from \( (\text{pstSCAB-phoU})_{\text{pAR011}} \) 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 \( \text{pstS} \) ORFs, four each of \( \text{pstC, pstA} \) and \( \text{pstB} \), and two \( \text{phoU} \) 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<sub>pAR011</sub> and PstS<sub>pCX4-10F</sub> 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<sub>pCX4-10F</sub> and PstS<sub>pAR011</sub> have a mutual amino acid identity of 60% and a similarity of 76%. PstS<sub>pAR011</sub> was found to have a 59% amino acid identity and 74% similarity to Y. pestis PstS while PstS<sub>pCX4-10F</sub> 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<sub>pAR011</sub> and PstS<sub>pCX4-10F</sub> with MtPstS). Aside from the overall sequence identity, PstS<sub>pAR011</sub> and PstS<sub>pCX4-10F</sub> 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<sub>pAR011</sub> and PstS<sub>pCX4-10F</sub>. 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<sub>pAR011</sub> and PstS<sub>pCX4-10F</sub> (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 LAK-DGLVQFPTVIGGVVPVVNIAIQPGQIKLTGQVLGDIYLGKITKWSDPALAALNPGLKLPDADI
pCX4-10FpstS 23 LKA-DGMIQFPAIIGGVVPVLNVEGLQPGQLLTGQLLADIYLGKVKKWNDDAPIAALNPGVKLPSTQI
YpPstS gi|218931093| 25 LAT-EGLFQFPTVIGGVVLAVNIPGLKSGELVLDGKTLGDIYLGTVKKWNDDAPIAALNPGVKLPDQNI
EcPBP gi|89110279| 25 LAQ-EGLFQFPTVIGGVVLAVNIPGLKSGELVLDGKTLGDIYLGKIKKWDDEAIAKLNPGLKLPSQNI
MtPstS gi|41353623| 26 SPVTLAETGMTLLYPLFNLWGPAFHERYPNVTITAQGTTGSGAGIAQAAAGTVNIGASDAYLSEG

......****

pAR011pstS  SPVRRADGSGTTFIFTNYLSKVNPD-WKAKVEGTAIVNWP---GAGGKGNEQVAAVFQLPSNISGY
pCX4-10FpstS  AVVRRSDGSGTSFLFTTYLSQVSAE-WKEKVGAGTAVSWPA---GVGGKNEGQASYQSQISIGSGY
YpPstS gi|218931093|  VEYAY---AKRNKMTYALLKNKDGNFVAPDDETFAAAAASPSTIDWAKAPAFGEVLTDEPGKN
EcPBP gi|89110279|  VEYAY---AKQNNLAYTKLISADGKPVSPTEHSFSSAAKGADW--SKSTQADLTPQKD
MtPstS gi|41353623|  IGISFLDQASQRGLGEAQLGNSSGNFLLPDAQSIAAAAAQFAKSTFANQAISMDGPAFD

......****

pAR011pstS  LEYAY---VKQNLKSLTLMKNDQGNNFFVPSDTAFAAAAGADW---KSSFYQITTQPGKD
pCX4-10FpstS  VEYAY---AKRRNMTYALLKNDQGNFVPSDTAFAAAASDWAFAAPAFGEVLTDEPGKN
YpPstS gi|218931093|  VEYAY---AKQNNLAYTKLISADGKPVSPTEHSFSSAAKGVDW---SKSFAQDLTNQKGD
EcPBP gi|89110279|  VEYAY---AKQNNLAYTKLISADGKPVSPTEENFANAAKAADW---SKFQTDLTNQKGD
MtPstS gi|41353623|  IGISFLDQASQRGLGEAQLGNSSFNPLDQAISAAAAGAFSHTFANQAISMDGPAFD

......****

pAR011pstS  AWPITNPTYIMKHKVQEFAPAATAAAMKFDWAYGNDKMA---DDDYVPL 318
pCX4-10FpstS  SWPIGTASFLMEMHANPEPKETERLEVLKSWAKNGKGLA---TELDSVP 321
YpPstS gi|218931093|  VWPIKTSTFVHLKHEQKNAANGETVLKFDGWIGHTGAKQ---NELYATL 321
EcPBP gi|89110279|  AWPISTTFTVILHDQKKEQGTEVLKFDWAYKGTGAKQ---NDLASL 321
MtPstS gi|41353623|  GYPIINYEAYINVNRRQKDAOATQTLQAFLHAWTDNKAFLQVHFQPL 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 \textit{pstA} (\textit{pstA}_{pCX6-13F} and \textit{pstA}_{pAR011}) and \textit{pstC} (\textit{pstC}_{pCX6-13F} and \textit{pstC}_{pAR011}) were identified while one of each \textit{pstA} and \textit{pstC} were identified from the metagenomes of pulp waste activated sludge (\textit{pstA}_{pCX4-10F} and \textit{pstC}_{pCX4-10F}) and soil (\textit{pstA}_{pCX9-45F} and \textit{pstC}_{pCX9-45F}). As previously illustrated in \textit{E. coli}, a functional PstC permease requires charged residues Arg 237 and Glu 241 in one of its transmembrane helices for phosphate transport (Cox \textit{et al.}, 1989). All PstC proteins identified in this study contain the equivalent of Arg 237 and Glu 241 with the exception of PstC\textsubscript{pCX9-45F} which has valine and leucine residues in place of the respective arginine and glutamate residues required for phosphate transport while PstC\textsubscript{pAR011} has a glycine instead of the required glutamine. Similarly it was shown in \textit{E. coli} that in order to be functional, one of the transmembrane helices of the PstA protein requires Arg 220 (Cox \textit{et al.}, 1989). Each of the metagenomic PstA have the required arginine residue, equivalent to Arg 220 of the \textit{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 \textit{pstB} ORFs, \textit{pstB}\textsubscript{pCX4-10F}, \textit{pstB}\textsubscript{pCX6-13F}, \textit{pstB}\textsubscript{pCX9-45F} and \textit{pstB}\textsubscript{pAR011} were identified from the pulp and municipal waste activated sludge and soil metagenomes. Each of the four \textit{pstB} ORFs is homologous to the \textit{E. coli pstB} and all but \textit{pstB}\textsubscript{pAR011} contain the conserved glycine and lysine residues, equivalent to Gly 48 and Lys 49, required for phosphate transport (Cox \textit{et al.}, 1989). In
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*<sub>pCX4-10F</sub>, *phoU*<sub>pAR011</sub> and *phoU*<sub>pCX9-45F</sub> were identified (Figures 4-6, 4-2b and 4-8). PhoU<sub>pAR011</sub> shares 85% amino acid identity with the putative PhoU from the *β*-proteobacterium *Leptothrix cholodnii* SP-6. However PhoU<sub>pAR011</sub> 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<sub>pCX4-10F</sub> is the putative PhoU from the *β*-proteobacterium *Lutielia 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<sub>pCX4-10F</sub> showed an amino acid identity of 25% and 24%, respectively. Similarly, PhoU<sub>pCX9-45F</sub> also shares a 41% amino acid identity with its nearest homologue, the *β*-proteobacterium *Dechloromonas 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
phosphate transport by PhoU<sub>pAR011</sub>, PhoU<sub>pCX4-10F</sub> and PhoU<sub>pCX9-45F</sub> differs from the known regulatory mechanism demonstrated in EcPhoU, PhoUAQUE and Tm1743.

A <i>phoB/phoR</i> two-component regulatory system controls a large number of genes that are part of the phosphate (<i>pho</i>) regulon. Genes of <i>pho</i> 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 <i>pho</i> regulon (Figure 4-5). A putative two component system <i>phoR<sub>pCX4-10F</sub>-phoB<sub>pCX4-10F</sub></i> was identified from pulp waste activated sludge metagenome (Figure 4-6). The genes <i>phoB<sub>pCX4-10F</sub></i> and <i>phoR<sub>pCX4-10F</sub></i> are found as an operon in which the stop codon of <i>phoB<sub>pCX4-10F</sub></i> overlaps with the start codon of <i>phoR<sub>pCX4-10F</sub></i>. The response regulator, <i>phoB<sub>pCX4-10F</sub></i> has an amino acid identity of 71% with the putative <i>phoB</i> of the β-proteobacterium Laribacter hongkongensis HLHK9 (accession no YP_002794170.1) while <i>phoR<sub>pCX4-10F</sub></i> shares a lower amino acid identity (44%) with the putative <i>phoR</i> 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 <i>phnGHIJKL</i> operon for C-P lyase. Previous work demonstrates the requirement of the <i>phn</i> operon by <i>S. meliloti</i> 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 <i>phnGHIJKL</i> are required to restore the Gly<sup>+</sup> 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.
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\textsubscript{pCX9-45F} and the sensor histidine kinase cztS_silS_copS\textsubscript{pCX9-45F} were found as two ORFs separated by 50bp (Figure 4-8).
<table>
<thead>
<tr>
<th>Cosmid clone (contig)</th>
<th>Gene identified</th>
<th>Similar protein</th>
<th>Organism</th>
<th>% identity</th>
<th>Region of identical amino acids</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCX4-10F (6-2)</td>
<td>orf4_pCX4-10F</td>
<td>PAS/PAC sensor hybrid histidine kinase</td>
<td><em>M. nodulans</em> ORS 2060</td>
<td>37</td>
<td>302</td>
<td>143</td>
</tr>
<tr>
<td>pCX9-45F (17)</td>
<td>cztR_silR_copRp&lt;sub&gt;cvX9-45F&lt;/sub&gt;</td>
<td>Two component transcriptional regulator, winged helix family</td>
<td><em>G. bemidjiensis</em></td>
<td>63</td>
<td>224</td>
<td>279</td>
</tr>
<tr>
<td>pCX9-45F (17)</td>
<td>cztS_silS_copS&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Heavy metal sensor signal transduction histidine kinase</td>
<td><em>G. bemidjiensis</em></td>
<td>31</td>
<td>205</td>
<td>90.5</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>adc_pCX6-13F</td>
<td>Adenylate cyclase</td>
<td><em>D. aromatica</em> RCB</td>
<td>45</td>
<td>487</td>
<td>408</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>A. aromaticum</em> EbN</td>
<td>37</td>
<td>506</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Azoarcus sp.</em> BH72</td>
<td>36</td>
<td>497</td>
<td>252</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>% identity</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------</td>
<td>------------</td>
<td>---------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>gdc&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Diguanylate cyclase/phosphodiesterase</td>
<td>D. aromatica RCB</td>
<td>51</td>
<td>592</td>
<td>571</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>J. sp. Marseille</td>
<td>49</td>
<td>576</td>
<td>568</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H. arsenicoxydans</td>
<td>48</td>
<td>577</td>
<td>555</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>draT&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Hypothetical protein-NAD+-dinitrogen-reductase ADP-D-ribosyltransferase activity</td>
<td>D. aromatica RCB</td>
<td>65</td>
<td>328</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. vinosum DSM 180</td>
<td>57</td>
<td>268</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Magnetococcus sp. MC-1</td>
<td>51</td>
<td>266</td>
<td>281</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>deoR&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Regulatory functions: DNA interactions</td>
<td>M. xanthus DK 16-22</td>
<td>56</td>
<td>211</td>
<td>235</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. dehalogenans 2-CP1</td>
<td>56</td>
<td>216</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Anaeromyxobacter sp. K</td>
<td>56</td>
<td>216</td>
<td>228</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>% identity</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>-------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX6-13F (4)</td>
<td>Orf12 pCX6-13F</td>
<td>Cyclic nucleotide-binding</td>
<td>P. pacifica SIR-1</td>
<td>69</td>
<td>185</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H. ocraceum DSM 14365</td>
<td>53</td>
<td>215</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. marina DSM 3645</td>
<td>33</td>
<td>202</td>
<td>148</td>
</tr>
<tr>
<td>pCX4-10F (13-1)</td>
<td>phoR&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Phosphate regulon histidine kinase</td>
<td>Methylovorus sp. SIP 3-4</td>
<td>45</td>
<td>408</td>
<td>287</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Azoarcus sp. BH72</td>
<td>43</td>
<td>409</td>
<td>278</td>
</tr>
<tr>
<td>pCX4-10F (13-1)</td>
<td>phoB&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Phosphate regulon response regulator</td>
<td>L. hongkongensis HLH-K9</td>
<td>71</td>
<td>170</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. aromaticum EbN1</td>
<td>69</td>
<td>170</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thauera sp. MZ1T</td>
<td>68</td>
<td>170</td>
<td>208</td>
</tr>
</tbody>
</table>
**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\textsubscript{pCX4-10F} protein of unknown function; Orf2\textsubscript{pCX4-10F}, probable transmembrane permease; Orf3\textsubscript{pCX4-10F} protein of unknown function; Orf4\textsubscript{pCX4-10F}, putative histidine kinase sensor protein; Orf5\textsubscript{pCX4-10F}, hypothetical protein; Orf6\textsubscript{pCX4-10F}, putative outer membrane protein.
Contig 6-2

1

\[\text{gst2}_{pCX4-10F}\quad \text{orf3}_{pCX4-10F}\quad \text{orf4}_{pCX4-10F}\quad \text{orf5}_{pCX4-10}\]

Contig 14

1

\[\text{pstS}_{pCX4-10F}\quad \text{pstC}_{pCX4-10F}\quad \text{pstA}_{pCX4-10F}\quad \text{pstB}_{pCX4-10F}\quad \text{phoU}_{pCX4-10F}\]
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 phosphoglucomutase; 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.
**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: Orf1pCX9-45F, protein of unknown function; Orf2pCX9-45F, protein of unknown function; Orf3pCX9-45F, unknown protein; Orf4pCX9-45F, unknown protein; Orf5pCX9-45F, ATP/GTP binding protein; Orf6pCX9-45F, TonB-dependent receptor plug domain; Orf7pCX9-45F, putative serine-threonine protein kinase; Orf8pCX9-45F, protein of unknown function.
Two additional genes involved in signal transduction were found in the metagenome of municipal waste activated sludge. A putative adenylate cyclase, \( \text{adc}_\text{pCX6-13F} \) and a putative diguanylate cyclase/phosphodiesterase \( \text{gdc}_\text{pCX6-13F} \) 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). \( \text{adc}_\text{pCX6-13F} \) and \( \text{gdc}_\text{pCX6-13F} \) are homologues of the respective putative adenylate cyclase and diguanylate cyclase/phosphodiesterase from the \( \beta \)-proteobacterium \( \text{Dechloromonas aromatica} \) RCB with adenylate cyclase homologues sharing 45% amino acid identity and the diguanylate cyclase/phosphodiesterase homologues share 51% amino acid identity. \( \text{gdc}_\text{pCX6-13F} \) 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 occurrence 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 \textit{nifHD}K\textit{fdxN}\textit{nifY} preceded by a putative upstream dinitrogenase-reductase ADP-D-ribosyltransferase (DRAT) negative regulator, \textit{draT}\textsubscript{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 \textit{nifK}\textsubscript{pCX6-13F} and \textit{nifD}\textsubscript{pCX6-13F}) and a dinitrogenase reductase homodimer (component II encoded by \textit{nifH}\textsubscript{pCX6-13F}). Nitrogenase activity is supported by an iron-molybdenum cofactor (\textit{nifY}\textsubscript{pCX6-13F}) that binds to the active site of the dinitrogenase enzyme and an indole pyruvate ferredoxin oxidoreductase (\textit{fdxN}\textsubscript{pCX6-13F}) which mediates electron transfer in metabolic reactions. The \textit{nif} genes identified here are homologues of the putative \textit{nif} genes of the \(\beta\)-proteobacterium \textit{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, \textit{glnD1}\textsubscript{pCX9-45F} and \textit{glnD2}\textsubscript{pCX9-45F}, were identified from the metagenome of soil (Figure 4-8). The product of the \textit{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 (\textit{ntrBC, nifA, glnA}) involved in nitrogen fixation and assimilation (Zhang et al., 2001). \textit{GlnD1}\textsubscript{pCX9-45F} and \textit{GlnD2}\textsubscript{pCX9-45F} are homologues of halophilic \(\gamma\)-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

<table>
<thead>
<tr>
<th>Cosmid clone (contig)</th>
<th>Gene identified</th>
<th>Similar protein</th>
<th>Organism</th>
<th>Identity (%)</th>
<th>Region of identical amino acids</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCX6-13F (1)</td>
<td>( nifH_{pCX6-13F} )</td>
<td>Nitrogenase iron protein subunit</td>
<td>(D. aromatica) RCB</td>
<td>93</td>
<td>287</td>
<td>515</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Azoarcus sp.) BH72</td>
<td>87</td>
<td>287</td>
<td>493</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(T. turnerae) T7901</td>
<td>83</td>
<td>283</td>
<td>473</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>( nifD_{pCX6-13F} )</td>
<td>Nitrogenase molybdenum-iron protein alpha chain: nitrogenase component I, alpha chain</td>
<td>(D. aromatica) RCB</td>
<td>90</td>
<td>288</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Azoarcus sp.) BH72</td>
<td>85</td>
<td>288</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(D. dadantii) Ech 703</td>
<td>78</td>
<td>288</td>
<td>159</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>( nifK_{pCX6-13} )</td>
<td>Nitrogenase molybdenum-iron protein beta chain</td>
<td>(D. aromatica) RCB</td>
<td>85</td>
<td>397</td>
<td>415</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Azoarcus sp.) BH72</td>
<td>76</td>
<td>397</td>
<td>355</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Azoarcus sp.)</td>
<td>77</td>
<td>397</td>
<td>355</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>( fdxN_{pCX6-13F} )</td>
<td>4Fe-4S ferredoxin, iron-sulfur binding</td>
<td>(D. aromatica) RCB</td>
<td>94</td>
<td>17</td>
<td>40.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(R. capsulatus)</td>
<td>70</td>
<td>17</td>
<td>35.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(R. phaeroides) KD131</td>
<td>70</td>
<td>17</td>
<td>33.9</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>Identity (%)</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------</td>
<td>--------------</td>
<td>--------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>nifY&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Dinitrogenase iron-molybdenum cofactor biosynthesis</td>
<td>D. aromatica RCB</td>
<td>77</td>
<td>243</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Azoarcus sp. BH72</td>
<td>52</td>
<td>211</td>
<td>193</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. vinelandii DJ</td>
<td>49</td>
<td>207</td>
<td>176</td>
</tr>
<tr>
<td>pCX9-45F (10-3-19)</td>
<td>glnD&lt;sub&gt;1pCX9-45F&lt;/sub&gt;</td>
<td>UTP-GlnB uridylyltransferase, GlnD</td>
<td>H. halophila SL1</td>
<td>55</td>
<td>620</td>
<td>662</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. ehrlichii MLHE-1</td>
<td>54</td>
<td>637</td>
<td>654</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N. mobilis Nb-231</td>
<td>53</td>
<td>634</td>
<td>637</td>
</tr>
<tr>
<td>pCX9-45F (12)</td>
<td>glnD&lt;sub&gt;2pCX9-45F&lt;/sub&gt;</td>
<td>UTP-GlnB (protein PII) uridylyltransferase, GlnD</td>
<td>C. salexigens DSM 3043</td>
<td>45</td>
<td>177</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. ehrlichii MLHE-1</td>
<td>46</td>
<td>166</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Beggiatoa sp. PS</td>
<td>48</td>
<td>171</td>
<td>151</td>
</tr>
</tbody>
</table>
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, \( \text{gst1}_{\text{pCX4-10F}} \) and \( \text{gst2}_{\text{pCX4-10F}} \), a glutathione-dependent formaldehyde dehydrogenase, \( \text{adhC}_{\text{pCX4-10F}} \) and an S-formylglutathione hydrolase, \( \text{fghA}_{\text{pCX4-10F}} \) 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, \( \text{dmm}_{\text{pCX6-13F}} \) was identified. Across 557 amino acids of the C-terminal domain of Dmm\(_{\text{pCX6-13F}}\), a 33% identity to the putative DNA modification methyltransferase protein of the radiation resistant \textit{Deinococcus radiodurans} R1, suggesting that Dmm\(_{\text{pCX6-13F}}\) is divergent from known DNA modification methyltransferases.
Table 4-4: Identified genes involved in cellular defense mechanisms

<table>
<thead>
<tr>
<th>Cosmid clone (contig)</th>
<th>Gene identified</th>
<th>Similar protein</th>
<th>Organism</th>
<th>% identity</th>
<th>Region of identical amino acids</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCX4-10F (13-1)</td>
<td><em>gst1</em>&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Glutathione S-transferase domain protein</td>
<td><em>Acidovorax</em> <em>sp</em>. JS 42</td>
<td>68</td>
<td>228</td>
<td>325</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Diaphorobacter</em> <em>sp</em>. TPSY</td>
<td>68</td>
<td>228</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>A.avenae subsp. citrulli</em> AAC00-1</td>
<td>66</td>
<td>229</td>
<td>316</td>
</tr>
<tr>
<td>pCX9-45F (6-2)</td>
<td><em>gst2</em>&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Glutathione S-transferase domain protein</td>
<td><em>V. paradoxus</em> S110</td>
<td>61</td>
<td>178</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Janinthobacterium</em> <em>sp</em>. Marseille</td>
<td>56</td>
<td>189</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>gst12</em> <em>S. meliloti</em> 1021</td>
<td>62</td>
<td>180</td>
<td>199</td>
</tr>
<tr>
<td>pCX4-10F (9)</td>
<td><em>adhC</em>&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Glutathione dependent formaldehyde dehydrogenase</td>
<td><em>N. spumigena</em> CCY9419</td>
<td>82</td>
<td>369</td>
<td>585</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Lyngbia sp.</em> PCC8106</td>
<td>81</td>
<td>369</td>
<td>582</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>P. lavamentivo rans</em> DS-1</td>
<td>83</td>
<td>369</td>
<td>582</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>P. zucineum</em> HLK1</td>
<td>61</td>
<td>256</td>
<td>280</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>M. maris</em> MCS10</td>
<td>63</td>
<td>256</td>
<td>278</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>γ proteobacterium</em> HTCC5015</td>
<td>64</td>
<td>244</td>
<td>276</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>% identity</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>------------</td>
<td>---------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX6-13 (1)</td>
<td>dmm&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>DNA modification methyltransferase-related protein</td>
<td><em>Rosieflexus</em> sp. RS-1</td>
<td>35</td>
<td>493</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>D. radiodurans</em> R1</td>
<td>33</td>
<td>489</td>
<td>236</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Lyngbia sp.</em> PCC8106</td>
<td>31</td>
<td>506</td>
<td>230</td>
</tr>
</tbody>
</table>
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 \( \beta \)-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 \( \gamma \)-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\textsubscript{pCX4-10F} is homologous to many HAD hydrolases of subfamily IB while hadH\textsubscript{1pCX6-13F} is homologous with many subfamily IIA HAD like hydrolase and hadH\textsubscript{2pCX6-13F} 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 \textit{E. coli} K12 NagD, a protein involved in N-acetylglucosamine metabolism, for which structural data is available (Tremblay et al., 2006). HadH\textsubscript{1pCX6-13F} 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>
<thead>
<tr>
<th>Cosmid clone (contig)</th>
<th>Gene identified</th>
<th>Similar protein</th>
<th>Organism</th>
<th>% identity</th>
<th>Region of identical amino acids</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCX4-10F (10)</td>
<td>hda&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Chromosomal replication initiator</td>
<td>A. aromaticum ebA1</td>
<td>49</td>
<td>67</td>
<td>57.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thauera sp. MZ1T</td>
<td>45</td>
<td>74</td>
<td>55.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Janintherbacterium sp. Marseille</td>
<td>49</td>
<td>73</td>
<td>53.1</td>
</tr>
<tr>
<td>pCX9-45F (18)</td>
<td>int&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Integrase, catalytic region</td>
<td>Thioalklyvibrio sp. HL-EbGR7</td>
<td>50</td>
<td>259</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R. leguminosarum bv trifolii WSM1325</td>
<td>46</td>
<td>254</td>
<td>219</td>
</tr>
<tr>
<td>pCX4-10F (13-1)</td>
<td>pcnB&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Poly(A)-polymerase</td>
<td>B. petri DSM 12804</td>
<td>63</td>
<td>354</td>
<td>432</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. duloso AUO158</td>
<td>56</td>
<td>447</td>
<td>429</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. avium 197N</td>
<td>63</td>
<td>354</td>
<td>426</td>
</tr>
<tr>
<td>pCX9-45F (12)</td>
<td>rpsB&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Ribosomal protein S2</td>
<td>Thioalkali-vibrio sp. K90mix</td>
<td>63</td>
<td>259</td>
<td>329</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thioalkali-vibrio sp. HL-EbGR7</td>
<td>62</td>
<td>244</td>
<td>328</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. aromaticum EbA1</td>
<td>64</td>
<td>264</td>
<td>324</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>% identity</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>----------</td>
<td>------------</td>
<td>-------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX9-45F</td>
<td>tsf&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Translation elongation factors Ts</td>
<td><em>M. thiooxidans</em> DMS010</td>
<td>55</td>
<td>293</td>
<td>290</td>
</tr>
<tr>
<td>(12)</td>
<td></td>
<td></td>
<td><em>F. philomiragia</em></td>
<td>57</td>
<td>285</td>
<td>283</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>N. oceanii</em> ATCC 19707</td>
<td>55</td>
<td>292</td>
<td>280</td>
</tr>
<tr>
<td>pCX9-45F</td>
<td>map&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Methionine aminopeptidase</td>
<td><em>Thioalkali-vibrio sp. HL-EbGR7</em></td>
<td>71</td>
<td>235</td>
<td>368</td>
</tr>
<tr>
<td>(12)</td>
<td></td>
<td></td>
<td><em>S. maltophilia</em> R551-3</td>
<td>69</td>
<td>242</td>
<td>363</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>S. maltophilia</em> K279A</td>
<td>68</td>
<td>242</td>
<td>363</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>% identity</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------</td>
<td>------------</td>
<td>--------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX4-10F (13-1)</td>
<td><em>hadH</em>&lt;sub&gt;pCX4-10F&lt;/sub&gt;</td>
<td>Haloacid dehalogenase like hydrolase, subfamily IB</td>
<td><em>B. phymatum</em> STM815</td>
<td>62</td>
<td>222</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>B. glumae</em> BGR1</td>
<td>63</td>
<td>224</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>R. metallo-durans</em> CH34</td>
<td>62</td>
<td>222</td>
<td>287</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td><em>hadH</em>&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Haloacid dehalogenase like hydrolase, subfamily IIA</td>
<td><em>M. mobilis</em> JW8</td>
<td>49</td>
<td>178</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>M. flagellatus</em> KT</td>
<td>46</td>
<td>171</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Methylo-vorus sp.</em> SIP 3-4</td>
<td>47</td>
<td>177</td>
<td>179</td>
</tr>
<tr>
<td>pCX6-13F (4)</td>
<td><em>hadH</em>&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Haloacid dehalogenase like hydrolase, subfamily IIG, 5’ nucleotidase 2-amino-4-hydroxy-6-hydroxymethyl yldihydropteridinediphosphokinase</td>
<td><em>P. pacifica</em> SIR-1</td>
<td>58</td>
<td>515</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>A. dehalo-genans</em> 2CP-C</td>
<td>45</td>
<td>493</td>
<td>408</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Anaeromyxobacter sp.</em> K</td>
<td>44</td>
<td>492</td>
<td>407</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>M. petro-liphilum</em> PM-1</td>
<td>58</td>
<td>150</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>V. paradoxus</em> S110</td>
<td>58</td>
<td>129</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>B. petrii</em> DSM 12804</td>
<td>57</td>
<td>153</td>
<td>142</td>
</tr>
<tr>
<td>Cosmid clone (contig)</td>
<td>Gene identified</td>
<td>Similar protein</td>
<td>Organism</td>
<td>% identity</td>
<td>Region of identical amino acids</td>
<td>Score</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------</td>
<td>------------</td>
<td>-------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>pCX6-13F (13)</td>
<td>folP&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Dihydropteroate synthase</td>
<td>D. aromatica RCB</td>
<td>63</td>
<td>134</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thauera sp. MZ1T</td>
<td>62</td>
<td>134</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. japonicus Ueda107</td>
<td>64</td>
<td>136</td>
<td>172</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>orf5&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Metal dependent phospho-hydrolase</td>
<td>D. acidovorans SPH-1</td>
<td>31</td>
<td>358</td>
<td>93.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. testosteroni KF1</td>
<td>36</td>
<td>280</td>
<td>89.4</td>
</tr>
<tr>
<td>pCX9-45F (10-13-1)</td>
<td>apbA&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>2-dehydropteroyl 2-dehydropteroate 2-reductase</td>
<td>A. delafieldii 2AN</td>
<td>32</td>
<td>365</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S. sviceus ATCC 29083</td>
<td>48</td>
<td>399</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. cavernae DSM 12333</td>
<td>43</td>
<td>278</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. acidiphila DSM 44928</td>
<td>42</td>
<td>282</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R. picketti 12D</td>
<td>53</td>
<td>266</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R. solana-ceanum GMI1000</td>
<td>52</td>
<td>266</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R. solana-ceanum UW552</td>
<td>52</td>
<td>266</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M. mobilis JLW8</td>
<td>31</td>
<td>296</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R. eutropha JMP134</td>
<td>28</td>
<td>306</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L. hong-kongensis HLHK9</td>
<td>31</td>
<td>302</td>
<td>134</td>
</tr>
</tbody>
</table>

182
<table>
<thead>
<tr>
<th>Cosmid clone (contig)</th>
<th>Gene identified</th>
<th>Similar protein</th>
<th>Organism</th>
<th>% identity</th>
<th>Region of identical amino acids</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCX6-13F (1)</td>
<td>orf6&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Putative myo-inositol 1-monophosphatase</td>
<td>Janibacter sp. HTCC 2649</td>
<td>52</td>
<td>231</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C. flavigena DSM 20109</td>
<td>50</td>
<td>256</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. cavernae DSM 1233</td>
<td>53</td>
<td>204</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Janibacter sp. HTCC 2649</td>
<td>45</td>
<td>351</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Putative DNA-binding protein</td>
<td>K. radio-tolerans SRS 30216</td>
<td>44</td>
<td>322</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. faecium DSM 4810</td>
<td>38</td>
<td>337</td>
<td>159</td>
</tr>
<tr>
<td>pCX6-13F (1)</td>
<td>orf7&lt;sub&gt;pCX6-13F&lt;/sub&gt;</td>
<td>Putative DNA-binding protein</td>
<td>Janibacter sp. HTCC 2649</td>
<td>45</td>
<td>351</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K. radio-tolerans SRS 30216</td>
<td>44</td>
<td>322</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B. faecium DSM 4810</td>
<td>38</td>
<td>337</td>
<td>159</td>
</tr>
<tr>
<td>pCX9-45F (17)</td>
<td>pabB&lt;sub&gt;pCX9-45F&lt;/sub&gt;</td>
<td>Parapaminobenzoate synthase subunit I</td>
<td>D. baculatum DSM 4028</td>
<td>72</td>
<td>42</td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. maxima CS328</td>
<td>65</td>
<td>42</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A. dehalo-genans 2CP-C</td>
<td>58</td>
<td>42</td>
<td>57</td>
</tr>
</tbody>
</table>
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 \( \delta \)-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, \( \text{folK}_{pCX4-10F} \) and \( \text{folP}_{pCX6-13F} \), encoding enzymes involved in folic acid biosynthesis were identified from pulp and municipal waste activated sludge metagenomes (Figure 4-6 and 4-7). \( \text{folK}_{pCX4-10F} \) 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 of antimicrobial agents in the way of inhibition of folic acid biosynthesis (Yan et al., 2001). The nearest homologue to FolK\textsubscript{pCX4-10F} is the putative HPPK of the \textit{\beta-Proteobacterium Methylibium petrophilum} PM1 (accession no YP_001022207) at 58%. The overall amino acid sequence identity of FolK\textsubscript{pCX4-10F} compared to the well-characterized \textit{E. coli} HPPK (accession no 21238998) is only 40%. However, of the nine residues involved in substrate binding, FolK\textsubscript{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 \textit{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 asparagine residue for the formation of H bond with the HP substrate (Yan et al., 2001). Of the nine residues involved in substrate binding in \textit{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\textsubscript{pCX4-10F} include only Gly 14 and Phe 128 (corresponding to Gly 8 and Phe 123 in \textit{E. coli} HPPK). Leu 45, Tyr 53 and Trp 89 are replaced in FolK\textsubscript{pCX4-10F} as follows: Val 51, Phe 59 and Asn 94.

The second gene, \textit{folP\textsubscript{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
bacteria such as *Mycobacterium tuberculosis* and *Staphylococcus aureus* (Baca *et al.*, 2000). FolP<sub>pCX6-13F</sub> 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<sub>pCX6-13F</sub>. An even lower amino acid sequence identity was observed between FolP<sub>pCX6-13F</sub> 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<sub>pCX6-13F</sub> protein is its small size (157 a.a.) compared to an almost double size (270-280 amino acids) *E. coli*, *M. tuberculosis* 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<sub>pCX9-45F</sub>* 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<sub>pCX9-45F</sub> 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, especially 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 occurrence 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\textsubscript{pCX4-10F}, Orf3\textsubscript{pCX9-45F} and Orf4\textsubscript{pCX9-45F} (Figures 4-6 and 4-8) have no homologues present in the NCBI database for bacterial isolates while Orf3\textsubscript{pCX6-13F}, Orf5\textsubscript{pCX6-13F}, Orf3\textsubscript{pCX9-45F} and Orf4\textsubscript{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.

<table>
<thead>
<tr>
<th>Contig designation</th>
<th>Orf designation</th>
<th>Best isolate homologue</th>
<th>Identity (%)</th>
<th>Best homologues from environmental samples</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCX4-10F</td>
<td>Orf1pCX4-10F</td>
<td><em>Herminiimonas aresenic oxydans</em> HEAR2639</td>
<td>90</td>
<td>JCVI_PEP_1112736227834 GS123</td>
<td>78</td>
</tr>
<tr>
<td>13-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pCX4-10F</td>
<td>Orf5pCX4-10F</td>
<td><em>Methylobacillus fagellatus</em> KT Mfla_1127</td>
<td>30</td>
<td>JCVI_PEP_1105110335817 GS000a</td>
<td>31</td>
</tr>
<tr>
<td>6-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pCX4-10F</td>
<td>Orf3pCX4-10F</td>
<td>-</td>
<td>-</td>
<td>JCVI_PEP_1105160931511 GS025</td>
<td>32</td>
</tr>
<tr>
<td>6-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pCX6-13F-1</td>
<td>Orf2pCX6-13F</td>
<td><em>Dechloromonas aromatica</em> RCB Daro_1407</td>
<td>74</td>
<td>JCVI_PEP_1113974750160 SLUDGE_AUS</td>
<td>45</td>
</tr>
<tr>
<td>pCX6-13F-1</td>
<td>Orf3pCX6-13F</td>
<td><em>Dechloromonas aromatica</em> RCB Daro_1409</td>
<td>79</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pCX6-13F-1</td>
<td>Orf4pCX6-13F</td>
<td><em>Dechloromonas aromatica</em> RCB Daro_1418</td>
<td>83</td>
<td>JCVI_PEP_1113368395762 SLUDGE_AUS</td>
<td>60</td>
</tr>
<tr>
<td>pCX6-13F-4</td>
<td>Orf5pCX6-13F</td>
<td><em>Plesiocystis pacifica</em> S-R1 PPSIR1_23339</td>
<td>40</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pCX6-13F-4</td>
<td>Orf6pCX6-13F</td>
<td><em>Candidatus Nitrospira defluvii</em> UPF0047</td>
<td>82</td>
<td>JCVI_PEP_1105154612789 GS032</td>
<td>64</td>
</tr>
<tr>
<td>pCX9-</td>
<td>Orf1pCX9-45F</td>
<td><em>Desulfococcus</em></td>
<td>49</td>
<td>JCVI_PEP_1113522379096</td>
<td>66</td>
</tr>
<tr>
<td>Contig designation</td>
<td>Orf designation</td>
<td>Best isolate homologue</td>
<td>Identity (%)</td>
<td>Best homologues from environmental samples</td>
<td>Identity (%)</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>--------------</td>
<td>-------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>45F 10-3-19</td>
<td>deovorans Hxd3 Dole_1205 Beggiatoa sp. PSZP_01999465</td>
<td>31</td>
<td>JCVI_PEP_1105138462911 GS020</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>CX9-45F 10-3-19</td>
<td>Orf2CX9-45F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CX9-45F 10-3-19</td>
<td>Orf3CX9-45F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CX9-45F 10-3-19</td>
<td>Orf4CX9-45F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CX9-45F 10-3-19</td>
<td>Orf8CX9-45F</td>
<td>Solibacter usitatus Ellin6076 Acid_4938</td>
<td>36</td>
<td>JCVI_PEP_1113521680492 FARM SOIL</td>
<td>46</td>
</tr>
</tbody>
</table>
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 phylogenetic 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 surprising 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 high-throughput 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 examining 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.chemlink.com
Bibliography


